NAGOYA UNIVERSITY

A Predictive State Representation Framework for General-Purpose Mobile Reasoning Agents

by

Robin Karlsson

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Graduate School of Informatics
Department of Intelligent Systems

August 2024

NAGOYA UNIVERSITY

Abstract

Graduate School of Informatics
Department of Intelligent Systems

Doctor of Philosophy

by Robin Karlsson

Fully autonomous intelligent mobile robots promise a mobility revolution, with an impact on society similar to the industrial revolution. Despite some progress, we are still far from achieving the full extent of a mobility revolution. The ansatz of this thesis is that the full autonomous driving problem is not solvable by iterative engineering improvements, but is in fact an artificial general intelligence agent problem.

This thesis presents conclusions from an investigation of limitations of common autonomous driving paradigms. A theoretical framework for future autonomous general-purpose mobile reasoning agent capable of spatio-semantic commonsense reasoning is proposed to overcome identified limitations. In particular, an open-vocabulary predictive state representations is presented as an artificial hippocampus implemented by a latent variable generative predictive world model capable of continual learning based on the principle of predictive coding. The mathematical theory of latent compositional semantics is used to form queryable spatio-semantic memories.

The predictive state representation is supported by rigorous theoretical grounding and experimental evidence in the autonomous driving domain. Experiments prove discoverablity of latent compositional semantics by vision-languge models (VLMs), learning to predict a diverse set of spatially and semantically accurate predictive environment states by the proposed open-vocabulary predictive world model (OV-PWM). The usefulness of the predictive state representation is demonstrated by enabling a self-supervised directional soft lane probability model to learn navigational patterns better than SOTA supervised models.

The proposed framework is well-grounded in the research literature and provides a multitude of promising new research directions. Limitations and future work include expanding the state representation from 2D to 3D, adding temporal dynamics, and improve diversity and accuracy of latent compositional semantic VLM inference.

Acknowledgements

I would like to express my deepest gratitude to Professor Kazuya Takeda, who has been an exceptional supervisor, mentor and friend during my Ph.D journey. His unwavering encouragement and support enabled me to balance the demands of pursuing a doctorate with the responsibilities of supporting my family. Our relationship transcends the boundaries of academia, and I consider him a person of trust who has significantly impacted both my personal and professional life.

I am also grateful for the guidance and wisdom provided by Professor Alexander Carballo throughout my Ph.D studies. His practical know-how in mobile robotics and sensors has been invaluable to my research, and I appreciate his consistent encouragement and support. Additionally, I would like to extend my thanks to Professor Keisuke Fuji for the feedback and suggestions he provided during our insightful discussions on mathematical aspects of machine learning.

I owe a debt of gratitude to Professor Kento Ohtani, who has always been present and ready to assist with practical matters throughout my years in graduate school. Furthermore, I would like to express my appreciation for Dr. Tomoki Hayashi, whose diligent maintenance of the lab computer cluster ensured smooth progress in my research. His expertise in ML devops and computing resource management know-how has been valuable.

I would also like to extend my heartfelt thanks to the dedicated secretaries of the Takeda Lab who have provided immense support throughout my Ph.D journey. In particular, I am deeply grateful to Mrs. Chika Ando for her warm and unwavering efforts in assisting me with administrative reports and a multitude of other tasks over the years.

I would like to acknowledge the Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System for providing me with financial support through their scholarship program during my Ph.D studies. This support has enabled me to focus on my research without worrying about daily expenses, making a significant difference in my academic journey.

Last but not least, I would like to express my thanks to my wife and little boy. Despite my constant stress and immense workload with little income, you were there to keep the house clean and the boy and our dogs fed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background: The Promise of a Mobility Revolution

The philosopher Immanuel Kant formulated the Categorical Imperative as the supreme principle of all moral beings to do what is "universally good" [9, 10]. A logical consequence is that pursuing universally good actions is an intrinsically valuable pursuit for moral beings like humans. Following the same reasoning, a collective of moral beings also possess an imperative to pursue what is universally good, and by doing so improve the existential circumstances of the collective as a whole. In modern socioeconomic science the general goodness of human existence is denoted quality of life.

The development of society as the collective of collective moral human beings is studied in the fields of political philosophy [11–13], social sciences [14, 15], and economics [16–18]. Modern economists typically measure quality of life by measurable indicators, such as Gross Domestic Product (GPD) per capita and Human Development Index [19]. The anthropologist Harari [15] and economist Piketty [18] present historical socio-economic evidence showing objective human quality of life has been correlated with the abundance of produce in societies throughout the world. It is evident that the degree of quality of life, and thus the advancement of society, is correlated with the amount of goods produced and available for the populace. However, the amount of available human labor is a limited and in most regions a costly resource. Consequentially, advancement of society is also correlated with the efficiency with which humans can produce goods and complete tasks. In other words, the quicker a task can be completed, or to the extent a task can be automated, the more surplus goods can be produce to advance society and quality of life.

The current exponential trend of societal progress originated during the 16-17th century with the *scientific revolution* [18]. The development of the scientific method of empirical observation, causal experimentation, and logical reasoning led to the development of new technological advancements including metallurgy, tooling, and machinery. The ensuing *industrial revolution* in the 17-18th century gave humanity automated production, resulting in a world with an abundance of goods. The *electronics revolution* in the late 19th century automated information transmission, greatly enhancing efficiency of human task completion by providing us with means of instantaneous communication across space. This thesis proposes that a future *mobility revolution* is the next production efficiency jump based on the promise of automated transportation of physical goods and humans. Fully autonomous transportation is expected to revolutionize society and further enhance production efficiency by freeing up human time spent on the task of moving about and transporting goods. The fundamental technology for automated transportation is autonomous vehicles powered by autonomous driving systems.

Autonomous driving (AD) systems are conceptually equivalent to AI agent systems. An agent is a defined as an intelligent system that exists in an objective external dynamic environment state x_t at time t. The environment is perceived by a sensor and perception or digital information processing system $f()$. The system produces imperfect partial observations or percepts z_t approximating the true environment state x_t. In general, an agent builds up an approximate environment state $\hat{x}_t$ based on a state estimator $h()$ conditioned one or all previous observations z and the previous estimated state $\hat{x}_{t-1}$. The goal of an agent is to perform an optimal actions a_t to satisfy a desired behavior as deemed by an utility function $h()$ conditioned on the estimated environment state $\hat{x}_t$ and previous actions

$$z_t := f(x_t) \tag{1.1}$$

$$\hat{x}_t := g(\hat{x}_{t-1}, z_t, \dots z_0) \tag{1.2}$$

$$a_t := \arg\max h(a_t|\hat{x}_t, a_{t-1}, \dots a_0). \tag{1.3}$$

The approach taken to implement the perception function (1.1), the state representation function (1.2), and behavior function (1.3) fundamental functions defines the agent.

The fundamental challenge of designing intelligent agents is how to perceive and represent the environment and its own state with sufficient sophistication, as well as produce sequences of actions resulting in optimal behavior over the short- and long-term time horizon. Recent advances in machine learning (ML) are rapidly enabling the realization increasingly sophisticated agents.

The deep learning revolution [20, 21] is believed to be the prime enabling technology for intelligent autonomous driving systems. Throughout the history of AI several emerging paradigms have enjoyed periods of widespread optimism. Prominent examples are expert systems [22, 23] reaching a peak popularity in the 1980s, and statistical modeling approaches [24, 25] developed in the 1990s. Eventually optimism was dampened by limitations in scope of practical applications of AI for general real-world problems. In contrast, the levels of optimism and resilience of deep learning models as practical and noise robust universal function approximators [26] and a versatile method for representation learning [27] has been on an increasing trend for over a decade. Current optimism is backed up by revolutionary practical applications of AI throughout society [28]. The distinguishing properties of the deep learning methodology are the capability to learn useful semantic abstractions from imperfect noisy data, and the empirical observation that prediction performance scales with amount of data, model size, and computational resources [29]. Combined with the onset of the big data revolution in 2020s to digitalize information throughout society, the robust and scalable properties of deep learning resulted in the hypothesis that any degree of intelligence could be reached simply from obtaining enough data and computational resources [30]. A hypothetical practically feasible path to achieving artificial general intelligence (AGI) was widely perceived for the first time [31].

The capability of deep learning models to automatically discover complex features and representations within the input data [27] makes them well-suited for core tasks in autonomous driving, such as perception [32] and motion planning [33]. This has led many researchers and engineers to believe that fully autonomous intelligent mobile systems, and therefore the mobility revolution, are within reach in the immediate future. The promised vision of fully automated transportation systems that can navigate from point A to B, while interacting safely with their environment and other traffic participants, at any given time, seemed like a principally solved problem based on deep learning technology. The focus shifted from being a fundamental AI problem to an engineering problem predominantly revolving around collecting enough data to fully cover the operational domain, annotate the data by human endowed semantic meaning, and possessing enough computational resources to adequately process the amassed annotation data.

However, despite the progress made in ML and mobile robotics, this thesis argues that we are still far from achieving a true mobility revolution as evident from the past decade of partial progress. While progress has been made towards creating semi-autonomous vehicles with advanced driver assistance features such as highway lane following [33] or autonomous parking [34], achieving safe full autonomy in unconstrained environments, and correctly handling unexpected long-tail events, remains a challenging problem. Currently research and development work is primarily done towards engineering refined

versions of semi-automated systems that ultimately depends on human supervision to ensure safe operation in uncontrolled dynamic environments. The design of these systems can navigate predefined routes and stop at specific locations without colliding with anything en route while limiting acceleration and jerk. The proliferation of such systems has the potential to provide social convenience, but the impact is limited compared to fully automated transportation as humans remain in the loop. Basing autonomous driving systems on uninterpretable deep learning models also have ethical and legal implications. The question of assignment of responsibility in the case a learned black box decision making function causes an accident or violates traffic laws is generally avoided by putting to the human being which supervises the system. Judging by empirical evidence, the ability of machines to learn correlation patterns from data by deep learning is not enough to create truly intelligent agents that can navigate any environment safely and efficiently [35, 36].

The ansatz for this thesis is that solutions that work for sub-problems might not necessarily scale up to solve the ultimate problem. The hypothesis is that region constrained autonomous shuttle buses and taxis are primarily a software engineering problem, while the creation of truly autonomous human-like intelligent mobile agents is very much still a research problem requiring solutions beyond correlative pattern learning approaches. It might even be the case that the full autonomous driving problem is an AGI problem. While current AI systems can perform specific tasks with human-level performance, the same narrow expert systems lack the ability to understand how to solve problems beyond the training data, or produce efficient behavior for interacting with other agents in a fully autonomous manner [35]. Creating truly intelligent transportation systems may require not only advancements in software engineering and hardware but also breakthroughs in AGI research.

This thesis analyzes the limitations of existing autonomous driving paradigms, and lays out a novel direction based on the requirements of a future general-purpose mobile reasoning agent. The hypothesis is that a general-purpose agent capable of reasoning over world knowledge, and learn to imagine the environment state beyond observations based on experience, will also overcome the limitations plaguing existing rule- and learning-based approaches to autonomous driving systems and finally realize the promised mobility revolution.

The following section presents a brief history of autonomous vehicles. The exposition clarifies how the functional components of modern commercial autonomous vehicle systems, such as remote operation, lane maps, and computer vision, where in fact originated decades ago. The historical analysis argues that progress since the 1970s are dominantly computer and sensor hardware, computer vision algorithms, and digitalization of lane

maps. In comparison, the conceptual approach has evolved little from the modular *perception-planning-control* methodology established more than 50 years ago.

1.1.1 History: 100 Years of Autonomous Vehicles

Understanding history provides context for interpreting current events and trends. To understand the progress of the current state-of-the-art (SOTA) in autonomous driving systems, a broader inter-generational perspective is useful. In fact, the quest for autonomous mobile robots capable of safely navigating public road environments without human oversight has a long history spanning over a century. Here follows an approximately chronological account of the pursuit of autonomous vehicles, showing how components of the current mainstream approach has naturally evolved over time.

Remote operation. The pursuit of self-driving vehicles began in 1925, when a driverless car rolled down the streets of New York City. This event was captivating to the public and marked the beginning of the driverless car era. Francis P. Houdiina, a former U.S. Army electrical engineer and founder of the Houdiina Radio Control Co., built what is believed to be the first radio-operated automobile. He set up a 1926 Chandler sedan with a transmitting antenna that operated small electric motors controlling the vehicle's speed and steering angle. A crew following behind in another car remotely controlled the vehicle. The radio controlled vehicle drove through heavy traffic in Broadway, managing to turn corners, accelerate, decelerate, and honking its horn. Unfortunately, the demonstration ended when it crashed into another vehicle. Despite this setback, variations of the vehicle were showcased years later on public roads in the US to the excitement of onlookers. The public has always held the prospect of autonomous driving in high regard and heralded its benefits with enthusiasm.

Physical lane maps. Autonomous vehicles have been an ongoing commercial pursuit since the early 1930s, with various visions and technologies being proposed and tested throughout the decades. In the early stages, Norman Bel Geddes envisioned a future where cars could drive themselves with electronic speed and collision control systems similar to those found in railroads. His Futura ride for General Motors at the 1939 World's Fair imagined grooves that would keep cars apart in their own "lane tracks." The idea was to engage automatic systems and relieve the driver from driving until reaching one's exit, with related visions involving magnetic trails, physical slots or troughs, or train-like rails engaging hidden steel wheels on the inside of each tire. Despite technological advancements over the years, the two basic ideas have remained largely unchanged: smart cars and smart roads. Primary goals for autonomous vehicles include safety, speed,

access, more cars sharing the road, intelligent intersection navigation, and reducing congestion. Early self-driving plans focused on special freeways to guide suitably equipped cars safely along them, but building public infrastructure proved challenging.

Navigation by AI and computer vision In the late 1960s, experimental robots began navigating through novel environments at Stanford Research Institute (SRI) and other US universities, testing out new artificial intelligence (AI) techniques [28]. By 1971, semi-autonomous space probes were landing on other worlds, spurring a visionary future of autonomous intelligent mobile agents. Early AI researchers began dreaming of cars smart enough to navigate ordinary streets on their own. However, the challenges were daunting for then popular expert systems. Focus lay in reverse engineering the relevant systems in a moving biological systems like cockroaches: sensing, environment processing, and reacting with appropriate behavior. In the 1980-1990s, German pioneer Ernst Dickmanns and his group at Univ. Bundeswehr Munich (UniBW) created three generations of an autonomous driving Mercedes sedan called VAmP. The autonomous vehicle would cover 1000 km in Paris traffic at up to 130 km/h, and 1700 km on the German autobahn driving up to 180km/h while passing other cars [37]. In Japan, the Tsukuba Mechanical Engineering Lab developed a computerized driverless car that could achieve speeds of up to 30 km/h using machine vision to track white road markings.

Hardware, implementation, and digital maps. Autonomous driving research began in earnest in the 1980s when the Defense Advanced Research Projects Agency (DARPA) launching the first Grand Challenge in 1984 to encourage widespread development of self-driving cars. The DARPA Grand Challenges held from 2004 to 2007, were revolutionary in advancing AV technology. The first competition, a 241 km course in the Mojave Desert, saw no vehicle complete the route due to technical difficulties. However, the event sparked interest and investment in self-driving technology at universities such as Stanford and Carnegie Mellon. In the subsequent years, universities and industry made significant strides towards developing reliable practical autonomous vehicles. Progress culminated in the DARPA Urban Challenge in 2007 that first demonstrated autonomous vehicles operating in realistic urban environments [38].

Commercialization. By the late 2000s, Google started researching and developing autonomous vehicle with the intent of future commercialization as part of a moon shot project. The resulting Google Self-Driving Car Project lead by the DARPA Grand Challenge winners Sebastian Thrun and Anthony Levandoowski. Under Thrun's leadership, the team recruited top researchers in the field and began refining the DARPA Challenge approach and technology into a commercial system. The goal was to create a reliable self-driving car capable of safely carrying passengers in real-world urban traffic conditions. Google's autonomous vehicle system, now commercially developed by Waymo,

has since guided a fleet of vehicles over at least 800,000 km without causing any accidents, making the company a leading advocate for fully autonomous vehicles. As of the 2020s, a multitude of companies and universities in all regions of the world are developing autonomous driving systems according to diverse approaches. The main focus is real-world safety and scalability, and reducing the need of human supervision. Practical problems include robust perception, interactive motion planning, system safety, and defining legislative regulations.

This thesis proposes that the conventional paradigms pursued by the majority of companies and universities are fundamentally limited. The following section explains why the conventional approaches may not sufficiently scale up to realize the full extent of autonomous mobility.

1.2 Autonomous Driving Paradigms and Limitations

Autonomous driving agents can be categorized based on their underlying design principle. This thesis proposes a categorization by one of three paradigms: First, rule-based agents based on defining explicit behavior by human programmed rules, execution structure, or algorithms. Secondly, learning based agents where behavior emerges implicitly from patterns in observational or example data. Finally, vision-language models (VLMs) based on perception grounded with world knowledge and reasoning capabilities of large language models. The limitations identified for each paradigm provide motivation for why the full autonomous driving problem is an AGI problem, as well as how the proposed predictive state representation based approach is a part of overcoming identified limitations.

1.2.1 Rule-based procedural instruction

The rule-based approach to autonomous driving agents is a methodology that relies on a set of predefined rules or heuristics to navigate and control self-driving vehicles [38–40]. These rules are typically derived from traffic laws, road regulations, and common driving practices, and they dictate how the vehicle should respond in different situations and environments. A rule-based system is typically separates the agent system (1.1)-(1.3) into separated task-specific modules or components to manage complexity [41]. Conventional systems partition components by task, such as perception, localization, path planning, and control. The rule-based approach is characterized by its simplicity and ease of implementation, and a capable working system can be designed without complex algorithms or extensive computational resources.

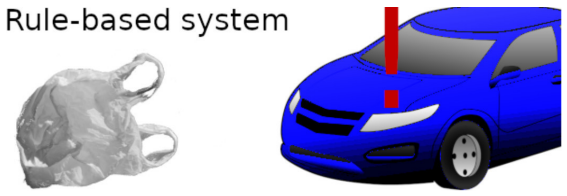

FIGURE 1.1: Rule-based autonomous driving agents have limited capability to make complicated decisions, like to break or not break for a flying plastic bag in a congested roadway, due to the challenge of programmatically defining "common sense" knowledge.

One of the main advantages of the rule-based approach is its ability to specify explicit behavior on how the vehicle should behave in various scenarios. For instance, rules can be defined for speed limits, lane changes, intersections, and overtaking maneuvers [42]. By adhering to these rules, autonomous driving agents can verifiably be ensured that they comply with traffic laws and regulations, which is crucial for safety and legal compliance. Moreover, the rule-based approach allows autonomous vehicles to make decisions quickly and efficiently, as there is typically no need for complex computations or extensive data processing. For example, rule-based approaches for motion planning include generating paths using Dijkstra's algorithm [43] or the A-star algorithm [44]. The simplicity of the rule-based approach makes it an attractive option for companies looking to develop self-driving vehicles and other mobile robots quickly and cost-effectively.

However, rule-based approaches have limitations as they rely on handcrafted rules and features that may not generalize well to all scenarios or environmental conditions [45]. One of the main challenges is that it can be difficult to anticipate all possible driving scenarios and develop human-defined rules that cover every eventuality. For example, if a self-driving vehicle encounters an unexpected obstacle, such as debris on the road, there may not be a predefined rule for how to respond. Specifying explicit rules for what debris that can and cannot be traversed is nontrivial. Another example is to break or not break for a flying plastic bag in a congested highway environment as illustrated in Fig. 1.1. Colliding with objects should in general be avoided, but common sense implies that hitting a plastic bag is better than causing a high-speed chain collision by panic breaking. Additionally, relying on human effort can lead to inconsistencies or errors in implementation, or interaction of multiple rules result in conflicts and unexpected outcomes [46]. Moreover, as road conditions and regulations change over time, the rules need to be updated regularly to ensure that they remain accurate and relevant.

Predefined rules can provide a reliable framework for behaving safely and efficiently in well-defined environments like highways where anomalies are relatively rare [39, 42, 47]. However, in general environments a system must be capable to learn from experience to adapt itself to ambiguous, complex and changing environments. Human beings rely on experience and "common sense" to overcome seemingly simple problems, like deciding to

drive or not drive over a fallen tree branch. Such common sense rules generally cannot be precisely defined by humans. Therefore an intelligent agent, like human beings, must possess a predictive world model which allows simulating expected outcomes of hypothetical actions based on common sense knowledge of the world.

While the rule-based approach allows for rapid proof of concept demonstrations, the approach does not scale gracefully to uncontrolled environments. In particular, rule-based systems lack "common sense" which humans generally rely upon to overcome unexpected situations [48–50]. The lesson learned is that setting out to a priori define correct behaviour to all encounterable situations in the real world is simply intractable, and that common sense is hard to specify programatically. Continuous learning during operation is therefore a required component of a fully autonomous mobile agent.

1.2.2 Supervised correlation pattern learning

The second autonomous driving agent paradigm is based on supervised learning for discovering useful correlation patterns in data. Learning-based autonomous driving has been a significant area of research for the past decade, with numerous studies focusing on developing systems that can navigate complex environments semi- or fully autonomously [33, 51, 52]. Learning-based methods use machine learning techniques such as supervised imitation learning [53], reinforcement learning (RL) [54] and inverse RL (IRL) [55] to enable an autonomous vehicle to make decisions based on the perceived current and past environment states. The learning-based approach allows the agent system to learn to integrate the perception (1.1), state representation estimator (1.2), and behavior function (1.3) components in a highly adaptable manner.

One of the most popular learning-based approaches for autonomous driving is deep reinforcement learning (DRL) [56, 57]. DRL combines deep neural networks with RL algorithms to create agents that can learn complex behaviors from high-dimensional sensory inputs, such as images or point cloud data. Several studies have demonstrated the effectiveness of DRL in various tasks related to autonomous driving [58, 59]. The hypothetical advantage of RL is that useful task semantics such as road, lane markings, and other vehicles, can be learned implicitly from raw data based solely on a task reward signal [31]. For instance, using Deep Q Network (DQN) [57] to train an agent to navigate through urban areas using camera images as input [60, 61]. The network learned image representations that detected the road successfully without being explicitly trained to do so.

However, training an autonomous driving agent in real-world scenarios presents several challenges [59]. RL methods typically require randomly exploring the action space to

FIGURE 1.2: Learning-based system generally lack explicit decision factors and a logical reasoning process, meaning the resulting behavior is not guaranteeably safe. The capability of explicit rule specification and following is likewise limited.

gather experience for learning a behavior maximizing reward. In the real world random exploration is not possible due to danger to the agent and environment. Moreover, it may not always cover all possible situations that the agent could encounter during deployment. Learning from a scalar reward is inefficient as methods generally lack a causal model for what decision factors cause the reward. The combination of learning from failure and inefficient sample learning efficiency is the primary challenge in the autonomous driving agent domain [62]. Human beings can learn and adapt quickly based on instruction or failure likely due to their capacity of performing commonsense reasoning based on a highly capable world model facilitating cause-and-effect mental simulations [63, 64] Another challenge is designing a robust reward function for real-world environments that precisely specify the intended behavior aligned with human interests [65]. The problems of learning-based agents are illustrated in Fig. 1.2 depicting the challenge of guaranteeing desirable and safe behavior of an agent traversing a complicated rule-constrained multiagent environment.

To address the issue of safety, researchers have proposed using simulation environments to train RL agents before deploying them in real-world scenarios [58]. Simulation allows for generating a vast amount of data quickly and cheaply, making it an attractive option for training autonomous driving policies [66]. Another approach is to bootstrap behavior from models trained by imitation learning [33, 51, 67]. For instance, [68] introduced a safe policy that learns to predict errors made by a primary policy trained initially with supervised learning without querying a reference policy. This approach ensures that the agent does not deviate from its learned behavior while still exploring new strategies safely. However, imitation learning does not fundamentally solve the exploration problem, and generating example data can be time-consuming and expensive.

Inverse reinforcement learning (IRL) is another popular learning-based approach for autonomous driving. IRL involves inferring an underlying reward function from expert demonstrations and using it to train RL agents. This method has been used in various tasks related to autonomous driving, such as intent prediction for traffic actors like pedestrians or vehicles [42]. However, estimating the cost function accurately is challenging, which can lead to sub-optimal policies being learned by the agent. To address

this issue, researchers have proposed using end-to-end learning approaches that directly map observations to actions without explicitly defining a cost function.

Another challenge associated with IRL and DRL methods is their reliance on i.i.d. (independent and identically distributed) data assumptions for data generation. This assumption does not hold in real-world scenarios where the environment's state changes dynamically due to factors such as weather conditions, traffic patterns, or pedestrian behavior. To overcome this issue, researchers have proposed using Data Aggregation (DAgger) methods [55, 69] that iteratively collect observation-action pairs during testing and retrain the agent with aggregated data from both training and testing phases.

Learning-based approaches such as DRL and IRL have shown promise in enabling autonomous vehicles to navigate complex environments independently. However, several real-world challenges remain unaddressed, including data collection limitations, exploration-exploitation tradeoffs, reward function specification, and non-i.i.d. sequential data considerations pose challenges to real-world application of learning-based approaches. This thesis proposes to overcome practical limitations of learning-based approaches by advancing beyond correlative pattern learning on the sensor observation level. The reliance on inefficiently learn from failure experience can be mitigated by incorporating cognitive world models facilitating reasoning based on commonsense knowledge. With commonsense world knowledge and a mental simulator, it becomes possible to predict detrimental outcome of hypothetical action sequences with limited experience of failure.

1.2.3 Vision-Language Models

Large language models (LLMs) trained on massive amounts of text data have proved to be a scalable method to learn and reason about commonsense knowledge of the world [70]. The capability of LLMs for common-sense reasoning and versatility in handling various inputs has inspired researchers to explore their potential as components in autonomous driving agents [71–76]. Vision-language models (VLMs) [77–81] adds a visual encoder trained to transform image content into the language embedding space. This agent paradigm enables end-to-end autonomous driving systems that leverage the power of both computer vision and natural language processing. Representing percepts by language provides a means to perform spatio-semantic commonsense reasoning by LLMs grounded in the external environment. VLMs enable more robust generalization by allowing the system to understand complex scenes better [75, 82], improve far-horizon planning [83, 84], support rich and precise human-machine communication [73], and plan safe trajectories [85–87].

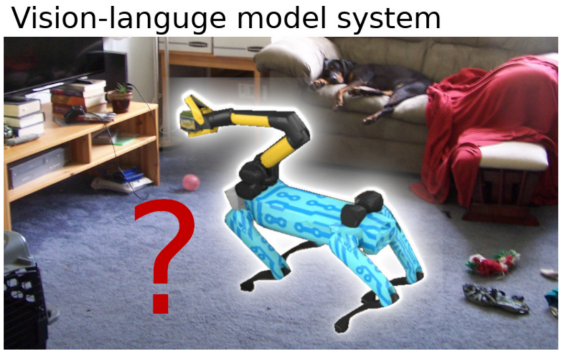

FIGURE 1.3: VLM-based agent systems typically lacks a principled spatio-semantic state representation. Most agents are limited to perceive and act on what is in immidiate view. Other agents leverage a explicit list of objects. Neither approach provides a adequate state representation for recalling how to clean up a messy room.

Various approaches have been proposed for integrating VLMs in autonomous driving systems that combines characteristics from rule- and learning-based systems by typically applying learned models in a structured modular structure. The perception and state representation functions (1.1)-(1.2) are modeled by a VLM outputting states estimates $\hat{x}$ functioning as both percepts and states. The behavior function (1.3) is typically implemented by a LLM finetuned on domain data for outputting actions or analysing the state $\hat{x}$. A typical approach is based on designing a language prompt or instruction for which a multimodal LLM completes by leveraging visual features, world knowledge, and task knowledge specified in the prompt itself [88].

The integration of vision and language in autonomous driving systems has also enabled the development of new applications, such as natural language-based user interfaces and task instructions for vehicle control [73, 75], visual question answering (VQA) systems for scene comprehension [89], and multimodal fusion techniques for improved perception [72].

While providing convincing proof of concept work, there are several limitations with current VLM approaches that need to be addressed for their wider application in real-world scenarios. One major limitation is the lack of generalization across unseen data and tasks. While LLMs have demonstrated superior common-sense capabilities and improved performance on various natural language processing tasks, incorporating these capabilities into real-world autonomous driving tasks remains a challenge. This is particularly relevant in addressing long-tail scenarios, where the model needs to handle rare or unusual situations that may not have been encountered during task-specific finetuning training [75].

Another limitation is the complexity of designing prompts for optimal performance. The effectiveness of VLMs relies heavily on the design and structure of the language prompt

used to guide the reasoning process [70, 90]. This introduces a level of subjectivity into the model's performance, as different prompts can lead to significantly different results. Additionally, it is not always feasible or practical to include all relevant environmental observation information in the text prompt, which may limit the model's ability to make accurate predictions and decisions [91]. The effectiveness of these models is contingent upon the quality and diversity of the training data, with a lack thereof potentially leading to reduced performance or an inability to generalize effectively [75]. Furthermore, VLM approaches are computationally intensive and require significant resources for training and deployment. This can be a barrier to their wider adoption, particularly in resource-constrained environments or applications with real-time requirements.

This thesis argues that the most limiting factor of VLM-based agents is their lack of a principled spatio-semantic memory as illustrated in Fig. 1.3. The state representation $\hat{x}+t$ in (1.2) is typically limited to a textual description of an input image [91, 92], or at best, an list of explicit semantic objects and their spatial location [91]. This thesis propose that the state representation of of VLM agents must go beyond an explicit listing of all possible observed objects and their spatial location. The state representation should instead be modeled as a 3D vector space or map containing the full geometric extent of semantic objects similar to the biological hippocampus [93, 94] Additionally, semantic information encoded into the map should querying of diverse visual and semantic attributes of objects as demonstrated by my theory of latent compositional semantics [95].

1.3 Hypothesis: Full Autonomous Driving is an AGI Agent Problem

The three mainstream approaches to designing autonomous driving agents each have inherent limitations and challenges as explained in Sec 1.2. The identified limitations and challenges serve as theoretical and experimental justification that revolutionary, not evolutionary, progress is needed. Accordingly, I propose the hypothesis that solving the full autonomous driving agent problem is akin to solving the embodied AGI agent problem. This section present how intermixing components of each paradigm compliments their weaknesses and can result in a prototypical general-purpose AGI-like agent design.

The primary issue with rule-based procedural instructions programmed by humans is scalability problems due to the vastness and complexity of the world, and the ambiguity of commonsense knowledge. This thesis proposes that continuous learning from observational experience and building a predictive world model that predicts expected outcomes of hypothetical actions based on world knowledge, learned from observational experience

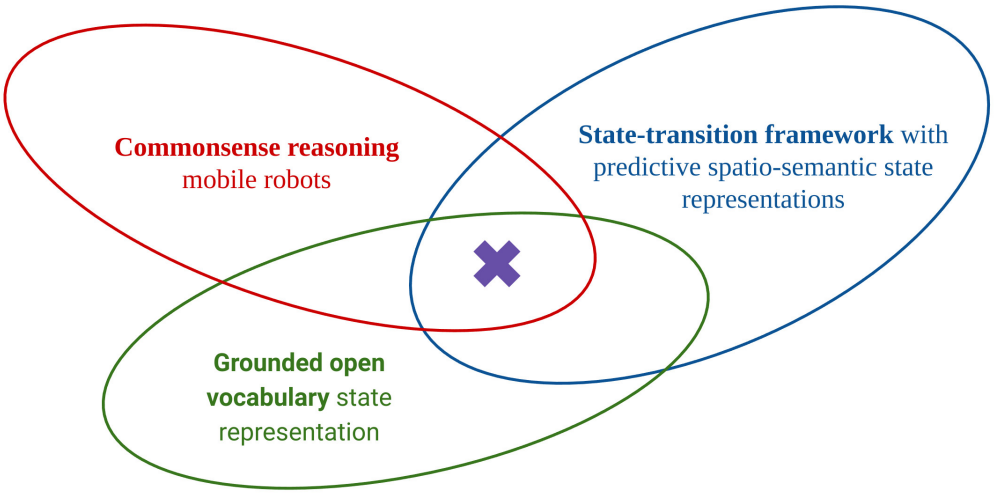

FIGURE 1.4: Elements of the proposed general-purpose mobile reasoning agent.

and non-observational knowledge source, must be an integral part of the solution. The realization of commonsense reasoning agents is further deliberated upon in Chapter 3.

The primary issue with conventional correlation-based supervised pattern learning is the unreliability of weakly structured connectionist models like neural network not discovering causal mechanisms governing the world [36, 96]. Models without causal structure and grounding in extensive world knowledge are bound to be fragile beyond the correlations observed within the training dataset [35, 97]. Even the hypothetical experiment supposing to capture and annotate data representing the entire world is principally flawed, as the world is a dynamic and evolving system that cannot be captured by a static data distribution [98]. Relying on correlation-based learning methods for decision making is therefore a fundamentally and principally flawed approach, and a system based on learning causal processes from observation, hypothesis formulation, and experiment must be pursued [63]. This thesis proposes implementing directional state-transition modeling based on semantically rich predictive world states natively compatible with multimodal LLMs capable of reasoning over commonsense reasoning encoded in natural language. A state-transition models bootstrapped with world knowledge allows learning and predicting causal outcomes of actions as the inference is directional and decisions are experimentally testable by the agent as an "artificial scientist". The proposed predictive state representation is deliberated upon on Chapter 5.

The fundamental issue with spatio-semantic reasoning by VLMs is their spatially and semantically limited representation of the environment state, preventing flexibly and efficiently mapping of "things" to "where". Another related challenge is enabling the representation and querying a diverse set of object semantics including affordances and general attributes [99–101]. This thesis proposes to implement a spatio-semantic environment representation capable of storing grounded memories of rich objects semantics

akin to the biological hippocampus [94]. The proposed predictive state representation both capture the spatial extent and rich object semantics as further deliberated on in Chapter 5 and Chapter 6. A related challenge is representing and querying diverse learnable attributes or semantics of objects. This thesis proposes to leverage open vocabulary predictive states [95] using latent compositional semantics [95]. The proposed predictive state representation goes beyond explicit listing of all possible observed object semantics and instead represents objects by learnable compositional semantics such as *surface-of-a-table-object* corresponding to the semantic set [*table*, *surface*, *wood*, *brown*, *furniture*, ...]. Additionally the thesis proposes the predictive state representation as an artificial hippocampus, expanding VLM-based agents from a primarily image-centric world state to a 3D spatio-semantic spatial memory similar to maps relied upon by conventional mobile robots.

Overcoming the limitations of the existing paradigms as proposed in this section will result in an AGI-like general-purpose mobile reasoning agent as explained in the following and visualized in Fig. 1.4. First, the agent is an AGI-like agent due to inheriting the capabilities of LLMs to perform commonsense reasoning, contextual information integration, and few-shot learning. Additionally, the state-transition framework supports mental simulations of sequences of abstract actions for discovering causal models of the environment and avoiding negative outcomes based on commonsense knowledge instead of trial-and-error. Secondly, the agent is general-purpose based on its ability to follow abstract instructions specified in natural language, as well as leverage a semantically rich and grounded open vocabulary predictive environment state representation facilitating querying of a priori unknown task semantics. Finally, the agent is an embodied mobile agent distinguished from a general-purpose multimodal LLM by the spatially grounded semantic environment state representation.

1.4 Definition: General-Purpose Agents

Throughout this thesis the term "general-purpose agent" is used to signify the working concept of *an agent designed with the capability to complete a priori unknown tasks requiring reasoning over an a priori unknown set of semantics*. The concept of general-purpose is closely aligned with AGI-like capabilities. A more throughout definition goes as follows.

A general-purpose robot is a machine designed to perform a wide range of tasks rather than being limited to a single specialized function. This versatility allows them to be adapt to various tasks across multiple domains, such as navigation, object manipulation,

question answering, and so on. Their flexibility is enhanced by hardware designs featuring microphones, speakers, articulated arms, grippers, and mobility, enabling them to work with various products or in diverse environments by leveraging multiple sensor and actuator modalities.

The motivation behind the development of versatile general-purpose machines lies in creating robotics systems that mimic the general capabilities and adaptability of humans, enabling them to operate effectively across many different situations and jobs. This approach addresses some of the limitations and challenges faced by mainstream autonomous driving agent approaches, such as scalability problems due to world complexity and ambiguity of common-sense knowledge.

This thesis propose realizing general-purpose mobile reasoning agents by continuous learning from observational experience, along with building a predictive world model that anticipates expected outcomes based on hypothetical actions. The thesis also suggests implementing causal state-transition modeling using predictive states compatible with reasoning over common sense knowledge encoded in natural language. The use of open vocabulary predictive states leveraging latent compositional semantics is proposed to overcome representational limitations of flexibly and efficiently mapping object semantics to a principled spatio-semantic spatial memory akin to an artificial hippocampus.

1.5 Scope of Thesis

This thesis presents conclusions from over six years of research into AGI-like frameworks for mobile robotics including autonomous driving systems. The investigation starts from a holistic analysis of strengths and limitations of common autonomous driving research and development paradigms, as well as functional faculties of biological intelligence as a proof of concept for the viability of general-purpose intelligent agents.

The scope of the thesis and related research literature is vast. The thesis contribution consists of two parts. The first part focuses on the theory of latent compositional semantics presented in Chap. 4, and open-vocabulary predictive state modeling framework presented in Chap. 5-6. These topics are parts of published works and presented with theoretical grounding together with extensive and rigorous experimental evidence. The experimentally evidenced results provides a foundational basis of the larger theoretical agent framework. The presented experimental results are for 2D top-down open vocabulary environment state representations suitable for navigation tasks. A full 3D predictive state representation for truly general-purpose agents is considered future work.

The second thesis contribution part presents theoretical proposals concerning the structure and faculties of AGI-like agent frameworks. The theoretical findings are based on extensive reading of inter-disciplinary research literature and deductive reasoning. In particular, chapter 3 proposes a reasoning framework grounded in spatio-semantic memories that identifies a set of core faculties as well as a computational structure combining the faculties into an autonomous general-purpose mobile reasoning agent. The resulting framework principally ties together real-world robot agents operating on partial observations, state transition modeling for long-term planning, and LLM-based reasoning over commonsense world knowledge as an internal mental simulator. A full implementation and experimental evaluation of the proposed theoretical internal mental simulator framework is considered future work.

The proposed predictive environment state representation and internal mental simulation is well-grounded in the broad research literature and provides a multitude of promising novel research directions for realizing a new generation of future versatile mobile agents capable of commonsense reasoning over principled spatio-semantic environment representations in the form of predictive states.

1.6 Thesis Overview

The thesis is organized as follows:

Chapter 2 presents a literature review of research directions related to state representations for general-purpose mobile reasoning agents. Topics include perception, open-vocbulary semantic representation learning, data-driven world modeling, and spatial cognition and navigation.

Chapter 3 provides background on biological general intelligent agents like humans and artificial intelligence agents. A theoretical framework presents core components of the proposed general-purpose mobile reasoning agent. The proposed open-vocabulary predictive state representation, associated predictive world model, and state-transition models incorporating commonsense world knowledge are the core component. Other topics include learning spatial navigation, task descriptions, planning by simulation, and low-level action execution.

Chapter 4 presents the theory of grounding latent compositional semantics as spatio-semantic memories, forming the base representation of the predictive state model. Latent compositional semantics is a principled learning-based knowledge representation and mathematical theory of unconditional open vocabulary semantic segmentation with proved mathematical properties and guarantees of optimally. Experiments show a single

latent compositional semantic embedding can represent a set of 100 random semantics for ideal uniformly distributed high-dimensional embedding spaces. Other experiments show that latent compositional semantics are discoverable from visual appearance and singular descriptions. A novel sufficient similarity semantic inference method overcomes fundamental limitations of conventional "most similar" semantic inference, achieving high-level overlapping semantic segmentation performance by 19.63 mIoU on average.

Chapter 5 explains the proposed open-vocabulary predictive world model as an artificial hippocampus that learns based on the principle of predictive coding similarity to the hippocampal formation in the brain. The chapter deliberates on the faculties of the biological hippocampus from a computational neuroscience perspective. The theory of predictive coding is proposed as a general learning framework grounded in biological intelligence. The internal mental simulator proposed in Chap. 3 is proposed as equivalent to an artificial hippocampus. The predictive world modeling methodology proposed in this thesis is proposed as an implementable model of such an artificial hippocampus by providing the required faculties, such as generating predictive spatio-semantic state representations and learning from observational experience. The chapter concludes with a mathematical derivation of the presented predictive world model. Derivations include hierarchical latent variable generative models, the novel posterior matching optimization algorithm, and the sampling-based inference algorithm.

Chapter 6 present an application of the latent compositional semantics and the open-vocabulary predictive world model as a predictive state representation for autonomous driving reasoning agents. The chapter explains how sensor observations are accumulated to form partially observed world states, and how the open-vocabulary predictive world model learns to generate a diverse set of plausible complete states. The usefulness of the predictive state representation is demonstrated by enabling a self-supervised directional soft lane probability model to learn navigational patterns better than SOTA supervised model while limited to partial observations only.

Chapter 7 concludes the thesis by summarizing the contributions and discussing limitations and future research directions.

Figure 1.5 provides a summary of the proposed core contributions. Each entry is categorized as an experimentally verified proposal, a theoretical proposal, or a synthesis of existing literature.

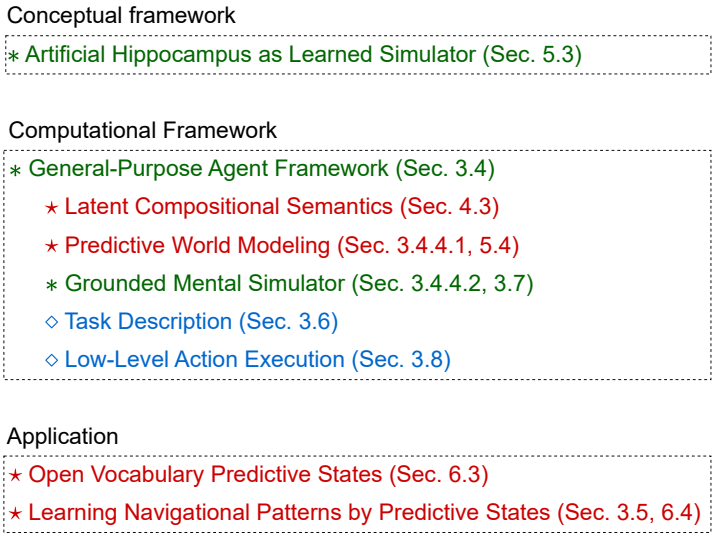

FIGURE 1.5: Structure of proposed core contributions presented in the thesis. Contributions are partitioned into groups based on degree of practical implementation and experimental verification. Red ($\star$) entries denote original proposals with experimental evidence. Green ($*$) indicates original theoretical proposals empirically supported by the literature. Blue ($\diamond$) proposals are synthesized from existing literature. Section numbers indicate central parts describing each contribution.

Chapter 2

Literature Review

2.1 Introduction

This chapter consolidates research relating to the realization of general-purpose mobile reasoning agents. The related research is partitioned into three main research groups approaching a common problem focused on in this thesis. Each group partitions relevant research into further refined sub-groups.

2.2 Perception and Semantic Representation Learning

The existing literature on self-supervised visual representation learning, dense representation learning, and open-vocabulary semantic segmentation suffers from several limitations. Current methods struggle to scale to high-resolution images, require large datasets with dense annotations, and often produce noisy or low-resolution feature maps. Additionally, common semantic interpretation methods like CLIP [77] lack spatial grounding, making such models unsuitable for tasks that require precise spatial information. Furthermore, unconditional open vocabulary semantic segmentation models like OVSeg [81] are not yet deeply understood, and their practical usefullness is limited by semantic inference requiring a complete partitioning of the image by a set of non-overlapping query semantics [95]. Knowledge representation frameworks, such as semantic networks and frames, are also limited by their inability to handle uncertainty, semantic vagueness, and incomplete data.

This thesis addresses these problems by proposing a novel method ViCE [102] that enables a model to efficiently learn to generate precise object-fitting semantic partitioning even for high-resolution images, and improves the state-of-the-art unsupervised semantic

segmentation benchmark on Cityscapes [103] and COCO [104]. Furthermore, the thesis proposes a mathematical framework for understanding the representations learned by unconditional open-vocabulary semantic segmentation models, showing that they can be interpreted as latent compositional semantic embeddings [95]. The proposed "sufficient similarity" open vocabulary semantic inference method overcomes the three fundamental flaws of conventional "most similar" open vocabulary semantic inference as identified in Sec. 4.3.4. See Table 2.1 for problem and proposed solution summaries. See the following sections for additional information and literature grounding.

TABLE 2.1: Perception and semantic representation learning problems and proposed solutions.

Problem Summary	**Proposed Solution**
Dense self-supervised visual representation learning lacks efficiency and effectiveness.	Introduce the method ViCE [102] that improves learning efficiency, allows training on larger feature maps, and reduce computational complexity by $\mathcal{O}(1000)$ (see Sec. 4.2.1).
Open-vocabulary semantic segmentation modeling lacks a mathematical theory and principled inference method for versatile real-world application.	Present the mathematical theory of latent compositional semantics [95] that proves unconditional open vocabulary semantic segmentation models learns latent embeddings representing sets of semantics (see Sec. 4.3). The proposed sufficient similarity inference [95] overcomes the three fundamental limitations of current open vocabulary inference methods (see Sec. 4.3.4).
Knowledge representation lacks a scalable method for learning and revising a diverse set of category associations from incomplete and noisy data.	Introduces compositional semantic embeddings [95] as a principled and scalable approach to learn compact and semantically diverse object descriptions from uncurated data (see Sec. 4.3.1).

Continued on next page

Table 2.1 – continued from previous page

Problem Summary	**Proposed Solution**
Cognitive psychology and philosophy argue that real-world objects are not perfectly described by a single category, and categories have fuzzy boundaries.	Provides a computational framework [95] to learn natural kinds from incomplete object descriptions, supporting the idea of semantic memories and hierarchical concepts Sec. 4.3.1).

2.2.1 Self-Supervised Visual Representation Learning

Early works experimented with pretext tasks as a substitute for human annotations [105–110]. Recent work demonstrates that image-level embedding classification with cross-entropy minimization on large datasets is a more effective approach capable of surpassing supervised pretraining [111, 112]. Contrastive methods [113–116] learn discriminative latent embedding vectors for images by "pulling together" views of the same image, and "pushing away" embeddings of different images. Recent non-contrastive methods [112, 117, 118] demonstrate approaches to avoid negative sampling to improve computational efficiency. Clustering methods [3, 119–124] simultaneously discovers a set of clusters or prototypes, and learns discriminative image embeddings. Contrary to contastive methods, the objective does not have to be approximated as optimizing over the entire set of negative representative clusters is tractable. DeepCluster [119] iteratively performs K-means clustering over the entire dataset and learns an embedding model and classification head to predict the cluster assignment. SeLA [120] presents a principled formulation for clustering and representation learning as a single optimization objective, by casting cluster assignment as an optimal transport problem [125, 126]. SwAV [3] and ODC [122] demonstrate that clustering can be done online per batch to increase learning efficiency.

2.2.2 Dense Representation Learning

Recent clustering-based methods approach dense representation learning as an instance segmentation problem [127–130] and regional feature correspondence [131–133]. These methods are purposed for pretraining backbones and generally output small feature maps (e.g. 7x7), in contrast to my method. Similarly to my method, VADeR [134] learns dense representations by contrasting pixel-level embeddings in augmented views. My method improves on VADeR by allowing training on larger feature maps (512x512 vs. 56x56 px), more views, optimization without a negative sample memory bank, and

contextual region masking. Self-supervised object detection [135–140] learns expressive embeddings for plausible object proposal regions sampled randomly or heuristically [141]. Masked image modeling (MIM) [142–145] demonstrates strong representation learning capability surpassing contrasting views. However, all these models output low-resolution feature maps. In contrast, my method ViCE generates precise object-fitting semantic partitioning even for high-resolution images.

Visual and semantic coherence is a useful inductive bias to enhance effectiveness of dense representation learning for images. By assuming that visually similar regions represent the same object semantics, redundant computation can be reduced by an $\mathcal{O}(1000)$ [102]. Existing works leverage self-supervised clustering approaches to learn coherent semantic groupings from mutual information [146, 147], geometric equivariance [5], and GAN-based approaches [148, 149]. Other works [150, 151] leverages self-supervised depth map estimation [152, 153] for enhancing semantic segmentation performance. Recently, DINO [112] demonstrated that attention maps for semantic objects naturally emerge for self-supervised Vision Transformer (ViT) models [154, 155]. STEGO [156] presents a method to distill features from DINO and achieve SOTA results. Karlsson et al. [102] introduce superpixelization as a natural hierarchical region decomposition for dense contrastive learning in unsupervised semantic segmentation of high-resolution images. Performing dense representation learning by decomposing images into a small set of visually coherent regions reduces the computational complexity of contrasting cluster assignment [3] by $\mathcal{O}(1000)$ while preserving detail. Experiments show that contrasting over regions instead of pixels improves the effectiveness of contrastive learning methods, extends their applicability to high-resolution images, improves overclustering performance, superpixels are better than grids, and regional masking improves performance.

2.2.3 Unsupervised Semantic Segmentation

Existing works leverage self-supervised clustering approaches to learn coherent semantic groupings from mutual information [146, 147], geometric equivariance [5], and GAN-based approaches [148, 149]. Other works [150, 151] leverages self-supervised depth map estimation [152, 153] for enhancing semantic segmentation performance. Recently, DINO [112] demonstrated that attention maps for semantic objects naturally emerge for self-supervised Vision Transformer (ViT) models [154, 155]. STEGO [156] presents a method to distill features from DINO and achieve SOTA results. My proposed model ViCE [102] improves learning efficiency also on high-resolution images by contrasting cluster assignment over superpixels. The expressiveness of the learned dense embeddings is demonstrated by improving the SOTA unsupervised semantic segmentation benchmark on Cityscapes, and for convolutional models on COCO.

2.2.4 Word embeddings and visual tokens

In natural language processing (NLP), the basic representation of words are categorical tokens or one-hot vectors. Learning semantic embeddings for words using unsupervised methods as a pretraining task [157–159] offers significant improvements for downstream tasks compared with learning word embeddings as part of the task [160]. Word embeddings is the de facto elementary representation used by all recent language models [161–164]. The metric of semantic similarity between words is co-occurrence in sentences [165]. Embedding models are optimized so that the embeddings for two words that often co-occur is close in vector space. A separate set of sampled word embeddings assumed to be unrelated are pushed away similarly to noise contrastive estimation [115, 166] to avoid degenerate solutions [167].

In computer vision, the bag of visual words model [168–170] decompose images into discriminative local image features typically extracted at keypoint locations by a SIFT detector [171]. Later works propose to discover mid-level visual elements or words with richer visual semantics in the form of discriminative patches [172] and mode seeking [173] based on learning through iterative clustering and classification similar to recent self-supervised clustering methods [119] but for representative HOG features [174] in pixel space. More recently, extraction of latent embeddings or tokens for image patches is demonstrated by prior GNN methods [175–177]. The Visual Transformer (VT) [178] adds recurrence to generate visual tokens from current and previous spatial attention maps. However, these methods require a separate transformation matrix to be learned through dense supervised learning task.

2.2.5 Natural language processing

The study of using natural language as an interface for human-machine communication, and how to enable machines to leverage human written knowledge, is called natural language processing (NLP). Natural language is ambiguous and sentence correctness is not perfectly decidable by rules [28]. Language models (LM) [161, 179] instead learn to predict the likelihood $p(\mathcal{X})$ of any sequence of text tokens $\mathcal{X}$ according to a natural language dataset.

Word embeddings [159, 180] substitute non-semantic word tokens by a semantic vector representing the meaning of the word. Word embeddings are discovered from maximizing similarity of embeddings of co-occurring words [181]. Contextual representations [182] extends word embeddings by encoding context from surrounding words. Large language models (LLM) [161, 179, 183] can generate semantic embeddings out of entire sentences.

My approach differs from word and sentence embeddings as I represent a set of semantic embeddings representing an object description by a single compositional semantic embedding.

Clustering [184] and mixture models [185, 186] in NLP discover groups of semantically similar text data. The Latent Dirichlet Allocation (LDA) model [185] parses documents into mixtures of discovered latent topics that allow a finer semantic similarity search. A generative probabilistic mixture model $p(z)$ approximates the distributions of semantic embeddings $z \in \mathcal{Z}$ by K mixture components $p_k(z)$ weighted by the probability π_k that each mixture component is sampled

$$p(z) = \sum_{k=1}^{K} \pi_k \, p_k(z). \tag{2.1}$$

The optimal model is a mixture of Dirac delta distributions $p_k(z) = \delta(z - z^{(k)})$ with number of mixtures K equaling the number of semantics in the distribution $\mathcal{Z}$. As the set of possible semantics in natural languages are unbounded, a common distribution approximation is the Gaussian mixture model (GMM) with $K \ll |\mathcal{Z}|$ components $p_k(z) = \mathcal{N}(\mu_k, \Sigma_k)$ representing the K best semantic clusters. However, this approximation has practical limitations. The required clusters K is generally not known. Optimizing the mixture model $p(z)$ is challenging. Storing the the mixture distribution parameters or all K semantic embeddings μ_k can be inefficient.

My latent compositional semantics approach instead leverage properties of high-dimensional hyperspheres to find an optimal semantic embedding z^* akin to clustering. The vector z^* defines $p(z \in \mathcal{Z})$ by similarity instead of approximating the entire distribution $p(z)$. My approach has mathematical guarantees of optimality, and can represent a large set of semantics by a single embedding while optimizable by gradient descent.

2.2.6 Open-vocabulary semantic segmentation

Open-vocabulary semantic segmentation is a computer vision task that leverages the power of vision-language models (VLMs) [77]. VLMs operate within a unified embedding space, enabling them to bridge the gap between visual and textual information. The core functionality of a global description VLM involves training a visual encoder $Enc_V()$ and a language encoder $Enc_L()$ in tandem. These encoders operate on a paired image x and text description t to generate semantically aligned visual embeddings z_v and textual embeddings z_t within a shared embedding space $\mathcal{Z}$. This alignment allows VLMs to act as an interface for querying visual data using natural language. Cosine

similarity is typically used to measure the semantic similarity between z_v and z_t. Training for these models often utilizes large-scale image-captioning datasets and contrastive learning techniques. While global description models hold promise for various applications including image-text matching, multimodal search, and visual question answering (VQA) [78, 79], their outputs lack spatial grounding within the input image. This limitation hinders their effectiveness in tasks that require precise spatial reasoning, such as navigation, manipulation, and environment mapping [85, 90, 187].

In contrast, dense vision-language models [6, 80, 81, 95, 188–193] produce aligned embedding maps. These embedding maps represent semantic information at the pixel level, allowing for a more precise fit to object boundaries within the image. One approach to achieve densification involves modifying pre-existing global description models. Techniques like removing the final global pooling layer, as employed in MaskCLIP [190], leverage the strong generalization capabilities of these models. While this approach offers the benefit of utilizing pre-trained global description models, the resulting outputs often exhibit significant noise levels. This noise can significantly hinder the practical application of such models in real-world robotics tasks requiring accurate segmentation information.

An alternative approach to achieving dense descriptions leverages pre-trained region proposal (RP) models [194]. These models predict a set of object-masked bounding boxes. Each bounding box is then fed into a pre-trained global VLM [81] to generate a semantic embedding. This embedding is subsequently projected onto all pixels encompassed by the corresponding masked region within the original image. While the object-crop approach demonstrates promising results for object-centric image inputs typical of small, controlled environments like kitchens or indoor spaces [80, 195], it exhibits limitations in handling large-scale and complex scenes. Road environments, for instance, require multi-scale object perception, which this approach struggles to achieve effectively. Furthermore, the computational cost associated with performing individual inference for each object can be significant.

In contrast to the previously discussed approaches, another research direction focuses on training a novel vision model specifically designed for dense feature representation. This model, denoted as $f_\theta()$, leverages an architecture and optimization scheme tailored for this task. One example of such an approach is LERF [196]. LERF integrates language embeddings within a neural radiance field (NeRF) [197], enabling semantic querying of 3D environment representations. This approach offers the potential for querying the environment based on semantic concepts. However, limitations exist. LERF may struggle with extrapolation tasks and potentially requires observing the entire environment before functioning effectively. Open-vocabulary object detectors bridge the gap between

semantic understanding and image regions by localizing predicted vision-language model (VLM) embeddings to bounding boxes [198]. Within the field of open-vocabulary semantic segmentation, two primary categories of models emerge: conditional and unconditional. Conditional models [189, 193, 199, 200] facilitate fine-grained semantic segmentation guided by additional text or image input during the forward pass. However, this approach has limitations in projecting general-purpose, open-set semantics into a broader representation encompassing both spatial and semantic information of the environment. In contrast, unconditional methods [6, 95, 188, 192] focus on predicting general-purpose embedding maps, enabling open-ended semantic querying after projection. Notably, unlike global embedding models [77], unconditional open-vocabulary semantic segmentation models require smaller datasets with dense annotations for training. The theory of latent compositional semantics [95] provides a valuable mathematical framework for understanding the representations learned by these unconditional models. This theory sheds light on the properties, guarantees, and representational capacity of these models.

The open-vocabulary predictive states proposed in this thesis leverages open-vocabulary semantic segmentation to achieve accurate semantic projection as environment representations. This projection is facilitated by the theory of latent compositional semantics [95]. This theory provides valuable insights into the mathematical properties and representational capacity of the modeled semantic embeddings.

2.2.7 Vision-language modeling

Multimodal models that semantically interpret images and text by a unified embedding space are called vision-language models (VLMs). Global description generating VLMs [77] consist of a visual $Enc_V()$ and language encoder $Enc_L()$. Both encoders are co-trained to generate a semantically similar visual and text embedding z_v and z_t for an input image x and text t in an aligned embedding space Z. Alignment enables VLMs to be used as an interface to query or express contents of visual data in natural language. Semantic correspondence between z_v and z_l is measured by cosine similarity. The encoders are typically trained on internet-scale image captioning datasets using contrastive learning. Global description models have many usages like image-text matching, multimodal search, multimodal generative modeling [78], and visual-question-answering [79]. However, outputs are not spatially grounded in the input image and therefore have limitations for tasks requiring precise spatial information such as navigation [85], manipulation [90], and mapping [187].

Dense description VLMs [6, 80, 81, 188–193] generate aligned embeddings for every image pixel for fitting semantics to object boundaries. MaskCLIP [190] aims to leverage the strong generalization power of global description VLMs by removing the global pooling layer. However, the output is considerably noisy and have limited practical usefulness for robotics tasks.

One approach to generate dense descriptions is to use a region proposal (RP) model [194]. The RP model predicts a set of object crops that are interpreted by a global VLM [81]. The resulting global embedding is projected onto all pixels covered by the region. The object crop approach works well for object-centered image inputs typical for indoor robotics environments, but less so for large and complex scenes requiring multi-scale object perception [80, 195]. Computational cost is high due to performing inference for every region separately.

Another direction of work instead trains a new vision model $f_\theta()$ with and architecture and optimization scheme designed for dense feature representation. LERF [196] grounds language embeddings in a neural radiance field (NeRF) [197], allowing querying semantics in 3D environment representations. Open-vocabulary (OV) object detectors localize predicted VL embeddings to bounding boxes [198]. Works related to open-vocabulary semantic segmentation can be categorized into two types. Conditional OV semantic segmentation models [189, 193, 199, 200] allows fine-grained query guided by additional text and/or image input. One drawback is that conditional inference require the original image. Unconditional methods [6, 188, 192] learns to predict general embedding maps such that the likelihood is maximized over the training dataset. Contrary to global embedding models [77], unconditional semantic segmentation models are trained on relatively small densely annotated datasets. The expressiveness of unconditionally predicted embeddings is not yet deeply understood.

This thesis presents an interpretation of unconditional OV semantic segmentation predictions as latent compositional semantic embeddings z^*. I show that the representation z^* combines the compactness of unconditional inference, the expressiveness of conditional inference, and the capacity to represent semantic object descriptions of length K.

2.2.8 Knowledge representation

A general-purpose intelligent agent needs to store information about the world in a practically useful form for reasoning and task completions. This problem is called knowledge representations. An ontology is a framework for organizing and representing knowledge into a hierarchy of categories or concepts. Philosophers and artificial intelligence scientists commonly recognize six types of knowledge [28]: concrete objects including

things and stuff, abstract categories for organizing objects in terms of similarity by shared properties, measurements for ordering of properties, and events, fluents, and time points specify temporally changeable statements.

First-order logic (FOL) [201] and extensions like fuzzy [202] and modal logic [203] traditionally express an object x being a member of a category $Category$ as $Category(x)$ or $x \in Category$. Semantic networks [204] is a subset of FOL designed to represent knowledge as a directed graph of objects and categories. Objects are associated to one or more categories by $MemberOf(\cdot, \cdot)$ relations. Categories are associated to other categories to form a taxonomic hierarchy. The hierarchy of categories allows objects to inherit semantic descriptions from higher-level categories, implying that an object that is a chicken is also a bird (but not the other way around):

$$Chicken(x) \Rightarrow Bird(x). \tag{2.2}$$

Frames [205] extend Semantic networks with inheritable default attribute values like $height = 1$ and properties $CanFly = True$ similar to object-oriented programming.

Semantic networks have several practical limitations. First, semantic vagueness is an inherent aspect of object descriptions as explained in Sec. 2.2.9. Expressing degree of membership is challenging in purely logical representations. Secondly, the problem of inferring correct and diverse category associations from perception is not addressed. Finally, a complete ontology encompassing the entire world does not exist. A scalable method for learning and revising a diverse set of category associations from incomplete and noisy data is needed.

This thesis presents compositional semantic embeddings as a principled and scalable approach to learn compact and semantically diverse object descriptions from uncurated data.

2.2.9 Cognitive Psychology and Philosophy

The proposed idea of latent compositional semantics has strong support in cognitive psychology, neuroimaging, and philosophy. Cognitive scientists believe there exist two types of long-term memory: declarative and nondeclarative [206]. Declarative memory involves the conscious recollection of events and facts, encompassing memories that can be explicitly articulated or recounted, while also including those that elude verbal description. It is also known as explicit memory. Declarative memory is further divided into two primary forms: episodic memory, which concerns personal experiences and specific events in particular places and times, and semantic memory, which encompasses

general knowledge about the world, concepts, and language [206]. Semantic memories are derived from an agent's experiences but is characterized by its abstract and conceptual nature, devoid of ties to any specific encounter [207]. My approach implements the idea of semantic memories into a computational learning framework.

Semantic concepts are organized into hierarchies [206]. These hierarchies comprise three levels: super-ordinate categories, situated at the top (e.g., items of furniture); basic-level categories, positioned in the middle (e.g., chair); and subordinate categories at the bottom (e.g., office chair) [208]. Concepts are rarely processed in isolation; instead, the processing is heavily influenced by the current context and environment [209, 210]. Qualities of observable objects are easier to contemplate [206], underscoring the intricate nature of how perceptual and cognitive systems interact when processing concepts and information about the world. My work shows how machines can learn hierarchical concepts from independent visual observations.

Philosophers argue that real world objects are generally not perfectly described by a single category, as categories themselves are not precisely specifiable. Categories with fuzzy boundaries are called natural kinds. To give an example, finding a perfect logical specification of a platonic "chair" is futile and is bound to results in unintended inferences [211]. Representing real world objects instead in terms of fuzzy semantic descriptions, and determining semantic membership through similarity, has strong support in philosophy. Wittgenstein [212] proposes that members of a category share family resemblance instead of necessary and sufficient characteristics. Lakoff [213] argues for categorization based on prototype similarity and analogies. Schwartz [214] writes that category membership is a matter of degree, meaning similarity to a cluster prototype is a useful measure of membership. In this work, I provide a computational framework to learn natural kinds from incomplete object descriptions.

2.3 Learning Predictive World Models

The problems addressed in this thesis relate to the challenges of learning predictive world models from partial observations. In various domains, such as image inpainting, stochastic prediction models, bird's-eye-view generation, and world models, there is a need to predict complete representations from incomplete or partially observed data. Existing approaches often rely on fully observed ground truth samples for training, assume a fixed set of semantic classes, or lack interpretability. Moreover, existing approaches may not be able to generate diverse plausible predictions, which is essential in real-world scenarios where uncertainty is inherent.

This thesis proposes a framework that addresses these limitations by learning to predict complete world states from partially observed states only [1, 2]. The approach is based on arbitrary conditional density estimation and extends prior missing data VAE approaches to high-dimensional representations. By learning a generative model that can predict diverse plausible completions, the framework provides a more robust and interpretable solution for predictive world modeling. The thesis also bridges the gap between game environments and partially observed real-world mobile robotics environments, opening a new research direction in application of world modeling techniques to real-world problems. See Table 2.2 for problem and proposed solution summaries. See the following sections for additional information and literature grounding.

TABLE 2.2: Predictive world model learning problems and proposed solutions.

Problem Summary	**Proposed Solution**
Stochastic state completion from partial observations.	Extends prior missing data VAE approaches to model high-dimensional representations without requiring fully observed ground truth samples for training (see Sec. 5.4.1).
Stochastic generation of plausible bird's-eye-view representations from perception inputs.	Learns to model $p_\theta(x^*\|x)$ for high-dimensional representations without requiring fully observed ground truth samples for training, and can generate diverse plausible predictions (see Sec. 5.4).
Learning a world model from partial observations.	Presents a predictive world model framework that learns to predict a 2D spatio-semantic representation from agent-centric partial observations, bridging recent SOTA world modeling approaches to partially observed real-world mobile robotics environments (see Sec. 5.4.3).
Spatio-semantic representation for mobile robots.	Learns to predict a 2D spatio-semantic representation from agent-centric partial observations, enabling top-down semantic partial observation as input and plausible explicit world representation as output (see Sec. 6.3 and Sec. 6.4).

2.3.1 Arbitrary conditional density estimation

The problem of arbitrary conditional density estimation[215–217] is about estimating the probability distributions $p(x_u|x_o)$, where the random variables x are expected to be partitioned into arbitrary plausible subsets of observed x_o and unobserved x_u random variables. In this section I present methods incorporating different application-specific presumptions on how x is partitioned into x_o and x_u.

Image inpainting methods predict unobserved pixels x_u from observed pixels x_o. The problem formulation is similar to the problem of predicting complete world states from partially observed states. The prototypical solution is to use an autoencoder (AE) [218] to compress partially observed images x_o into constrained latent codes z encoding similar visual patterns as learned from reconstructing complete images x by matching global contextual clues.

However, optimizing models simply by pixel-wise reconstruction is afflicted by the marginalization problem, resulting in blurry outputs as missing regions can be filled by many plausible pixel configurations. The Context encoder [219] attempts to address the blurriness problem by introducing an adversarial objective. Furthermore, GLCIG [220] introduces a course-to-fine generation scheme with diluted convolutions and two adversarial objectives. The global objective ensures the image remains coherent as a whole, while the local objective improves detail. Yeh et al. [221] find the closest sample in an image database and use its latent code for prediction. Contextual attention [222] adds an attention mechanism for long-distance information crossover. My framework similarly applies an adversarial objective for learning to predict texture-like content such as lidar reflectance intensity from road surface (henceforth, road surface intensity).

Other approaches focus on learning mask-aware convolutional filters. Liu et al. [223] introduces a special convolution filter and a observed element mask update rule for propagating information about which elements provide information. Yu et al. [224] introduces gated convolutions for learned mask updating. While I add an observed element mask to the model input following the missing data VAE approach, explicitly convoluting over masks is an interesting future direction.

Another line of image inpainting works focuses on pluralistic stochastic state completion methods based on generative models. GAN-based methods [225, 226] generate multiple plausible completions by conditioning on a random vector. VAE-based methods [227] replace the deterministic latent code generated by the AE to allow stochastic sampling of multiple plausible predictions. Previous methods improve training stability by constraining the latent distribution of partially observed images by matching the distribution for

fully observed images. PIC-Net [228] trains separate encoders for observable and unobservable image regions and matches the distributions between the two. UCTGAN [229] adds a cross-attention module to mix latent representations of partially and fully observed images. DSI-VQVAE [230] applies VQVAE to stabilize training. Concurrently to my work, Posterior Matching [231] presents arbitrary conditioning based on HVAEs by optimizing a secondary partially observed encoder to match the latent distributions of a fully observed encoder. I extend prior VAE work by introducing a two-stage training paradigm to allow learning to predict complete images from partially observed images only.

Another approach frames predicting unobserved state variables from observed variables as the missing data VAE problem. HI-VAE [232] derives an evidence lower bound (ELBO) for missing data by masking out contributions from unobserved data. EDDI [233] introduces an alternative Partial VAE model which processes observable data only by encoding elements by a positional encoding and processed by permutation invariant operations similar to PointNet [234]. VAEM [235] is a hierarchical VAE that operates on heterogeneous data by first transforming all input variables into a common latent space by a type-specific transformation. HH-VAEM [236] is a recent hierarchical VAE demonstrating effective sampling using the Hamiltonian Monte Carlo algorithm. Collier et al. [237] demonstrate results on high-dimensional image data. My work extends prior missing data VAE approaches by learning to model $p(x_u|x_o)$ for high-dimensional representations without requiring fully observed ground truth samples for training.

Video prediction methods aim to model a stochastic state transition process where a sequence of future images x_u are predicted conditioned on a sequence of past fully observed images x_o. Babaeizadeh [238] presents a sequential stochastic variational video prediction model based on predicting a latent code explaining away the stochasticity of the sequence. Denton [239] presents an end-to-end framework to explain away stochasticity by a frame-to-frame latent code and a learned prior to improve training robustness. The predictive world model framework presented in this thesis reformulates the stochastic latent variable video prediction approach of Denton to the problem of predicting complete world states from partially observed world states only.

2.3.2 Bird's-eye-view generation

Mobile robotics, and in particular AVs, pursue the problem of generating top-down bird's-eye-view (BEV) representations from perception inputs as a substitute or complement to human annotated maps [240].

Camera-based methods receive much attention because of affordability and motivation by human vision. However, lifting 2D perspective images to 3D is fundamentally an ill-posed problem. Inverse perspective mapping (IPM) [241–243] proposes to overcome the problem by assuming the ground plane is flat. However, the flat plane assumption is generally not true. Stereo cameras propose to solve the lifting problem by inferring depth maps based on physics. However, the resulting depth maps tend to be noisy for far-away objects, object borders, and objects covered with non-distinct textures. Learning-based methods are proposed to overcome the weaknesses of stereo-based depth map estimation. Cam2BEV [244] presents an approach that projects semantic features using IPM and corrects the projection by a spatial transformer module learned from synthetic ground truth BEVs. Many works are based on using monocular depth estimation [245–250] to lift images to a 3D point cloud before projection to a top-down 2D grid. Schulter et al. [251] proposes an adversarial objective relying on ground truth maps to refine the resulting BEV representation. MonoLayout [252] learns the view transformation from self-supervised targets by integrating projected observations while still relying on ground truth maps for BEV refinement. Later works introduce probabilistic depth projection [253], categorical depth distribution network [254], and multi-task learning [255]. VED [256] is a variational encoder trained from stereo vision to predict low-dimensional (64x64 px) semantic BEV representations from forward-view monocular images. Other methods lift images using multilayer perceptrons (MLP) trained on ground truth maps [257–259]. Recently, cross-attention based transformer modules [260, 261] and Transformers modules [262] are applied to model view transformations motivated by the global attention mechanism not being limited to processing neighboring pixel information like CNNs. However, due to lacking inductive biases attention-based models tend to require more data, effort, and compute to train as well as for inference. While my framework in principle is compatible with depth estimation, I choose to leverage lidar for substantial improvements in representation accuracy and observation integration performance. Additionally, my generative model can generate diverse plausible predictions, unlike view transformation models which typically are deterministic functions.

Lidar-based BEV generation methods have a significant advantage from explicitly measuring distance though deemed prohibitively expensive for mass deployment by some. Fishing Net [259] utilizes lidar information to improve spatial accuracy of BEVs generated by sensor fusion. MP3 [263] uses a learned module for generating map elements from lidar observations and ground truth map supervision. HDMapNet [264] also includes image information. In contrast to these methods, the predictive world model framework presented in this thesis not rely on preexisting ground truth maps for supervised training. My method is also generative and can provide diverse predictions, which

is fundamentally necessary as the correct prediction for occluded regions are generally indeterminable.

2.3.3 World models

The idea of learning a predictive model of the world in machine learning was first introduced by Schmidhuber [265–267]. A common approach is to learn latent state representations from images using a VAE [268–270], and use the learned latent code as a compact representation of the world state for planning actions. Other works use adversarial learning to optimize the latent code [271, 272], or contrastive learning with latent variables to model stochastic transition processes [273].

Another line of work focuses on inferring a set of object encodings from images. Watters et al. [274] uses a variational encoder to infer a fixed set of latent object encoding vectors from a sequence of images. Later works apply a VAEs to learn semantically richer object embeddings [275, 276]. MONet [277] is a prominent model for learning to extract a variable amount of semantic object encodings using a recurrent attention module. Recent works leveraging MONet demonstrate the merits of explicit object discovery for future state prediction using compositional reasoning [278], and for reinforcement learning [279, 280] surpassing the performance of SOTA model-free RL models [281, 282].

This thesis approache the world modeling problem of learning to predict a 2D spatiosemantic representation from agent-centric partial observations. The method bridges recent SOTA world modeling approaches from game environments to partially observed real-world mobile robotics environments.

2.3.4 Spatio-semantic representations

Mobile robots typically perform planning for spatial tasks by localizing its pose within a map [283]. ICP [284] or SLAM [285] by modern implementations [286, 287] is the conventional approach to map 3D environments by matching sequential point clouds and accumulating them into a common vector space. Semantic SLAM not only estimate the geometry but also the semantics of the environment or an object [288]. The 3D representation can be projected onto a 2D birds-eye-view (BEV) map convenient for navigation tasks [289]. The image-like 2D map representation is suitable for predictive generative modeling [1]. Until recently, semantic mapping approaches were limited to predefined sets of semantic classes, and thus to narrow tasks.

Open-vocabulary spatial representation methods encode maps by VL embeddings instead of class embeddings. The VL embeddings are typically generated by a pretrained

global VLM [290], open-vocabulary object detector [291], or a dense VLM [187, 292, 293]. The open-vocabulary approach in principle allows querying any task-relevant semantics stored in the VL embeddings measuring cosine similarity with a text query embedding. NeRFs implicitly represents 3D objects and environments by a neural network [197, 294, 295] and have recently been extended represent open-vocabulary semantics [296]. Integrating LLMs opens up new possibilities for spatio-semantic reasoning based on a top-down perceptual feedback loop [291, 297] similar to the human vision-for-perception system [298–300].

2.4 Spatial Cognition and Navigation

The existing approaches in spatial cognition and navigation for autonomous vehicles suffer from several limitations. Many path prediction methods rely on a limited dataset of ground truth maps for supervised learning which may not be correspond to particular encountered real-world environments. Additionally, these methods often focus on specific observable actors or agents, neglecting the complexity of navigating in dynamic environments with multiple agents and taking into consideration potential unobserved agents. Lane graph and map prediction methods, on the other hand, are often limited by their dependence on ground truth lane maps, annotated data, or specific road topologies such as highways. Furthermore, end-to-end learning approaches for autonomous vehicles generally lack explainability and modularity, making it challenging to incorporate prior knowledge or adapt to new environments.

This thesis addresses these limitations by proposing a novel approach by introducing a predictive state representation [1, 95]. The usefulness of the predictive state representation is demonstrated by enabling learning to predict all plausible navigational patterns in the environment independently of observed agents and without relying on ground truth maps for supervision [301]. The proposed method facilitates learning explicit agent-agnostic navigational patterns, analogous to the function of an artificial hippocampus. By doing so, the thesis presents a theoretical framework based on grounding commonsense reasoning on the predictive state representations as a alternative approach to end-to-end learning, potentially enabling modularization and incorporation of prior knowledge and adaptation to new environments. The proposed method's ability to predict navigational patterns from partial observations without requiring ground truth lane map annotations makes it a potentially scalable and robust solution for autonomous vehicles. See Table 2.3 for problem and proposed solution summaries. See the following sections for additional information and literature grounding.

TABLE 2.3: Spatial cognition and navigation problems and proposed solutions.

Problem Summary	**Proposed Solution**
Predicting multimodal paths for specific actors.	Learning to predict all plausible navigational patterns in the environment independently of observed agents without depending on ground truth maps for supervision [45, 240, 301] (see Sec. 6.4).
Lane graph and map prediction relying on ground truth annotations.	Predicting lane graphs from partial observations without requiring ground truth lane map annotations using the directional soft lane probability (DSLP) method [301] leveraging predictive environment states [1, 2] (see Sec. 6.4).
End-to-end learning for autonomous vehicles lacking explainability and out-of-distribution robustness.	Present a theoretical general-purpose mobile robotics framework based commonsense reasoning over world model grounded in the experimentally verified open vocabulary predictive environment states [1, 2, 95] and learned navigational patterns [301] (see Sec. 3.4 and Sec. 6.4).

2.4.1 Path prediction

Recent works present methods to predict multimodal paths for specific actors. Salzmann et al. [302] and Baumann et al. [303] trains a convolutional neural network (CNN) on bird's-eye-view (BEV) environment representations to predict a dense map representing valid ego-vehicle paths using a weighted dense classification error and future ego-vehicle trajectories. Barnes et al.[304] trains a CNN on perspective images with self-supervised labels generated from driving data. Ort et al. [305] fuses high-level navigational guidance from a coarse map with path generation reflecting the observed environment. Casas et al. [263] optimizes a model to predict an environment map and possible paths for the ego-agent based on images and point clouds using a ground truth lane map as supervision. Prez-Higueras et al. [306] trains a CNN model to predict a

multimodal path affordance map between any two points to be used as a prior for an RRT* path planner [307]. Kitani et al. [308] trains a Hidden Parameter Markov Decision Process (HiPMDP) model using inverse reinforcement learning and observation data. Ratliff et al. [309] presents an imitation learning approach that maps input features to a cost map based on example paths. My approach expands on prior works by learning to predict all plausible navigational patterns in the environment independently of observed agents without depending on ground truth maps for supervision.

2.4.2 Lane graph and map prediction

Homayounfar et al. [310] trains a recurrent neural network (RNN) model to predict polylines as road lanes in highway road scenes using ground truth lane maps. An extension [311] introduces forking and merging lane topologies. Guo et al. [312] predicts 3D road lanes from perspective images using ground truth annotations. Zürn et al. [313] trains a Graph-RCNN model to predict lane anchors and edges using images and point clouds with ground truth lane map supervision. Can et al. [314] trains a transformer model to detect lane segments from images and subsequently connected into lane graphs. Zhang et al. [315] trains a three-stage network using ground truth map supervision to predict a dense lane map and subsequently predict keypoints used to generate the graph. Mi et al. [316] presents a hierarchical coarse-to-fine approach to train an attention graph neural network to generate road lane graphs. Karlsson et al. [45] presents a self-supervised method to train a directional soft lane affordance (DSLA) map from single trajectories. A follow-up work [240] shows how to generate discrete road lane graphs by searching for connected paths in the DSLA map using the A* algorithm. The directional soft lane probability (DSLP) method [301] is a scalable approach to predict lane graphs from partial observations without requiring ground truth lane map annotations and yet achieve better performance than supervised baselines [313, 314]. DSLP extends [45, 240] by introducing a principled regularizer, a sampling-based maximum likelihood graph generation method, and demonstrates the approach on real-world data.

2.4.3 End-to-end learning for autonomous vehicles.

Originally proposed by Pomerleau [317] and more recently repopularized by Bojarski et al. [33], the end-to-end learning paradigm aims to learn a driving model or policy mapping perception to control by optimizing for an extrinsic goodness objective. Imitation learning approaches [33, 318, 319] learn a policy that results in similar behavior as expert examples. Reinforcement learning (RL) approaches [54] optimize a policy to maximize an extrinsically defined reward such as time-to-human-override. Recently, approaches

learning an explicit predictive world model [320, 321] show that robust policies can be learned from expert observation only. This thesis demonstrates how the proposed predictive state represenatation facilitates learning explicit agent-agnostic navigational patterns analogously to the function of an artificial hippocampus. Learning explicit navigational patterns is an alternative approach to enhance explainability of end-to-end learning, or incorporate an end-to-end learning aspect into the conventional modularized mobile robotics system [322].

2.5 Summary

The first section presents related research involving perception and representation of semantics. The review starts with presenting research for self-supervised representation learning where machines discover generalizable semantics occuring across a dataset. Semantic networks and compositionality are important concepts in cognitive psychology and philosophy for representing knowledge about the world. However, semantic networks face limitations such as uncertainty, semantic vagueness, inferring category associations from perception, and learning from incomplete, noisy data. Research about methods for learning semantics from data are presented, including word embeddings, visual tokens, and natural language processing. Finally, research about learning to infer dense semantic embedding maps are introduced, including open-vocabulary semantic segmentation and recent SOTA vision-language models. This thesis propose latent compositional semantics as a principled representation for forming queryable spatio-semantic memories as a basis for predictive state representations. Additionally, latent compositional semantics is presented as a mathematical model of unconditional open-vocabulary semantic embeddings.

The second section presents literature on learning latent state representations from images using Variational Autoencoders (VAEs) for world modeling and planning actions in reinforcement learning (RL) tasks. Some works infer object encodings from images to extract semantic meaning, while others use adversarial or contrastive learning with latent variables to model stochastic transition processes. Recent researach shows that model-based RL can surpass the perfomance of SOTA model-free RL methods. The thesis approaches the world modeling problem by predicting 2D spatio-semantic representations from agent-centric partial observations, bridging recent SOTA world modeling approaches to partially observed real-world mobile robotics environments.

The third section introduces literature of recent advancements in spatial cognition and navigation for mobile robots like autonomous vehicles. Presented topics include Methods to predict multi-modal paths for specific actors from bird's-eye-view or perspective

images, and inverse reinforcement learning to map input features to cost maps based on example paths. Lane graph and map prediction techniques for predicting navigational patterns or road lanes are presented. End-to-end navigation learning methods aim to learn a driving model or policy mapping perception to control by optimizing for an extrinsic goodness objective. Other approaches include imitation learning and RL methods. Approaches with explicit predictive world models show robust policies can be learned from expert observations. This thesis presents how to leverage predictive environment state represention to learn navigational patterns from observation only. The proposed approach is based on parallels with the biological hippocampus and facilitates interpretable end-to-end learning from observation and is compatible with conventional modularized mobile robotics systems.

Chapter 3

General-Purpose Mobile Reasoning Agents

3.1 Introduction

This chapter introduces the concept of general-purpose mobile reasoning agents as proposed in the thesis. The chapter starts by introducing biological general intelligence agents and their characteristic faculties, as a proof of concept that creating adaptable and generally intelligent agents are possible. Next, the artificial counterpart of biological agents are introduce in general terms. A proposal for what faculties general-purpose artificial intelligence agents require are given. The following section gives an overview of the proposed general-purpose predictive agents, including open-vocabulary state representation, open-vocabulary predictive world models, and state transition modeling. The chapter concludes by deliberating on learning to navigate constrained environments, means of providing task descriptions for general-purpose agents, as well as the potential of realizing high-level planning incorporating world knowledge by mental simulation.

3.2 Faculties of Biological General Intelligence Agents

Naturalism is a philosophical perspective that supposes the universe is purely materialistic and governed by natural laws and forces [323]. Naturalism rejects the existence of supernatural or spiritual entities like non-material human souls. In the naturalist perspective, the existence of intelligent biological humans is therefore proof that creating generally intelligent artificial agents is theoretically possible. Analyzing the difference

between natural biological intelligence and current artificial machine intelligence may reveal insights into what missing constituents are necessary to achieve artificial generally intelligent systems [93].

Human intelligence is a complex and multifaceted phenomenon that has long been a subject of study in various scientific disciplines, including philosophy [324, 325], psychology [326], neuroscience [327, 328], and cognitive science [93, 329]. According to the predictive coding theory [330], one of the key features of human intelligence lies in the extensive and versatile model of the world, which allows humans to establish useful causal models and make sense of an inherently uncertain environment [331]. This ability is thought to be supported by a range of cognitive processes, including sequence learning and prediction, formation of memory-based predictions, processing of prediction errors, and integration of multimodal sensor information [94].

At the heart of these cognitive abilities lies the hippocampus [332], a brain structure that plays a crucial role in several aspects of predictive coding. The hippocampus is known to be essential for learning and representing sequences of events or experiences, and the formation and retrieval of episodic and spatial memories, which can be used to generate predictions about sensory inputs based on past experiences in similar contexts or environments [94, 330]. Furthermore, some theories suggest that the hippocampus is involved in detecting and processing prediction errors, which are discrepancies between predictions and actual sensory inputs. These prediction errors are believed to be crucial for updating internal predictive world models and the general mechanism for learning [333].

However, it is important to note that human intelligence cannot be reduced solely to the hippocampus or any other single brain region. Instead, it emerges from a complex interplay of multiple brain regions, including the neocortex, prefrontal cortex, and other subcortical structures [328]. The primary advantage of human intelligence over that of other animals lies in our ability to harness abstract concepts, analogies, metaphors, and stories to express our experiences and make sense of new information [334]. This is thought to be supported by the hippocampus's encoding of semantic knowledge in ways that transcend the physical properties of the referenced item [93].

Having established the basis of human intelligence, the question about what is general and non-general intelligence is still not a clear cut case. The concept of general intelligence has been a topic of interest in both psychology and computer science [93, 326, 335, 336], with different definitions and measurements being used to understand its nature. The notion of general intelligence agents implies systems that can accomplish a variety of complex and novel tasks in complex and novel environments, possibly including improvement by learning [337]. However, the definition of what is

a complex task and a simple task is also not clear cut. Measuring and optimizing performance for particular task losses the task's meaningfulness as general intelligence indicator [93, 338].

In psychology, traditional definitions of intelligence have been passed down from Western philosophy drawing on standardized tests of abstract reasoning and pattern discovery [338]. However, it is commonly considered that these definitions may be influenced by social constructs and biases [339]. The measurement of intelligence in humans faces the same dilemma as in AI research, where the nature and structure of mental ability depend on how it is measured.

In computer science, researchers have been building AI systems that can perform specific tasks with varying levels of success. However, there is debate about whether simple algorithmic solutions will suffice for building AGI, or if there is a special ingredient or combination of factors unique to natural general intelligence. A general intelligence is likely more than the sum of many narrow intelligences [337, 340, 341].

This thesis takes a pragmatic approach to general intelligence and emphasize on particular faculties instead of degree of capability. Perhaps the most decisive faculty is the ability to learn and apply pre-existing knowledge in service of solving novel tasks [28, 336]. Another major faculty is the ability to communicate and store information by highly or infinitely expressive language like human natural language [334]. Grounding a general intelligence in human natural language provides two benefits: the possibility to bootstrap learning from preexisting written human knowledge, and align the intelligence with human values through the means of direct communication of abstract thought.

3.3 Artificial Intelligence Agents

This section presents a brief historical account of the scientific pursuit of artificial intelligence (AI). The remainder of this section provides an analysis of how the conceptual AI agent framework presented in Sec. 3.3 can be adapted to emulate key faculties of biological general intelligence.

The field of Artificial Intelligence (AI) has undergone significant developments since its inception. One of the early pioneers Aristotle discussed the concept of mechanisation of logical reasoning in his works, suggesting an algorithm for decision-making that involves deliberation about means rather than ends and that the conclusion resulting from two premises is an action [342]. Alan Turing made significant contributions to modern AI. His seminal paper introduced concepts like the Turing test, machine learning, genetic algorithms, and reinforcement learning [343]. Turing also proposed it might be easier

to create human-level AI by learning algorithms rather than explicitly programming knowledge and intelligence in machines. The work of Herbert Simon and others in the late 1960s proposed that finding approximate decisions that are "good enough" better represent actual human behavior compared with finding optimal solutions as in theorem proving [344, 345]. The importance of domain knowledge was emphasized in the development of expert systems, leading to the first commercialization of AI in the 1980s [28]. Expert systems are based on encoding human knowledge into a system designed to make intelligent decisions [22, 23, 346]. However, the scope of human knowledge that can be manually encoded is severely limited, which in turn limit the practical versatility and applicability of expert systems [347, 348]. Recently, AI has undergone a reunification of subfields such as computer vision, robotics, speech recognition, multiagent systems, and natural language processing due to renewed interest in learning from data by statistical modeling, optimization, and machine learning [28, 349]. The shift towards big data and data-driven learning approaches has led to significant advancements in the field since the 2000s [30].

Unlike biological brains, which are characterized by their complexity and versatility, conventional AI systems are typically based on relatively simple architectures that rely on statistical pattern recognition and optimization methods [28]. While some modern AI systems have achieved impressive performance in tasks such as image classification [20], natural language processing [70], game playing [21], and scientific discovery [350], they still fall short of human-level intelligence in many respects. One key limitation of AI systems up until recently is their lack of generalizability and adaptability to new situations or contexts. Unlike humans, who can apply abstract concepts and knowledge to a wide range of problems and domains, most AI models are highly specialized and require large amounts of labeled data for training. This limits their ability to transfer learning from one task to another and makes them vulnerable to overfitting and catastrophic forgetting [93]. Another limitation is the lack of common sense reasoning and background knowledge that humans naturally possess. While AI systems can be trained on vast amounts of data, they still struggle with tasks that require understanding of context, causality, or commonsense inferences [36]. For example, a recent study showed that a modern language models such as GPT-3 [70] could generate convincing text based on given prompts but failed to answer basic questions about the world [35]. This thesis presents an AI agent framework that is capable of commonsense reasoning and planning by mental simulation based on a predictive environment state with semantics grounding world knowledge in the perceived environment analogously to the biological hippocampus.

The concept of an intelligent agent is central to the field of artificial intelligence (AI). An agent is defined as a program that can perceive its environment, create a useful

state representation, and perform actions, with the goal of maximizing a performance measure. Here "performance" refers to how well the agent achieves its goals, which may include maximizing utility or minimizing costs. [28] The driving challenge to realize useful artificial intelligence agent is to create programs that produce rational behavior for completing tasks in a manner aligned with the instruction and implicit human values. A formal definition of agents is provided in Sec. 1.1 as a particular implementation of a perception function (1.1), a state representation function (1.2), and behavior function (1.3). This section reflects on correspondence between the agent components (1.1-1.3) and faculties of biological general intelligence as explained in Sec. 3.2.

Agent perception functions (1.1) implemented by computer vision systems based on hierarchical feature representation learning have strong correspondence with the biological visual cortex system [351]. Recent vision-languange models (VLMs) have provided a means to implement language grounded bottom-up [81] and top-down [79, 193] visual processing akin to the biological vision system. The human visual system comprises two subsystems [352–354]. The vision-for-perception system located in the ventral stream processes information in a slow, top-down manner to create perceptual representations from ambiguous or incomplete visual input by leveraging visual and semantic memory [353]. These representations support conscious mental processes such as recognition, visual thought, and planning. The vision-for-action system located in the dorsal stream processes information in a real-time, bottom-up manner to perceive the entire environment and infer behaviorally-relevant visual affordances, including cues for spatial navigation [355]. This thesis presents the theory of latent compositional semantics as an agent perception framework for learning to infer and represent visual percepts by diverse sets of semantic attributes as required for general-purpose agent systems including humans.

Agent state representation functions (1.2) based on explicit representations of space, like latent feature maps and grid maps, are believed to have correspondence with the biological hippocampus [332, 356]. However, unlike conventional task-specific grid maps used in robotics to represent free space or specific semantics like "road", the hippocampus stores spatially grounded general-purpose semantic representations allowing querying of any task-related semantic known to the human agent. Real-world agents can generally only perceive a partial glimpse of the true environment state through its sensors. Partially observed environment states naturally contain a high degree of ambiguity, which in turn complicates learning optimal decision making [357]. In contrast, the hippocampus provides predictive state representations based on predictive coding and observational experience [94, 330, 331] and improve learning optimal decision making by disambiguating environment states. This thesis proposes a framework for an artificial hippocampus

based on the theory of latent compositional semantics as general-purpose semantic representation. Additionally, the thesis propose a predictive world model to disambiguate partially observed environment states with a potential to robustify state-transition modeling akin to an artificial hippocampus.

Agent behavior functions (1.3) generally involve frameworks for planning and control. AI agents tasks with complex goal-driven problems like games [21] and robotics [358] generally depend on a planning step based on traditional search algorithms or model-based reasoning. Search algorithms are based on modeling the abstract problem into a known current state, a deterministic state-transition model, and a goal state. Algorithms like A* [359], breadth- or depth-first search [360], and Monte Carlo Tree Search [361] explores the state space to find a path or sequence of actions from the current state to the goal state. Model-based reasoning algorithms enhances the complexity of the state-transition model by incorporating stochasticity and predictions over partially observed states. Techniques like Markov Decision Processes (MDPs) [362], Partially Observable MDPs (POMDPs) [363], and hidden Markov models (HMMs) [364] efficiently searches for an optimal sequence of actions while taking into consideration uncertainty by efficient algorithms like the Viterbi algorithm [365]. Current theories propose that behavior of biological intelligent agents emerge from higher-order cognitive processes in the prefrontal cortex called denoted executive functions [366]. Mental simulations are critical for predicting optimal actions for goal-driven behavior [63, 64]. Particular faculties include episodic future thinking for imagining specific future events, counterfactual reasoning for considering "what-if" scenarios, and prospection for grounding cognition in hypothetical future states [367–369]. AI agents transform high-level actions into low-level motor and actuator command signals by control methods. Common techniques include inverse kinematics [370] to find a set of joint angles for a target manipulator pose. Trajectory optimization methods [371] output control signals for smooth and efficient movement. Feedback control systems [372] ensures actions are performed as intended by adjusting control signals during execution. Biological intelligence utilize an equivalent architecture, with the motor cortex executes sequences of voluntary movements as planned by the premotor cortex and supplementary motor area [373, 374]. Muscle fibre contractions are controlled by signals originating from motor neurons and passing through neuromuscular junctions [375]. This thesis presents a predictive state representation simultaneously providing a compact latent state encoding and a spatially grounded semantically rich open vocabulary representation. Predictive state representation is analogous to an artificial hippocampus as explained in Chap. 5. The thesis provide theoretical justification for why the hippocampus-like predictive state representation can facilitate higher-level commonsense reasoning and abstract mental simulation [63, 64, 367]. The proposed

predictive state representation is backed by rigorous experimental evidence. The commonsense reasoning and abstract mental simulation based on the proposed predictive state representation is theoretically justified in Chap.4- 5. However, a proof of concept implementation and experimental verification of the full general-purpose mobile reasoning agent is out of scope of for this thesis.

The following sections propose fundamental cognition components needed to endow future general-purpose mobile robot reasoning agents with faculties in line with those of general-purpose biological intelligence agents like humans.

3.4 General-Purpose Predictive Agents

Throughout this thesis, several key requirements for general-purpose mobile agents capable of commonsense reasoning are identified and addressed. Requirements include the ability to perform tasks defined by weakly specified goals [90], interpret the world using a priori unknown task-relevant semantics [81], represent the environment by a spatio-semantic memories supporting querying of spatially grounded task-relevant semantics [95], predict plausible states by a predictive world model continually optimized by surprise minimization over observational experience [333], and implement problem solving by state-transition modeling leveraging reasoning over commonsense knowledge [84].

This section expands on the required cognitive faculties listed in Sec. 3.3 with concrete and implementable components. Section 3.4.2 *General-Purpose Agents* deliberates on the components enabling a commonsense reasoning AI agent to complete tasks without being explicitly programmed for the task, or depending on excessive task-specific demonstration data. Section 3.4.3 *Open-Vocabulary Predictive State Representation* presents how open vocabulary predictive state representations enable AI agents to represent the environment by grounded semantics satisfying the requirements of an embodied general-purpose reasoning AI agent. Section 3.4.4 *Predictive Agents* explains the evolution from reactive to predictive agents capable of predicting the outcome of a sequence of actions, as well as evaluate the goodness or utility of the predicted outcomes. The remainder of this section provides a holistic overview of the proposed structure of general-purpose mobile reasoning agents. The presentation includes references to similarities with general biological intelligence as presented in Sec. 3.2-3.3.

See Fig. 1.5 for a diagram delineating what components of the proposed agent framework is experimentally verified or remain a theoretical proposal supported by literature evidence.

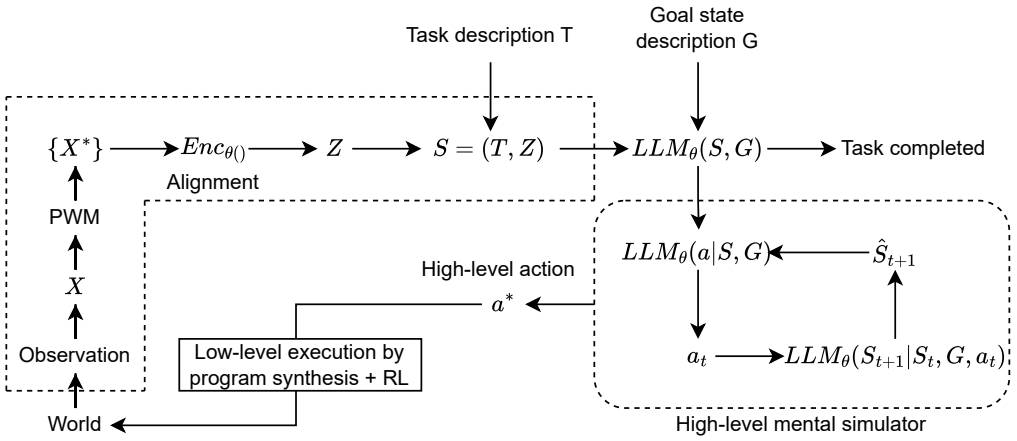

FIGURE 3.1: Conceptual overview of the proposed general-purpose predictive agent framework. The agent perceives the world through sensor observations. The observations are accumulated into spatio-semantic memories. A predictive world model generates a set of diverse plausible worlds as explicit and compact latent open-vocabulary states. On theoretically derived grounds supported by experimental evidence in the literature, the states can be furthermore be interpreted by a multimodal LLM. Rational actions are inferred by reasoning over future outcomes. High-level actions are transformed into low-level actuations by program synthesis and internal simulation. Autonomous goal state detection allows the agent to know when a task is completed.

3.4.1 Agent Framework Overview

The presentation of the proposed general-purpose mobile reasoning agent framework structure is based on the visualization in Fig. 3.1. The presentation is split into two parts: the first part explains how the environment is transformed into predictive environment states natively readable by multimodal large language models (LLMs). The second part explains how the world knowledge and contextual information understanding is leveraged to do commonsense reasoning and high-level mental simulation over sequences of abstract actions. First, the external world or environment is perceived by camera and depth sensors. The resulting precepts are RGB images and point clouds with known correspondence by projection calibration parameters [1]. The RGB image percepts are processed into observations by a vision-language model (VLM) into dense latent compositional semantics embedding maps [95] as explained in Chap. 4. Point clouds are used to ground semantic information in the embedding maps to a 3D vector space. Sequential observations are integrated into a common 3D vector space by a simultaneous localization and mapping (SLAM) framework [287, 376]. The integrated observations are denoted partially observed world states X [1] due to being limited to only the perceived environment region. A predictive world model (PWM) maps the partially observed world state X into a latent distribution, from which a set of diverse

plausible complete world states $\{X^*\}$ is sampled. Sampling complete states reduce ambiguity and thus enhance the accuracy of predicted states while supporting predictive diversity through sampling [1]. Predicted states X^* is aligned with the language embedding space of a multimodal large language model (LLM) by an encoder $Enc_\theta()$. The encoder is optimized by a novel method based on latent compositional semantics [95] and computational geometry bridging the spatio-semantic and textual representation of semantic objects with geometric extent. The optimization objective theoretically ensures that all geometric and semantic detail of the explicit state representation X^* is retained in the aligned representation Z consisting of K latent state embedding token vectors. See App. C for further details regarding the optimization method.

The latent state embedding token vectors Z and natural language task description T represent the goal-driven predictive world state S aligned with the multimodal LLM's language embedding space hypothetically with negligible information loss. A LLM-based task completion detection module can infer if the current state S matches the goal state defined by a natural language goal state description G. A LLM-based high-level mental simulator takes the state S and goal description G to search for the optimal high-level action a^* to take based on traversing a tree of many potential high-level action sequences. The mental simulator can leverage world knowledge and commonsense reasoning to predict consequences of actions. Leveraging world knowledge and commonsense reasoning allows the agent to avoid detrimental actions without requiring negative experience through exploration. The optimal high-level action a^* is given to a low-level action execution function generating low-level action or control signals for realizing the abstract action a^*. The hypothetical function is based on program synthesis leveraging simulation and reinforcement learning with LLM generated reward functions [377, 378]. The resulting low-level actions are actuated and cause a world state transition getting the agent closer to reaching the goal state.

The proposed general-purpose agent framework in Fig. 3.1 have similarities with biological general intelligence in all parts of the system. The sensor to predictive environment state information processing pipeline is analogous to how the biological visual cortex processes retinal signals into spatio-semantic predictive representations in the hippocampal cortex [94, 330, 332, 351, 379]. The subsequent commonsense reasoning and mental simulation processing pipeline correspond to the higher-order executive functions [366] and mental simulation [64] faculties of the biological brain.

As mentioned in Sec. 1.5 and Fig. 3.1, empirical evidence for the open vocabulary predictive state representation and related predictive world model is provided in Chap. 4-5. The following sections explain how the proposed predictive state representation satisfies

the theoretically deduced requirements for a rich spatio-semantic state presentation supporting grounded commonsense reasoning. The remaining components in Fig. 3.1 are theoretically deduced with empirically verified properties in their isolation as explained throught this chapter. As explained in Sec. 1.5 and Chap. 7, the practical demonstration of integrating all presented components into a general-purpose mobile agent is considered out of scope for this thesis but is pursued as future work.

3.4.2 General-Purpose Agents

This section deliberates on core faculties required by a general-purpose agent, and how this thesis propose to implement the identified faculties. Cognitive scientists believe that adaptable intelligent agents like humans represent the world using a small set of foundational cognitive faculties for perceiving inanimate objects, external agents, numeric concepts, social relations, and spatial environments [99]. These cognitive abilities allow intelligent agents to perform commonsense physical reasoning and imagine counterfactual scenarios to facilitate task accomplishment [380]. A key capability is predictive world modeling [265–267, 381]. See Sec. 1.4 for a definition of general-purpose in contrast to conventional specialized robot designed for a specific or a priori known narrow range of tasks.

Weakly specified goal understanding. General-purpose mobile robots promise machines capable of safely completing tasks in novel environments without relying on exact human programmed instructions [90]. Performing tasks defined by weakly specified goals is crucial for general-purpose mobile agents, as it enables them to adapt to a wide range of scenarios without being limited by predefined instructions. This contrasts, conventional agents behavior functions defined by highly exact programming languages lack the ability to express ambiguous instruction relying on commonsense, as so often is required in complex real-world tasks. Behavior functions defined by learning-based methods likewise lack ability to perform ambiguously defined tasks without large amount of example data and thus severely limit practical usefulness. See Sec. 1.2.1 and Sec. 1.2.2 for further details about the limitations of rule-based and learning-based agent approaches. A promising approach to realize general-purpose robots is to leverage large language models (LLMs) [85, 90, 92, 382–387] trained on internet scale information about the world. LLMs as contextual information integrator functions enable previously infeasible capability to synthesize a natural language task instructions, extensive world knowledge, and commonsense reasoning, to produce an output according to the instruction [70, 79, 388]. LLMs are thus a core enabling technology for general-purpose agents due to comprehending and completing instructions specified in fluid natural language. Furthermore, multimodal LLMs allows communicating information to humans by natural language,

potentially providing explicit causal decision factors [89] and intent of future actions or predictions [75]. The explicit and specific nature of language-based feedback overcomes limitations of implicit feedback in the form of attention masks [389–391] and other ad-hoc explanability methods requiring human interpretation [392–394].

Open world semantics. Another required faculty is the ability to interpret the world in terms of a priori unknown task-relevant semantics. To complete a novel task in a new environment, a general-purpose mobile robot need to comprehended the environment by an a priori unknown set of semantics. However, learning useful semantic representations from correlation patterns in raw observational data by self-supervised learning [3, 112] by itself does not result in semantic representation grounded in the innate world knowledge and contextual information integration capabilities of LLMs [70, 79]. Open vocabulary Vision-language models (VLMs) [6, 77, 80, 81, 95, 188–193, 297, 395] align self-supervised semantic feature extraction models [144, 396] with natural language semantics. Interpreting percepts by VLMs and transforming the resulting open vocabulary semantic embeddings into the LLM embedding space and connects the internal reasoning processes of LLM agents with the external world. This thesis presents latent compositional semantics as a mathematical theory of how open vocabulary semantics are represented. A new sufficient similarity semantic inference method is proposed based on the mathematical theory. The new inference method overcomes principal limitations of the conventional most similar semantic inference method. See Chap. 4 for further details.

Spatio-semantic memory. However, spatio-semantic reasoning tasks may require information beyond what is currently observed. Efficiently fetching an item out of view require a spatio-semantic memory of where the item is located [91]. Inferring navigational patterns like road lanes may require predictive assumptions for the unobserved environments behind obstructions [301]. A spatio-semantic memory [397] or scene representations [398] allows an agent to query semantics from observational memory [290, 292, 293], to navigate [187], and plan by reasoning [92]. Common representations for spatio-semantic include 3D reconstruction [399], object-centric, topological maps [400], scene graphs [401], and top-down metric grid maps [381]. This thesis propose open vocabulary predictive state representations as a principled means to represent spatio-semantic memories for general-purpose mobile reasoning agents.

Spatially grounded visual reasoning. All required task semantics may not correspond to visual object semantics. Inferring what visual object semantics to query to complete a task is part of the visual reasoning problem [86, 91, 383, 402, 403]. For example, to complete the task “fetch a cold beverage high in vitamin C”, the general-purpose

mobile agent may need to reason beyond its explicit spatio-semantic memory about object location, and reason that cold beverages are generally located in the fridge, and reason about which of the available beverages have the highest vitamin C content [297]. If the general-purpose agent is tasked to navigate a cluttered parking lot to find a specific car model. The agent must be able to reason about semantics relevant to safe navigation including other vehicles, pedestrians, signs, and implicit rules of cooperative driving in rule-constrained environment [75, 404].

Commonsense knowledge. General-purpose agents need to be capable of commonsense reasoning due to handle the vastness and complexity of real-world environments and task descriptions. Commonsense reasoning aids in resolving ambiguities that arise from lack of explicit information, contradicting data points, and contextual integration of ambiguous task descriptions and world knowledge. Commonsense knowledge enhance the agent system's real-world scalability by enabling more efficient utilization experience by learning with knowledge [28]. Concrete advantages include adaptation to new situations by generalizing from limited observation or example data by efficient few-shot learning [405, 406]. Learning from observational experience in the real world is feasible if leveraging reasoning over commonsense knowledge to avoid negative outcome experience during exploration [73, 75, 84, 89, 386]. Commonsense reasoning can serve as a good heuristic for discovering causal mechanisms by leveraging high-level semantic knowledge [70, 348, 407–409]. Finally, commonsense knowledge is strongly related with spatially grounded visual reasoning as the primary means to infer what semantics to infer as well as facilitating the reasoning process itself [410–412].

The proposed agent framework is primarily optimized by surprise minimization similar to the predictive coding hypothesis in the biological brain. Learning via surprise minimization is hypothesized to be the fundamental principle of biological intelligence for continuously evolve and improve based on existential experience [331]. The continual learning capability of the biological brain is replicated in the proposed general-purpose mobile agent. The latent compositional semantic vision encoder model is primarily optimized by self-supervised learning (SSL) objective [144]. The predictive world model and alignment encoder is entirely optimized on observational experience by a SSL objective [2]. LLMs are primarily optimized by next token prediction [413] and can leverage self-reflection for improving predictive performance with observational experience [414–416]. Continual learning enables general-purpose agents to maintain a high-level of competence and adapt to various domains while continuously improving their capabilities over time.

3.4.3 Open-Vocabulary Predictive State Representation

This section deliberates on state representations, and how open-vocabulary predictive state representations relate to general-purpose predictive agents is deliberated on in this section. A state representation is a data structure enabling agents to store relevant information and represent the external environment as perceived by the agent [417]. The state may also contain information about the robots internal state. The state representation must be computationally implementable, efficient, and support inference to be useful for an agent. A good state representation should be expressive yet compact and ideally be both machine and human readable in order to improve human-machine communication.

Sensors for state estimation. The philosophical idea underlying state representations is that there exists an objective and predictable external world that can be modeled [102, 325]. Environment states for real-world mobile robot agents fundamentally cannot be perfectly known, but only estimated to a certain degree of accuracy by sensor observations and world modeling [265]. Conventional sensors for mobile robots are primarily composed of two types of physical light-sensing mechanisms with complementary strengths and weaknesses. First, active sensing lidar (light detection and ranging) sensors accurately represent metric space using point clouds. Secondly, passive sensing cameras captures rich semantic information about the perceived environment. Sensor fusion approaches aim at leveraging the complementary strengths of both vision modalities [259]. Image content can be projected to a 3D point cloud if the pixel-wise depth and camera calibration parameters are known [418]. In principle, monocular [248, 419–421] or stereo vision [422–425] can provide depth maps and enable a vision-only perception configuration equivalent to biological vision systems [351]. Observations are semantically interpreted by perception or computer vision models. Significant progress has been made in recent decades in terms of increased semantic expressiveness and generality due to exponential increase in computing power and data [20, 30]. Semantic point clouds are the natural data structure for representing both spatial and semantic information. Other modalities have been studied in robotics, such as auditory (hearing) and tactile (touch) perception. Auditory perception has seen significant research in speech recognition, music perception, machine learning of music, and general sounds [426, 427]. Tactile perception is important in robotics with manipulators. Automated perception of smell has seen less research but deep learning models have been shown to predict smells based on molecular structure [428]. This thesis does not however consider sensing modalities beyond vision and lidar observations.

Requirements for general-purpose mobile robots. The spatio-semantic environment representation for general-purpose mobile robots needs to satisfy three properties: First, the representation needs to encode rich open-set semantic object descriptions [206, 209, 210]. For narrow problems like object avoidance in constrained environments, it may suffice to detect and represent an object by one of a fixed set of classes like *table*. A general-purpose agent [92] however, requires a richer compositional representation of the object including alternative names like *desk*, properties like *rigid*, and affordances like *flat surface*, all of which cannot be manually annotated during the system development phase. Secondly, the representation needs to support querying of overlapping semantics, such as a *dog* also being an *animal*. Overlapping semantics must be learnable from independent observations or datasets without relying on human customization effort limiting scaleability [429]. Third, the representation must be efficient in terms of storage. Spatio-temporal accumulations of raw observations rapidly grow into an unreasonable amount of data [398]. To keep the environment representation compact, observations need to be abstracted into declarative semantic memories [206–208]. An additional practically beneficial property is explicit environmental representation. Explicit representations can communicate to humans robots' environmental understanding, intended plan of actions, along with interpretable factors for decision making. Explicit representations also allow humans to provide precise spatially grounded instructions to robots [85]. This thesis introduces semantically rich metric state representations by open-vocabulary semantic embeddings grounded in a metric spatial representation. The proposed state representation functions as rich spatio-semantic memory for general-purpose mobile robots. The theory of latent compositional semantics [95] provides a theoretical background in the form of a mathematical model for the semantic queryability and representational capacity of the embeddings making out the state representations.

Symbolic and metric state representations. State representations can be categorized into symbolic and metric data structures. Symbolic state representations representations, or knowledge bases, use symbolic, logical, or relational descriptions or formal languages to represent the state of the robot and its environment [28]. Representational structures include propositional logic, first-order logic, and semantic networks. Constructing knowledge-based agents involves a declarative approach where agents work by combining assertions of sentences in the knowledge base with logical inference [430]. The fundamental concepts of logic are independent of any specific form, enabling representation languages to specify syntax rules for forming sentences. Proposition logic is a simple example of this concept, where the truth values of symbols determine the truthfulness of sentences with respect to models in the real world [431]. However, proposition logic has limitations and cannot express some statements such as those involving quantifiers like *all* ($\forall$) or *some* ($\exists$). More complex languages like first-order and temporal logic enhances

the generality of expressible states, such as quantifiers, though bring their own challenges in terms of computational complexity and practical implementability [432, 433]. Symbolic state representation have proven to be a useful and human interpretable tool for building reliable intelligent agents capable of reasoning about the world around them and take intended actions according to their current state or knowledge base. However, Symbolic states principally do not support representing uncertainty, handling noise, and representing continuous valued information or degree of truth [24, 96, 434]. The inherent weakness to imperfect data limits learning symbolic knowledge from raw data without sophisticated contextual information processing and world knowledge.

Metric state representations leverage numerical or continuous numerical values to represent the robot's state and environment in terms of "degree" of some semantic [417]. Data structure examples include feature vectors or semantic embeddings, occupancy grids, and topological maps. Metric representations can capture spatially detailed information about the environment, but they may require significant processing power and memory resources. The amount of semantics expressible by homogeneous state representations like conventional grid maps is limited as much memory and computation is required as one metric represent ion is required to specify the degree of each semantic. The state representation approach presented in this thesis is based on accurate metric environment information [2]. Lidar measurements are significantly less noisy and accurate than vision-based depth estimation and thus the preferable spatial sensing modality for the proposed approach. Furthermore, lidar sensor devices are becoming increasingly affordable thanks widespread adoption resulting in sustainable mass production, leading into a virtuous cycle.

Spatio-semantic memory. Spatio-semantic representation of the environment is essential for general-purpose mobile agents [85, 187, 291]. By employing spatio-semantic memory as state representations, an agent can query spatially grounded rich semantics by aligned multi-modal vision-language models (VLMs) [6, 77, 80, 81, 188–193, 395] compatible with human communicated instructions and support reasoning over commonsense world knowledge incorporated in large language models (LLMs) [92]. Spatial grounding of VL embeddings in 3D can be done by projecting 2D dense VL embedding maps to point clouds [290, 292, 293] or neural radiance fields (NeRF) [296]. The spatio-semantic environment state representation for general-purpose agents thus requires the following properties: encode an open-ended vocabulary of semantic concepts, represent and allow querying of overlapping semantics (e.g. a *couch* is also a *furniture*), and store observations compactly. Learning open-set semantic concepts as embeddings z existing in a common semantic embedding space $\mathcal{Z}$ instead of K predefined classes as unit vectors $\hat{e}_k$ is a more scalable approach to increasingly understand the world with sufficient sophistication to complete a wide range of tasks and environments, and adapt quicker

to new unfamiliar situation. This thesis propose grounding latent compositional semantics [95] using lidar observations into a common 3D vector space. See Chap. 4-6 for further details.

Predictive state representations. Completing tasks generally requires information beyond what is currently observed. A spatio-semantic cognitive memory [397], or semantic scene representations [398], enables a mobile robot to query semantic information about prior observations [290, 292, 293], to navigate [187], and do planning by language-based reasoning [92]. Common spatio-semantic representations for mobile robots are 3D reconstruction [399], object-centric, topological maps [400], scene graphs [401], and top-down metric grid maps [1, 2]. Predictive state representations go beyond representing past and current observations of the environment and agent state [2]. Predictive states leverage a world model [1, 265] that allows predicting relevant information about the environment and future agent states based on current observations and actions taken. The world model is central to predictive states can be learned directly from data by maximizing the likelihood of predictions given observational experience of the world. World models effectively piece together what is unseen from what is and has been seen, and can predict outcomes of actions before they are taken. Examples include Predictive State Representations (PSRs) for controlled dynamical systems [435] and Observable Operator Models (OOMs) generalizing hidden Markov models (HMMs) [436]. This thesis introduces semantically rich predictive state representations of environments based on an open-vocabulary world model [2]. Together with the proposed world model, the predictive state representation functions as an artificial hippocampus for general-purpose mobile robots as explained in Chap. 5. The core advantage of the proposed open-vocabulary predictive state and world model approach is the dual explicit and latent state representation. The explicit representation $x \in \mathbb{R}^{H \times W \times D}$ encodes the 2D environment represented by a grid map of height H and width W, by grounded rich semantics of dimension D in a human interpretable representation supporting grounded semantic querying. Simultaneously, the compact latent state representation $z \in \mathbb{R}^{D}$ necessarily captures the same information as the explicit representation x while being compatible with the learned state-transition modeling paradigm.

3.4.4 Predictive Agents

This section explains how predictive agents based on combined state-transition modeling and LLM commonsense reasoning over world knowledge relates to general-purpose mobile reasoning agents. The concept of an agent is based on the existence of an external environment and an entity with agency that interacts with the environment. A useful agent has agency, meaning it is driven to complete a particular task, possibly involving

completing sub-tasks by task decomposition [409, 437]. Typically goals are specified by inherently goal-driven entities like humans, or intrinsically developed agency by the machine agent [438] A formalization of completing a task is to find a sequence of actions taking the agent from an initial state to a goal state [439]. Completing complex tasks typically involve long and complex sequences of actions. Defining the agent behavior functions or program that predicts or determines which actions are taken in each state is one of the core challenges to solve in a any agent system. A predictive model is a function or a mapping that takes input data representation and produces an output prediction or estimate. Inputs are commonly referred to as features or independent variables. Outputs are called as target variables or dependent variables. Here follows an analysis of requirement and implementable components of the proposed general-purpose mobile reasoning agent framework as a predictive agent.

Agent frameworks. Agent programs embody the principles underlying intelligent systems. As explained in Sec. 1.1 and formalized in (1.3), all agent programs can be mathematically modeled as sequential state-transition models based on perceiving states and attempting to predict rational action that completes the task. Particular implementations of state representations and state-transition models covers a wide spectrum of complexity. Some models simply reacts on the current state being equivalent to the percepts, while others integrates observations over time to elucidate a more complex environment state to base actions on. Typically agents are organized into four types according to abilities [28]:

1. Reflex or reactive agents [440] are the simplest type of intelligent agent, selecting actions based solely on the current percept and ignoring the rest of the percept history. A simple vacuum cleaner is a typical example of a simple reflex agent that behaves according to pre-programmed instructions based on current sensor readings. Reactive agents simply react to the current environment observations without any form of internal memory, environment state, or world model.

2. Model-based reflex or reactive agents [417, 441] are more sophisticated than simple reflex agents, as they use an internal environment model to keep track of its state and history. This allows them to make decisions based on both current percepts and previous experience.

3. Goal-based agents [442, 443] are designed to achieve specific goals by choosing actions that lead towards desired outcomes or follows an instruction. They may use a hierarchy of goals to prioritize different objectives and select the most promising action based on their current state and available options. Goal-based agents can handle non-deterministic or partially observable environments better than simple

and model-based reflex agents, as they can reason about future states and choose actions accordingly. Goal-based agents are also deliberative agents, as they posses an explicit internal model of the world and can reason about their actions beforehand.

4. Utility-based agents [90, 357, 362, 444] are a more advanced type of agent program that aims to not only reach a goal state, but also maximize expected utility by considering the long horizon outcome of their actions. They use decision theory to evaluate different options based on their potential outcomes and select the one with the highest expected value. Utility-based agents can handle uncertainty and risk effectively, but they require accurate models of the environment and precise estimation of utilities.

Each kind of agent program combines particular components in particular ways to generate actions. The performance measure evaluates the behavior of the agent in an environment, and a rational agent acts so as to maximize the expected value of the performance measure. A mixed or hybrid agent combine elements from both reactive and deliberative agents, allowing them to respond quickly to environment stimuli if necessary while still planning ahead and reasoning. The same dichotomy of fast and slow cognition is well-studied in human intelligence [445].

State-transition modeling with commonsense reasoning. Conventional utility-based agent framework are structured as state-transition models initialized as blank slates and learned by trial-and-error experience. This thesis propose to instead leverage LLM powered commonsense reasoning as part of a state-transition framework based on spatially grounded and semantically rich predictive state representations. The proposed framework can therefore hypothetically plan long sequences of high-level actions without requiring to experience negative outcomes in order to avoid unwanted states. Eliminating the need to experience negative outcomes in order to learn to avoid such is a prerequisite for safe learning from observation experience in the real world. Core LLM agent abilities include perform hierarchical planning by task decomposition [409, 437] and program synthesis [385–387], and reason with commonsense world knowledge [92]. The proposal is discussed in detail in Sec. 3.4.4.2.

Complete vs. partially observed environment states. The distinction between fully observable and partially observable environments is important in the design of real-world mobile robot intelligent agents [1]. In a fully observable environment, the agent can see the entire state of the world at any given time, while in a partially observable environment, some aspects of the state may be hidden from view. This thesis provides a

solution for the partially observable environment problem with potential to unify "game AI" [21, 57] and mobile robotics approaches [2]. The solution is described in Sec. 3.4.4.1.

3.4.4.1 Predictive world modeling

The objective of the general-purpose predictive agent is to solve novel task by reasoning over commonsense world knowledge. Predictive world modeling is the core component bridging the spatio-semantic environment state as perceived by the agent with the world knowledge and reasoning capability of LLMs as explained throughout this section.

Spatio-semantic representation requirements. Cognitive scientists believe that adaptable intelligent agents like humans represent the world using a small set of foundational cognitive components for perceiving inanimate objects, external agents, numeric concepts, social relations, and spatial environments [99]. These cognitive abilities allow intelligent agents to perform commonsense physical reasoning and imagine counterfactual scenarios to facilitate task accomplishment [380]. A key capability is predictive world modeling [265–267, 381]. In contrast, mobile robots are conventionally designed and programmed for performing a priori specified tasks in known environments. General-purpose mobile robots on the other hand, aim to be flexible intelligent agents that can understand novel situations and complete a wide variety of tasks in new environments by leveraging world knowledge. Large language models (LLMs) have emerged as a promising direction to achieve general-purpose agents [85, 90, 92, 382–387]. Core LLM agent abilities include understanding weakly specified goals defined in natural language [90], perform hierarchical planning by task decomposition [409, 437] and program synthesis [385–387], and reason with commonsense world knowledge [92].

To complete a novel task in a new environment, a general-purpose mobile robot need to comprehended the environment by an a priori unknown set of semantics. Vision-language models (VLMs) [6, 77, 80, 81, 188–193, 395] is a common approach to ground rich open-vocabulary (OV) semantics in the observed environment and connect the internal reasoning processes of LLM agents with the external world. However, spatio-semantic reasoning tasks may require information beyond what is currently observed. Efficiently fetching an item out of view require a spatio-semantic memory of where the item is located [91]. Inferring navigational patterns like road lanes may require predictive assumptions for the unobserved environments behind obstructions [301]. A spatio-semantic memory [397] or scene representations [398] allows an agent to query semantics from observational memory [290, 292, 293], to navigate [187], and plan by reasoning [92]. Common representations for spatio-semantic include 3D reconstruction [399], object-centric, topological maps [400], scene graphs [401], and top-down metric grid maps [381]. The

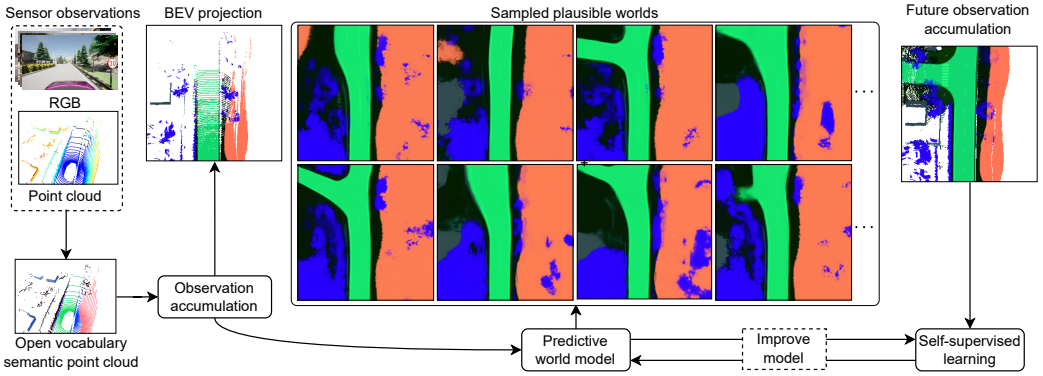

FIGURE 3.2: The framework integrates open-vocabulary semantic point cloud observations into a common vector space. A predictive world model samples a set of diverse plausible complete world states from the partially observed state. The model improves through continual learning from experience by comparing predicted and observed future states based on predictive coding. High-dimensional semantic embeddings are projected to RGB color values for visualization. [1, 2]

spatio-semantic environment state representation for general-purpose agents thus requires the following properties: encode an open-ended vocabulary of semantic concepts, represent and allow querying of overlapping semantics (e.g. a *couch* is also a *furniture*), and store observations compactly. The theory of latent compositional semantics [95] satisfies the above requirements.

Open-Vocabulary Predictive World Model. This thesis proposes an Open-vocabulary Predictive World Model (OV-PWM) [2] as a spatio-semantic memory and internal simulator for general-purpose mobile robots. World models are abstract representations of the environment that enable planning over latent structures decoupled from particular observable appearances (i.e., pixels) [265–270, 273]. The advantage of world models is demonstrated by recent model-based reinforcement learning [279, 280]. This thesis presents a framework for implementing a self-supervised predictive world model that generates a diverse set of explicit plausible complete open vocabulary world states trained from partially observed states only [1, 2]. The experimental results verify the feasibility of this approach in realistic real-world environments as shown in Fig. 3.2. The OV-PWM is a latent variable generative model [446–449] that learns from egocentric partial observations to predict complete environment states represented by grounded open-vocabulary semantics. The OV-PWM functions as an implementation of an artificial hippocampus that learns a distribution of compact latent codes capturing the structure of observed environments. See Chap. 5 for further details about the predictive world model as an artificial hippocampus.

The explicit open-vocabulary environment representations enabled by OV-PWMs provides several potential advantages to implicit representations and conventional offline

map-based mobile robots with human-annotated semantics [2]. First, the OV-PWM can disambiguate the observed state by substituting unknown regions with plausible predictions based on prior observational experience. Committing to a particular complete state simplifies learning policies by removing the implicit marginalization over many plausible underlying states for state transition modeling. Secondly, OV-PWMs can bridge conventional map-based and perception-based planning and control methods. For example, safer motion planning may be achieved by sampling diverse plausible structures of unobserved regions and account for worst-case scenarios. Additional potential advantages include improving localization by densifying observations, verifying offline map consistency with the actually observed environment, and leverage the highly expressive but compact latent state for planning in latent space [273]. Thirdly, learning a world model based on grounded open-vocabulary semantics allows optimizing a single general OV-PWM for multiple tasks requiring different semantic perceptual information. Fourthly, leveraging unconditional open-vocabulary semantics supports inferring overlapping semantics by sufficient similarity inference [95].

The proposed framework uses an HVAE model [449] to encode and reconstruct future world states generated by the plausible state completion module. This process does not substitute for the predictive world model, as it leverages future observations that are not available at inference time. The original partially observed world representations are transformed by data augmentation techniques, including random rotation, translation, and warping operations, when training the predictive world model. The geometric augmentations are essential to achieve geometric invariance for generalization [2].

The predictive world models framework proposed by this work aims at generating plausible complete worlds from partially observed states only via self-supervised learning using sensor observations [1]. This approach is particularly useful for real-world spatial environments, as it does not rely on maps or predefined representations. Instead, the model learns to represent and sample general and spatially complex structures in a probabilistic manner. The experimental results demonstrate that this framework effectively closes the gap between perfect prediction and partially observed worlds by 61.7% on average when evaluating over both past and future observations [2]. The process of generating plausible complete world states involves multiple stages, including encoding and reconstructing pseudo ground-truth world states using a regular hierarchical VAE (HVAE) model with learned latent variable prior $p_\theta(z)$ and posterior $q_\theta(z|x)$, as well as performing data augmentation on the original partially observed world representations when training the predictive world model.

The experimental results presented in Chap. 5 show that the proposed framework can

generate high-quality partially observed world states and accurately predict future observations with an average IoU of 98.73%. See Fig. 3.2 for visual examples of generated plausible worlds.

Summary and limitations. Predictive world modeling is critical in creating general-purpose mobile reasoning agents due to its role in producing an explicit spatially grounded and semantically rich environment representation. The proposed approach leverages recent advances in deep generative models and provides a promising alternative to traditional map-based navigation methods. The predictive environment representation can be aligned with LLMs to enable common sense reasoning grounded in the observed environment as explained in Sec. 3.4.1. Facilitating commonsense reasoning over world knowledge promises to make general-purpose agents adaptable to various tasks and environments while overcoming limitations of existing paradigms as explained in Sec. 1.2. The predictive world modeling approach also has potential applications in narrower agent tasks based on conventional autonomous driving systems explained in Sec. 1.2. Adding spatio-semantic representations of the environment as perceived by the sensors can compliment predefined maps by robustify against unsafe behavior or failure to operate due to map errors [45, 240, 301].

A limitation of the presented OV-PWM implementation is the top-down 2D grid representation. 2D embedding maps do not represent vertical information and multi-layered environments as required for general 3D representations. Extending the OV-PWM approach to 3D representations using voxel grids or neural radiance fields is a promising direction of future work to enable spatial reasoning in fully general complex 3D structures. Learning can be improved by reducing degenerate accumulated observation samples resulting from inaccurate and erroneous ICP scan matching steps by implementing a robust SLAM-based observation accumulation framework. Allowing the OV-PWM model to learn to recreate actual representations of the world instead of overfitting to noisy samples is expected to improve both generative accuracy and learning efficiency.

3.4.4.2 State-Transition Modeling

This section presents an account of state-transition modeling as a mental simulation framework. State-transition models are a type of probabilistic model used to represent systems where entities can transition between different states over time. These models have been widely applied in various fields, including AI, robotics, and operations research. In the context of AI, state-transition modeling has played a significant role in areas such as planning, decision making, and control. The basic idea behind state-transition models is that the system under consideration can be represented using a

set of states and transitions between these states. This thesis presents state-transition models as a sequence of abstract high-level actions by reasoning over commonsense world knowledge grounded in the predictive environment state representation. The presentation starts with explaining the conventional state-transition modeling approach. The later part explains how recent LLM-based reasoning frameworks are theoretically adapted for mental simulation grounded in the aligned predictive world model states.

Search problems. Problem-solving agents leveraging search are the prototypical approach to solve problems by planning [28, 443]. Search problems are formalized by defining a state space as a finite or infinite set of atomic states x as an abstraction of the environment. The agents starts in an initial state x_i. The objective is to reach one out of potentially many goal states $x' \in X_g$. A goal state function $\texttt{is\_goal}(x)$ determines the set of goal states

$$X_g = \{x | \texttt{is\_goal}(x)\}. \tag{3.1}$$

The agent can do actions a. A state transition model $f_\theta(x, a)$ specifies the new state x' the agents will be in if the agent does action a being in state x

$$x' := f_\theta(x, a) \;\; s.t. \;\; a \in \texttt{actions}(x). \tag{3.2}$$

Only valid actions defined in $f_\theta(x, a)$ and represented by the set $\texttt{actions}(x)$ are doable. The state transition model $f_\theta(x, a)$ represents an approximation of the real world dynamics relevant to the problem being modeled. A problem-solving agent solves a problem by searching for a sequence of actions $\{a_1, \ldots, a_N\}$ taking the agent from x_i to $x' \in X_g$ based on $f_\theta(x, a)$. An action cost function $\texttt{action\_cost}(x, a, x')$ represent the cost of transitioning to x' from x by a. An optimal sequence of actions or plan is the one with smallest accumulated cost. The iterative search process results in a search tree growing from the initial state x_i outwards in search of a goal state $x' \in X_g$ [450]. Typical search algorithms for problem-solving agents include uninformed algorithms like breadth-first search and depth-first search [360]. Uninformed search algorithms expands the search tree without knowledge of which nodes are more likely to lead towards a goal state. Informed search algorithms like greedy best-first search [451] and A* [359] search with variants like weighted A* [452] leverage a heuristic function $h(x)$ to decide which states to search. Heuristic functions $h(x)$ estimates the optimal path cost from state x to a goal state $x' \in X_g$ [451]. The search problem formulation is widely used in various areas of AI, including constraint satisfaction algorithms, propositional logic, planning, Bayesian networks, and machine learning algorithms.

Traditional problem-solving agents solves problems defined in fully observed, discrete, deterministic, and known environments. Solving problems in partially observed, continuous, stochastic, and unknown infinite real-world environments require additional

considerations explained in the reminder of this section.

Environment state representation. The real-world environment of general-purpose mobile reasoning agents is far from atomic. Actions like actuation and sound generation are continuous at the most fundamental level. The amount of inferable semantics in the environment state is practically unbounded. A suitable level of abstraction is needed to make the search problem tractable for real-world general-purpose agent problems. The predictive state representation proposed in this thesis is such an abstraction.

The predictive state representation x can be seen as an approximate open vocabulary spatio-semantic snapshot of the world at a particular time t. The state x is inferred based on integrating past and current observational experience following the prototype The state x is a dual latent and explicit representation, supporting both latent state transition modeling using compact but semantically expressive latent states z with equivalent discriminability as the explicit states x. See Chap. 5 for more details about the predictive state representation. The dimensionality for z and x for results presented in this thesis is $\mathbb{R}^{16}$ and $\mathbb{R}^{H\times W\times 768}$, respectively. The open vocabulary state x can represent a large set of semantics by latent compositional semantic embeddings. See Chap. 4 for more details about latent compositional semantics. The explicit spatio-semantic state representation x can be aligned with the embedding space of LLMs by the computational geometry based encoder optimization method presented in Appendix C.

Partially observed to predictive states. Real-world mobile agent environments are typically expected to operate on partially observed environment states perceived by the agent. Conventional state-transition modeling methods struggle with such incomplete environment states. The naive probabilistic state-transition model approach is to marginalize out uncertainty by predicting state transitions as probability distributions. The stochastic nature of probabilistic state transition modeling does not allow long prediction sequences as uncertainty builds up in each prediction step

$$\hat{x}' = \int_x p(x) f_\theta(x'|x,a) dx. \tag{3.3}$$

The conventional approach to state-transition modeling under uncertainty include Markov decision processes (MDPs) and partially observable Markov decision process (POMDPs). MDPs model optimal decision making by extending traditional Markov chains through incorporating both uncertain action outcomes specified by probabilistic transitions between states, as well as rewards associated with those transitions [453]. MDPs have been widely applied in various domains, including game playing, resource allocation, and robot control. MDPs can be solved by value iteration and policy iteration algorithms. POMDPs extend MDPs problems to also include uncertain initial states. The

solutions to POMDPs is based on maintaining a probabilistic distribution

$$p(X) = \{p(x_1), \ldots, p(x_N)\} \quad , \quad \sum_{n=1}^{N} p(x_n) = 1 \tag{3.4}$$

over the set of plausible states X the agent may be in. The set X is called a belief state. In real-world environments the number of possible state in X is infinitely large, meaning the approach does not scale well beyond simple discrete environments.

This thesis presents a predictive world model $PWM()$ capable of generating a set of discrete plausible worlds $\{x_1^*, \ldots, x_K^*\}$ originating the partially observed world x. Representing the set of plausible worlds as a learned latent distribution is more scalable than maintaining a set of explicit plausible words as in the belief state approach. The predictive world modeling approach can equivalently be viewed as a learning-based application of belief state prediction as a probabilistic distribution of latent states learned from observational experience [2]. Basing state-transition modeling on explicit predictive states x^* sampled from a $PWM()$ instead of the marginal distribution over all possible states resulting from a partially observed state x is hypothesized to enable long prediction horizons in real-world environments. State estimation uncertainty is eliminated by replacing propagation of uncertainty with exploring a search tree of sampled plausible explicit states akin to Monte Carlo Tree Search (MCTS) over fully observable game environment states as demonstrated by AlphaGo [21]

$$\hat{x}^{*\prime} := f_\theta(x^*, a) \;\; s.t. \;\; a \in \texttt{actions}(x^*) \quad , \quad x^* \sim PWM(x). \tag{3.5}$$

The MCTS approach is visualized in Fig. 3.3. The agent perceives the world as a partially observed environment state x. The predictive world model $PWM()$ samples plausible complete predictive states x_i^* without uncertainty. Multiple search trees originating from the predictive states explores the state space by a state transition model $f_\theta()$ until a goal state $\texttt{is\_goal}(x^*)$ is found.

The proposed predictive state MCTS problem-solving agent approach is based on a theoretically sound state space of explicit and unambiguous environment states generated by the predictive world model $PWM()$. However, the model does not address how to represent actions, what actions are allowed in each state, and how to model the state-transition function $f_\theta()$. The advantages of learning prediction models based on explicit predictive states is demonstrated in Sec. 3.5 and Chap. 6. The remainder of this section explains how integrating LLM-based reasoning solves the remaining limitations and transforms the predictive state MCTS framework into a mental simulator leveraging abstract actions and predicted state transitions based on commonsense reasoning over world knowledge.

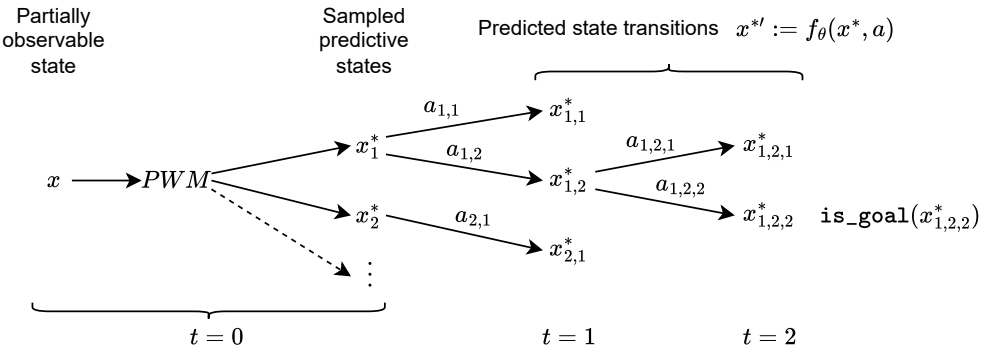

FIGURE 3.3: A state-transition model allows an agent to search the space of action sequences likely reaching a desired goal state.

LLM-based reasoning as mental simulator. Large language models (LLM) have demonstrated a transformational capability to learn and represent commonsense world knowledge from internet-scale text information [70, 161]. Furthermore, LLMs have also demonstrated a capability to perform multi-step reasoning over the assimilated commonsense world knowledge [70, 409, 454–456]. This thesis propose a principled method to ground LLM-based commonsense reasoning to the perceived environment and facilitate a MCTS search over abstract actions and state transitions. The versatility and leverage of commonsense world knowledge is a promising principled direction to enable future general-purpose mobile reasoning agents to complete complex tasks by search-based planning grounded in the perceived environment. The remainder of this section explains how to transform the predictive state MCTS approach visualized in Fig. 3.3 into a mental simulator leveraging reasoning over world knowledge.

The sampled set of predictive states x^* in Fig. 3.3 represent explicit open vocabulary plausible worlds underlying the partially observed environment state x. The predictive states x^* is hypothesized to be aligned into a LLM embedding space by the proposed encoder optimized by computational geometry explained in Appendix C. Leveraging a tree-of-thought (ToT) reasoning algorithm allows the LLM agent to predict doable high-level actions a for each predictive state x^*. The state-transition function is implemented by a LLM that predicts a high-level textual description of the environment state $x^{*\prime}$ based on commonsense world knowledge. The branching step-by-step reasoning algorithm runs until a goal state is found and a sequence of high-level actions $\{a_1, \ldots, a_K\}$ is returned. The goal state is identified by a LLM function evaluating the high-level state description and the task description [457–461]. Goal state detection methods allows the AI agent to itself determine if a goal state has been reached. Goal state detection enables versatile autonomous task optimization and learning from experience due to the fact agents can autonomously set up and evaluate completion of causal hypotheses by

experiment. Additionally, this direction opens up for causal discovery by formulating a hypothesis that can be tested and evaluated in the real world. See Sec. 3.6 and Sec. 3.7 for further details about task description representations and planning by mental simulation. The high-level actions can be transformed into low-level control programs by LLM-based program synthesis [377, 378, 387, 462, 463]. See Sec. 3.8 for further details about program synthesis for low-level action execution.

The grounded LLM-based reasoning approach hypothetically enables versatile, efficient and safe exploration of high-level action sequences. Versatility stems from semantically grounding the environment representation in the LLM embedding space and predicting state transitions based on commonsense world knowledge. The approach is efficient because the level of state and action abstraction is aligned with human decision making due to LLMs being trained written information based on the human world perspective. The returned sequence of high-level actions are safe as undesirable state transitions like collisions are identifiable by commonsense world knowledge inherent in LLMs. The mental simulator can also be optimized by self-reflecting on predicted and experienced high-level state outcomes in line with the predictive coding learning paradigm [331]. The ability to learn from future observations and update the state transition model is particularly important in dynamic real-world environments where conditions may change rapidly, requiring constant adaptation on the part of the agent [98]. The natural language-based mental simulator supports versatile human-machine communication and explainability by providing explicit sequences of actions, state transitions, and decision factors, making it easier for humans to understand and evaluate how an agent arrived at a particular conclusion or action.

The proposed mental simulator approach is limited in practice by the generality and accuracy of commonsense world knowledge within LLMs, the correctness of reasoning steps done by LLMs, accuracy of goal state detection, and achieving sufficient inference speed for timely decision making. The practical investigation of the proposed mental simulation based on predictive state representations is proposed as future work.

Another advantage of state-transition modeling lies in causal discovery, which involves using structured models as mental simulations that can learn from real-world outcomes. By creating such models, agents are better able to understand the underlying causes of events and make more accurate predictions about future states based on current observations.

The following sections provide a deeper contextual grounding and explanation of several practical aspects of general-purpose mobile reasoning agents: navigation, task description, mental simulation, and low-level action execution.

3.5 Navigational Patterns in Constrained Environments

General-purpose mobile agents perform tasks that involve traversing an environment. To navigate rule-constrained structured environments robots are required to correctly perceive and interpret the environment. This problem is called scene understanding. Navigational patterns, or directional pathways, are a core component of understanding how to traverse structured environments [322]. In particular, efficient and safe multi-agent navigation depends on each agent following mutually known navigational patterns. The patterns can be defined by explicit rules or be derived from social conventions and emergent behavior. However, learning to infer navigational patterns for complex environments based on observable features is difficult due to regional variation and noise including varying or missing surface markings, geometries, and materials.

Current methods for spatial navigation can be categorized into mapping- and learning-based approaches. The mapping approach [464] avoids the problem of automatized understanding of environments by encoding human knowledge in the form of lane maps and localizing the system within these maps. Creating a priori navigation maps is a conceptually simple, interpretable, and predictable way to safely navigate environments. In practice, this approach is difficult to scale up, as map creation, maintenance, and verification are costly in terms of human labor, typically limiting application to small predetermined environments. Additionally, dynamic navigational behavior like correctly avoiding parked cars or debris cannot be a priori encoded in static maps.

The learning approach involves training a model to infer navigational patterns based on environmental context. Some methods learn implicit patterns as part of accomplishing the primary task [33, 318, 319]. Other methods learn explicit patterns but require ground truth lane maps for training [313, 314]. Methods learning from observational data alone are promising scalable solutions to infer navigational patterns, as driving data can be obtained at a low cost. However, the real-world performance of existing methods is fragile and unpredictable in complex environments and lacks interpretability.

The human visual system comprises two subsystems [352–354]. The vision-for-perception system located in the ventral stream processes information in a slow, top-down manner to create perceptual representations from ambiguous or incomplete visual input by leveraging visual and semantic memory [353]. These representations support conscious mental processes such as recognition, visual thought, and planning. The vision-for-action system located in the dorsal stream processes information in a real-time, bottom-up manner to perceive the entire environment and infer behaviorally-relevant visual affordances, including cues for spatial navigation [353, 355].

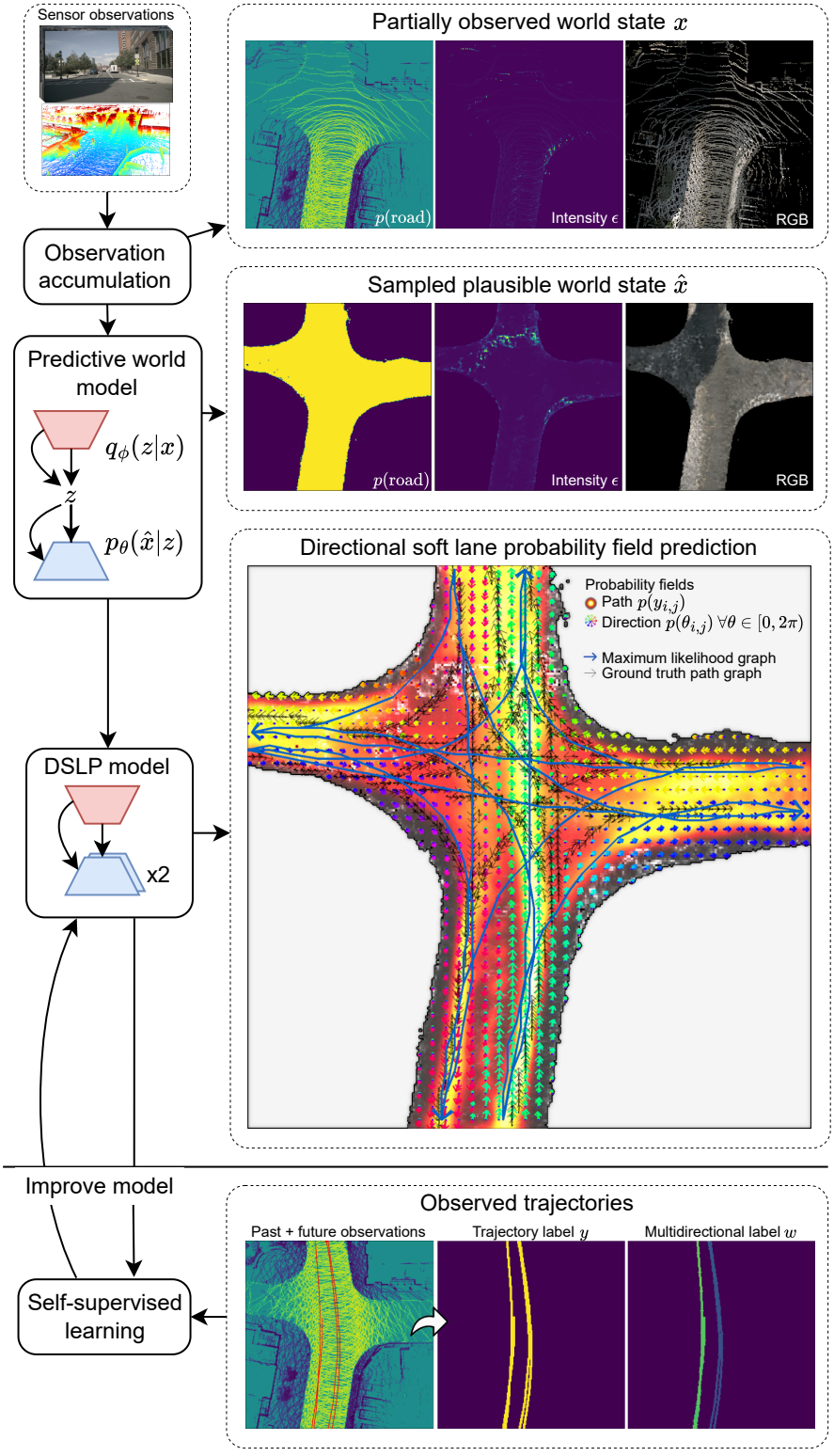

FIGURE 3.4: The method accumulates sensor observations into a common metric vector space representing the partially observed world state x. A predictive world model samples a set of diverse plausible complete world states $\hat{x}$. The directional soft lane probability (DSLP) model predicts two probability fields; the agent traversal probability $p(y_{i,j})$ and a multimodal directional probability distribution $p(\theta_{i,j})$ for each point (i, j). A fitted maximum likelihood graph corresponds to global navigational patterns. The DSLP model can learn navigational patterns from observed trajectories representing only a subset of all plausible trajectories.

In this thesis I present a self-supervised method for learning to infer navigational patterns from real-world partial observations as required for traversing unmapped real-world environments. My approach is inspired by the biological dorsal visual pathway [353] and endows artificial intelligent agents with a functionally similar self-improving system that learns to infer visual affordances for spatial navigation [465].

The model learns general contextual environment features that explain observed trajectories, and can thus infer navigational patterns for newly encountered environments. Learning from observed trajectories means learning from only a subset of all plausible trajectories. I propose an information-theoretic regularizer to overcome the problem of false negative traversal observations resulting from partial observations. My model combines complementary aspects of mapping- and learning-based approaches. It also produces an interpretable representation akin to maps. Lastly, this model improves with additional experience akin to continual learning [466] while avoiding catastrophic forgetting by retaining a replay buffer of past experiences [57].

I identify the navigational pattern prediction problem based on static environmental context as a sub-problem of the general dynamic agent behavior prediction problem. The main difference is that I do not consider the influence of dynamic objects such as parked cars and red traffic lights, or predict the movement of particular agents. While both problems can be solved through the same framework, I choose to remove dynamic object information from the input representation in order to objectively compare performance against ground truth lane graph methods.

While I perform experiments in a real-world urban road environment my method is applicable in any general structured environment.

3.6 Task Descriptions

This section discusses different paradigms established in the literature, and their inherent limitations of specifying task descriptions for general-purpose agents.

Formal specification. The concept of formal task descriptions for intelligent agents involves creating structured representations to define tasks in a precise, unambiguous manner that can be processed by computational systems. Formal language approaches were the dominant approach between 1950s-1980s culminating in expert system technology. A formal task description typically represent the environment, action space, and agent state in terms of logical rules, constraints, and Markov decision processes (MDPs). A logical rule is a formal representation that captures the relationship between different predicates or variables within an environment. Logical languages are precisely defined by

syntax or a set of formation rules. Popular languages include propositional, first-order, and temporal logic. These rules typically follow the form of *IF-THEN* statements, where the antecedent or condition leads to specific consequences if met. In the context of AI, these rules can be used to describe complex tasks and decision-making processes for intelligent agents. While logical rules primarily focus on describing relationships between variables, constraints are used to limit the possible values or combinations thereof within a given domain. Constraints can take various forms, such as equality, inequality, set membership, or mathematical expressions. By enforcing these restrictions, an agent can narrow down its search space and improve learning efficiency. Optimization problems are typically defined by constraints. Markov decision processes (MDPs) are mathematical models consisting of states, actions, transition probabilities between states based on chosen actions, and rewards associated with transitions [453]. Encoding tasks as an MDP problem enables efficient dynamic programming algorithm like the Viterbi algorithm [467] to find the optimal action sequence or plan that maximizes the expected cumulative reward. Formally specified task descriptions solvable by theorem proving or optimization are guaranteed to be optimal solutions.

The scope of problems that can be defined by formal languages are limited by real-world complexities as degree of truth, ambiguous semantics, and practical infeasibility of encoding the problem and sufficient background knowledge by non-conflicting statements. General-purpose agents should benefit from formally specified tasks when adequate, but agents must also be able to comprehend tasks involving ambiguous, probabilistic, and conflicting instruction statements.

Input-output examples. An implicit way of specifying a task is to exemplify it with desired input-output examples. In other words, if a function $f()$ successfully completed task based on an input x, then the output should be y. This approach is prevalent for narrow learning-based AI task solvers like computer vision models and machine translation, where a large dataset of input-output examples approximate the intended task and is used to find a model $f\theta()$ that best complete the intended task. The approach adequately addresses the problem of expressing ambiguity and conflicting or noisy instructions limiting approaches leveraging formal specifications.

However, the inherent implicit nature of specifying tasks do not provide a mean for precise information as expressible by formal or natural languages. Another limitation is the amount of examples required to adequately specify a task is generally unfeasible for the use case of general-purpose agents. A related problem is the challenge of providing extensive background world knowledge into the example-based task completion model which may exist in a different representation modality than the input-output examples defining the task.

Reward functions. Another indirect means to specify a task is by a degree of goodness expressing how an agent's behavior contributes towards completing an intended task in a desired manner. The goodness value is generated by a reward function based on one or many components including environment state, agent state, and human-created task heuristics. A reward function assigns scalar values to actions taken by an agent in specific states or state-action pairs within the environment. These rewards provide feedback on how well the agent is performing and help it learn optimal policies through trial and error [56]. Specifying a task by a reward function, together with structured search methods, is a powerful means to achieve superhuman task performance as has often been demonstrated in games [21] and simulation environments [350].

The reward function approach has considerable challenges when it comes to specifying complex real-world tasks. Specifying the reward function is non-trivial for typical real-world tasks for many reasons. The action sequence length required to complete a task might be excessively long, meaning the agent will have significant challenge to learn to complete the task. While simulation-based agents can afford millions of try-and-error experiments to find the desirable behavior, time efficiency is an important consideration in real-world task execution. Searching for the intended behavior by maximizing reward outside simulation may result in damage to the environment and the physical agent itself. Reward hacking is when an agent may discover an unintended means to accumulate reward and result in undesirable behavior. Even ignoring the challenge of reward hacking, specifying the reward function itself is a practical concern for non-trivial real-world tasks. While specifying tasks by reward functions provides means to achieve superhuman performance on precisely definable and simulatable tasks, real-world applications face substantial challenges. Fundamentally the challenges stem from lacking semantically rich means to specify reward functions and background world knowledge to enable a desired interpretation of the intended behavior, or in other words, aligning the agent.

Natural language. Task specification communicated between human beings are predominately in the form of natural language instructions. Natural language is considered the core enabler of the success of homo sapiens [15] and provides unlimited means to express any thought [334]. Deep learning has emerged as a powerful tool for NL processing tasks, including understanding NL instructions [70, 468]. The success of deep learning can be attributed to its ability to learn meaningful representations from raw data without relying on hand-engineered features. This is particularly beneficial in the context of NL processing, where the complexity and variability of natural languages pose significant challenges for traditional rule-based approaches. Traditionally natural language instruction was considered an intermediate representation to be translated into

formal machine interpretable language like first-order logic. However, with the advancement of huge transformer-based[155] large language models (LLMs) [70, 161] trained on massive amounts of natural language data and examples of instruction following [468], it is now possible for machines to interpret and follow task instructions specified directly in natural language. The capability to complete tasks specified directly in natural language provides significant advantages in practical usefulness and versatility of machine learning models as general-purpose AI systems. Additionally, tasks can be completed while leveraging extensive world knowledge abstracted into the network parameters.

Natural language instructions interpreted by LLMs have limitations in terms of fully specifying tasks for embodied agents operating in a physical 3D world. To complete an embodied task, a semantically rich and spatially grounded environment representation needs to be provided in order to ground the abstract task in the particular 3D environment. Expressing the environment plainly by textual representation is limited due to the inadequacy to concisely represent space and all possible semantic interpretations of objects by text alone.

Natural language instruction interpretation as consolidator. This thesis proposes an dual predictive environment state and task description representation that compliments the strengths of each paradigm in order to overcome each individual limitation. Natural language instruction can provide a fuzzy logic and approximate reasoning by instruction [454] to enhance the applicability of formally defined tasks to broader real-world problems solvable by general-purpose agents. The input-output example approach is enhanced by integrating prior world knowledge, resulting in a more sample-efficient approach to specify a task by example [70]. Note that the approach is compatible with non-textual modalities like images when leveraging multi-modal LLMs [79]. Empirical evidence shows that natural language instruction interpreted by LLMs as program synthesis is a promising means to define effective and versatile reward functions by iterative feedback-based improvement [377]. The proposed predictive world model framework, capable of continuous improvement from observational experience, is a promising direction to learn internal world simulators that facilitate learning policies by optimizing a reward function and subsequently transfer learned skills back to the real world [378].

3.7 Planning by Mental Simulation

Integrate Hierarchical planning by task decomposition [386, 387] is another critical component of designing general-purpose mobile agents capable of generating effective plans for accomplishing various tasks in new environments. Generation of effective

plans include producing high-level goals with intermediate sub-goals down to the level of actuation through program synthesis [385].

Planning by simulation, also known as forward search or model-based reasoning [28], is a method used to find a sequence of actions that achieves a goal. It involves creating an internal representation of the world or environment and simulating outcomes of potential actions within this approximative model of the world. This approach has been widely studied in AI research for various applications, including general-purpose problem solvers.

The idea of planning by simulation can be traced back to the work "Planning and Acting" [439] based on AND–OR trees. Prior work related to simulation-based planning include application to propositional theorem proving [469] and the AO* graph search algorithms [470]. Astrom [471] proposed approaching partially observed problems based on imperfectly estimated states as belief state problems solvable by optimal control methods over Makrov decision processes. This concept has been applied to robotics manipulation without sensors, as demonstrated by [472]. Further advancements include search heuristics [473] and incremental solution by belief state subsets [474]. The problem of continual learning of dynamically changing environments is tackled by algorithmic solutions like D*-Lite [475] and LifeLong Planning A* [476]. The fundamental limitation of classical search-based methods is the limited state space and transition model employed to approximate the world. While classical general problem solvers are guaranteed to find an optimal solution if a solutions exits, the approach does not scale up to ambiguous task descriptions and solutions requiring reasoning over broad commonsense world knowledge as required by general-purpose agents.

Large language models (LLMs) provides a unique capability to leverage commonsense world knowledge for concrete and abstract planning problems. The recent SOTA Tree of Thought (ToT) LLM-based algorithm [454] achieves versatile reasoning using world knowledge by iteratively exploring a causal tree structure of actions and hypothetical outcomes. The use of ToT enables the exploration of complex scenarios involving multiple actions and events that are not explicitly stated but can be logically deduced based on common sense. The process begins with an initial thought prompt, where the LLM generates plausible thoughts or actions related to the specified problem. The generated thoughts are then evaluated using another set of prompts, which assesses their feasibility, relevance, and impact on subsequent actions. The evaluation results determine whether a particular thought is selected as the next step in planning by simulation. This iterative process continues until a logical path to the solution is found.

ToT allows LLMs to reason over commonsense world knowledge through two main mechanisms. First, implicit reasoning refers to the model's inherent understanding of the

relationships between various entities, actions, and events in the real world according to the vast textual general information data used to train the LLM. This emergent grasp of causality enables LLMs to generate thoughts that are logically consistent with each other and the given scenario. Explicit reasoning involves using prompts specifically designed to elicit logical deductions based on common sense. These prompts guide the model through a series of questions or tasks that require it to apply its knowledge of cause-and-effect relationships in order to generate relevant thoughts for further evaluation.

The flexibility and effectiveness of ToT have been demonstrated across various tasks, including creative writing, puzzle solving, and crossword puzzles [454]. In each case, LLMs were able to reason about commonsense world knowledge using both implicit and explicit reasoning mechanisms in order to generate coherent plans or solutions for the given problem. Moreover, ToT outperforms alternative methods such as Chain-of-Thought (CoT) linear reasoning prompting [409] for complex scenarios requiring deliberate reasoning, highlighting its potential utility in real-world applications where sophisticated general-purpose planning capabilities are required.

This thesis propose a framework for spatio-semantic predictive world states which are compatible with multi-modal LLMs. The semantically rich and spatially explicit states represent a principled environment information source from which to do general-purpose abstract reasoning grounded in the mobile robot agent's environment.

3.8 Low-level Action Execution

Low-level action execution is a crucial aspect of robotics and artificial intelligence that involves generating specific motor commands for achieving precise movements or manipulation required to complete a given physical task. This process often requires programming, simulation, and the use of reward functions to guide the learning and execution of actions.

Program synthesis plays an essential role in recent state-of-the-art (SOTA) performance for low-level action execution. Program synthesis involves automatically creating explicit code or programs that control robots or other machines analogously to the conventional human written control code in robotics. Program synthesis plays an essential role in recent SOTA performance for low-level action execution [385]. It involves using large language models (LLMs) to generate code that can be executed directly by robots or other machines, allowing them to perform complex tasks with greater precision and efficiency.

Reward functions are used to guide the learning process by providing feedback on how well an agent is performing a given task. In reinforcement learning scenarios, agents receive rewards based on their actions and learn through trial and error to maximize their cumulative reward over time. Rewards can be defined for specific goals or sub-goals within a task, enabling the agent to break down complex problems into smaller, more manageable components. Such task decomposition is another important aspect of low-level action execution. It involves breaking down a complex task into smaller, more manageable sub-tasks that can be executed sequentially or in parallel [387]. This approach allows for greater flexibility and adaptability in the design of reward functions and has been shown to improve the efficiency of low-level action execution.

One of the key challenges in designing reward functions for low-level action execution is creating a function that accurately reflects the desired outcome while also being easy to optimize [28, 56, 477]. Recent advancements in LLMs have opened new possibilities for automating low-level action execution by program synthesis and reward function generation. LLMs can generate code or provide guidance on designing effective rewards based on input prompts describing desired tasks or scenarios. Recent works have explored using coding LLMs [462] and free-form white-box reward code generation [463], which has shown promise in producing more interpretable, adaptable, and human aligned rewards than traditional scalar rewards.

Simulation is an essential tool for low-level action execution. It allows researchers to test and refine reward functions in a controlled environment before deploying them on physical robots or other machines [92]. Simulations allows safe faster than real time trial-and-error optimization and environment generalization. Sim-to-real transfer [378] allows learned policies to be implemented on real-world mobile robots and environments with additional complexities like imperfect actuation and non-modeled dynamics.

Learned low-level action execution by LLM derived reward functions presented by recent work [377, 378, 385, 462, 463] provides a promising and sound approach to implement high-level actions represented by natural language instructions resulting from the proposed theoretical framework based on the experimentally verified predictive environment states.

3.9 Summary

The chapter starts with stating that the existence of biological humans is a proof that creating generally intelligent agents is possible, with natural biological intelligence providing insights into the necessary constituents for artificial generally intelligent systems.

Human intelligence is a complex phenomenon characterized by extensive and versatile models of the world, supported by various cognitive processes such as sequence learning, prediction formation, memory-based predictions, processing of prediction errors, and integration of multi-modal information. The hippocampus plays a crucial role in these abilities, but human intelligence emerges from a complex interplay of multiple brain regions that enable abstract concepts, analogies, metaphors, and stories to express experiences and make sense of new information. However, the definition of general intelligence is not clear cut, with different definitions and measurements used across psychology and computer science. This thesis takes a pragmatic approach by emphasizing requirements of particular faculties instead of degree of capability.

Biological general intelligence is contrasted with AI agents. AI agents are programs that perceive their environment, create a useful state representation, and perform actions with the goal of maximizing task performance. The driving challenge is to produce rational behavior for completing tasks in line with instructions and implicit human values. An AI agent possesses sensors or information input for perceiving the environment's current state, actuators or software commands to modify the environment, specific goals or targets to achieve through actions, and can make decisions autonomously without direct human intervention. The history of AI has seen significant developments since its inception, with early pioneers like Aristotle discussing practical reasoning and Alan Turing introducing concepts such as machine learning and reinforcement learning. More recent times have witnessed a shift towards big data-driven approaches leading to advancements in various subfields like computer vision, robotics, speech recognition, multiagent systems, and natural language processing. However, AI agents still fall short of human intelligence due to their simplicity compared to biological brains characterized by complexity and versatility. They also struggle with generalizability, adaptability, common sense reasoning, and background knowledge that humans naturally possess. This thesis propose components of general-purpose AI systems required to approach faculties of biological intelligent agents.

This thesis proposes the open-vocabulary predictive world modeling framework as a core component of general-purpose mobile agents. The novel learning method based on the principle of predictive coding allows learning to generate spatially grounded and semantically rich plausible complete environment representations from partially observed states from observational experience only. The predictive world model enables versatile new research directions for state-transition modeling using explicit predictive state representations learned from observational experience. State-transition modeling integrating reasoning over commonsense world knowledge is an essential component of designing general-purpose mobile reasoning agents due to its ability to support efficient exploitation, planning, causal discovery, explainability, and continual learning through

predictive coding. The dual latent and explicit open vocabulary predictive environment states present the possibility for multimodal LLMs to do grounded spatio-semantic reasoning.

The predictive environment states facilities learning navigational patterns from experience of observed trajectories only. The chapter covers how the means to describe tasks including by multimodal prompts interpretable by multimodal LLMs, how to perform abstract planning via mental simulation leveraging commonsense world knowledge in LLMs. Additionally, efficient learning of low-level action execution by LLM-derived reward functions and mental simulation is presented.

The proposed predictive state representation has two limitations. The current latent variable generative model is implemented for a top-down 2D grid map representation of the environment. The 2D representation is adequate for mobile robot navigation operating inn planar environments like autonomous vehicles but does not encode vertical information. Future work could explore using alternative latent variable generative 3D representations like voxel grids or neural radiance fields in order to better represent spatio-semantic memories as predictive states of general 3D environments.

The proposed mental simulator approach is limited in practice by the generality and accuracy of commonsense world knowledge within LLMs, the correctness of reasoning steps done by LLMs, accuracy of goal state detection, and achieving sufficient inference speed for timely decision making. The practical investigation of the proposed mental simulation based on predictive state representations is proposed as future work.

Chapter 4

Grounded Latent Compositional Semantics as Spatio-Semantic Memories

4.1 Introduction

This chapter introduces the grounded latent compositional semantics as a principled representation for forming queryable spatio-semantic memories as a basis for predictive state representations. Additionally, latent compositional semantics is presented as a mathematical model of unconditional open-vocabulary semantic embeddings.

The chapter starts by explaining the theoretical foundation of unsupervised dense representation learning as a scalable method for discovering useful semantics from vast amounts of raw observations. The usefullness of visual similarity as an inductive bias in the form of superpixel region partitioning is presented as part of this thesis.

Next, the foundation of open-vocabulary semantics is presented as a method for assigning human interpretable semantics to machine-discovered semantics. The remainder of the chapter presents the mathematical theory of latent compositionality and the proposed sufficient similarity semantic inference method.

4.2 From Sensor Observations to Semantic Representations

4.2.1 Unsupervised Dense Representation Learning

Deep learning is recognized as the most potent modelling tool available for representation learning on unstructured data [27]. The universal approximation theorem theoretically proves that deep neural networks (DNN) unbounded in either depth [478] or width [26] can approximate any function arbitrarily well.

Progress in state-of-the-art (SOTA) performance on general computer vision tasks in the last decade has been based on supervised learning using relatively large datasets annotated with semantic information by human labelers [20]. Despite a decade of progress, arguments are made that the original promise of generalizable and robust computer vision deep learning models has not yet been achieved and that the necessity of increasing the order of magnitude of labelled data is unsustainable in practice [35, 479]. Additionally, arguments can be made that learning from top-down categorization (i.e. "what it is") from semantically vague and inconsistent human annotation could be limiting our pursuit of robust computer vision [480], and that instead learning through bottom-up association (i.e. "what it is like in a context") is more akin to how visual concepts emerge for humans as supported by cognitive science [481–484] and similar to how word embeddings are learned in natural language processing (NLP) [157, 158, 165] as well as a motivation for capsule networks [485].

In NLP, self-supervised learning on massive unlabeled datasets approximating the real-world distribution of natural language sentences [486] is recognized as the de facto approach for leveraging the universal function approximation properties of DNNs, leading to the recent breakthrough training massive language models [70, 161, 487]. The crucial component that makes self-supervised learning successful in NLP is the fact that probabilistic enumeration of possible configuration spaces of natural sentences is computationally tractable and proved to be a highly useful learning signal source [158, 488, 489].

On the other hand, in the case of computer vision, high-resolution visual images consist of millions of high-dimensional and semantically meaningless pixels, making probabilistic enumeration over all possible configuration spaces computationally intractable and therefore limit transferability of contextual predictive self-supervised approaches known to be highly successful in NLP [490]. I propose that obtaining a means to partition an image into a small set of distinct regions encoded by a set of distinct and expressive semantic visual concept embeddings, analogous to how words in sentences are represented, is a necessary first step for unifying computer vision with NLP.

This work presents a novel method inspired by transferring principles for learning word embeddings [157–159] to the image domain. I devise how to train a model to represent images as a semantically rich embedding map partitioned into distinct, coherent regions, represented by a latent visual concept embedding (ViCE), similarly to how semantically rich word embeddings are discovered for words in the context of natural sentences. Essential aspects of my method are illustrated in Fig. 1, along with an embedding map visualization. My working hypothesis is that there exists a strong analogy between how image context defines the meaning of individually semantically meaningless pixel regions and how sentence context defines the meaning of individually semantically meaningless categorical word tokens [165]. Viewing the generation of natural images as a stochastic process where a set of latent visual concepts give rise to observable pixel appearances, I formulate my method to learn the inverse mapping from observed pixels to latent visual concepts through self-supervised learning.

The contextual supervisory signal for learning word embeddings in NLP have been mentioned before as a conceptual motivator for pretext tasks for self-supervised computer vision pretraining methods [105]. However, to the best of my knowledge, my method is the first to consider learning dense visual embedding maps with the explicit intent to be used as input representations for downstream task models.

By demonstrating the feasibility of representing images in terms of a small set of regions encoded by a set of distinct semantic visual concept embeddings, similarly to how semantic words embeddings partition sentences, I contribute towards realizing tractable probabilistic enumeration of configuration spaces for images and as a practical solution to the symbolic grounding problem [491] in vision. I hope my contribution will inspire further effort towards increasing the transferability of successful probabilistic methods from NLP to the visual domain and ultimately result in a similar breakthrough in self-supervised computer vision as the one experienced in NLP.

The concept of "the thing in itself" in Kantian philosophy denotes the existence of objects as they are independent of observation. Similarly, one can view natural images perceived by a photometric sensor to be generated from a set of latent semantic visual concepts. I model this process by a model $f(x|z)$ that generates the observable sensor measurements x of semantic entities represented by a set of latent semantic concepts $C = (c^{(1)}, \ldots, c^{(K)})$.

The purpose of perceiving the world is to provide information for completing tasks. A particular task requires interpreting the world in terms of a set of task-relevant semantics. The goal of semantic interpretation methods is to learn a function f_θ to approximate the inverse mapping

$$f_\theta(\tilde{X}^{(m)}) \simeq Z \quad \forall m \in (1, \ldots, M) \tag{4.1}$$

while simultaneously discovering the set of latent semantic concepts C. The problem of finding the inverse mapping is called vision as inverse graphics [492–494].

I relate my approach to discovering semantic meanings for pixels to discovering semantic meanings for words in NLP similar to recent MIM works [143–145, 495]. Methods to learn semantically rich word embeddings [157–159] are based on co-occurrence [165] and context [161, 496] of individually meaningless tokens. Each visual concept vector c corresponds to a distinct visual concept primitive or basis vector, and visual concepts are linear combinations of these primitives. The set of concepts C is known and finite, ensuring tractable probabilistic enumeration over possible configuration akin to successful probabilistic language modeling approaches in NLP [161, 497]. I choose to demonstrate my method with the recent SOTA self-supervised learning method SwAV [3] to learn both f_θ and C, though in principle any cluster-based self-supervised method can be used. Fig. 4.2 shows an overview of my method.

Two concrete examples of semantic interpretations are image and point cloud sensor observation representations. For images or vision sensor observations, the conventional approach is to learn a mapping f_θ that predicts the same visual concept embedding map $z \in \mathbb{R}^{D \times H \times W}$ with the same spatial resolution as the input image $x \in \mathbb{R}^{3 \times H \times W}$.

For point clouds or 3D sensor observations, the conventional approach is to learn a mapping g_θ that predicts a semantic embedding $z \in \mathbb{R}^{D \times N}$ for each point in the input point cloud $P \in \mathbb{R}^{3 \times N}$, where N is the number of points in the cloud. The mapping g_ϕ is typically implemented as a deep neural network that takes the 3D coordinates of the points as input and outputs a high-dimensional feature vector for each point, capturing its semantic properties.

4.2.1.1 Superpixels: Visual Coherence as Inductive Bias

A high-resolution image contains millions of individually meaningless and mostly redundant pixels. However, it is known that training on high-resolution images is beneficial for learning to segment small objects such as poles and pedestrians [498]. Nevertheless, naively applying self-supervised representation learning methods based on vector comparison on high-resolution embedding maps is inefficient. To solve this problem, I propose to decompose the image into a small set of visually coherent regions using superpixelization [499] and apply representation learning methods to this greatly reduced set of elements. Superpixel methods like Simple Linear Iterative Clustering (SLIC) [500] reduce elements by $\mathcal{O}(1000)$, transforming an image from millions of pixels into less than a thousand regions. I choose SLIC because of advantages [501] such as more uniform

region distribution compared to graph-based methods [502]. In contrast to grid decomposition, which is the standard for ViT models [112, 154], superpixels can preserve detail by representing thin and small patches like poles as distinct regions while requiring 75% fewer elements on average with the same base element size. While in this thesis my objective is to show that even the simplest form of region decomposition is useful, it is likely that leveraging learning-based superpixelization methods [503–505] can further improve performance.

View generation and region masking. I generate augmented views for discerning the latent semantic visual concepts through photometric invariance [114] and geometric equivariance [5]. I introduce region masking as an additional augmentation for contextual invariance shown to improve performance. To generate views with different contexts, I first sample a center point $(x, y)^*$ in the image. Sampling is done in content-rich regions to better satisfy the equipartitioning of concepts assumption [3, 120] for each training batch. I found that probabilistic sampling from a Gaussian filtered Canny edge detection map[4] is a useful measure of image content. Views $\tilde{X}^{(m)}$ are generated by sampling M view centers $(x, y)^{(m)}$ around $(x, y)^*$ while ensuring a mutual image subregion exists. I generate geometrically equivariant views by first sampling a resize coefficient $\beta^{(m)}$ for each view m. β determines the size of the cropped view region as exemplified by the red and blue crop regions in Fig. 4.2. All view crops are resized to the common view size, thus enforcing the model to learn resolution invariant representations. All views are randomly flipped horizontally. All views are augmented by random color distortion and Gaussian blurring before normalization to learn appearance invariant visual concepts [114, 506, 507]. A ratio of superpixel regions is masked with noise as a means to learn robust features and alleviate the shortcut learning problem [508]. I provide the view generation algorithm as pseudocode in Appendix D.

Learning algorithm. The objective $\mathcal{L}_{cl}$ is designed to simultaneously learn the mapping function f_θ in Eq. 4.1, and optimize the distribution of latent visual concepts C. The algorithm can be viewed as an extension of SwAV [3] to the problem of learning dense embedding maps. I refer to prior work for an explanation of SwAV [3, 120, 126, 509]. The rest of this section explains the flow of a training iteration as visualized in Fig. 4.2. I provide pseudocodes in the Appendix.

A training iteration starts by partitioning an image $X^{(n)} \in \mathbb{R}^{3\times H\times W}$ with height H and width W into a superpixel region map $A^{(n)} \in \mathbb{R}^{H\times W}$, with integer values specifying every pixel's region index. Next, a set of M augmented views $\tilde{X}^{(n)} = \{\tilde{X}^{(1,n)}, \ldots, \tilde{X}^{(M,n)}\}$ and corresponding superpixel map crops $\tilde{A}^{(n)} = \{\tilde{A}^{(1,n)}, \ldots, \tilde{A}^{(M,n)}\}$ of size h and w are generated for each image. $\tilde{A}^{(n)}$ is processed to contain only mutual regions existing in all views. The learned function f_θ transforms $\tilde{X}^{(n)}$ into a normalized visual embedding

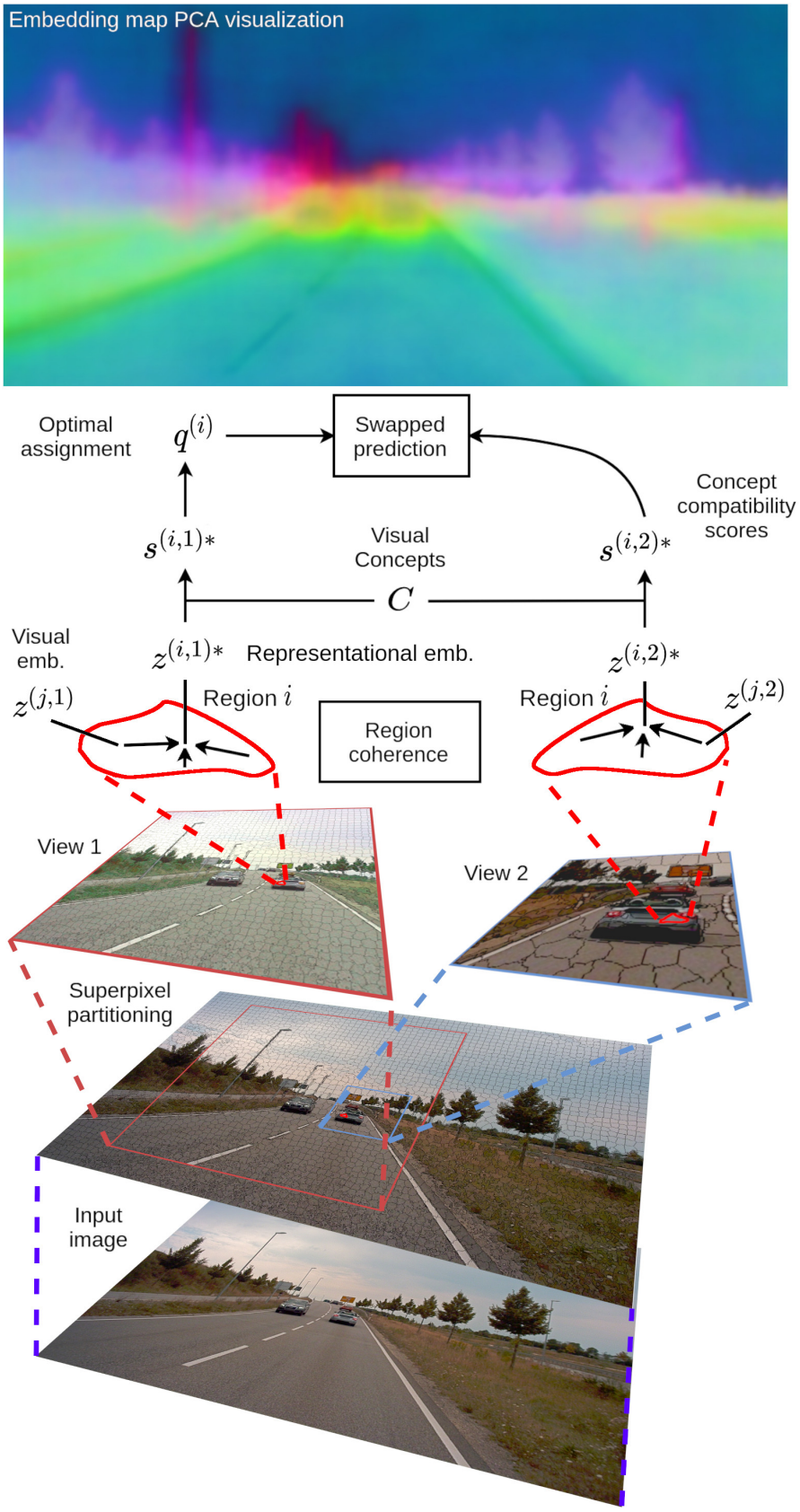

FIGURE 4.1: A hierarchical decomposition into visually coherent superpixel regions represented by a representational embeddings $z^{(i)*}$ increases the effectiveness of self-supervised methods for learning dense embedding maps. Learning $z^{(i)*}$ is posed as a swapped prediction problem [3]. All embeddings $z^{(j)}$ are optimized to equal $z^{(i)*}$ for regional coherence.

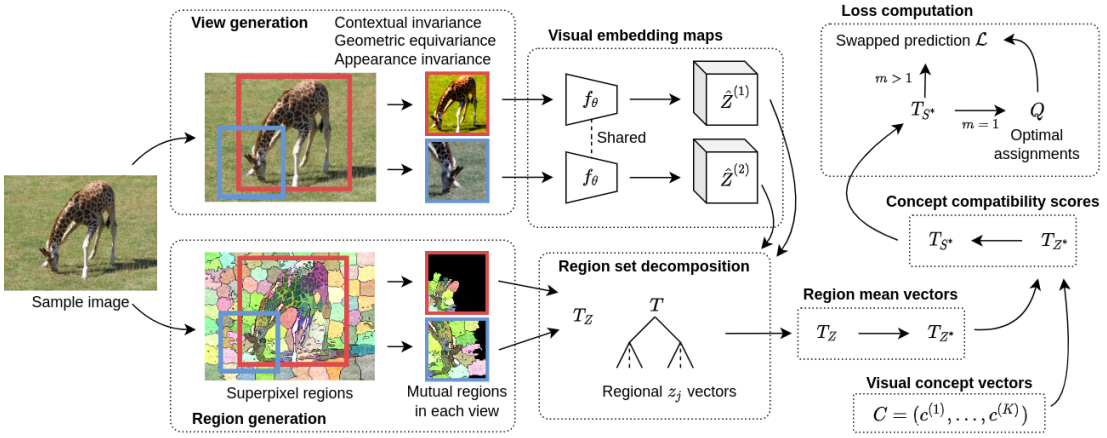

FIGURE 4.2: Overview of ViCE. A training iteration starts by generating M augmented views. First, I partition the image into I mutually common superpixel regions. The model f_θ transforms view images into visual concept embedding maps $\hat{Z}^{(m)}$. All vectors z_j are arranged in a tree structure T_Z used to conveniently organize indices of corresponding regions. A mean vector z_i^* is computed for each region. Next, I score each z_i^* in terms of closeness to each concept vector $c^{(k)}$, resulting in region-specific score vectors s_i^*.

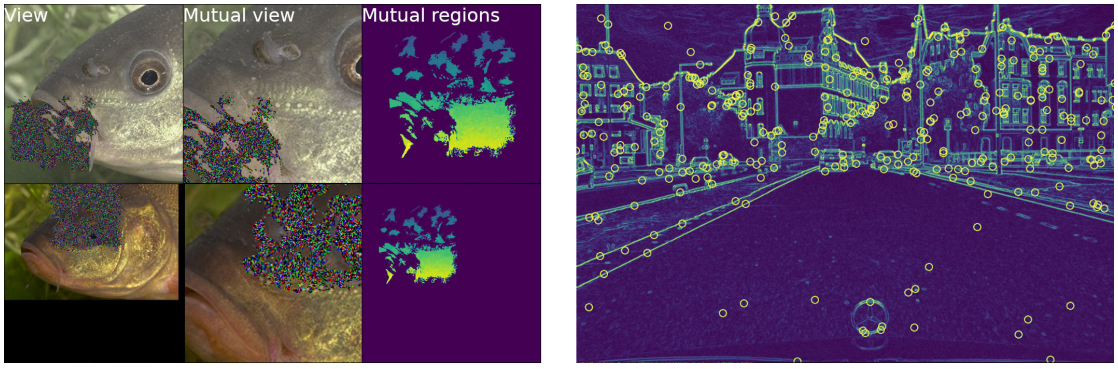

FIGURE 4.3: (Left) Examples of two generated view pairs. The first image displays the actual view feed to the model. The second image illustrates the mutual image region. The third image shows mutual superpixel regions colored by region index. (Right) View generation centers sampled from a probability mask representing image complexity measured by the Canny edge detection algorithm [4].

tensor $\hat{Z}^{(n)} \in \mathbb{R}^{D \times h \times w}$. Next $\hat{Z}^{(n)}$ is decomposed region-wise into row vectors $z_j \in \mathbb{R}^D$ and stored in a tree structure T_Z used to conveniently organize indices of corresponding regions i in view m of image n. Vectors of non-mutual regions are discarded. A single mean vector $z^{(i,m,n)*}$ is computed to represent each region i and stored in T_{Z^*}. Each vector $z^{(i,m,n)*}$ is scored in terms of compatibility or closeness to each visual concept vector $C = (c^{(1)}, \ldots, c^{(K)})$ by computing the following matrix product

$$s^* = (z^*)^T C \tag{4.2}$$

with $C \in \mathbb{R}^{D \times K}$ represented as an optimizable weight matrix. Note that the dot product

$z \cdot c$ equals the cosine distance as both vectors are normalized. All regional score vectors $s^{(i,m,n)*}$ are stored in a tree structure T_{S^*}. The concept assignments $q^{(i)}$ are determined by optimally distributing $s^{(i,m,n)*}$ uniformly over all concepts $c^{(k)}$ so that the overall compatibility between all $s^{(i)}$ and $c^{(k)}$ are maximized for regions in the primary view $m = 1$ [3]. I compute $q^{(i)}$ efficiently by the Sinkhorn-Knopp algorithm [120, 126]. A FIFO queue of accumulated $s^{(i,1,n)*}$ vectors is used to improve the empirical approximation of a uniform distribution of concepts [3, 120]. The swapped prediction learning objective [3] is

$$\mathcal{L}_{cl} = -\frac{1}{N(M-1)} \sum_{n=1}^{N} \sum_{m=2}^{M} \frac{1}{I} \sum_{i=1}^{I} q^{(i)} log\, \sigma \left(\tfrac{1}{\tau} s^{(i,m)*} \right) \tag{4.3}$$

where $\sigma()$ is the softmax function and τ is temperature. Two normalized embeddings $z^{(a)}$ and $z^{(b)}$ are compared for semantic similarity using the dot product. This operation is equivalent to comparing two word embeddings by cosine distance [157, 158].

Experiments. I implement ViCE in the self-supervised learning framework VISSL [510] based on PyTorch [511]. The quality of learned embeddings are evaluated on the COCO-Stuff164k [512, 513] reduced to 27 classes [146] and the Cityscapes [103] benchmark datasets. The reduced COCO-Stuff164k coarse dataset [146] has 118,000 train 4172 test images. The Cityscapes dataset has 2975 train and 500 test images. and I use the framework MMSegmentation [514] for evaluation and visualization. The comparative baseline for dense representation learning is the SOTA unsupervised semantic segmentation CNN model PiCIE [5] based on DeepCluster [119]. I experiment with ResNet 18 and 50 backbones [515] and two decoder architectures; the SOTA model DeepLabV3+ (DLV3+) [516] for high-resolution images, and the Feature Pyramid Network (FPN) [517] used in the baseline.

I evaluate the semantic richness and spatial accuracy of the resulting embedding maps using clustering and linear models. For unsupervised semantic segmentation I compute a set of K clusters based on output embeddings using FAISS [518]. Each cluster is greedily assigned the majority label class, or optimally assigned by the Hungarian matching algorithm [519] to cover all classes. For linear model evaluation, I train a 1×1 convolution layer without a nonlinear activation function. All models are trained and evaluated on separate train and validation sets. Note that the visual concepts learned by ViCE during training are not used for evaluation, and it is therefore fair to compare ViCE and baseline performance as long as the number of clusters is the same in both evaluation models.

I conduct experiments on 32 V100 32 GB GPUs. Each GPU loads four images, and generates five augmented views. High- and low-resolution views correspond to 512 $\times$ 512 pixels and 256 $\times$ 256 pixels, respectively. The resulting total batch size is 128 images with 640 views. To generating superpixels, I use SLIC [500] implemented in

FIGURE 4.4: ViCE learns dense semantic embeddings from raw image data. Here I visualize the output of a linear model interpreting the embeddings. The left and center images display output for low- and high-resolution images. The right image shows output from my comparative SOTA baseline PiCIE [5].

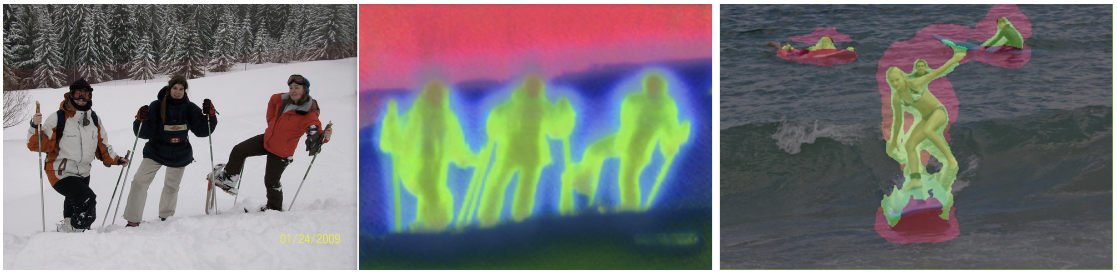

FIGURE 4.5: The center figure show output embeddings visualized in RGB colors. The right figure shows output of ViCE with the clustering-based evaluation model.

OpenCV [520] with average region size 20 px. Maximal mask coverage is 25 %. The view resize coefficients β are sampled between 0.5 to 2. The embedding dimension D and the number of visual concepts C are 128. I use the same set of hyperparameters in all experiments. Parameters for the objective $\mathcal{L}_{cl}$ are the same as SwAV [3]. The FIFO queue consists of 5K score vectors s^* per GPU. The model is optimized using the LARS optimizer [521] with weight decay 10^{-6}. The learning rate (LR) schedule is linear warmup followed by cosine decay [522, 523]. I set the peak LR using the linear LR scaling rule [524] with a base LR 0.04 for a single 4 GPU node. I initialize models with the default PyTorch pretrained weights obtained by training on ImageNet [525] for 600 epochs. However, my method can learn from random initialization as shown in Table 4.1.

Results. Table 4.1 presents results on low-resolution image experiments. C K denotes evaluation with K clusters, $\diamond$ denotes reproduced results with optimal cluster assignment, $\star$ denotes greedy assignment, and $*$ denotes ViT-based models. The best CNN-based cluster and linear model results are written in bold. Both ViCE (low-res) and PiCIE [5] use the same ResNet 18 backbone, FPN decoder, and 320×320 px image downsampling procedure for fair comparison. All ViCE models are trained for 4 epochs for COCO, and 24 epochs for Cityscapes, respectively. I trained and evaluated my PiCIE models using the official code [5]. The high-resolution and overclustered model achieves SOTA results on Cityscapes, and on COCO for convolutional models. The

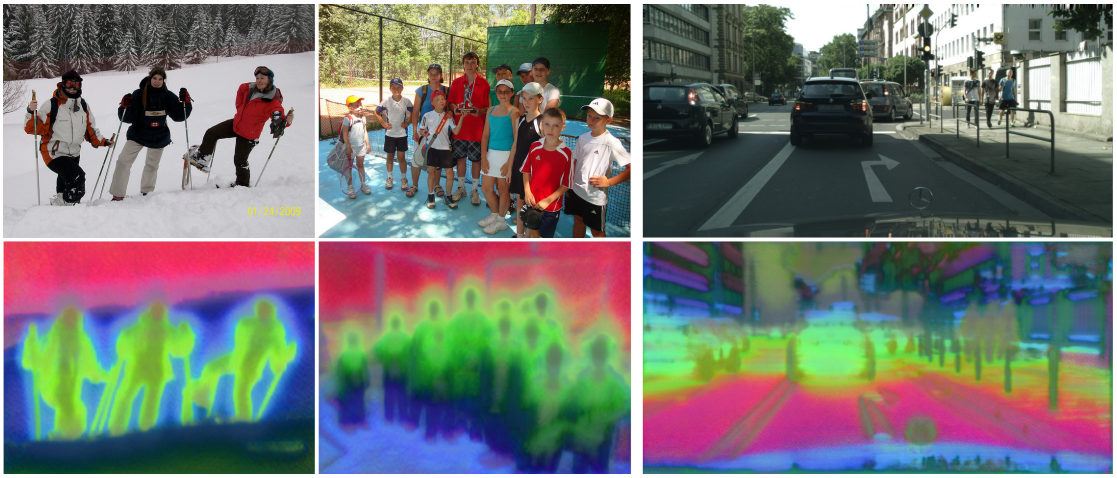

FIGURE 4.6: Dense embedding maps visualized as RGB images.

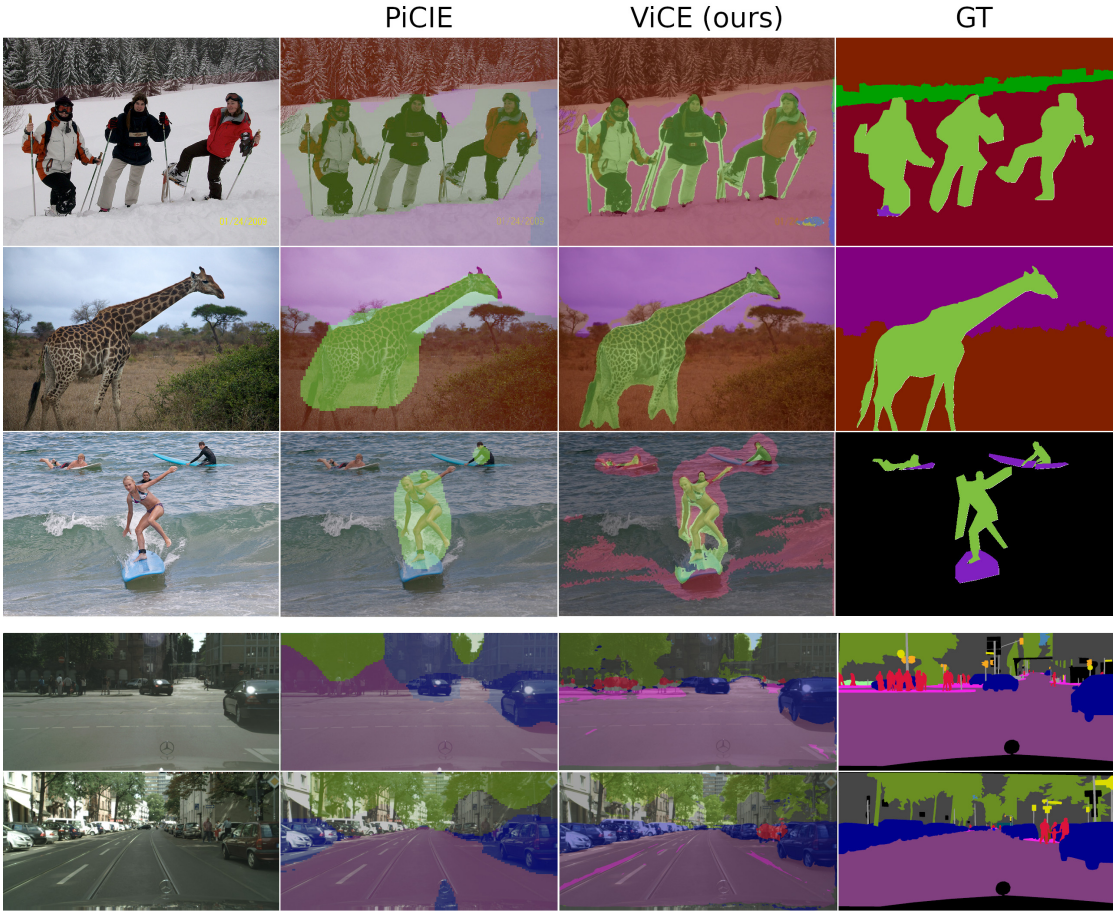

FIGURE 4.7: Output cluster visualizations on COCO (top) and Cityscapes (bottom).

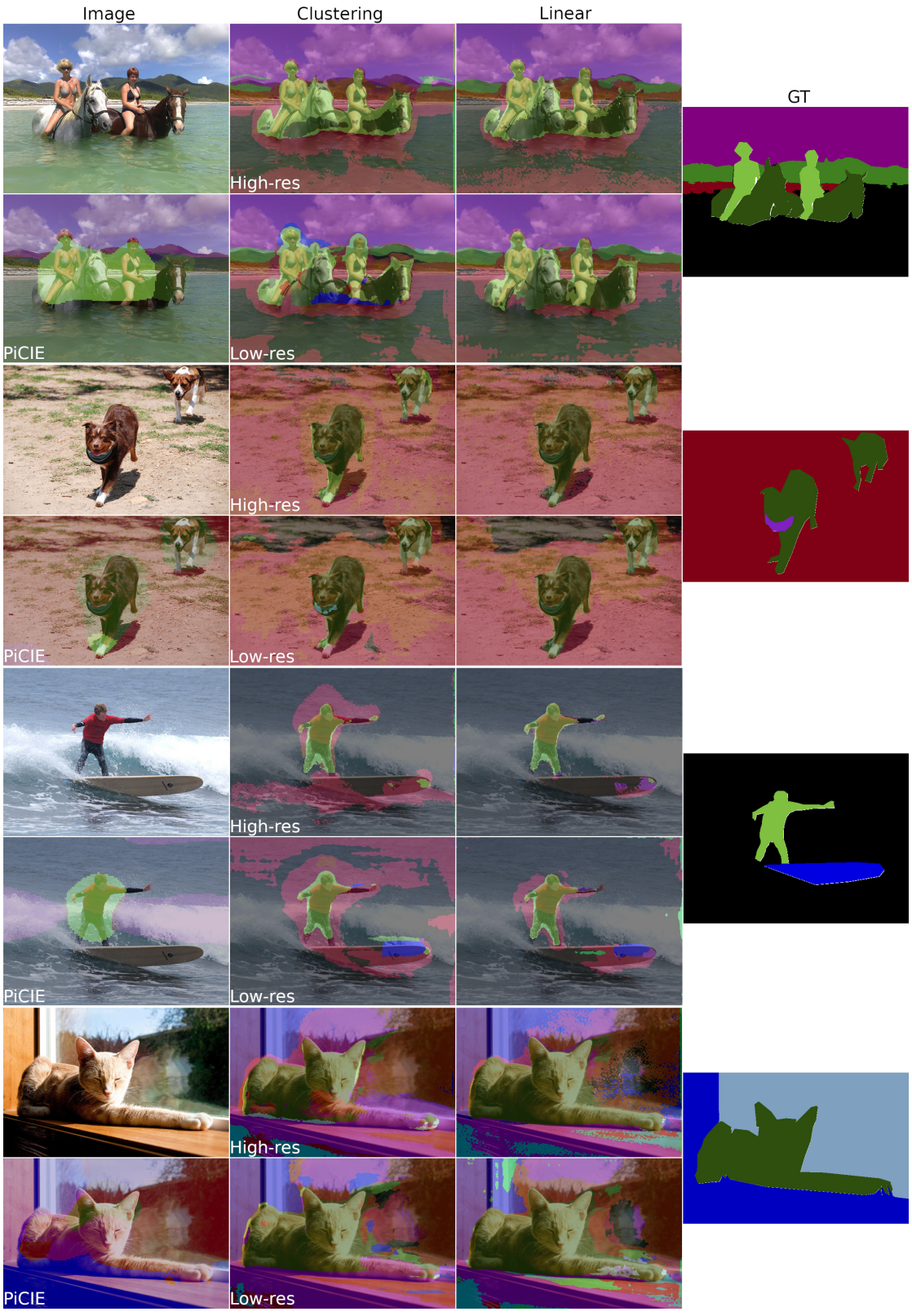

FIGURE 4.8: Output visualizations of cluster and linear evaluation models trained on low- and high-resolution COCO images.

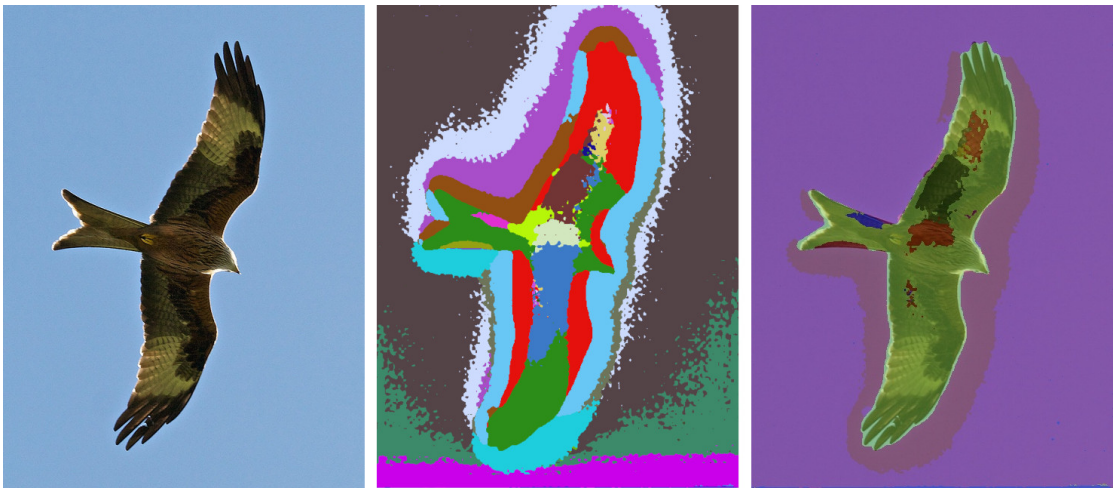

FIGURE 4.9: Visualization of output clustering. The center image shows clusters with random colors. The right image shows how clusters are mapped to semantic classes.

generic image COCO results show that ViCE is adept at discovering concepts using overclustering [527]. I believe this property stems from online clustering being more stable than offline clustering methods [3, 122]. The Cityscapes results show ViCE improving on PiCIE in all experiments. ViCE performs better than the SOTA ViT-based model STEGO [156] on Cityscapes with high-resolution and overclustering. I trained the best high-resolution C 256* COCO model in 64 h and the equivalent PiCIE model in 52 h. Fig. 4.4 and 4.9 shows clustering output visualizations. Table 4.2 shows that the best high-resolution models improves on the best low-resolution models evaluated on high-resolution images. Note that effectively training on high-resolution images is made possible by superpixelization.

The upper section of Table 4.3 provides an ablation study for low-resolution images evaluated by a linear model. The first column represents the baseline ViCE model using an RN18 backbone and FPN decoder [517] without region decomposition. The second columns indicate gains from random masking. The third and fourth column shows gains from applying grid and superpixel region decomposition. The final column indicates that utilizing the more complex DLV3+ decoder [516] is detrimental in the case of low-resolution images. I speculate this is because atrous convolutions in high-resolution decoders skip relevant neighboring information in tiny feature maps. The first column in the bottom section of Table 4.3 is empty, as learning dense embeddings for high-resolution images without superpixelization is computationally intractable. The second column showcase the radical difference in using superpixelization. The third column demonstrates the importance of utilizing a high-resolution decoder. The final column shows how superpixels are better than grids with equivalent base element sizes.

TABLE 4.1: Representation quality experiment results on low- and high-resolution images.

Model		mIoU	Acc.	Model		mIoU	Acc.
	COCO				*Cityscapes*		
ResNet50 [515]	C 27	8.9	24.60	ResNet50 [515]	C 27	-	-
MoCoV2 [526]	C 27	10.40	9.60	MoCoV2 [526]	C 27	-	-
DINO* [112]	C 27	9.60	30.50	DINO* [112]	C 27	-	-
IIC [146]	C 27	6.71	21.79	IIC [146]	C 27.	6.35	47.88
PiCIE [5]	C 27	13.84	48.09	PiCIE [5]	C 27	12.31	65.50
	C 27$^{\diamond}$	14.60	48.37		C 27$^{\diamond}$	11.85	64.29
	C 27*	9.27	38.31		C 27*	8.80	82.48
	C 128*	10.75	49.81		C 128*	7.97	56.52
	C 256*	12.42	**66.02**		C 256*	12.71	**89.86**
	Linear	14.77	54.75		Linear	-	-
PiCIE+H [5]	C 27+100	14.40	50.0	PiCIE+H [5]	C 27+100	-	-
ViCE (low-res)	C 27	11.40	28.91	ViCE (low-res)	C 27	12.81	31.87
	C 27*	11.55	50.49		C 27*	19.52	80.34
	C 128*	16.66	52.33		C 128*	21.48	81.55
	C 256*	17.98	54.92		C 256*	21.24	81.72
	Linear	25.49	62.78		Linear	**31.55**	86.33
				No pretrain	Linear	24.84	82.99
ViCE (high-res)	C 256*	**21.77**	64.75	ViCE (high-res)	C 256*	**25.23**	84.28
	Linear	**29.38**	**68.16**		Linear	30.40	**87.0**
STEGO* [156]	C 27	28.20	56.90	STEGO* [156]	C 27	21.00	73.20
	Linear	41.00	76.10		Linear	-	-

TABLE 4.2: Performance of best models trained on high- and low-resolution images

Dataset	Resolution	Configuration	Cluster mIoU	Linear mIoU
COCO	Low	RN50, FPN	19.37	27.63
	High	RN50, DLV3+	**21.77**	**29.38**
Cityscapes	Low	RN18, FPN	21.48	**31.55**
	High	RN18, DLV3+	**25.23**	30.40

In Table 4.4 I show how ViCE benefits when learning from a large general visual domain. Training on COCO and evaluating on Cityscapes with a linear model increases performance from 30.40 to 34.14 (+3.74) mIoU by improving the distinctiveness of complex classes like "Traffic sign". My findings show that general vision models can learn more useful features compared to narrow vision models even when applied in the narrow domain. The recent SOTA model STEGO [156] similarly uses a backbone trained on ImageNet only.

Fig. 4.6 visualizes dense embedding maps to demonstrate how ViCE discovers distinct

TABLE 4.3: Representation quality ablation study on low- and high-resolution images.

	Low-resolution Cityscapes				
	FPN 1px	Masking	Grid 10px	Super 10px	DLV3+
mIoU	29.66	30.42	31.30	**31.55**	11.56
Time	34h 4min	31h 6min	5h 31min	**5h 31min**	5h 37min

	High-resolution Cityscapes			
	FPN 1px	FPN super 20px	DLV3+ grid 20 px	DLV3+ super 20px
mIoU	-	8.98	25.53	**29.38**
Time	92h 20min (est.)	4h 55min	10h 1min	**6h 16min**

TABLE 4.4: Domain generalization performance

Training data domain	Evaluation data domain	mIoU	aAcc
Cityscapes	Cityscapes	30.40	**87.00**
COCO	Cityscapes	**34.14**	86.10

semantic visual entities or concepts from natural images without human supervision or proposals heuristics [135, 141]. For example, persons are represented differently from the ground surface, and human faces and bodies are semantically similar. I visualize embedding maps by PCA dimensionality reduction [528] and scale each z to the RGB range.

4.2.2 Open-vocabulary Semantic Segmentation

As explained in Sec 4.2.1, unsupervised dense representation learning models discovers semantics from commonalities in visual appearance within large sets of vision data. The discovered semantics in the resulting embedding map fits image boundaries and allows discriminating useful objects after identifying the semantic representation of said object by linear modeling or clustering [102]. The resulting semantic information contained in these embedding maps is rich but often remains implicit until it can be mapped into another embedding space grounded human world. The human world grounding problem is important to facilitate human-machine communication, including understanding human provided instructions and allow humans to understand the machine's decision process and possibly output itself. Additionally, grounding an agent's semantics allows for connecting the perceived environment with existing world knowledge recorded in human written textual information.

Open-vocabulary semantic segmentation is a computer vision task that aims to learn dense maps of open vocabulary semantic embeddings. Conventional closed set semantic segmentation models which maps pixels to one basis vector $\mathbf{e}_k$ in a $\mathbb{R}^K$ dimensional embedding space. In contrast, open-vocabulary semantic segmentation models instead maps pixels to a point on a unit hypersphere spanned by a fixed set of D orthogonal basis vectors $\mathbf{e}_1 \dots \mathbf{e}_D$ representing primitive latent semantics. Open vocabulary semantic embeddings are generally distributed over all basis vectors, and the cosine similarity of two embeddings specify their semantic similarity

$$\text{sim}(h_1, h_2) = \frac{h_1 \cdot h_2}{||h_1||||h_2||} = (h_1)^T h_2. \tag{4.4}$$

Segmentation models typically bootstrap learning by finetuning unsupervised models on supervised data containing annotations of human world semantics. As such open vocabulary semantic segmentation models addresses the challenge of mapping machine-discovered semantics with the human world semantic embedding space. Once mapped, the open vocabulary semantic embedding maps enable querying semantics as spatially precise regions in images. The mapping is essential, as it allows computer vision systems to semantically interpret sensor observation in a manner that connects with human interpretable decision making processes and world knowledge.

While allowing for querying of any semantics, the models are nevertheless trained on a closed set of semantics and example images. Open vocabulary models should therefore not be considered truly open as in allowing query any semantics, but open in the sense that they readily allow representation and training of very large and diverse set of classes in a common embedding space. Creating boundary fitting dense semantic annotations consumes far more effort than creating caption or object detection annotations. Open world semantic segmentation modeling therefore faces challenges such as limited training data for novel object classes that result in lower recall rates for these objects.

Promising directions for future research include enabling open-vocabulary capabilities on other scene understanding tasks, unifying open vocabulary detection and open vocabulary segmentation, and using multi-modal large language models to enhance perception abilities by incorporating user intent reasoning within a linguistic context. Additionally, exploring means to reduce computational cost and inference time is important to enhance usability of open vocabulary segmentation models for practical real-world general-purpose mobile agents.

4.2.3 Spatially Grounded Semantics

Semantic information alone does not enable embodied agents to accomplish physical tasks. By grounding semantics, information describing "what" and "where" is unified to represent the semantics of spatial locations in the environment. The process involves creating a spatial representation and map a semantic interpretation of the external environment as it is perceived by sensor observations. The resulting spatio-semantic environment representation consisting of grounded semantics facilitates spatio-semantic reasoning by the agent.

Another perspective is that the gap between abstract task specifications and the external environment is closed by grounding semantics perceived by the agent. Typically tasks are described using natural language, or formal symbolic specifications, that may not directly correspond to an agent's sensory inputs or actuators. By grounding semantics in spatial representations, a connection between these abstract representations and the physical reality of the agent is established.

The process of grounding semantics in spatial representations typically consists of several steps. First, the agent perceives the environment by multimodal sensors such as cameras and lidars. Generally passively sensing cameras provide the best information to infer semantics of the environments, while active sensing lidars captures an accurate measure of the spatial structure of the environment. Secondly, the semantic representation is mapped onto the spatial representation by projection or equivalent techniques [529]. The resulting information can be represented as a semantic point cloud that jointly provide a means to query semantic information with spatial precision. Finally, the resulting spatio-semantic representations are projected into a common vector space by simultaneous localization and mapping (SLAM) [530–532]. SLAM works by computing the translation and rotation transformation to optimally match sequential point clouds. Knowing the transformation allows accumulation of point clouds in a common reference frame or vector space accumulated over a sequence of observations.

From the other direction, task reasoning and planning involve mapping abstract task specifications to action sequences executable in the physical world while considering spatial constraints and affordances of the environment. This mapping is enabled by the grounded semantic representation of knowing precisely "what" is "where", as well as the 3D geometric extent of objects. The spatio-semantic representation supports monitoring how an action execution is proceeding by the same semantic and spatial query mechanism, and thus bridge the gap between abstract task specifications, conceived plan, and the embodied action execution in the physical world.

Grounding semantics in spatial representations is particularly important for general-purpose mobile robot agents, such as service robotics, where robots must understand and operate within dynamic environments and leverage a priori unknown semantics in order to accomplishing tasks specified by humans. A general spatio-semantic memory with rich semantics grounded in a precise 3D spatial location is a principled representation for general-purpose robots to interpret and execute tasks in the real world requiring versatile semantic understanding, precise spatial representation of objects, as well as a spatio-semantic memory of previously seen objects.

This thesis propose spatio-semantic memories as a means to unify conventional robotics primarily relying on map-based representations for navigation and action planning, and recent vision-language models (VLMs) which operate primarily based on ungrounded semantic information about the currently perceived environment (e.g. a forward-facing camera image)

4.3 Latent Compositional Semantics

In this section, I first present the idea of compositional semantics, and how a single vector $z^* \in \mathbb{R}^D$ implicitly represents a diverse set of semantic object descriptions $\mathcal{Z}$. In Sec. 4.3.2-4.3.3 I derive properties of compositional semantic embeddings z^* for uniform and non-uniform embedding distributions based on mathematical analysis of high-dimensional hyperspheres. Finally, in Sec. 4.3.5.2 I analyze practical discoverability of compositional semantics in real world VL embeddings spaces by iterative gradient descent.

This section presents an investigation into latent compositional semantics as a means to compactly represent objects by rich semantic descriptions within explicit environment representations. I prove that mathematical properties of high-dimensional hyper-spheres enable a single compositional semantic embedding z^* to define a set of semantic text descriptions encoded into semantic embeddings $\mathcal{Z} = \{z^{(1)}, z^{(K)}\}$. The experiments verify that a single embedding z^* can robustly represent 10 semantically related real-world embedded text descriptions, and up to 100 randomly sampled embeddings for ideal uniformly distributed embedding spaces. Based on my knowledge, this thesis propose a new perspective on unconditioned dense VL embedding prediction models [6] as a scalable, robust, and learnable neural approximations of semantic networks [204] for knowledge representation.

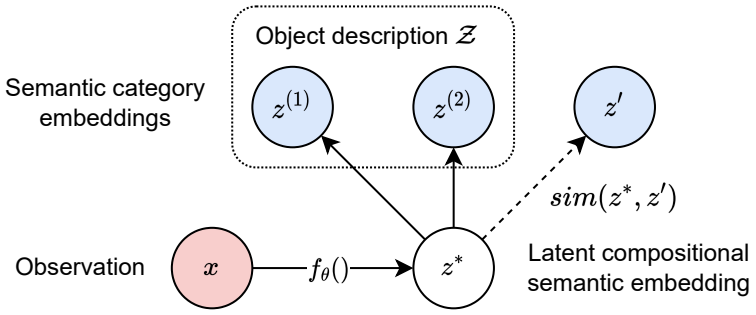

FIGURE 4.10: The compositional semantics framework. An observation x is mapped into an embedding z^* that specifies an object description $\mathcal{Z}$ in terms of interpretable semantic categories $z^{(k)}$ through fuzzy membership by similarity.

4.3.1 Compositional Object Representations

Knowledge representations aim to describe concrete objects by membership to abstract semantic categories. Semantic networks are a common object description representation encumbered by practical limitations. I propose compositional semantics as an efficient and practical vector space representation for describing objects by a potentially large set of semantic categories by a scaleable learning-based method. Compositionality means that complex expressions, such as sentences or functions, can be determined or understood based on the meanings of their individual parts [533].

My proposed framework for compositional semantics is shown in Fig. 4.10. The objective is to find a hyperspherical latent compositional semantic embedding $z^* \in S^{D-1}$ for an object which is similar to all semantic embeddings in the set $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(K)}\}$ that broadly describe the object. Semantic similarity is defined in terms of separation distance in the embedding space S^{D-1}. Distances between embeddings on unit hyperspheres in Euclidean vector spaces are conveniently represented by cosine similarity

$$\cos\omega = \frac{\langle z^*, z^{(k)} \rangle}{||z^*||\ ||z^{(k)}||} = (z^*)^T z^{(k)}. \tag{4.5}$$

An observation x is mapped into a compositional semantic embedding z^* discovered by a learned one-to-one mapping function $f_\theta(x)$. The optimal embedding z^* is found by maximizing the mean cosine similarity (4.5) over all describing semantics $z \in \mathcal{Z}$. I presume the distribution $p(z)$ is approximate uniformly distributed. In this thesis I show that contrastive optimization by minimizing (4.5) over negative samples z' is not required if $\mathcal{Z}$ is known. The mathematical properties of high-dimensional hyperspheres ensure that any other randomly sampled embeddings is very likely to be dissimilar to z^*. The optimal compositional semantic embedding z^* thus separates the set of describing semantics $\mathcal{Z}$ from all other semantics $z' \sim U(S^{D-1})$. Observations x denote any observable representation including image pixel regions.

During inference, the representation z^* for an observation x, implicitly encodes $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(T)}\}$ concatenated from past independent learning samples $(x^{(t)}, z^{(t)})$. From the perspective of knowledge representation, z^* implicitly encodes the degree of membership for any queried semantic z by semantic distance or equivalently cosine similarity (4.5) :

$$MemberOf(x, z) \propto sim(z^*, z) \;\; z^* := f_\theta(x). \tag{4.6}$$

The set of inferred $\hat{\mathcal{M}}$ and original $\mathcal{M}$ set of object descriptions are approximately equal

$$\hat{\mathcal{M}} = \{MemberOf(x, \hat{z}) \mid \hat{z} \in \hat{\mathcal{Z}}\} \tag{4.7}$$

$$\mathcal{M} = \{MemberOf(x, z) \mid z \in \mathcal{Z}\} \tag{4.8}$$

$$|\hat{\mathcal{M}} \cup \mathcal{M}| \simeq |\mathcal{M}| \tag{4.9}$$

as the set of inferred sufficiently similar semantic description embeddings are approximately equal

$$\hat{\mathcal{Z}} = \{\hat{z} | sim(z^*, \hat{z}) > \tau \;\; \forall \hat{z} \in S^{D-1}\} \simeq \mathcal{Z}. \tag{4.10}$$

The degree of membership by similarity (4.6) reflects the fact that real world objects rarely have a single, clear-cut semantic specification [212, 213]. The threshold of sufficient semantic membership τ is subjective and needs to be optimized in respect to a purpose or task [214]. Note that the mapping $f_\theta(x)$ discovers z^* from independent samples $(x^{(t)}, z^{(t)})$ by iterative gradient descent.

4.3.2 Compositional properties of Uniformly Distributed Semantics

VL embeddings are typically located on the surface of a high-dimensional unit hypersphere. In this section I analyse the compositional properties of VL embeddings spaces based on mathematics for high-dimensional probability distributions [534].

I begin the analysis by formally defining latent compositional semantic embeddings z^*.

Definition 4.1. A vector $z^* \in \mathbb{R}^D$ on the unit hypersphere S^{D-1} is a compositional semantic embedding for a set of semantic embeddings $z \in \mathcal{Z}$ if

$$\mathbb{E}\, sim(z^*, z) > \mathbb{E}\, sim(z^*, z') \quad \forall z \in \mathcal{Z}, z' \sim U(S^{D-1}) \tag{4.11}$$

where $U(S^{D-1})$ is the uniform distribution over S^{D-1}.

The following theorem specify the theoretically optimal z^* embedding is simply a centroid.

Theorem 4.2. *[Discoverability I] It is always possible to find the optimal compositional semantic embedding* $z^* \in \mathbb{R}^{D \gg 1}$ *satisfying Definition 4.1 as the centroid of the set of semantics* $\mathcal{Z}$

$$z^* = \frac{1}{K} \sum_{i=1}^{K} z^{(i)} \;\; \forall z^{(i)} \in \mathcal{Z}. \tag{4.12}$$

Proof. See Appendix B.3. □

The proof is based on finding the z^* maximizing cosine similarity by partially differentiating the equivalent minimum square distance.

A property of high-dimensional vector spaces is that any two random variable vectors are expected to be approximately orthogonal. The following lemma is used to prove Theorem 4.2

Lemma 4.3. *[Expected similarity] For two independent random vectors* $Z^{(i)}$, $Z^{(j)}$ *sampled from an isotropic high-dimensional distribution* $Z \in \mathbb{R}^D$ *with* $D \gg 1$

$$\mathbb{E}\, sim(Z^{(i)}, Z^{(j)}) = \frac{1}{\sqrt{D}}. \tag{4.13}$$

Proof. See Appendix B.1 □

The proof involves recognizing Z as an isotropic distribution and computing the expectation of a dot product for two random vectors $Z^{(i)}$ and $Z^{(j)}$.

Next I derive a probabilistic bound defining the separability of a set $\mathcal{Z}$ of object descriptions and random descriptions z' by similarity with the latent compositional semantic embedding z^* for $\mathcal{Z}$.

Theorem 4.4. *[Probabilistic bound] The probability* P *a compositional semantic embedding* z^* *is more similar to all its semantic members* $z \in \mathcal{Z}$ *than any unrelated semantic embedding* $z' \sim U(S^{D-1})$ *is*

$$P\left(sim(z^*, z) > sim(z^*, z')\right) = 1 - \tfrac{1}{2} I_{\sin^2(\theta_{min})}(\tfrac{D-1}{2}, \tfrac{1}{2}) \tag{4.14}$$

where $I_x(a, b)$ *is the regularized incomplete beta function and*

$$\theta_{min} = \arccos(sim(z^*, z_{min})) \tag{4.15}$$

is the angle θ_{min} *defined by the least similar member*

$$z_{min} = \arg\min(sim(z^*, z)) \; \forall z \in \mathcal{Z}. \tag{4.16}$$

Proof. See Appendix B.4 □

The proof is based on noting that the probability P a randomly sampled unrelated embedding z' falsely in the set of semantic members $\mathcal{Z}$ is proportional to the area ratio of the hyperspherical cap S_{cap}^{D-1} spanned by z^* and z_{min}. The proof builds upon Lemma 4.3 and 4.5.

Lemma 4.5. *[Hyperspherical cap] The compositional semantic embedding z^* and all semantic member embeddings $z \in \mathcal{Z}$ lie in a hyperspherical cap S_{cap}^{D-1}*

$$\{z^*\} \cup \mathcal{Z} \in S_{cap}^{D-1} = \{z \in \mathbb{R}^D : \|z\| = 1, \theta_z \leq \theta_{min}\}. \tag{4.17}$$

Proof. See Appendix B.2 □

I conclude that latent compositional semantic embeddings z^* can always be found for VL embeddings. The goodness of z^* can be measured by the probabilistic estimate (4.14)

4.3.3 Compositional Properties of Open-Vocabulary Semantics

The mathematical properties for latent compositional semantic embeddings z^* in Sec. 4.3.2 are derived for uniformly distributed embeddings. In this section, I analyze the validity of the results for non-uniform hyperspherical distributions.

Proposition 4.6 (Discoverability II)**.** *It is always possible to find an optimal compositional semantic embedding $z^* \in \mathbb{R}^D$ for any non-uniform distribution $z \in p(z|z \in \mathbb{R}^{D>1}, \|z\| = 1)$ that is not singular.*

Proof. See Appendix B.5 □

The proof is based on showing that Definition 4.1 holds also when expected similarity is higher than for uniformly distributed embeddings spaces as given by Lemma 4.3.

The shape of the non-uniform density $p(z)$ of common VLMs is a product of optimization by contrastive learning with random negative sampling [77]. Few general properties can be inferred for non-uniform densities. So et al.[535] finds that vision and text CLIP embeddings are distributed in separate modality-specific hyperspherical caps. Wang et al. [536] identifies the uniformity-alignment dilemma stating that perfect uniformity and alignment cannot be simultaneously achieved due to semantically similar but randomly sampled false negatives.

I found that using the probabilistic bound (4.14) for highly non-uniform VL embedding densities $p(z)$ results in poor estimates. The reason is that unrelated embeddings are far more similar than those for uniform distributions. Instead I propose a statistical sampling-based approach to obtain a probabilistic estimate for (4.11) in Definition 4.1 without requiring to estimate the non-uniform density $p(z)$. The probability in (4.14) is estimated by sampling N random semantic embeddings $z \sim p(z)$ and counting the number of samples being within the hyperspherical cone S^{D-1}_{cap} spanned by z^* and z_{min} (4.17) such that

$$P\left(sim(z^*, z) > sim(z^*, z')\right) \simeq \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{S^{D-1}_{cap}}(z^{(i)}). \tag{4.18}$$

The empirical results show that latent compositional semantic embeddings z^* are useful for all tested non-uniform VL embedding distributions. Additionally, the empirical estimate (4.18) provides an accurate measure of goodness.

4.3.4 Sufficient Similarity Inference

Conventional semantic segmentation presume an input image can be sensibly partitioned into a set of K fixed hand-crafted semantic classes $\mathcal{E}_K$. Each class k is represented by a one-hot embedding $e^{(k)} \in \mathcal{E}_K$. The embeddings $\mathcal{E}_K$ span different dimensional axes on the positive quadrant of the unit hypersphere S^{K-1}. The partitioning is computed by assigning class k^* represented by the most similar embedding $e^{(k)}$ to each predicted embedding $\hat{z}$

$$k^* = \arg\max_k \left[sim(\hat{z}, e^{(k)})\right] \quad \forall e^{(k)} \in \mathcal{E}_K. \tag{4.19}$$

Open world semantic segmentation likewise partition the image by assigning the most similar semantic k^* in a set of word semantics $\mathcal{Z}_K$ distributed over S^{K-1}. The semantics of $\mathcal{E}_K$ defines the orthogonal basis of S^{K-1} and thus limit queryable semantics to $\mathcal{E}_K$. In contrast, learning word semantics results in a semantically meaningful orthogonal basis, allowing any $\mathcal{Z}_K$ to be defined and queried at inference time.

Boyi et al.[6] identifies two weaknesses of the most similar partitioning approach: First, any object such as a *window-on-a-building-facade* can both be described as a "window" as well as part of a "building" at a higher-level. Hard partitioning by highest similarity haphazardly predicts one or the other. Secondly, hard partitioning assigns a semantic to every image element even if all queried semantics have low similarity with the image content. An example is a dog queried by the two semantics "grass" and "toy" is interpreted as "toy". The use of abstract word semantics like "other" as a substitute for unspecified semantics is not a principled solution as there is no guarantee that the

similarity between z^* and queried but unrelated semantic $z^{(k)}$ is less similar than the ambiguous semantic meaning of "other"

$$sim(z^*, z_{other}) \overset{?}{>} sim(z^*, z^{(k)}) \;\; \forall z^{(k)} \in \mathcal{Z}_K. \tag{4.20}$$

I propose sufficient similarity as a principled inference method that allows semantic overlap and empty query results by a single compositional semantic embedding z^*. To evaluate semantic membership by sufficient similarity, I first compute a set of similarity threshold values $T = \{\tau_1, \ldots, \tau_K\}$ for each known semantic $k \in \{1, \ldots, K\}$. The value of τ_k is found by maximizing the likelihood that $sim(z^*, z^{(k)}) > \tau_k$ for true elements in past observations. At evaluation time, instead of selecting the most similar semantic k^* in (4.19), any similarity with semantic $z^{(k)}$ higher than the threshold τ_k are deemed sufficiently similar to be a member of the semantic group k

$$sim(z^*, z^{(k)}) > \tau_k \Rightarrow MemberOf(z^*, k). \tag{4.21}$$

I view (4.21) as a practical probabilistic approach for finding the mathematically derived hyperspherical cap S^{D-1}_{cap} (4.17) defining the membership set $\mathcal{Z}$ (4.10) that maximizes the likelihood over past observations. For simplicity, I estimate a single maximum likelihood value τ_k for each semantic k by a logistic regression model. To fit the model, a set of similarity values $sim(z^*, z)$ are sampled from positive and negative elements of k using annotations y. The optimal τ_k is the the decision boundary or $sim(z^*, z)$ value that best separates positive and negative elements according to the model

$$p\Big(MemberOf\left(sim(z^*, z), k\right)\Big) = 0.5. \tag{4.22}$$

However, the method is not fundamentally limited to estimating only single constant values τ_k. To the best of my knowledge, the similarity thresholding method proposed by Cui et al. [537] is closest to my approach. While Cui et al. uses thresholding for uncertainty estimation, I propose thresholding to determine category membership (4.6).

4.3.5 Latent Compositional Semantics from Data

4.3.5.1 Discovery from Semantic Sets

In the following sections I set out to verify the properties and discoverability of latent compositional semantic embeddings z^* derived in Sec. 4.3.1-4.3.3. I perform experiments on embedding spaces for four representative models: the VLMs CLIP [77], OpenCLIP [395], X-Decoder [193], and the language model SBERT [179]. Additionally, I do

experiments on ideal uniformly distributed embedding spaces $U(S^{D-1})$.

Experiments 1. The first set of experiments investigates the lower bound capacity for z^* to represent an arbitrary set of K randomly sampled VL embeddings. I estimate the lower bound capacity of z^* by sampling K embeddings z forming an object description set $\mathcal{Z}$ of random semantics. Next I compute the optimal z^* by (4.12) and measure the separation between 100,000 randomly sampled embeddings Z' and the set Z represented by z^*. Separability is measured by (4.18) approximating (4.14) for uniform and nonuniform distributions. High separability means it is highly unlikely any non-related random semantic is closer to z^* than the least close related semantic $z_{min} = \arg\min(Z)$. In other words, z^* has high cosine similarity (4.5) only with semantics z of the object description $\mathcal{Z}$. See Fig. 4.10 for a visualization.

To generate embeddings, I sample words from the English lexical database WordNet [538]. Sampled words gets transformed into a semantic embedding z by the models' language encoders. Ideally distributed embeddings are sampled uniformly on the hypersphere $U(S^{D-1})$. CLIP experiments use the largest available *ViT-L/14@336px* model generating 768 dimensional embeddings. For OpenCLIP I use the largest *ViT-bigG-14* model, pretrained on the *laion2b_s39b_b160k* dataset, generating 1280 dimensional embeddings. I use the largest available *Focal-L* model for X-Decoder outputting 512 dimensional embeddings. SBERT uses the *all-mpnet-base-v2* model generating 768 dimensional embeddings. I measure performance of object descriptions $\mathcal{Z}$ of varying length K to estimate maximum representation capacity of z^* for each embedding space. Two additional experiments for higher dimensional embeddings explore the theoretical limits of z^* for large object descriptions $\mathcal{Z}$. Each experiment is repeated 1000 times for statistical estimation.

Results 1. Here I provide results and findings for z^* representing sets $\mathcal{Z}$ of randomly sampled semantics z. Table 4.5 shows the expected similarity between optimal z^* (4.12) and object description semantics z is always higher than for unrelated semantics z'. The results verifies that z^* for all embedding distributions and object description sizes K satisfy Definition 4.1 and Theorem 4.2 for finding the optimal z^*. Table 4.6 shows lower bound separability of related $z \in \mathcal{Z}$ and non-related semantics $z' \in \mathcal{Z}'$ by z^*, verifying Theorem 4.4. All embedding spaces allow reliable separability for small object descriptions $K \leq 3$, verifying Proposition 4.6 for finding z^* for non-uniform distributions. For intermediate descriptions $K \leq 5$ separability of CLIP embeddings reduces to chance. SBERT maintains strong separability. Only ideal uniform distributions achieve perfect separability for large descriptions $K \leq 10$. Table 4.7 shows that sufficiently

TABLE 4.5: Compositional semantics expectation delta

	$\Delta\,\mathbb{E} = \mathbb{E}\,sim(z^*, z) - \mathbb{E}\,sim(z^*, z')$		
Distribution	$K = 3$	$K = 5$	$K = 10$
CLIP [77] [b]	0.135 (0.043)	0.083 (0.032)	0.043 (0.024)
OpenCLIP [395] [c]	0.245 (0.032)	0.156 (0.027)	0.083 (0.020)
X-Decoder [193] [a]	0.236 (0.045)	0.150 (0.037)	0.080 (0.027)
SBERT [179] [b]	0.397 (0.040)	0.273 (0.035)	0.156 (0.026)
$U(z)_{D=768}$	0.577 (0.012)	0.447 (0.010)	0.316 (0.080)
$U(z)_{D=1280}$	0.577 (0.010)	0.447 (0.080)	0.316 (0.006)
$U(z)_{D=2048}$	0.577 (0.008)	0.447 (0.006)	0.316 (0.005)
$U(z)_{D=4096}$	0.577 (0.005)	0.447 (0.005)	0.316 (0.003)

a: $D = 512$, b: $D = 768$, c: $D = 1280$

TABLE 4.6: Separation of related and nonrelated random semantics

	$P\,(sim(z^*, z) > sim(z^*, z'))$		
Distribution	$K = 3$	$K = 5$	$K = 10$
CLIP [77] [b]	0.954 (0.117)	0.533 (0.261)	0.187 (0.143)
OpenCLIP [395] [c]	1.000 (0.001)	0.907 (0.115)	0.400 (0.180)
X-Decoder [193] [a]	0.990 (0.0223)	0.750 (0.1682)	0.301 (0.156)
SBERT [179] [b]	1.000 (0.002)	0.977 (0.043)	0.647 (0.188)
$U(z)_{D=768}$	1 (0)	1 (0)	1 (0)
$U(z)_{D=1280}$	1 (0)	1 (0)	1 (0)
$U(z)_{D=2048}$	1 (0)	1 (0)	1 (0)
$U(z)_{D=4096}$	1 (0)	1 (0)	1 (0)

a: $D = 512$, b: $D = 768$, c: $D = 1280$

high-dimensional uniformly distributed embedding spaces can represent very large object descriptions of size $K \leq 100$ with perfect separability. Note that largest 4096 dimension embedding space equals the ResNet output embedding map dimension [539]. The probabilistic bound (4.14) accurately predict the empirical separation probability result for uniform distributions. The bound fails for highly non-uniform distributions as expected. Figure 4.11 visualizes embedding similarity distributions for different embedding spaces and object descriptions sizes K. Figure 4.12 shows how increasing dimensionality gradually improves separability.

I find that the object description size K representable by z^* is only constrained by embedding space dimensionality D and degree of uniformity. The OpenCLIP embedding space provides better separability than the popular CLIP and SOTA multi-task optimized X-Decoder models. The pure language model SBERT has better embedding space than all VLM models. I propose to learn unconditional dense VLMs on language model embeddings instead of global description VLMs like CLIP as the pre-trained vision encoder is not used. The findings motivate further work towards increasing

TABLE 4.7: Large object description expectation delta and separation

	$K = 100$		
Distribution	$\Delta\,\mathbb{E}$	$P\,(sim(z^*, z) > sim(z^*, z'))$	
		Empirical (4.18)	Bound (4.14)
CLIP [77] [b]	0.004 (0.008)	0.011 (0.011)	1.0*
OpenCLIP [395] [c]	0.009 (0.007)	0.013 (0.012)	1.0*
X-Decoder [193] [a]	0.008 (0.009)	0.013 (0.012)	1.0*
SBERT [179] [b]	0.018 (0.009)	0.015 (0.013)	1.0*
$U(z)_{D=768}$	0.010 (0.001)	0.605 (0.148)	0.612
$U(z)_{D=1280}$	0.010 (0.002)	0.838 (0.116)	0.863
$U(z)_{D=2048}$	0.010 (0.002)	0.967 (0.040)	0.988
$U(z)_{D=4096}$	0.010 (0.001)	1.000 (0.001)	1.000

a: $D = 512$, b: $D = 768$, c: $D = 1280$, *: Error from non-uniformity

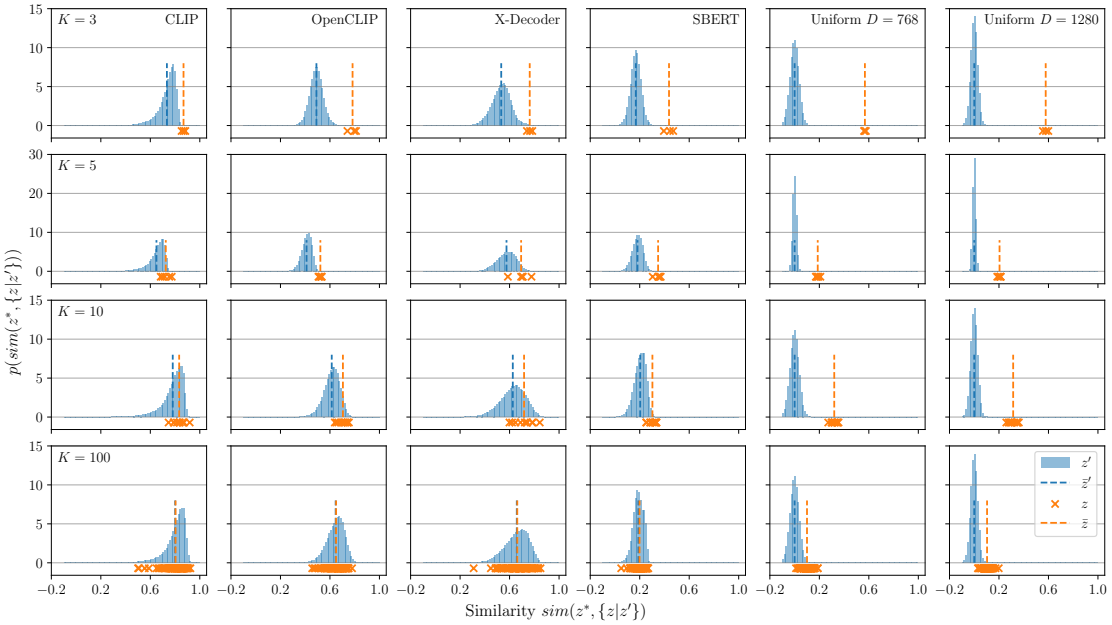

FIGURE 4.11: Similarity distributions between a latent compositional semantic embedding z^* and all object description embeddings $z \in \mathcal{Z}$ it represent (orange) and randomly sampled unrelated word embeddings z' (blue). Columns show different embedding spaces. Each row shows object descriptions of different size K. A z^* is useful if it separates the distribution of z and z' by cosine similarity (4.5).

uniformity of existing VLM embedding distributions to better leverage the capacity of high-dimensional embedding spaces and to improve discriminatability of compositional semantic embeddings [535, 540].

Here I provide results and findings for z^* representing sets $\mathcal{Z}$ of randomly sampled semantics z. Table 4.5 shows the expected similarity between optimal z^* (4.12) and object description semantics z is always higher than for unrelated semantics z'. The results verifies that z^* for all embedding distributions and object description sizes K satisfy Definition 4.1 and Theorem 4.2 for finding the optimal z^*. Table 4.6 shows lower bound separability of related $z \in \mathcal{Z}$ and non-related semantics $z' \in \mathcal{Z}'$ by z^*, verifying Theorem 4.4. All embedding spaces allow reliable separability for small object descriptions $K \leq 3$, verifying Proposition 4.6 for finding z^* for non-uniform distributions. For intermediate descriptions $K \leq 5$ separability of CLIP embeddings reduces to chance. SBERT maintains strong separability. Only ideal uniform distributions achieve perfect separability for large descriptions $K \leq 10$. Table 4.7 shows that sufficiently high-dimensional uniformly distributed embedding spaces can represent very large object descriptions of size $K \leq 100$ with perfect separability. Note that largest 4096 dimension embedding space equals the ResNet output embedding map dimension [539]. The probabilistic bound (4.14) accurately predict the empirical separation probability result for uniform distributions. The bound fails for highly non-uniform distributions as expected. Figure 4.11 visualizes embedding similarity distributions for different embedding spaces and object descriptions sizes K. Figure 4.12 shows how increasing dimensionality gradually improves separability.

I find that the object description size K representable by z^* is only constrained by embedding space dimensionality D and degree of uniformity. The OpenCLIP embedding space provides better separability than the popular CLIP and SOTA multi-task optimized X-Decoder models. The pure language model SBERT has better embedding space than all VLM models. I propose to learn unconditional dense VLMs on language model embeddings instead of global description VLMs like CLIP as the pretrained vision encoder is not used . The findings motivate further work towards increasing uniformity of existing VLM embedding distributions to better leverage the capacity of high-dimensional embedding spaces and to improve discriminatability of compositional semantic embeddings [535, 540].

Experiments 2. The second set of experiments estimates the separability for 500 realistic object descriptions consisting of related semantics. Each object description is generated by an LLM [1] and consists of K descriptive semantics including names,

[1] Claude 2 provided by Anthropic (`claude.ai`)

TABLE 4.8: Separation for realistic object descriptions

	$P\left(sim(z^*, z) > sim(z^*, z')\right)$		
Distribution	$K = 3$	$K = 5$	$K = 10$
CLIP [77] [b]	1 (0)	0.976 (0.059)	0.745 (0.192)
OpenCLIP [395] [c]	1 (0)	0.996 (0.015)	0.877 (0.146)
X-Decoder [193] [a]	1 (0)	0.998 (0.018)	0.917 (0.035)
SBERT [179] [b]	1 (0)	1.000 (0.001)	0.981 (0.056)

a: $D = 512$, b: $D = 768$, c: $D = 1280$

properties, and affordances. The results represent expected representational capacity of z^* in practical real-world application.

Results 2. Here I provide separability results for z^* representing sets $\mathcal{Z}$ of realistic object descriptions composed of related semantics z. Table 4.8 shows that realistic sets of related semantics have better separability than the lower bound of random semantic descriptions presented in Table 4.6. All VLMs achieve strong separability for $K \leq 5$, and SBERT allows large object representations of $K \leq 10$.

Figure 4.13 visualizes similarity distributions for three particular object descriptions of varying lengths K. The top row shows distributions for the short object description of a "medium-sized utility vehicle" $\mathcal{Z}_1 = \{$truck, van, vehicle$\}$. All related $z \in \mathcal{Z}_1$ are perfectly separable from the distribution of non-related $z' \notin \mathcal{Z}_1$ by z^* and θ_z given by 4.15. The middle row shows the separability of a medium sized description for a "patch on a drivable flat asphalt road with painted lane markings" $\mathcal{Z}_2 = \{$road, lane marking, drivable, asphalt, flat$\}$. All models achieve above 99 % separability. The bottom row visualizes the distribution of a large description of a "white wooden table surface" $\mathcal{Z}_3 = \{$ 'table', 'wood', 'counter', 'solid', 'surface', 'white', 'static', 'flat', 'furniture', 'static' $\}$. The VLMs do not reach reliable separability. In contrast, the language model SBERT achieves 98 % separability, demonstrating that SBERT embedding are practically useful up to about 10 semantics. I consider the results as upper bounds for visually learned representations by VLMs.

4.3.5.2 Discovery from Visual Appearance

The mathematical properties in Sec. 4.3.2-4.3.3 are derived while presuming all member semantics $z \in \mathcal{Z}$ are known. In this section I verify the possibility of finding latent compositional semantic embeddings z^* by iterative optimizing z^* one z at a time, instead of averaging the set $\mathcal{Z}$ as in (4.12).

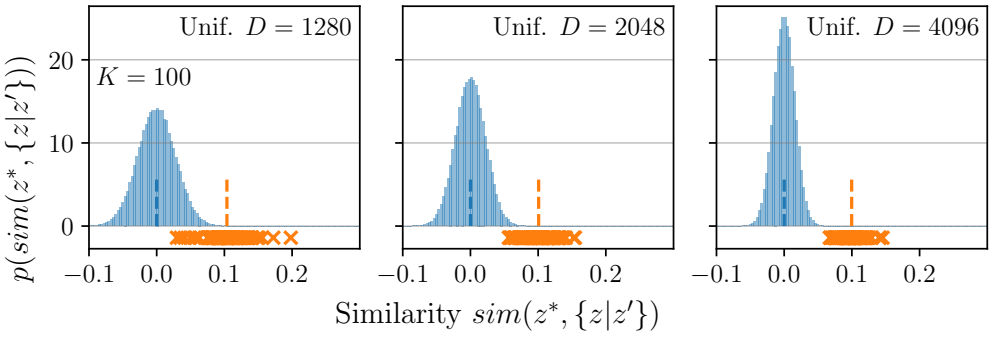

FIGURE 4.12: Similarity distributions for large object descriptions $\mathcal{Z}$ in very high-dimensional uniformly distributed embedding spaces.

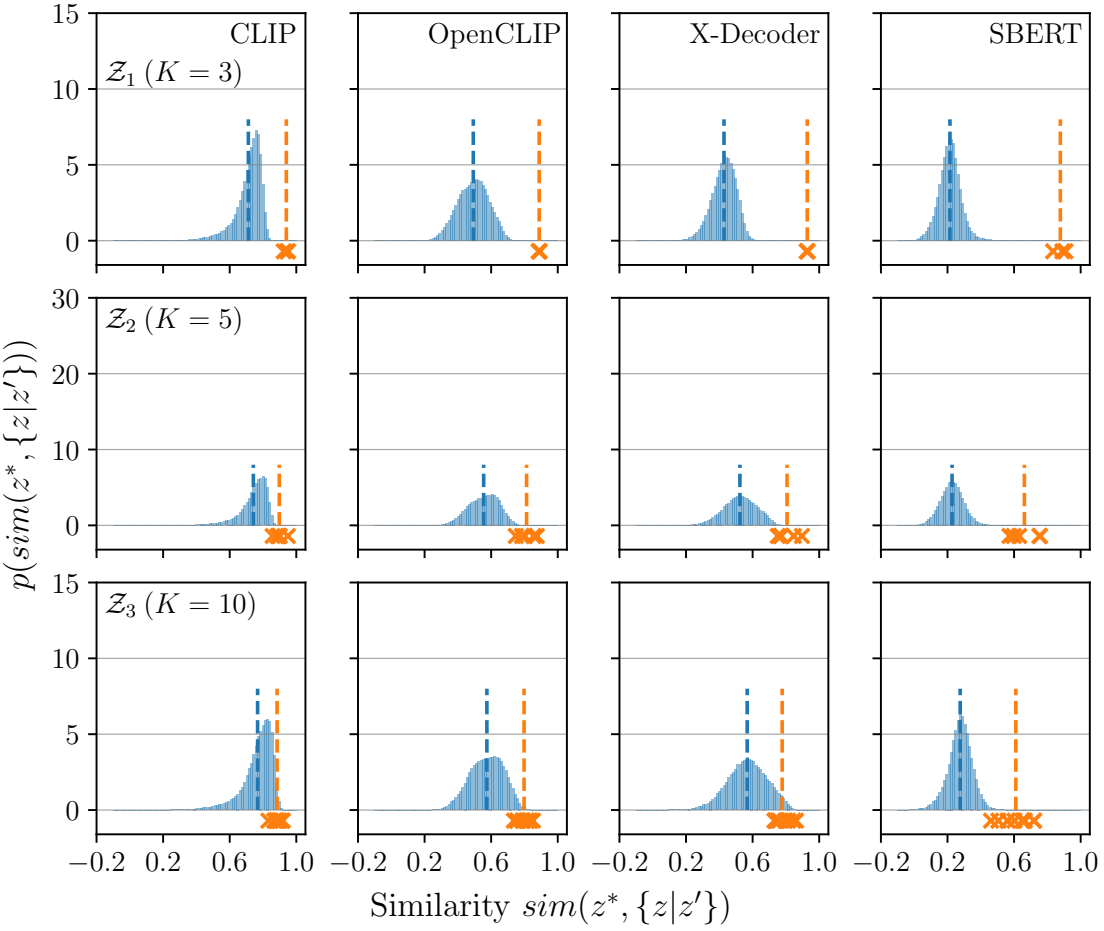

FIGURE 4.13: Similarity distributions for three realistic object descriptions $\mathcal{Z}_i$ of varying sizes K (orange) and randomly sampled word embeddings z' (blue).

Proposition 4.7 (Discoverability III)**.** *It is always possible to find an optimal compositional semantic embedding* $z^* \in \mathbb{R}^D$ *by iterative gradient descent optimization*

$$z^{*(t+1)} = z^{*(t)} - \lambda \nabla_{z^*} \left[\sum_{i=1}^{L} sim(z^{*(t)}, z^{(i)}) \right] \tag{4.23}$$

over random subsets $\tilde{\mathcal{Z}}^{(t)} \subseteq \mathcal{Z}$, $|\tilde{\mathcal{Z}}^{(t)}| = L$ *given a sufficiently small learning rate* λ*.*

Proof. See Appendix B.6. □

The proof is based showing that the cosine similarity optimization objective is convex, and noting that all convex problems have global convergence guarantees.

Figure 4.14 illustrate how an unconditional open vocabulary semantic segmentation model discovers latent compositional semantic embeddings from independent visual examples. From visual examples of a "couch-object" being a *couch*, and from other visual examples being *furniture*, I prove mathematically that the conventional unconditional open vocabulary semantic segmentation objective results in learning a latent compositional semantic embedding z^* from which both semantic properties can be inferred by sufficient similarity inference [95]. The bottom row illustrates how the conventional most similar inference objective fails to infer that a "couch-object" is simultaneously a *couch* and *furniture*. See Sec. 6.2 for further information and experimental results of inferring latent compositional semantics from examples of visual appearance. See Sec. 4.3.4 for a detailed explanation of the proposed sufficient similarity inference method.

4.4 Discussion and Limitations

While the proposed grounded latent compositional semantics breaks new theoretical ground in directions for queryable spatio-semantic memory representations, practical limitations to work out exist.

One limitation is the relative few point observations outputted by low-cost lidar sensors. The sparsity of points results in a sparsity of mapped image-semantics to the accumulated semantic point cloud. Accurate dense depth estimation methods would enable mapping all image-observation semantics to a semantic point cloud representation. However, the technical inadequacy of dense depth estimation methods limits the practical usefulness of this approach. Physics-based stereo camera approaches have limited long-range accuracy, while learned depth estimation methods have limited generalization capability outside the training data domain. Another alternative is to investigate NeRFs

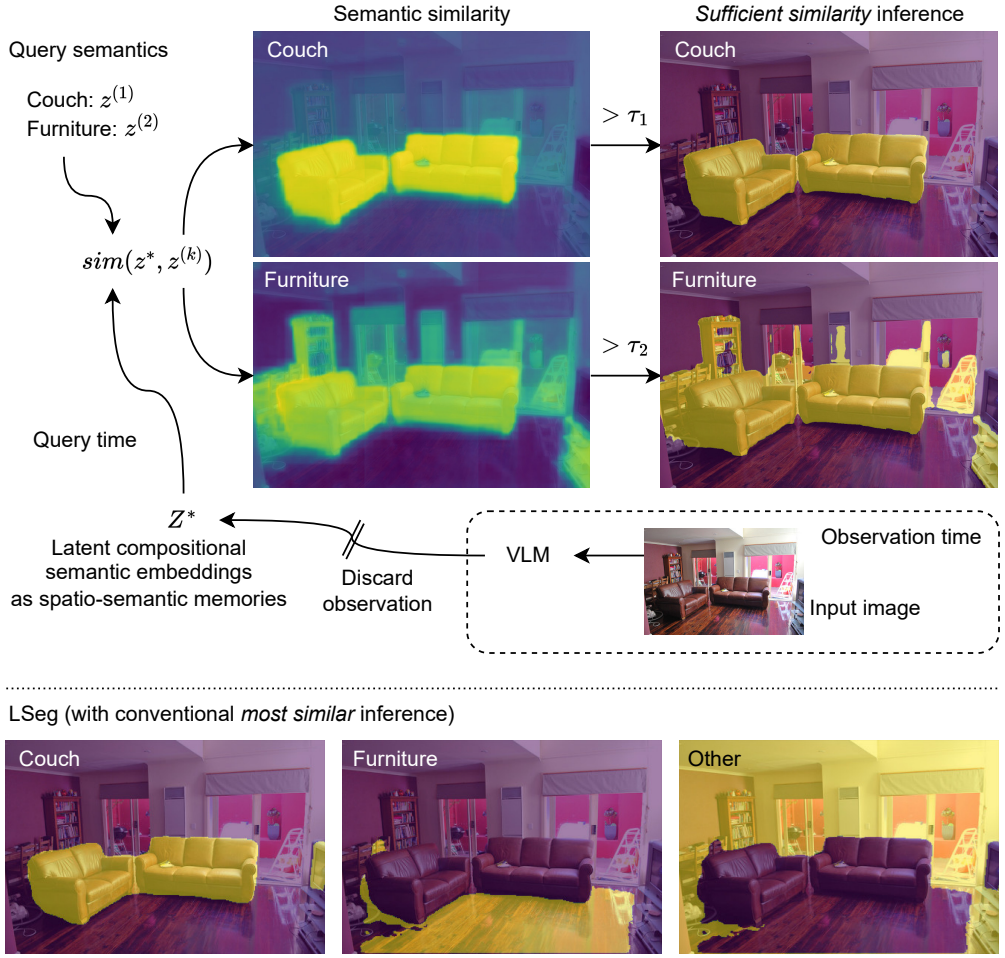

FIGURE 4.14: I show that unconditional open vocabulary semantic segmentation VLM models learn to map images into latent compositional semantic embedding maps Z^*. The sufficient similarity inference method allows predicting overlapping semantics for any set of queried semantics $\{z^{(k)}\}$ by similarity with z^*, without requiring original input images. Conventional unconditional models like LSeg [6] fail at inferring semantic overlap (*couch* is also *furniture*) and incomplete partitionings (*other* is a flawed substitute for unspecified semantics). Projecting Z^* to spatial coordinates result in accurate and rich open-vocabulary spatio-semantic memories.

capable of real-time optimization and generation as an alternative means to spatially encode image semantics into a dense 3D spatio-semantic memory representation.

The mathematical foundation of latent compositional semantics established, and the experimental results shows proimising results in discovering latent compositional semantics from examples of visual appearance and semantics by bootstrapping from pretrained dense representation learning computer vision models. Further experimental investigation is needed to investigate sample efficiency for discovering latent compositional semantics. For example, what is the relation between number of visual examples and generalization of a semantic concept? Another example is investigating how to more efficiently align discovered latent semantics to their ideal counterparts on the unit hypersphere. Further investigation into how to optimize the distribution of semantics in

the encoding space of typical learned semantic embedding decoders like CLIP [77] and SBERT [179] to cover a larger region of the hyperspherical shell embedding space [301]. Experimental results shows that a uniform distribution over the entire hypersphere would greatly improve the capacity and discriminitability of sets of semantics by latent semantic embeddings by an order of magnitude. The presented sufficient similarity semantic inference method only allows inferring semantics based on one or many prior example of said semantic. Further investigation into how the sufficient similarity method can be extended to allow inference of novel semantics would greatly extend the usefulness of the proposed method. Proposed future work includes investigating the relation between number of example semantics and generalization performance when determining sufficient similarity threshold values τ. Another direction is investigating how a set of priorly known semantics are related to a novel semantic in order to infer a suitable similarity threshold without explicit examples. For example, to infer a novel semantic like *dog* from known semantics like furry, mammal, legged, *pet*, and so on.

4.5 Summary

Unsupervised dense representation learning to discover distinct semantic visual entities from natural images without human supervision or proposal heuristics. This thesis presents visual similarity as an inductive bias in the form of superpixel region partitioning to improve effectiveness of dense representation learning. Open-vocabulary semantic segmentation models learn to map machine-discovered semantics with human interpretable semantic embedding spaces, enabling querying of any semantics by relative similarity in embedding space. Methods based on point projection and point cloud accumulation are presented as means of grounding inferred semantics into a spatial representation.

The remainder of the chapter presents the mathematical theory of latent compositional semantics as a means to discover and represent discriminable sets of semantics by a single latent embedding. The sets of latent semantics are equivalent to compositional object representations. The properties and representational capacity of latent compositional semantics are presented with mathematical proofs. Experiments on four embedding spaces including CLIP and SBERT shows that latent compositional semantics can represent up to 10 semantics encoded by SBERT, and up to 100 semantics for ideal uniformly distributed high-dimensional embeddings. Experiments with VLMs show that latent compositional semantics are discoverable from visual appearance by iterative gradient descent. The proposed sufficient similarity semantic inference method overcomes fundamental limitations of conventional inference, and improves higher-level overlapping semantic inference performance by 19.63 mIoU in the presented experiments.

Chapter 5

Predictive State Representation as Artificial Hippocampus

5.1 Introduction

The primary foundation of the intellectual powers of humans lies in their extensive and versatile model of the world, including knowledge structures and concepts that humans harness to relate objects, events, and words [93]. According to the predictive coding theory, the brain continuously generates predictions or hypotheses about the causes of sensory inputs based on prior knowledge and experience. These predictions are then compared with the actual sensory inputs, and any discrepancies or prediction errors are used to update the brain's internal models and refine future predictions. The hippocampus is believed to be the central orchestrator for such high-level abstraction involving several aspects of predictive coding [94].

The hippocampus is known to be essential for learning and representing sequences of events or experiences. This ability to encode and predict sequences is thought to be a key component of predictive coding, as it allows the brain to anticipate upcoming sensory inputs based on learned patterns.

The hippocampus plays a critical role in the formation and retrieval of episodic and spatial memories. These memories can be used to generate predictions about sensory inputs based on past experiences in similar contexts or environments. Some theories suggest that the hippocampus is involved in detecting and processing prediction errors, which are the discrepancies between the brain's predictions and the actual sensory inputs. These prediction errors are thought to be crucial for updating the brain's internal models and driving learning. The hippocampus receives inputs from various sensory

modalities, such as vision, audition, and spatial information. This integration of multimodal information is thought to contribute to the formation of coherent predictions and the updating of internal models based on prediction errors. It's important to note that the hippocampus does not act in isolation but is part of a larger network of brain regions involved in predictive coding, including the neocortex, prefrontal cortex, and other subcortical structures [333].

5.2 Predictive Coding as a Continual Learning Framework

Predictive coding from theoretical neuroscience and variational autoencoders (VAEs) [7] from machine learning and identify their common origins and mathematical frameworks. This work is inspired by previous works at this intersection, including hierarchical probabilistic models in predictive coding and machine learning [541] and implementation of predictive coding techniques in deep probabilistic models [542].

Predictive coding emerged within neuroscience as a theory that neural circuits are engaged in estimating probabilistic models to make predictions about incoming sensory input. This is done through an iterative process of prediction error minimization and updating the internal model based on new information [330]. On the other hand, VAEs are a type of deep generative model that uses variational inference to approximate the posterior distribution over latent variables given observed data. They consist of an encoder network that maps input data to a lower-dimensional latent space and a decoder network that reconstructs the original data from the latent representation [446].

The commonality between predictive coding and VAEs lies in their use of probabilistic models for making predictions about sensory input or generating new samples. Both approaches rely on minimizing prediction errors to update internal representations and optimize model parameters. However, there are also key differences between the two frameworks. For example, while predictive coding is a biologically plausible theory that has been supported by empirical evidence from neurophysiological studies [330, 331, 333], VAEs are based on mathematical principles and do not necessarily reflect biological mechanisms [94].

5.3 Artificial Hippocampus as Learned Simulator

The hippocampus plays a vital role in various aspects of natural intelligence, including learning associations, building causal models, encoding episodic memories, forming spatial representations, and facilitating high-level abstractions. Its complex neuronal

design principles offer valuable insights for developing artificial intelligence systems that can mimic human cognition's sophistication and adaptability [93]. The function of the hippocampus in relation to biological intelligence has been extensively studied across various disciplines, revealing its critical role in learning, memory, and abstract concept formation. According to research, the human hippocampus acts as a nexus for high-level abstractions, housing "concept cells" that encode semantic knowledge beyond the physical properties of referenced items This neural structure allows the first vertebrates to learn associations between stimuli, actions, and outcomes [93], playing a crucial role in building causal models that help make sense of an uncertain world. The hippocampus is involved in various aspects of human intelligence, including declarative memory [95], spatial navigation [301], and facilitates the encoding, consolidation, and retrieval of episodic memories, enabling individuals to recall past experiences. The hippocampus also contributes to spatial cognition by forming cognitive maps that facilitates spatio-semantic reasoning.

This thesis propose that the presented predictive state representation generated by a predictive world model is analogous to an artificial hippocampus. Both predictive world models and the hippocampus store cognitive maps or models of the world and generates predictive spatio-semantic representations of future sensory input based on past observational experience. Like the hippocampus, the predictive state representation presented provides an internal model supporting spatio-semantic memory formation and reasoning including navigation. The presented predictive world model framework is capable of continual learning from observational experience based on the principle of predictive coding, like the hippocampus. The plasticity of the hippocampus is high as one observation is sufficient to update the predictive model. The proposed machine learning model is also shown to be capable of updating the predictive posterior based on a single semantic observation.

However, a certain discrepancy exist. The hippocampus is adapt at learning temporal sequences of events, while the currently presented predictive world modeling approach does not incorporate temporal dynamics.

5.4 Predictive World Models

This work proposes an Open-vocabulary Predictive World Model (OV-PWM) as a spatiosemantic memory and internal simulator for general-purpose mobile robots. The OV-PWM is a latent variable generative model that learns from egocentric partial observations to predict complete environment states represented by grounded open-vocabulary semantics. The OV-PWM functions as an implementation of an artificial hippocampus

that learns a distribution of compact latent codes capturing the structure of observed environments.

Predictive World Models (PWM) aim to learn latent representations capturing the underlying structure of the environment. PWMs having learned the structure are able to supplement perception by predicting unobserved regions. Prediction generation follows the two-staged variational autoencoder (VAE) [7] latent variable approach: First, an encoder predicts a latent distribution $p(z|x)$ for the objectively real world x^* partially observed by sensors as x. Secondly, a particular latent variable z is sampled from $p(z|x)$. Finally, a decoder maps z into the most likely world x^*. The process is abstracted as the arbitrary conditioning latent variable generative model $p(x^*|x)$. In this thesis I demonstrate how to learn $p(x^*|x)$ to sample diverse and plausible complete worlds x^* from partially observed worlds x represented by open-vocabulary semantic embeddings $h \in \mathbb{R}^D$ with dimension $D >> 1$.

The Open-Vocabulary Predictive World Model (OV-PWM) is implemented by the SOTA hierarchical VAE (HVAE) model VDVAE [449] with an additional posterior matching encoder [231, 381]. HVAEs [447–449] are capable of learning hierarchical latent variable distributions expressing a high degree of structure at different abstraction levels. HVAEs generalizes autoregressive models [449] and can achieve higher likelihoods than SOTA autoregressive models like PixelCNN [543] using fewer learned parameters and generate samples thousands of magnitudes quicker [449].

The explicit open-vocabulary environment representations enabled by OV-PWMs provides several potential advantages to implicit representations and conventional offline map-based mobile robots with human-annotated semantics. First, the OV-PWM can disambiguate the observed state by substituting unknown regions with plausible predictions based on prior observational experience. Committing to a particular complete state simplifies learning policies by removing the implicit marginalization over many plausible underlying states for state transition modeling. Secondly, OV-PWMs can bridge conventional map-based and perception-based planning and control methods. For example, safer motion planning may be achieved by sampling diverse plausible structures of unobserved regions and account for worst-case scenarios. Additional potential advantages include improving localization by densifying observations, verifying offline map consistency with the actually observed environment, and leverage the highly expressive but compact latent state for planning in latent space [273]. Thirdly, learning a world model based on grounded open-vocabulary semantics allows optimizing a single general OV-PWM for multiple tasks requiring different semantic perceptual information. Fourthly, leveraging unconditional open-vocabulary semantics supports inferring overlapping semantics by sufficient similarity inference [95].

The following sections presents detailed description of the model and how it is trained and used for inference.

5.4.1 Learning complete states from partial states

The primary challenge is to learn a generative model predicting complete worlds by predictive coding [94, 542] from a set of partially observed incomplete worlds as "ground truth" data only. In general, learning to predict "nothing" or "unknown" is an easier solution than predicting plausible structures when lacking a complete ground truth learning signal to enforce commitment to a particular prediction. I employ the novel posterior matching latent variable generative model as a solution introduced in my prior work [381]. In this work I extend the approach to model high-dimensional open-vocabulary semantic embeddings and in the process simplify the previous two-stage approach into a single-stage end-to-end paradigm.

5.4.2 Latent variable generative modeling

The goal of generative modeling is to approximate the distribution $p(x)$ by a learned model $p_\theta(x)$ maximizing the likelihood of finite empirical dataset $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$.

A latent variable generative model $p(x, z)$ approximates the joint distribution of observed variables or data x and compact latent variables or codes z. The problem can be factorized into a conditional model

$$p(x, z) = p(x|z)p(z) \tag{5.1}$$

representing the process generating observed variables x from z as well as the distribution of z. The problem is that learning $p_\theta(x)$ and $p_\theta(x|z)$ is computationally intractable for high-dimensional data when using naive methods due to the unknown interactive structure of x and z.

A solution is to reformulate the problem of learning $p_\theta(x)$ is approximate variational inference. Approximate variational inference propose to simultaneously learn an amortized inference function $q_\theta(z|x)$ approximating the true latent representation distribution $p(z|x)$ and the generative process $p_\theta(x|z)$.

The variational inference scheme used to optimize the likelihood of the generative model $p(x)$ is derived as follows. The generative model $p(x)$ is the marginal distribution of the joint distribution of the latent variable generative model

$$p_\theta(x) = \int p_\theta(x,z)dz = \int p_\theta(z|x)p_\theta(x)dz = \mathbb{E}_{z\sim p_\theta(z|x)}p_\theta(x). \tag{5.2}$$

Taking the logarithm of both sides and leveraging the amortization factorization

$$p_\theta(x,z) = p_\theta(z|x)p_\theta(x) \tag{5.3}$$

$$p_\theta(x) = \frac{p_\theta(x,z)}{p_\theta(z|x)} \tag{5.4}$$

allows for a convenient decomposition

$$\log p_\theta(x) = \mathbb{E}_{z\sim p_\theta(z|x)} \log p_\theta(x) \tag{5.5}$$

$$= \mathbb{E}_{z\sim p_\theta(z|x)} \log \frac{p_\theta(x,z)}{p_\theta(z|x)} \tag{5.6}$$

$$= \mathbb{E}_{z\sim p_\theta(z|x)} \log \frac{p_\theta(x,z)q_\phi(z|x)}{p_\theta(z|x)q_\phi(z|x)} \tag{5.7}$$

$$= \mathbb{E}_{z\sim p_\theta(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)} + \mathbb{E}_{z\sim p_\theta(z|x)} \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \tag{5.8}$$

$$= \left[\mathbb{E} \log p_\theta(x,z) - \mathbb{E} \log q_\phi(z|x)\right] + \left[\mathbb{E} q_\theta(z|x) - \mathbb{E} p_\theta(z|x)\right]. \tag{5.9}$$

The optimization objective is derived by denoting the first RHS term as $L_{\theta,\phi}(x,z)$ and identifying the second RHS term as the KL divergence $D_{KL}(q_\phi(z|x), p_\theta(z|x))$ and rearranging terms

$$\log p_\theta(x) = L_{\theta,\phi}(x,z) + D_{KL}(q_\phi(z|x), p_\theta(z|x)) \tag{5.10}$$

$$L_{\theta,\phi}(x,z) = \log p_\theta(x) - D_{KL}(q_\phi(z|x), p_\theta(z|x)). \tag{5.11}$$

As $D_{KL}(q_\phi(z|x), p_\theta(z|x)) \geq 0$ it follows from (5.11) that

$$L_{\theta,\phi}(x,z) \leq \log p_\theta(x). \tag{5.12}$$

The optimization goal is to maximize $p_\theta(x)$, that is, the likelihood of data x according to the model $p_\theta(x)$. It follows from (5.11) that maximizing $L_{\theta,\phi}(x,z)$ must necessarily maximize $p_\theta(x)$ as $L_{\theta,\phi}(x,z)$ is a lower bound of $L_{\theta,\phi}(x,z)$, giving $L_{\theta,\phi}(x,z)$ the name variational or evidence lower bound (ELBO). The computable optimization objective for

maximizing $L_{\theta,\phi}(x,z)$ is derived by equivalently minimizing the negation of $L_{\theta,\phi}(x,z)$

$$\max_{\theta,\phi} L_{\theta,\phi}(x,z) = \min_{\theta,\phi} -L_{\theta,\phi}(x,z) \tag{5.13}$$

$$= \min_{\theta,\phi} -\left[\mathbb{E}\log p_\theta(x,z) - \mathbb{E}q_\phi(z|x)\right] \tag{5.14}$$

$$= \min_{\theta,\phi} -\left[\mathbb{E}\log p_\theta(x|z) - \mathbb{E}\log p_\theta(z) - \mathbb{E}q_\phi(z|x)\right] \tag{5.15}$$

$$= \min_{\theta,\phi} -\mathbb{E}\log p_\theta(x|z) + \mathbb{E}q_\phi(z|x) - \mathbb{E}\log p_\theta(z) \tag{5.16}$$

$$= \min_{\theta,\phi} -\mathbb{E}\log p_\theta(x|z) + D_{KL}(q_\phi(z|x), p_\theta(z)). \tag{5.17}$$

The lower bound $L_{\theta,\phi}(x,z)$, and indirectly the model likelihood $p_\theta(x)$, is therefore optimized by increasing $p_\theta(x|z)$ and decreasing $D_{KL}(q_\phi(z|x), p_\theta(z))$.

The variational autoencoder (VAE) is a deep generative model that implements approximate variational inference. Both the amortized inference function $q_\phi(z|x)$ and generative model $p_\theta(x|z)$ are implemented by neural network function approximations. The VAE simultaneously learns $q_\phi(z|x)$ and $p_\theta(x|z)$ by inferring a distribution of latent variable z and subsequently reconstruct a sampled z back into the observable variable x. The distribution of latent variables $p_\theta(z)$ is assumed to be a known distribution like the Normal distribution. The $D_{KL}(q_\phi(z|x), p_\theta(z))$ term constrains the learned posterior distribution $q_\phi(z|x)$ to match the prior $p_\theta(z)$ so that new samples can be generated by simply sampling from the known distribution $p_\theta(z)$.

Vanilla VAEs suffer from constrained expressiveness due to being limited to a single set of latent variables z. The limitation is characterized by generation of low-fidelity high-dimensional data like blurry high-resolution images.

The hierarchical VAE (HVAE) overcomes this limitation by introducing layers of latent variables $Z = (z^{(1)}, \ldots, z^{(K)})$. Each layer k models structure of different levels of abstraction. The hierarchical order of latent variables naturally results in a decoupling of overall structure and visual appearance. The HVAE prior distribution, posterior distributions, and generative model can be factorized as

$$p_\theta(Z) = p_\theta(z_1|z_2)\ldots p_\theta(z_{K-1}|z_K)p_\theta(z_K) \tag{5.18}$$

$$q_\phi(Z|x) = q_\phi(z_1|z_2,x)\ldots q_\phi(z_{K-1}|z_K,x)q_\phi(z_K|x) \tag{5.19}$$

$$p_\theta(x|Z) = p_\theta(x|z_1)\ldots p_\theta(z_{K-1}|z_K)p_\theta(z_K) \tag{5.20}$$

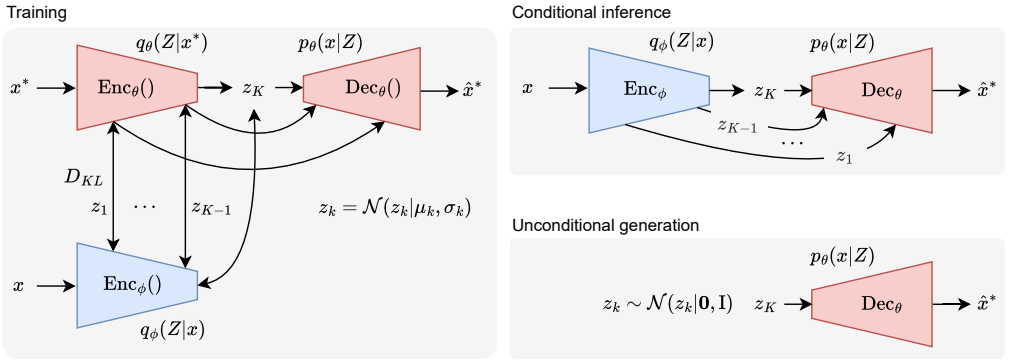

FIGURE 5.1: Predictive world model. The encoder $\text{Enc}_\theta()$ learns a hierarchical latent variables Z representing the environment $\hat{x}$ conditioned on the *past-to-future* partially observed state x^*. The posterior matching encoder $\text{Enc}_\phi()$ learns to predict the same distribution Z from the *past-to-present* state x. The decoder Dec_θ learns to reconstruct diverse and plausible complete states $\hat{x}$ from Z.

where all random variables z are modeled by Normal distributions $\mathcal{N}(z|\mu, \sigma)$. Deeper or more abstract codes (i.e. z_K) encode the global structure, while shallow codes (i.e. z_1) encode the visual appearance of elements in x. The deepest latent variable prior $p_\theta(z_K)$ is a known distribution like the Normal distribution like a VAE. However, subsequent priors $p_\theta(z_{K-1}) \dots p_\theta(z_1)$ are learned priors for increased model expressivity.

5.4.3 Model implementation and training

I implement the OV-PWM based on the recent SOTA HVAE architecture called Very Deep VAE (VDVAE) [449]. The HVAE model has 48 layers of 16 dimensional latent variables (e.g. $K = 48$) with incrementally increasing feature map resolution and decreasing intermediate feature dimension throughout the layers.

I use two inputs to train the model. The first input is the presently observed world $x \in \mathbb{R}^{H \times W \times D}$ (e.g. *past-to-present* accumulated observations). The second input is the future observed world $x^* \in \mathbb{R}^{H \times W \times D}$ (e.g. *past-to-future* accumulated observations). x and x^* are high-dimensional grid map with elements representing normalized open-vocabulary semantic embeddings $h \in S^{D-1}$ with dimension D. Unobserved elements are represented by the zero vector $\mathbf{0}$.

The two inputs are processed by two structurally identical but separate encoders. The future observed world x^* is processed by the encoder $Enc_\theta(x^*)$, approximating $q_\theta(Z|x^*)$, into intermediate feature maps $Y^* = \{y_1^* \dots y_{K-1}^*\}$ and a latent feature vector y_K^*. The presently observed world x is processed by a the posterior matching encoder $Enc_\phi(x)$, approximating $q_\phi(Z|x)$, into $Y = \{y_1 \dots y_{K-1}\}$ and a latent feature vector y_K.

A single decoder generates a sample $\hat{x}^*$ by first sampling the latent variable z_K from a distribution conditioned on y_K^*. The intermediate reconstruction $\tilde{x}_K^*$ is computed from z_K and learned bias variables. Subsequent latent variables z_k are sampled by the corresponding intermediate feature maps y_k^* from the encoder and the previous intermediate reconstruction $\tilde{x}_{k-1}^*$. Subsequent intermediate reconstructions $\tilde{x}_k^*$ are computed based on the sampled z_k and $\tilde{x}_{k-1}^*$. The features $Y = \{y_1 \dots y_K\}$ outputted by the posterior matching encoder $Enc_\phi(x^*)$ are optimized to predict the same latent distribution $q_\theta(z_k)$ as the $q_\theta(z_k)$ distribution outputted by the future observation encoder $Enc_\theta(x^*)$.

The final intermediate feature map $\tilde{x}^* \in \mathbb{R}^{H,W,D'}$ is mapped into an open-vocabulary semantic embedding map $\hat{x}^* \in \mathbb{R}^{H,W,D}$ by a linear projection. Forcing the output to lie on the hypersphere S^{D-1} and thus represent the latent compositional semantic denoting the set of most likely membership semantics

$$\forall x_{i,j} \in x \sim p_\theta(x|Z) \Rightarrow x_{i,j} \in S^{D-1}, \tag{5.21}$$

resolves the problematic tendency of the previous semantic probability approach [381]. The prior probabilistic closed set semantics approach represents membership semantics by K probabilities that element (i, j) is a member of semantic $k \in K$. Forcing the model to predict a latent compositional semantic embedding h naturally allows inferring overlapping semantics while overcoming the maximum likelihood shortcut learning problem of readily predicting "unknown" instead of penalizing commiting to a miss-prediction. Uncertainty can instead be estimated by stochastic variation from repeatedly sampling the posterior [544]. The prior two-stage approach with intermediate pseudo ground truth states are not needed for OV-PWMs, and thus simplifying the method to a single-stage end-to-end learning process.

The $Enc_\theta()$ and $Dec_\theta()$ components of the dual encoder HVAE is optimized by maximizing the hierarchical ELBO

$$\max_{\theta,\phi} L_{\theta,\phi}(x, Z) = \min_{\theta,\phi} \mathbb{E}\left[-\log p_\theta(x|Z) + D_{KL}(q_\phi(Z|x)||p_\theta(Z))\right] \tag{5.22}$$

where $\log p_\theta(x|Z)$ is the likelihood of the sample x^* reconstructed from Z, and a KL divergence term that measures the separation between the learned posterior and prior distributions

$$D_{KL}(q_\theta(Z|x)||p_\theta(Z)) = \sum_{k=2}^{K} \mathop{\mathbb{E}}_{q_\theta(z_{\geq k}|x)} \left[D_{KL}(q_\theta(z_{k-1}|z_k, x)||p_\theta(z_{k-1}|z_k))\right] + D_{KL}\left(q_\theta(z_K|x)||p_\theta(z_K)\right). \tag{5.23}$$

I simultaneously train the secondary posterior matching encoder $Enc_\phi()$ to predict latent distributions Z for partially observed environments x which are similar to Z inferred from the regular encoder $E_\theta(x^*)$ with future observed worlds x^*. The second posterior matching encoder is optimized by minimizing

$$D_{KL}(q_\phi(Z|x^*)||q_\psi(Z|x_{po})) = \sum_{k=1}^{K} \mathop{\mathbb{E}}_{q(z_{>k}|x)} \left[D_{KL}(q_\phi(z_k|z_{>k}, x^*)||q_\psi(z_k|z_{>k}, x_{po})) \right]. \quad (5.24)$$

Maximizing the likelihood of $p_\theta(x|Z)$ in (5.22) is equivalent to minimizing the cosine distance for normalized OV semantic embeddings modeled by the OV-PWM model

$$\min \mathbb{E} - \log(p|Z) = \min \mathbb{E}(1 - \text{sim}(x, \hat{x})) = \min \mathbb{E}(1 - x^T \hat{x}). \quad (5.25)$$

The practical formulation of the hierarchical ELBO (5.22) used for optimizing the OV-PWM model is therefore

$$\max_{\theta,\phi} L_{\theta,\phi}(x, Z) = \min_{\theta,\phi} \mathbb{E} \left[(1 - x^T \hat{x}) + D_{KL}(q_\phi(Z|x)||p_\theta(Z)) \right]. \quad (5.26)$$

See Appendix A for a derivation of (5.25).

5.4.4 Model Inference

At inference time the model uses the posterior matching encoder $Enc_\phi()$ to generate a latent distribution Z that can be decoded by $Dec_\theta()$ into a predicted complete world state $\hat{x}^*$. The model can be used for unconditional generation by incrementally sampling latent variables Z from the learned prior distribution $q_\theta(Z)$. The regular encoder $Enc_\theta()$ trained on future observations x^* is not used during inference.

5.5 Discussion and Limitations

A limitation of my current approach is the top-down 2D grid representation. 2D embedding maps do not represent vertical information and multi-layered environments as required for general 3D representations. Extending the OV-PWM approach to 3D representations using voxel grids or neural radiance fields is a promising direction of future work to enable spatial reasoning in fully general complex 3D structures. While the model already demonstrates promising generalization capability in new environments,

the modeling of finely detailed semantics like *road markings* display room for improvement. Given that the original VDVAE model was trained on 32 V100 GPUs for 2.5 weeks (*we: 6 A6000 GPUs for 4 days*) on a large dataset of 70,000 samples [449] (*we: 7000 samples*), and the OV-PWM training performance trend indicate further improvement, it is reasonable to expect additional training time and diverse observational experience to further boost performance. Reducing degenerate samples resulting from inaccurate and erroneous ICP scan matching steps by implementing a robust SLAM-based observation accumulation framework may further improve training efficiency. Despite the limited computational resources, training set size, and degenerate samples, my method learns to generate outputs with intricate details emerging even from the unconditional prior. Other directions include incorporating agents and temporal dynamics into predictive world model, as well as demonstrating the advantages of learned simulators in practical embodied task planning and decision making problems using large-scale, real-world data.

5.6 Summary

The human brain's extensive model of the world, which includes knowledge structures and concepts, is the foundation for human intelligence. The hippocampus plays a crucial role in this process as it continuously generates predictions about sensor inputs based on prior knowledge and experience, based on the theory of predictive coding. The hippocampus learns sequences, forms memory-based predictions, processes prediction errors, and integrating multi-modal information. Predictive coding is presented as a principled continual learning framework that shares commonalities with variational autoencoders in machine learning. This thesis proposes that an artificial hippocampus can be developed using predictive state representation generated by a predictive world model, which can store cognitive maps and generate spatio-semantic representations of future sensory input based on past observational experience.

The presented open-vocabulary predictive world model (OV-PWM) is presented as an artificial hippocampus. The OV-PWM model learns hierarchical distributions of compact latent representation of directly from raw observations using dual-encoder HVAE with posterior matching optimization. The proposed predictive world model is capable of predicting diverse complete environment states for unobserved environment regions by iteratively sampling from the learned hierarchical prior distribution. The predictive world model forms the basis of the dual explicit and latent predictive state representation.

Chapter 6

State Representation for Autonomous Driving Reasoning Agents

6.1 Introduction

This chapter presents a practical implementation of the proposed predictive state representation formalism as presented in previous chapter. The focus is on future for autonomous driving agents capable of spatio-semantic reasoning based on the spatially grounded and semantically rich predictive environment states.

The chapter starts by presenting how sensor observation forms partially observed environment states. Practical implementation of unconditional open-vocabulary semantic segmentation models, or dense vision-language models (VLMs), is explained. The sufficient similarity semantic inference method for inferring overlapping semantics is presented, along with experimental results for learning latent compositional semantics from examples of visual appearance.

The following sections presents the how to accumulate observations over time into partially observed open-vocabulary environment state representations, as well as present experimental results for the proposed open-vocabulary predictive world model (OV-PWM) applied in the autonomous driving domain. The chapter concludes by explaining how the predictive state representation is used to learn navigational patterns from observation only.

6.2 Sensor Observations to Partial World States

This section describes how to generate open-vocabulary partial environment states from multimodal sensor observations. I leverage recent advances in unconditional open-vocabulary semantic segmentation based on the theory of latent compositional semantics [95] as my semantic representation. The partial world state representations serve as the input representation for learning Open-Vocabulary Predictive World Models (OV-PWM) described in Section 5.4.

6.2.1 Sensor Observation Processing

Mobile robots perception systems typically fuse complementary sensor modalities. Passively sensing RGB cameras provide rich semantic understanding. Actively sensing lidars or depth sensors provide accurate metric spatial perception. Sensor fusion approaches aim at leveraging the complementary strengths of both vision modalities [259].

Semantic point clouds are the natural unified data structure for representing both spatial and semantic information. A semantic point cloud is created by grounding semantic embedding maps extracted from 2D image pixels into spatial coordinates. The grounding is performed as follows: first, a point cloud is projected onto the image frame by a transformation specified by camera calibration parameters. Predicted open-vocabulary semantic embeddings are mapped to all points coinciding with the respective image coordinates. All points outside the image frame are discarded. The remaining set of points thus contain spatial information in the form of $(x, y, z) \in \mathbb{R}^3$ coordinates and semantic embedding $z \in \mathbb{R}^D$ with dimensionality D, resulting in a a semantic point cloud $P \in \mathbb{R}^{N \times 3+D}$ where N is the number of semantically annotated points. See Figure 6.4 for visualized high-dimensional open-vocabulary semantic point clouds projected to RGB values.

6.2.2 Interpreting Observations as Open-Vocabulary Semantics

I propose to map unconditional open-vocabulary, or latent compositional semantic embeddings to point clouds. Here follows a brief explanation starting from conventional class semantics. A set of K class semantic embeddings are defined by separate basis vectors $\mathbf{e}_k$ in a $\mathbb{R}^K$ dimensional embedding space. Each semantic represented by $\mathbf{e}_k$ is orthogonal to every other semantic $\mathbf{e}_{l \neq k}$, meaning every semantic is equally similar or dissimilar to every other semantic. Conventional class semantics therefore do not encode semantic similarity.

Open-vocabulary instead has a fixed embedding space spanned by D orthogonal basis vectors $\mathbf{e}_1 \dots \mathbf{e}_D$ representing primitive latent semantics. All vectors $\mathbf{e}_d$ defines a latent prototypical semantic. All vectors in the embedding space are normalized and thus lie on the unit hypersphere S^{D-1}. A projection function $f_\theta()$ maps any visual or text semantic h onto S^{D-1}. As h are generally distributed over all basis vectors, the cosine similarity of two normalized embeddings measures the relative semantic similarity. The equation for computing cosine similarity is repeated bellow:

$$\text{sim}(h_1, h_2) = \frac{h_1 \cdot h_2}{||h_1||||h_2||} = (h_1)^T h_2. \tag{4.4}$$

My predictive world modeling approach is based on interpreting RGB images by an unconditional open-vocabulary semantic segmentation model [396]. The segmentation model outputs a dense embedding map $H = \mathbb{R}^{H \times W \times D}$ representing open-vocabulary semantics with a one-to-one pixel correspondence. A mathematical theory of unconditional open-vocabulary semantics [95] explains how models learns to output latent compositional semantics h^* representing discriminable sets of membership semantics $\mathcal{H} = \{h_1, \dots h_K\}$ as a hyperspherical cap S^{D-1}_{cap} defined by h^* and a sufficient similarity threshold τ. To compute if an observation i in a semantic point cloud represents query semantics (e.g. if a point is *road*), the cosine similarity between the latent compositional semantic h^*_i mapped to point i and the embedded query semantic h_q must be higher than the sufficient similarity threshold of the query semantic τ_q and thus in S^{D-1}_{cap}:

$$\text{sim}(h^*, h_q) > \tau_q \Rightarrow MemberOf(\text{point } i, \text{query semantic}). \tag{6.1}$$

Conventional "most similar" open-vocabulary inference approaches [81] forgo knowing a sufficient similarity threshold τ and thus seemingly allow querying any never encountered semantic. Nevertheless, the "most similar" inference approach has two fundamental flaws [6, 95]: First, every point i can be a member of only one of the queried semantics. For example, a point on window-on-a-building-facade should be simultaneously inferrable as both "window" and as part of a "building" at a higher-level. Naively hard-coding rules such as stating "window" are also "building" is not generally true. Secondly, the set of query semantics is presumed to constitute a complete partitioning of all points, as even unrelated points will be mapped to one of the query semantics. For example, a *dog* queried by "grass" and "toy" is interpreted as "toy". Naively using abstract word semantics like "other" as a substitute for unspecified semantics is not a principled solutions as the similarity between the predicted semantic h and the unrelated

query semantic h_q is not guaranteed to be lower than the ambiguous meaning of "other"

$$\text{sim}(h, h_{other}) \overset{?}{\geq} \text{sim}(h, h_q). \tag{6.2}$$

Sufficient similarity inference is a principled solution to the flaws of "most similiar" inference by allowing overlapping semantic inference (e.g. semantic membership with "window" and "building" can be simultaneously inferred) and inferring only true semantics irrespective of the set of query semantics (e.g. *dog* is neither "grass" or "toy"). In this work I follow the theory of latent compositional semantics interpretation of unconditional semantics [95] and demonstrate applying the sufficient similarity inference method for OV-PWMs.

In this work, I investigate whether or not high-dimensional open-vocabulary embeddings can be modeled by the predictive world model approach. I therefore do not consider the perception problem of inferring unconditional open-vocabulary semantics from images, and instead leverage point clouds annotated with CARLA ground truth semantics [545] for experiments. I design a taxonomy where each ground truth semantics is provided two additional high-level semantics (ex: a "road" is also a "drivable" and a "static" object). A single optimal latent compositional semantic embedding h^* is computed as the mean centroid of the three associated semantics [95] and appended to each point to form an open-vocabulary semantic point cloud. I refer to prior work for in-depth investigations concerning learning and inferring open-vocabulary semantic embeddings from visual data [6, 81, 95, 396].

6.2.2.1 Vision-Language Model

An unconditional dense VLM is a learned one-to-one function $f_\theta()$ that maps images $x \in \mathbb{R}^{3\times H\times W}$ to dense embedding maps $Z \in \mathbb{R}^{D\times H\times W}$ consisting of aligned VL embeddings $z_{(i,j)} \in \mathbb{R}^D$ at point (i, j) in the image frame. The output Z represents observations abstracted into declarative semantic memories [206–208] which maximizes the predictive likelihood over past observations given x, without conditioning on an input text [188] or an image query [200]. Unconditional prediction [6, 77] is necessary for efficient open vocabulary spatio-semantic memory representations as explained in Sec. 3.4.3. However, since objects have not one but several semantic descriptions [206, 209, 210], a single semantic embedding $z_{(i,j)}$ must simultaneously encode a multitude of task-relevant semantics.

I investigate the feasibility of discovering compositional semantics by f_θ as an image encoder-decoder dense prediction deep neural network architecture. To maximize

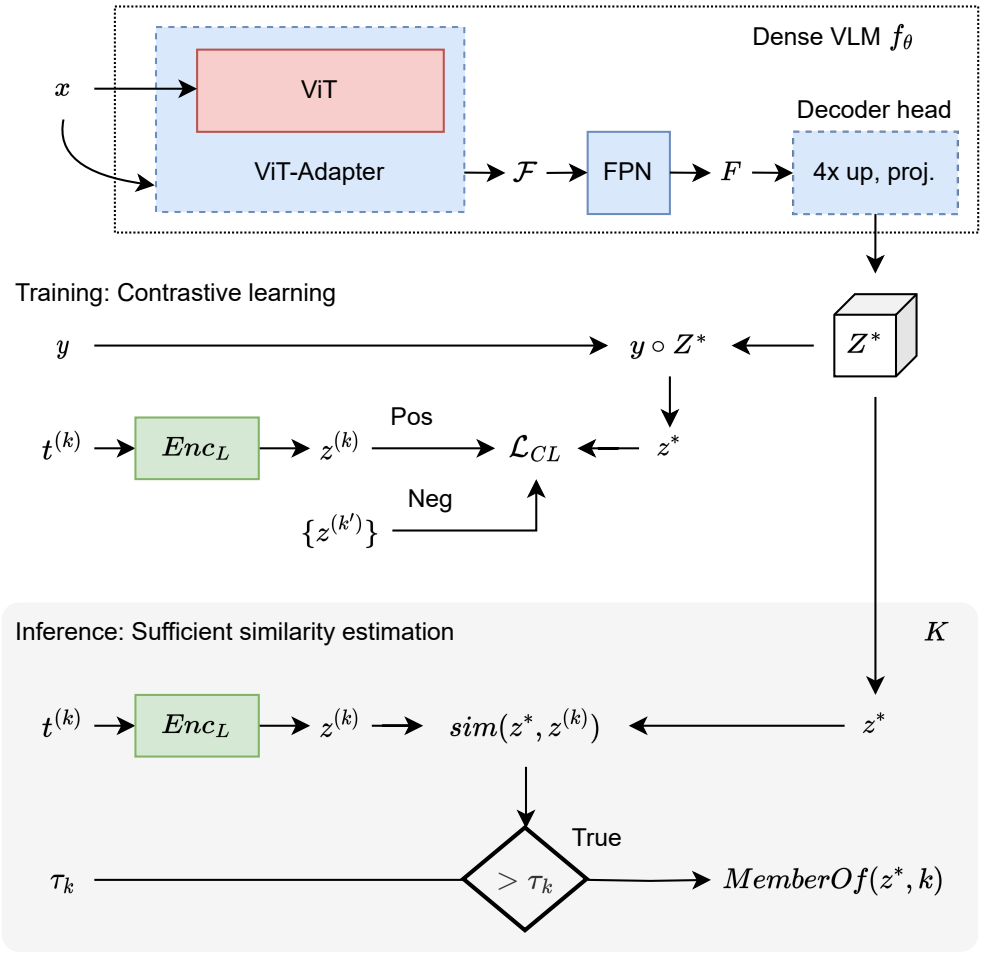

FIGURE 6.1: The unconditional dense VLM f_θ transforms an image x into an embedding map Z^* representing compositional semantics z^* for every pixel. During training, predictions z^* for elements masked by y are optimized to be similar to targets $z^{(k)}$ and dissimilar to all other semantics $z^{(k')}$ generated from text descriptions $t^{(k)}$ by a language encoder Enc_L. During inference, z^* allows querying multiple semantics K by similarity. All elements above the similarity threshold τ_k are members of the semantic group k. τ_k is set to maximize likelihood of predicting past observations.

the generality of my findings, f_θ is implemented by conceptually simple, general, and well-performing SOTA modules as shown in Fig. 6.1. A vision transformer (ViT) backbone [154] extracts visual features from image observations x. I use the ViT-Adapter [396] as a dense prediction task adapter to enhance the ViT backbone with vision-specific inductive biases. The adapter outputs a set of multi-scale feature maps $\mathcal{F} = \{F_1, F_2, F_3, F_4\}$. A Feature Pyramid Network (FPN) [517] integrates $\mathcal{F}$ into a single feature map F. A simple decoder head bilinearly upsamples F into the input image resolution and do a final 1×1 convolution to project features into normalized semantic embedding maps Z.

In the remainder of this section, I explain how in fact dense latent compositional semantic embedding maps Z^* are discovered by an unconditional dense VLM f_θ when trained to predict Z.

The model f_θ is initialized with pretrained backbone parameters and trained end-to-end to predict semantic embedding maps Z from images x and dense annotations. Annotations consists of K types of paired semantic text descriptions $t^{(k)}$ encoded into semantic embeddings $z^{(k)}$, and boolean image masks $y \in \mathbb{B}^{H \times W}$ specifying which image elements $x_{(i,j)}$ are associated with t. I denote an observation n as a tuple $(x, t, y)_n$.

I use the contrastive learning objective

$$\mathcal{L}_{CL} = \mathbb{E}\left[-\log \frac{e^{sim(z,z^{(k)})/\tau}}{e^{sim(z,z^{(k)})/\tau} + \sum_{k'} e^{sim(z,z^{(k')})/\tau}}\right] \tag{6.3}$$

with temperature τ to optimize f_θ to predict z similar to $z^{(k)}$ for elements specified by y and negative samples $z^{(k')}$. The set of negative samples $\mathcal{Z}' = \mathcal{Z} \setminus \{z^{(k)}\}$ consists of all known annotated semantics $\mathcal{Z}$ in the dataset except the current sample annotation $z^{(k)}$. I optimize over all $\mathcal{Z}'$ for every batch instead of randomly sampling negatives as the number of semantics are tractable. I note that the general objective (6.3) is equivalent to the previously proposed cross-entropy over softmax normalized embedding similarity objective [6]

$$\mathcal{L}_{CE} = \mathbb{E}\left[-(c^{(k)})^T \log \sigma\left(sim(\hat{z}, z^{(k)})/\tau\right)\right] \tag{6.4}$$

with $c^{(k)}$ denoting one-hot class or description type vectors, $\sigma()$ as the softmax function. The equivalence is apparent by zeroing out all but the one-hot true target embedding resulting from the dot product sum and expanding the softmax function

$$\mathcal{L}_{CE} = \mathbb{E}\left[0 - \ldots - \log \frac{e^{sim(\hat{z},z^{(k)})/\tau}}{\sum_{k'=1}^{K} e^{sim(\hat{z},z^{(k')})/\tau}} - \ldots - 0\right]. \tag{6.5}$$

Next I verify that the objective (6.3), and equivalently (6.4), can learn latent compositional semantic embeddings z^* from independent nonoverlapping descriptions. Proposition 4.7 proves that z^* can be learned by gradient descent. I can therefore presume without loss of generality, that two descriptions $z^{(k_1)}$ and $z^{(k_2)}$ appear simultaneously in a batch for two independent but visually similar objects x_1 and x_2 mapping to the same latent semantic z. The combined loss is

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\left(\mathcal{L}_{CL}(z, z^{(k_1)}) + \mathcal{L}_{CL}(z, z^{(k_2)})\right) \\
&= \frac{1}{2}\left(-\log \frac{1}{c} e^{sim(z,z^{(k_1)})} - \log \frac{1}{c} e^{sim(z,z^{(k_2)})}\right) \\
&= -\frac{1}{2}\left(\log e^{sim(z,z^{(k_1)})} + \log e^{sim(z,z^{(k_2)})} - 2\log c\right) \\
&= -\frac{1}{2}\left(sim(z, z^{(k_1)}) + sim(z, z^{(k_2)})\right) + \log c
\end{aligned} \tag{6.6}$$

As the optimal z minimizing (6.6) equals the centroid of $z^{(k_1)}$ and $z^{(k_2)}$, the optimal z is the optimal latent compositional semantic embedding z^* as proved by Theorem 4.2. I conclude that the iterative optimization by objective (6.3) enable f_θ to learn z^* from visual similarity and nonoverlapping descriptions.

6.2.3 Experimental Results: Latent Compositional Semantics From Visual Appearance

The following experiment investigates if z^* can be discovered from independent observations of visual appearance paired with nonoverlapping annotations. I present two experiments to answer this question.

First, I evaluate how well four representative SOTA unconditional open vocabulary semantic segmentation models can infer overlapping compositional semantics. Each model is trained on conventional non-overlapping annotations. ZSSeg [81] generates region proposals by SAM [546] and uses CLIP [77] to predict semantic embeddings z. X-Decoder [193] is a conditional VLM that predicts N object mask proposals and match masks with the most likely query semantic. I convert X-Decoder into an unconditional model by integrating all N VL mask semantics by the mask probability at each pixel location. I use largest available *Focal-L* model trained on COCO captions and dense labels. LSeg [6] is a dense VLM trained to output unconditional VL embedding maps. I use the released *ViT-L/16* model weights trained on seven datasets including COCO-Stuff [104], ADE20K [547], and Mapillary [548]. ViT-Adapter [396] is a recent general-purpose dense computer vision architecture I implement as my trainable model. The ViT backbone is initialized with self-supervised BEiT model weights [144]. The model is trained with SBERT embeddings on the same seven dataset as LSeg for 160K iterations on four A6000 GPUs with a total batch size 4 and 0.75e-4 learning rate. I create three modified datasets with overlapping semantics following the three level label hierarchy proposed in the COCO-Stuff dataset [104] (e.g. a *car-object* is described as either "car", "vehicle", or "outdoor"). I emphasize that none of the models have been explicitly trained on the additional overlapping semantics.

Secondly, I estimate the performance gained by directly training a model with overlapping annotations on the COCO-Stuff dataset [104] as an upper performance bound. I train four ViT-Adapter models using CLIP or SBERT embeddings with two dataset variants. The first variant uniformly samples annotations from one of the three label hierarchy levels. The second variant weights sampling so all annotation classes are equally likely. Uniform and weighted sampling represent the long-tail distribution over low- and high-level semantics, respectively. Each image annotation is sampled only once for each

sample, meaning compositional semantics must be learned by generalizing from independent observations of visual appearance. Additionally, I estimate separability (4.18) and distance between the learned z^* embeddings with the optimal z^*_{opt} computed as the centroid of the ground truth overlapping semantics (4.12).

I evaluate compositional semantics by mIoU computed using the conventional *most similar* partitioning and my proposed *sufficient similarity* method introduced in Sec. 4.3.4. To use sufficient similarity I precompute τ_k for every semantic category k from 2000 samples from the training dataset covering all annotation semantics.

My results show common VLM models trained on conventional nonoverlapping annotations discover compositional semantics z^* as specified by Definition 4.1 and Theorem 4.4. Discovering z^* enables inferring overlapping semantics by my proposed sufficient similarity inference method.

Table 6.1 presents segmentation performance for original non-overlapping annotations (e.g. COCO) and my novel compositional semantics (e.g. COCO CS) dataset variants with overlapping annotations. Each model is evaluated by the conventional most similar (MS) and my proposed sufficient similarity (SS) method. Levels (denoted $^{(2)}$) specify which hierarchical semantics are being evaluated (e.g. level 1 *cat*, level 2 *animal*, and level 3 *outdoor*). The region proposal method ZSeg [81] underperforms other models explicitly trained on dense annotations despite the promise of highest generality. The mask-based conditional method X-Decoder [193] modified to output unconditional dense embedding maps performs worse than the inherently unconditional pixel-level prediction model LSeg [6]. My ViT-Adapter [396] based model implementation with a general-purpose SOTA architecture for dense vision tasks performs even better on both my novel overlapping and non-overlapping semantic inference tasks. My proposed sufficient similarity inference method improves inference performance of second level overlapping semantics across all models by 19.63 mIoU on average. Conventional most similar (MS) inference has a performance advantage over the sufficient similarity (SS) inference on level 1 semantics. The reason is that MS inference overfits level 1 semantics due to their prevalence in training data. Additionally, MS inference is fundamentally limited to predicting a single semantic, unlike SS inference which can theoretically achieve a perfect overlapping segmentation score.

Table 6.2 shows evaluation results for ViT-Adapter [396] models trained directly on overlapping COCO compositional semantics using different embedding spaces and annotation sampling strategies. Learning from a weighted sampling (WS) annotation distribution results in a uniform exposure of semantics from all levels and the best overall performance despite rarity of higher-level semantics. The performance gap between the best ViT-Adapter model in Table 6.1 and Table 6.1 on overlapping conventional

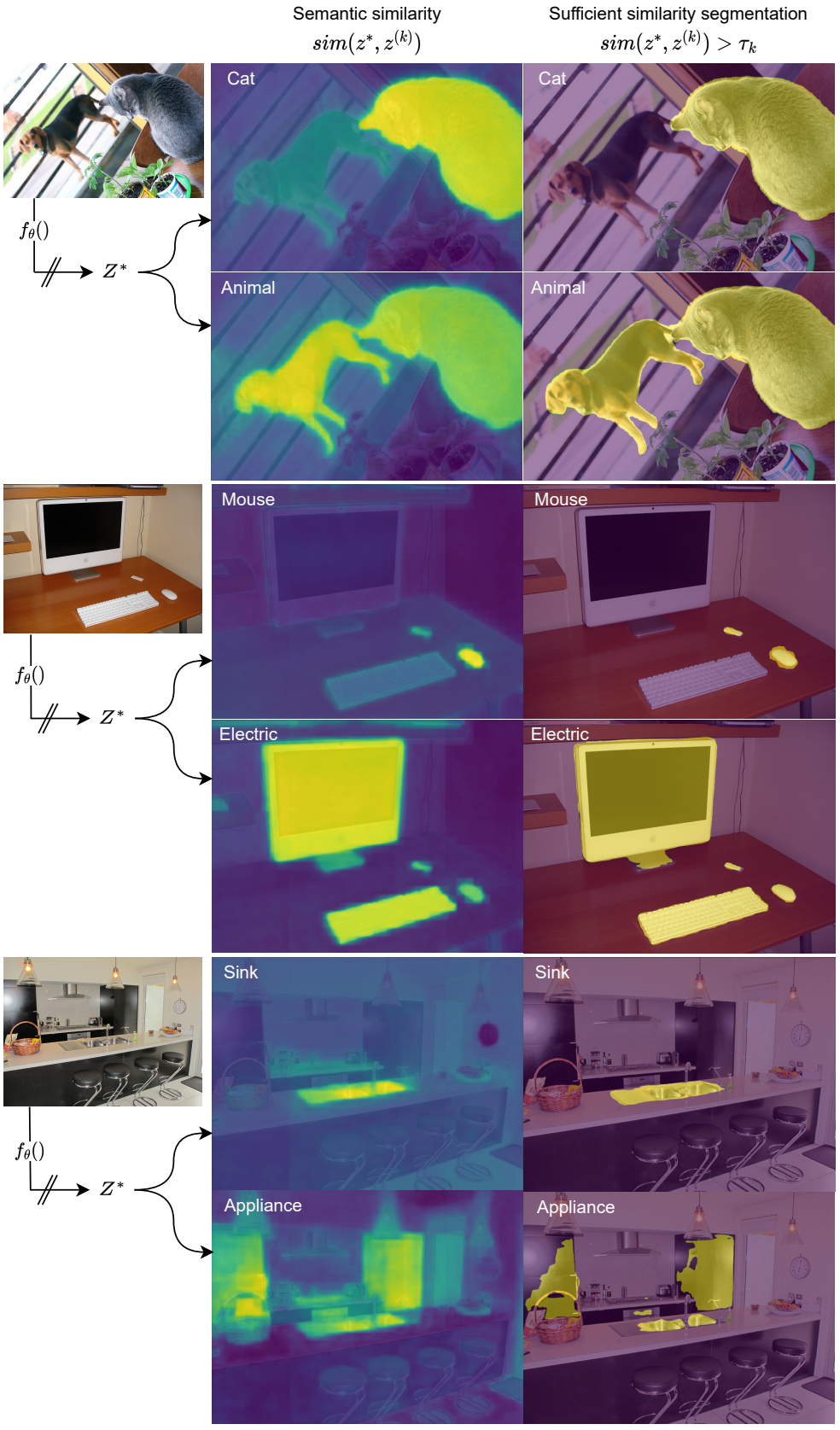

FIGURE 6.2: Examples of overlapping semantics inferrable from latent compositional semantic embeddings z^* representing learned object descriptions $\mathcal{Z}$. The 3rd and 4th examples illustrate failure cases related to sufficient similarity threshold τ_k estimation for low- and high-level semantics, respectively.

TABLE 6.1: Unconditional open vocabulary segmentation and overlapping segmentation performance

		mIoU			
Model	COCO	COCO CS			
	MS	MS	SS	$MS^{(2)}$	$SS^{(2)}$
ZSSeg [81]	11.23	10.87	2.24	3.21	8.28
X-Decoder [193]	28.57	25.33	28.14	10.19	22.52
LSeg [6]	38.42	37.10	20.41	14.59	14.66
ViT-Adapter [396]	**48.12**	**46.97**	39.16	12.55	**54.19**
Model	ADE	ADE CS			
	MS	MS	SS	$MS^{(2)}$	$SS^{(2)}$
ZSSeg [81]	9.93	9.14	3.05	6.35	5.86
X-Decoder [193]	6.48	5.88	12.77	4.85	14.39
LSeg [6]	27.40	25.11	11.37	6.35	16.23
ViT-Adapter [396]	**47.47**	43.21	30.29	28.99	**31.63**
Model	Mapillary	Mapillary CS			
	MS	MS	SS	$MS^{(2)}$	$SS^{(2)}$
ZSSeg [81]	6.51	5.39	3.39	0.67	13.75
X-Decoder [193]	11.52	9.07	9.39	2.03	20.26
LSeg [6]	30.08	24.51	11.81	0.01	22.97
ViT-Adapter [396]	**46.92**	**37.94**	24.69	0.00	**40.15**

MS: Most similar evaluation, SS: Sufficient similarity evaluation, $^{(2)}$: Level 2 semantics evaluation only, CS: Compositional semantics

TABLE 6.2: Learning compositional semantics by overlapping annotations

Model	$p(\mathcal{D})$	COCO CS [mIoU]					
		MS	SS	$MS^{(2)}$	$SS^{(2)}$	$MS^{(2,3)}$	$SS^{(2,3)}$
CLIP	US	25.90	32.99	34.19	57.29	**33.27**	55.93
	WS	**45.94**	37.89	12.71	50.18	12.18	48.82
SBERT	US	24.95	38.19	33.92	**58.55**	32.39	**57.38**
	WS	45.67	**42.23**	12.77	56.55	12.26	55.57

US: Uniform sampling, WS: Weighted sampling, $^{(2,3)}$ Level 2 and 3 semantics, SS: Sufficient similarity evaluation, MS: Most similar evaluation

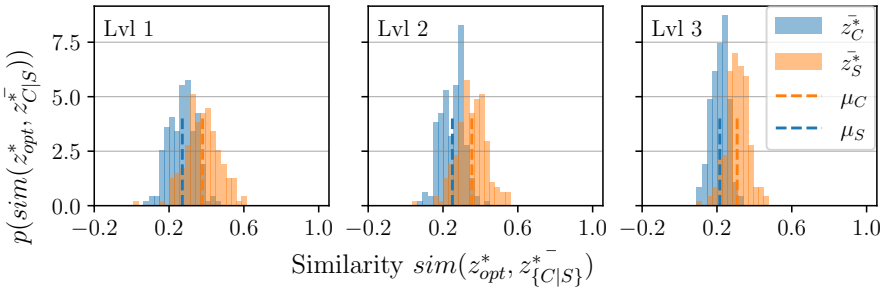

FIGURE 6.3: The distribution of mean similarities between optimal z^*_{opt} and learned z^* CLIP (blue) and SBERT (orange) embeddings for three semantic levels.

semantics is only 2.63 mIoU. The small gap indicate that learning z^* from existing single non-overlapping annotations is an effective approach. See Figure **??** and 6.2 for overlapping semantic inference visualizations.

In Figure 6.3 I visualize the mean similarity distribution between learned z^* and optimal z^*_{opt} embeddings by Theorem 4.2. Learned z^* are far from optimal z^*_{opt} for both CLIP and SBERT embedding models, similarly to how learned VL embeddings have a similarity or alignment gap with the encoded text annotations [536, 549]. However, the results in Table 6.1-6.2 proves that learned z^* have adequate similarity with z^*_{opt} for sufficient similarity segmentation of small semantic sets $\mathcal{Z}$. Increasing alignment between learned z^* and z^*_{opt} will enable z^* to represent larger $\mathcal{Z}$ and approach the theoretical capacity of the text embedding space investigated in Sec 4.3.5.

6.2.4 Observation accumulation

The agent accumulates a sequence of unfiltered semantic point clouds $P^{(1)}, \ldots, P^{(T)}$ centered in the agent's reference frame over time $t = 1 \ldots T$ into a single semantic point cloud $\bar{P}^{(T)}$. This task is called point cloud registration or scan matching problem [417]. I use the Iterative Closest Point (ICP) algorithm [284] to estimate the sensor motion and align sequential observations into the same reference frame. ICP takes the previous and latest point cloud and computes the transformation matrix $T_{t \to t+1}$ which best aligns the previous point cloud $P^{(t)}$ to the latest one $P^{(t+1)}$. The matrix $T_{t \to t+1}$ corresponds to the agent motion between the two observations as shown in (6.7). Multiplying the accumulated point cloud $\bar{P}^{(t)}$ with $T_{t \to t+1}$ as in (6.8) transforms all points into $\tilde{P}^{(t+1)}$ in reference frame of the newest observations. This step is done recursively every timestep as new observations are perceived. Finally I add the new observations $P^{(t+1)}$ to the transformed accumulated observations $\tilde{P}^{(t+1)}$, resulting in a new set of accumulated observations $\bar{P}^{(t+1)}$ as in (6.9)

$$T_{t \to t+1} = ICP(P^{(t)}, P^{(t+1)}) \tag{6.7}$$

$$\tilde{P}^{(t+1)} = T_{t \to t+1} \bar{P}^{(t)} \tag{6.8}$$

$$\bar{P}^{(t+1)} = concatenate(\tilde{P}^{(t+1)}, P^{(t+1)}). \tag{6.9}$$

A visual example of accumulated semantic point clouds is shown in Fig. 6.4.

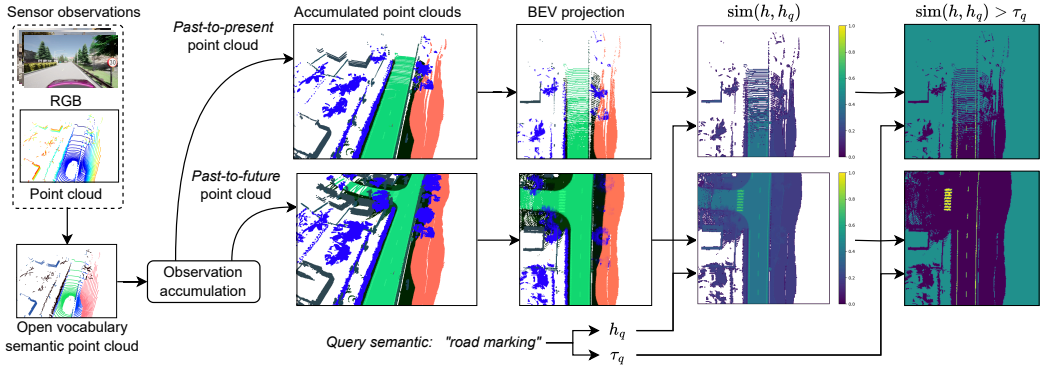

FIGURE 6.4: Process transforming sensor observations into open-vocabulary partial world states. A semantic segmentation model interprets images. The inferred semantic embedding map is attached to the point clouds. Sequential semantic point clouds are accumulated into an ego-centric reference frame. Top-down projection creates BEV representations. BEVs can be measured for similarity and sufficient similarity with a query semantic. High-dimensional semantic embeddings are projected to RGB color values for visualization

6.2.5 Partial World State Representation

The accumulated open-vocabulary semantic point cloud $\bar{P}$ encodes the agent's observable environment in a sparse spatio-semantic 3D representation. However, conventional perception and planning methods benefit from a top-down 2D representation for computational efficiency. 2D discrete grids can be processed by convolutional neural networks (CNN) [550] and visual transformers (ViT) [154] forming the backbone of SOTA latent variable generative models for images [449, 543, 551, 552].

I generate partial open-vocabulary semantic world state $x \in \mathbb{R}^{H \times W \times D}$ by projecting $\bar{P}$ into a 2D top-down bird's-eye-view (BEV) grid map spanning the region of size $(H \times W)$ around the agent. Let (i, j) index a grid cell in x. For each point $p \in \bar{P}$ with coordinates (x, y, z), I compute the grid cell indices (i, j) and append $x_{i,j}$ with the semantic embedding h of p. Set set of appended semantics $\mathcal{H} = \{h^{(1)}, \ldots, h^{(K)}\}$ of all points p coinciding with a grid cell (i, j) are averaged into the centroid h^* of $\mathcal{H}$. The theory of latent compositional semantics provides mathematical guarantees of optimally retaining the original semantics of $\mathcal{H}$ [95]. A key advantage of open-vocabulary semantic embedding representations is the inherent discrimination of unobserved or unknown information by the zero vector $\vec{0}$. In contrast, observed information is represented by unit vectors h laying on the hypersphere S^{D-1}. This naturally encodes ignorance in the model and enables distinguishing unknown from empty regions during inference.

Leveraging the theory of latent compositional semantics with sufficient similarity inference [95] allows seamlessly representing and inferring multiple overlapping semantics in the same grid cell (i, j). For example, a grid cell corresponding to a *road marking*

may also possess to *road* and *drivable* semantics, which is not principally achievable by conventional "most similar" inference as explained in Sec 6.2.2.

The presented open-vocabulary partial environment state x forms the input and learning signal to the open-vocabulary predictive world model described in the following section.

6.3 Open-Vocabulary Predictive States

Predictive World Models (PWM) aim to learn latent representations capturing the underlying structure of the environment. PWMs having learned the structure are able to supplement perception by predicting unobserved regions. Prediction generation follows the two-staged variational autoencoder (VAE) [7] latent variable approach: First, an encoder predicts a latent distribution $p(z|x)$ for the objectively real world x^* partially observed by sensors as x. Secondly, a particular latent variable z is sampled from $p(z|x)$. Finally, a decoder maps z into the most likely world x^*. The process is abstracted as the arbitrary conditioning latent variable generative model $p(x^*|x)$. In this thesis I demonstrate how to learn $p(x^*|x)$ to sample diverse and plausible complete worlds x^* from partially observed worlds x represented by open-vocabulary semantic embeddings $h \in \mathbb{R}^D$ with dimension $D >> 1$.

6.3.1 Experiments

In this section I describe the experiments conducted to measure how well an open-world predictive world model (OV-PWM) can learn a compact latent representation of environments represented by high-dimensional open-vocabulary embeddings.

I set up my experiments using the open source autonomous driving simulator CARLA [545]. The simulator provides a set of realistic 3D environment, a traffic manager, and supports accurate rendering of synchronized sensor data streams like RGB images, depth maps, and lidar point clouds. I used the latest 0.9.15 release.

The experimental set up is explained next. I run the simulator and collect approximately 20 minutes of observational experience from environments *Town05*, *Town06*, *Town07*, and *Town10* as observational experience or training data. A separate environment *Town04* is used for evaluation. The environments are chosen based on providing road marking semantics. I compute and append ideal latent compositional semantics to the point cloud according to a three level taxonomy with overlapping semantics as explained in Sec 6.2.2. Semantics are encoded as 768 dimensional SBERT embeddings [179]. Next I process the sequential observations into accumulated semantic point clouds as

explained in Sec 6.2.4. All points 2m above ground are filtered. Dynamic objects are filtered by sufficient similarity inference. From accumulated point clouds I generate BEV partial world representations as explained in Sec 6.2.5. I use the same translation and warping data augmentation technique as detailed in prior work [381] on the model training samples to improve generalization. The evaluation samples are not augmented. The resulting number of training and evaluation samples are 7145 and 178 samples, respectively.

The HVAE model is trained on the generated training samples for 180K iterations for four days using six A6000 GPUs. See the public code repository for hyperparameter details. The trained HVAE model is evaluated on the separate evaluation set of unaugmented samples.

The following two metrics are employed to measure the goodness of the OV-PWM model. First, semantic similarity between the predicted embedding maps $\hat{x}^*$ and future observed worlds x^* is measured as the mean cosine distance between the predicted and observed open-vocabulary embeddings $x^*_{i,j} \in S^{D-1}$ and $\hat{x}^*_{i,j} \in S^{D-1}$ covered by the observed element mask M

$$\text{sim}(x^*, \hat{x}^*) = \frac{1}{|M|} \sum_{(i,j)\in M} \text{sim}(x^*_{i,j}, \hat{x}^*_{i,j}) = \frac{1}{|M|} \sum_{(i,j)\in M} (x^*_{i,j})^T \cdot \hat{x}^*_{i,j}. \tag{6.10}$$

Secondly, semantic accuracy is measured by intersection over union (IoU) of queried semantics. I compute IoU based on sufficient semantics interpretation of unconditional open-vocabulary semantics according to the theory of latent compositional semantics [95]. The OV embedding maps $\hat{x}^*$ and x^* are first checked element-wise for membership with the query semantic by an a priori computed sufficient similarity threshold value τ_{sem}

$$b_{i,j} = \begin{cases} \mathbb{T}, & \text{if sim}(x_{i,j}) > \tau_{sem} \\ \mathbb{F}, & \text{otherwise} \end{cases}$$

resulting in the boolean maps b and $\hat{b}$ with elements represented as true $\mathbb{T}$ and false $\mathbb{F}$. The query semantic IoU is computed as

$$\text{IoU}(x^*, \hat{x}^*) = \frac{\sum_{(i,j)\in M} b_{i,j} \cap \hat{b}_{i,j}}{\sum_{(i,j)\in M} b_{i,j} \cup \hat{b}_{i,j}} \tag{6.11}$$

with the boolean map $\hat{b}$ obtained from $\hat{x}^*$ considered as ground truth target. The mean IoU (mIoU) is used to quantify the performance over a set of query semantics H

$$\text{mIoU} = \frac{1}{H} \sum_{h\in H} \text{IoU}_h. \tag{6.12}$$

I estimate optimal sufficient similarity threshold values for query semantics τ_q by logistic regression models maximizing likelihood over the train split observations following prior work [95]. The optimal τ_q is the decision boundary or $(sim)(x, x_q)$ separating true positive and negative points with least error according to the model

$$\tau_q = \max \left[\text{MemberOf}(x, q) \; p(\text{MemberOf}(\text{sim}(x, x_q) \geq \tau_q, q)) \right]. \tag{6.13}$$

I provide a set of unconditionally sampled world states $\hat{x}^*$ to asses the robustness of the learned open-vocabulary world model. Unconditional generation starts by randomly sampling the deepest latent variable $z_K \in \mathbb{R}^{16}$ in (5.20) and generate $\hat{x}^*$ without conditioning on a partially observed world x as input.

6.3.2 Results

In this section I present the CARLA simulator experiment results. The results shows that environments represented by high-dimensional open-vocabulary semantic embeddings can be accurately modeled by the predictive world modeling approach. Additionally, I analyze the results with a perspective on potential real world, large-scale application.

Table 6.3 shows semantic IoU prediction accuracy for an urban environment sequence not in the training sample distribution. I apply a "best of N samples" evaluation approach [381] to demonstrate how sampling of diverse structures improves the likelihood of predicting the actual world from partial observations. The mean IoU prediction over all elements (i, j) and semantics is 65.13 mIoU with 1 sample, and increases to 69.19 mIoU with 32 samples. Modeling and predicting fine spatial patterns like *road markings* is challenging and reaches only 22.99 IoU over 32 samples. The advantage of generative modeling is most apparent in less predictable large semantic structures like *vegetation* and *sidewalk* as sampling increases accuracy by 9.45 and 9.10 IoU points, respectively. Over all semantics sampling increases the mean IoU by 4.06 IoU points.

Table 6.4 shows IoU prediction accuracy on a highway sequence not in the training distribution. Predictive performance for highway environments is generally higher than urban environments due to higher determinism. However, *road marking* predictability is lower due to lacking localized contextual cues such as intersection and narrow road structures.

Table 6.5 shows performance on a random subset of 200 samples from the training distribution. The results indicate that model training is not yet saturated on the limited training dataset as semantics like *road marking*, *side walk*, and *vegetation* has margin to improve. Comparison with test set performance given in Table 6.3 shows comparable

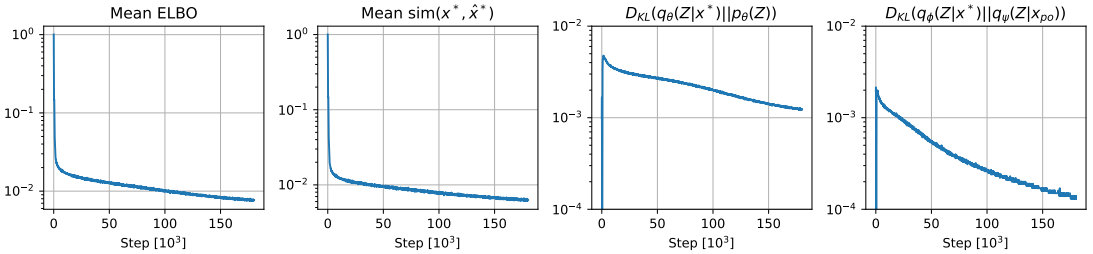

FIGURE 6.5: Training plots. The mean ELBO (5.22), cosine distance (5.25), posterior (5.23) and posterior matching (5.24) distribution separations metrics continue to decrease with additional compute.

TABLE 6.3: World model prediction accuracy by "best of N samples" on the urban test sequence.

#Samples		IoU 1	2	4	8	16	32
road	All	92.75	93.36	93.61	93.89	94.20	94.33
	Unobs.	84.07	85.70	86.10	86.69	87.54	87.74
road marking	All	21.02	21.21	21.95	22.31	22.91	22.99
	Unobs.	12.85	13.77	14.24	14.84	15.47	16.00
side walk	All	51.39	53.45	56.49	57.38	59.57	60.49
	Unobs.	41.50	45.51	48.72	50.33	52.07	52.53
vegetation	All	34.91	37.25	40.54	41.67	43.42	44.36
	Unobs.	28.11	31.97	35.08	36.27	37.96	40.02
static	All	97.61	97.61	97.85	98.08	98.12	98.23
	Unobs.	97.73	97.88	98.15	98.22	98.35	98.40
drivable	All	93.10	93.69	93.94	94.25	94.63	94.71
	Unobs.	84.89	86.60	87.00	87.55	88.52	88.72
mIoU	All	65.13	66.10	67.40	67.93	68.81	69.19
	Unobs.	58.19	60.24	61.55	62.32	63.32	63.90

performance with the training set, meaning generalization is achieved. As the training performance continues to improve log linearly as shown in Figure 6.5, it is reasonable to conclude that generalization performance will continue to improve with additional training.

My proposed OV-PWM framework lacks direct comparative baselines. To the best of my knowledge, only my prior work leverages lidar point clouds with generative modeling to predict spatial environments without requiring ground truth map data [381]. The prior closed set predictive world model trained on KITTI-360 data [553] is quantitatively evaluated only for *road* semantics and achieves 98.73 IoU. I consider my open-vocabulary urban environment result of 94.33 IoU to be of comparable quality and thus conclude learning open-vocabulary world models performs equivalently to closed set world models

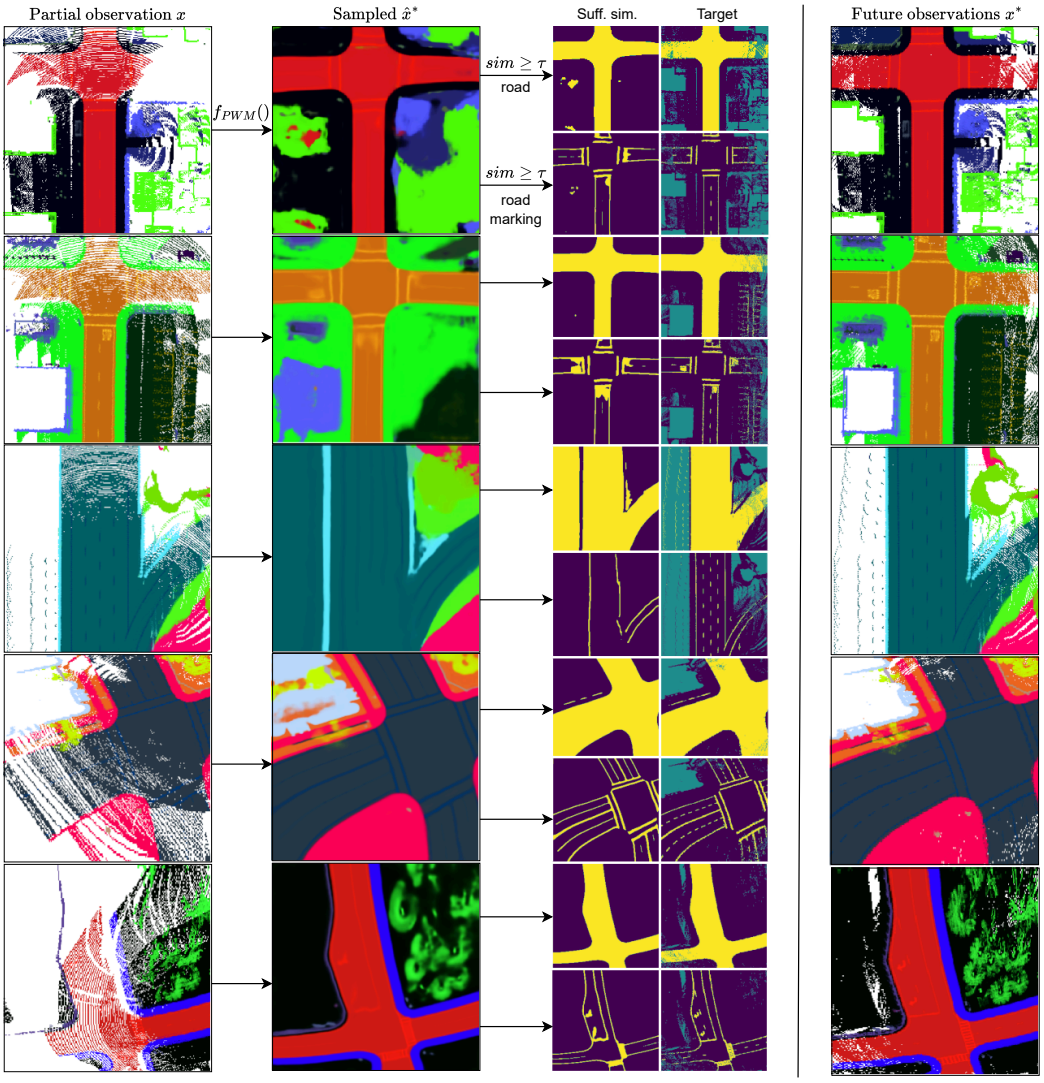

FIGURE 6.6: Conditional sampling visualizations. The high-dimensional open-vocabulary partial observation input x and sampled predictive world model output $\hat{x}^*$ are projected into RGB images by PCA. Semantic inference by sufficient similarity are shown in the third column. The actual worlds perceived by future observations are shown in the forth column. The first three rows shows evaluation samples. The remaining two rows shows samples from the training distribution.

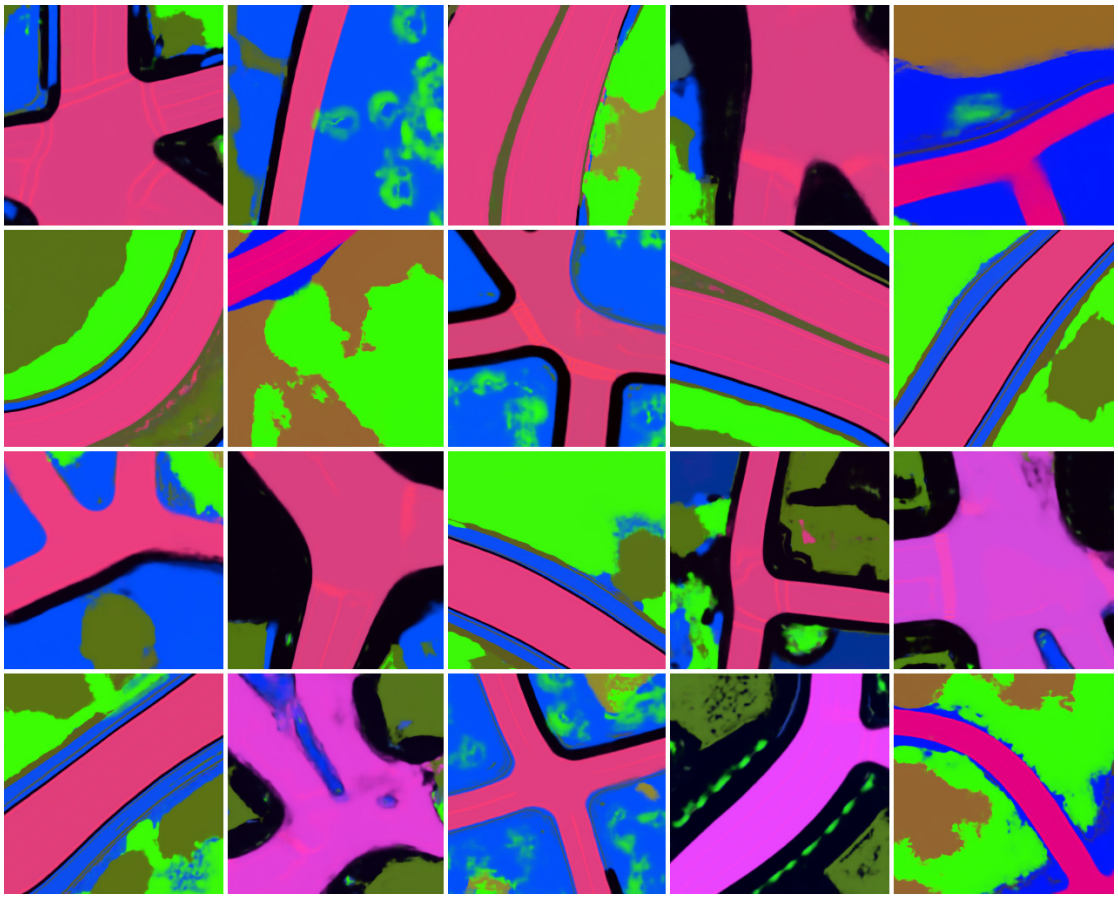

FIGURE 6.7: Unconditional sampling visualizations. High-dimensional open-vocabulary embedding maps are generated by the predictive world model $p_\theta(x|Z)$ through sampling from the learned prior distribution $p_\theta(Z)$. The embedding maps are visualized as RGB images by PCA projection.

TABLE 6.4: World model prediction accuracy by "best of N samples" on the highway test sequence.

#Samples		IoU 1	2	4	8	16	32
road	All	98.01	98.15	98.20	98.29	98.31	98.34
	Unobs.	95.93	96.68	96.96	97.14	97.28	97.44
road marking	All	9.90	11.15	11.19	12.02	12.19	13.20
	Unobs.	9.51	10.61	10.67	11.53	12.09	12.49
vegetation	All	38.29	38.64	39.22	40.01	40.23	40.45
	Unobs.	43.11	43.68	44.27	44.83	45.37	45.33
static	All	98.54	98.73	98.79	98.83	98.88	98.90
	Unobs.	95.10	96.40	96.64	97.10	97.36	97.49
drivable	All	98.02	98.15	98.21	98.29	98.31	98.34
	Unobs.	95.91	96.62	96.92	97.07	97.23	97.41
mIoU	All	68.55	68.96	69.12	69.49	69.58	69.85
	Unobs.	67.91	68.80	69.09	69.53	69.87	70.03

TABLE 6.5: World model prediction accuracy by "best of N samples" on the training distribution.

#Samples		IoU 1	2	4	8	16	32
road	All	97.16	97.20	97.30	97.35	97.38	97.42
	Unobs.	95.00	95.27	95.45	95.63	95.79	95.89
road marking	Obs.	34.14	34.33	34.62	34.83	35.04	35.24
	Unobs.	26.79	27.06	27.57	28.13	28.37	28.53
side walk	All	58.23	58.36	58.56	58.93	59.10	59.11
	Unobs.	55.03	56.17	56.20	57.12	57.54	57.69
vegetation	All	75.44	76.06	76.71	76.98	77.21	77.57
	Unobs.	66.33	68.01	68.67	70.03	71.29	71.89
static	All	98.79	98.81	98.81	98.82	98.83	98.84
	Unobs.	98.47	98.56	98.56	98.59	98.61	98.62
drivable	All	97.27	97.32	97.41	97.47	97.51	97.55
	Unobs.	95.47	95.83	96.09	96.25	96.27	96.39
mIoU	All	76.84	77.01	77.24	77.40	77.51	77.62
	Unobs.	72.85	73.48	73.76	74.29	74.65	74.84

while greatly simplifying the learning method to a one-stage end-to-end paradigm, as explained in Sec 6.3.

Other comparative baselines include image-based methods which generally are not generative models and trained and evaluated on the same ground truth data domain (e.g. within the same city). One such baseline is a recent SOTA image-based monocular model [260] achieving 68.34 road IoU on the KITTI Raw dataset [554]. The performance difference exemplify the advantage of leveraging lidar point clouds as done in my method.

Figure 6.6 provides visual examples of plausible world samples $\hat{x}^*$ generated from partial observations x. Examples of semantic inference by sufficient similarity are shown. The actual world perceived in future observations are included for comparison. The examples illustrates how large structures like *road* are accurately learned. Finer semantic details like *road markings* are comparatively challenging to represent and predict. However, training samples display improved granularity of fine semantics, indicating that further training on a larger training distribution covering additional pattern may enhance performance.

Figure 6.7 display a set of randomly sampled environments from the learned prior distribution $p_\theta(Z)$. The sampled environments showcase intricate details like road markings and semantically plausible configurations. Some generated samples are partially degenerate. Additional optimization of the learned prior $p_\theta(Z)$ and generative model $p_\theta(x|Z)$

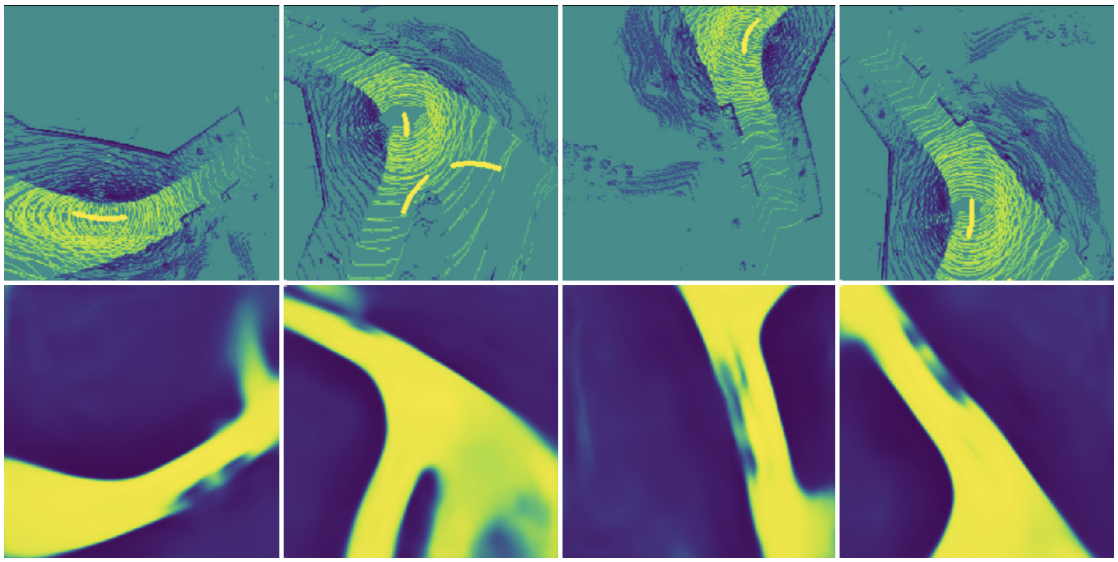

FIGURE 6.8: A visual example of how non-hierarchical VAEs [7] have limited capacity to represent high-dimensional structured data with high fidelity. The top row represent observed "road" semantics. The bottom row show predicted fuzzy "road" structures. The filled lines in the upper row are observed vehicle trajectories which presumably indicate "road".

is expected to reduce the likelihood of degenerate samples. Figure 6.5 shows that both $p_\theta(Z)$ and $p_\theta(x|Z)$ are likely to improve with additional training.

Figure 6.8 shows the representational capacity for a non-hierarchical vanilla VAE [7] with 128 dimensional latent variable z trained to represent boolean "road" structures as a one-dimensional Bernoulli distribution $p(x_{i,j} = \text{road}|z)$. It is clear that vanilla VAEs struggle to represent high-resolution structured data with high fidelity, let alone modeling a latent distribution of 768 dimensional open vocabulary semantic embeddings as possibly by HVAE models [447–449].

The predictive world model mean inference time is 0.175 sec or 5.71 Hz on an RTX 4090 GPU. My method is thus applicable for real-time application given a modern SLAM implementation [286, 287, 555] capable of operating faster than sensor frame rates.

6.4 Learning Navigational Patterns by Predictive States

6.4.1 Predictive state representation

I generate partial world states based on accumulated sensor observations following the method described in prior work [1]. The method shares similarities with a hierarchical biological model of human representation and processing of visual information [556]. The agent is initialized within an unknown metric vector space. Sensor observations are

projected onto this common vector space at discrete timesteps. Semantic information is inferred from images using a pretrained semantic segmentation model and appended to coincident 3D points to form semantic point clouds. Past semantic point clouds are integrated with new observations by scan matching using the ICP algorithm [557] and SLAM [558] for loop closure. The accumulated semantic point cloud is reduced to a five-layered 2D probabilistic BEV representation $x \in \mathbb{R}^{I \times J \times C}$ with dimension $I \times J$ elements, and C denoting the number of semantic information channels. In this work, C consists of five channels representing the semantic attributes of a spatial point (i, j); I represent road probability $p(\text{road})$ by a beta distribution, lidar reflection intensity ϵ as a scalar value, and visual appearance by RGB values.

Dynamic objects are detected by a pretrained object detection model and represented by 3D bounding boxes. Trajectory observations are generated by temporally tracking detected objects. Dynamic objects are considered "moving" if motion is observed or "static" otherwise. This classification allows filtering away observations associated with moving dynamic objects while keeping observations of static dynamic objects for training, as they may influence how other agents navigate the environment such as swerving out of the lane to avoid a parked car. The static dynamic objects can be removed at inference time to provide an agent-agnostic prediction of navigational patterns akin to a lane map.

The predictive world model [1, 2] samples diverse and plausible complete world states $\hat{x}$ conditioned on partially observed world states x as exemplified in Fig. 3.4. The world model is functionally similar to the biological ventral cortical pathway as the model disambiguates the partially observed environment by leveraging past experience [353]. The world model is computationally conceptualized as an arbitrary conditioning generative model and implemented by the recent SOTA hierarchical VAE (HVAE) model VDVAE [449] with the encoder module replaced by a posterior matching encoder [1]. In this work, the HVAE models the joint distribution of observable variables $p(r, \epsilon, R, G, B)$ factorized as the conditional distribution

$$p(r, \epsilon, R, G, B) = p(R|G, B, r)p(G|B, r)p(B|r)p(\epsilon|r)p(r) \tag{6.14}$$

using hierarchical latent variables z. Here r and ϵ denote road and lidar reflection intensity, and RGB are image color channels. The latent variable prior $p(z)$ and posterior $q(z|x)$ distributions are factorized as

$$p(z) = p(z_1|z_2) \dots p(z_{K-1}|z_K)p(z_K) \tag{6.15}$$

$$q(z|x) = q(z_1|z_2, x) \dots q(z_{K-1}|z_K, x)q(z_K|x) \tag{6.16}$$

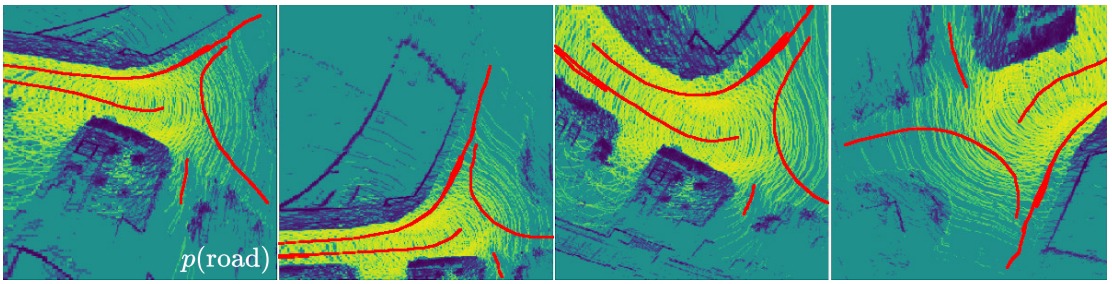

FIGURE 6.9: Geometric data augmentation generates diverse sample variations from a single real sample. Spatial information (dense maps) and observed trajectories (red lines) are transformed by the same function.

with random variables z modeled by normal distributions.

The world model learns to approximate the prior and posterior distributions by the parameterized models $q_\theta(z|x)$ and $p_\theta(x|z)$ using variational inference [446] and trained using self-supervised learning to predict future observations from present observations akin to the predictive coding problem [94]. Note that the vanilla HVAE cannot learn to generate diverse complete representations from partially observed representations only. I follow the posterior matching optimization method visualized in Fig. 5.1 and presented in prior work [1] to overcome this limitation. The method trains a regular HVAE using pseudo ground-truth world states x^*_{full}, and a secondary encoder $q_\phi(z|x)$ to predict a similar hierarchical latent distribution $z = \{z_1, \ldots, z_K\}$ as the primary encoder $q_\theta(z|x^*_{full})$ from x.

At inference time the model uses the partially observed encoder to generate a latent distribution $q_\phi(z|x)$ that can be decoded by $p_\theta(\hat{x}|z)$ into a completely observed plausible world state $\hat{x}$ similar to a pseudo ground-truth world state x^*_{full} without the need to observe the future.

6.4.2 Data Augmentation

I leverage geometric data augmentation [45] on all training samples to improve model generalization performance by learning geometric invariance. By learning geometric invariance, learned models are able to generalize beyond particular observed road scene geometries within the dataset. The same augmentation is applied on both the dense predicted environment state and observed trajectories.

Data augmentation is performed by random rotation and applying component-wise polynomial warping [559] to the road scene context and trajectory label. In the following ξ is a substitute for spatial coordinates i and j, and ξ' denotes warped coordinates. The

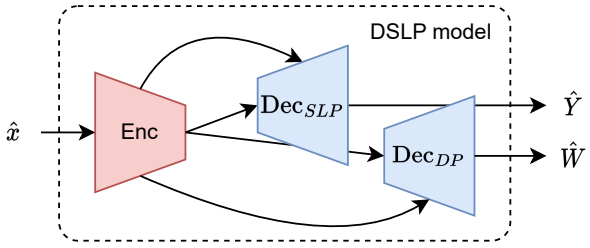

FIGURE 6.10: The Directional Soft Lane Probability (DSLP) model uses a dual decoder U-Net [8] model to transform a plausible world state $\hat{x}$ into a soft lane probability (SLP) map $\hat{Y}$ and directional probability (DP) tensor $\hat{W}$.

warping is specified by the following nonlinear function

$$a_0(\xi')^2 + a_1(\xi') + a_2 = \xi \tag{6.17}$$

with the following boundary conditions: $\xi' = 0 \wedge \xi = 0$, $\xi' = \xi_{max} \wedge \xi = \xi_{max}$, $\xi' = \xi'_0 \wedge \xi = \xi_0$. The input dimension is denoted ξ_{max}. The warp is defined by setting ξ'_0, ξ_0). The coefficients in 6.17 are derived using the previous boundary conditions

$$a_0 = \frac{1 - a_1}{\xi_{max}}, a_1 = \frac{\xi_0 - (\xi'_0)^2/\xi_{max}}{\xi'_0(1 - \xi'_0/\xi_{max})}, \tag{6.18}$$

where (i_0, j_0) are set to the input state mid-point, and the warping location (i'_0, j'_0) is sampled from a radial Normal distribution $\mathcal{N}(\xi|\mu, \sigma^2)$ with a mean μ centered at radius 0.15 ξ_{max} with values above 0.3 ξ_{max} clipped. I create dense warp maps by using the inverse function of (6.17) to map each warped coordinate ξ' to an original coordinate ξ. Fig. 6.9 shows visual examples of a sample augmentation.

6.4.3 Directional Soft Lane Probability Model

Here I present a method to train a model to predict unbiased probability maps of local directional traversability. The model input is the plausible world state $\hat{x}$ described in Sec. 5.4. I also present a method for inferring global navigational patterns from the local probability maps. See Fig. 3.4 and Fig. 6.13 for output visualizations.

The model is implemented by a U-Net neural network [8] with a single encoder and two decoders as illustrated in Fig. 6.10. The first decoder outputs a probability map $Y \in \mathbb{R}^{I \times J}$ representing soft lane probabilities for elements in a grid map of size $I \times J$. The second decoder outputs a map of categorical distributions $W \in \mathbb{R}^{M \times I \times J}$ representing M direction interval probabilities for each location (i, j). The methods for optimizing both probabilistic outputs are explained below.

6.4.3.1 Soft Lane Probability (SLP) Modeling

The likelihood of each environment location (i, j) being traversed by an unspecified agent is modeled by the predicted probability value $\hat{y}_{i,j} \in \hat{Y}$ and is called soft lane probability (SLP). Learning to predict an unbiased $\hat{Y}$ from partial observations is nontrivial, as the self-supervised learning signal contains false negative traversal observations (i.e. lacking an observed trajectory where traversals are probable). I formalize the problem as follows. Ideally I want to learn a distribution $q(y)$ that approximates the true distribution $p(y)$. However, optimizing $q(y)$ according to the learning signal results in learning the distribution of partially observed samples $\tilde{p}(y)$. A principled solution is to use a regularizer to decrease bias and make $q(y)$ better match $p(y)$.

In this thesis I present a semi-supervised objective that enables learning an unbiased probabilistic prediction of traversability based on an information-theoretic regularizer derived from balancing the information contribution from positive and negative partial observations in Y.

In information theory, the entropy $H(y)$ of a distribution $p(y)$ is considered a quantity that measures information content. The cross-entropy

$$H(p, q) \triangleq -\sum_{k=1}^{K} p(y = k)\ \log(q(y = k)) \tag{6.19}$$

measures the information overhead to compress a sample $y \sim p(y)$ using a code based on $q(y)$ [560].

Each partial observation Y contains two distinct groups of traversal information; a set of true positives representing certain information, and a set of true and false negatives representing uncertain information. The contributed information of the set of positive and negative observations are

$$H(Y_{pos} \subseteq Y, \hat{Y}) = -\sum_{i,j \in Y_{pos}} y_{i,j}\ log(\hat{y}_{i,j}) \tag{6.20}$$

$$H(Y_{neg} \subseteq Y, \hat{Y}) = -\sum_{i,j \in Y_{neg}} (1 - y_{i,j})\ log(1 - \hat{y}_{i,j}). \tag{6.21}$$

I devise a regularizer based on balancing the information contribution provided by (6.20) and (6.21) according to the ratio of observations

$$\alpha_{IB} = |Y_{pos}|\ /\ (|Y_{pos}| + |Y_{neg}|) \tag{6.22}$$

where $|Y_*|$ denotes the number of positive and negative observed elements (i, j). The balanced information contribution $H^*(Y|\hat{Y})$ is obtained by linearly interpolating the information contributions according to the ratio of observations

$$H^*(Y|\hat{Y}) = \alpha_{IB}\, H(Y_{neg}|\hat{Y}) + (1 - \alpha_{IB})\, H(H_{pos}|Y). \tag{6.23}$$

Linear interpolation is a monotonic function that balances the information contributions while preserving the total information quantity

$$0 \leq H^*(Y|\hat{Y}) \leq \max(H(Y_{pos}|\hat{Y}), H(Y_{neg}|\hat{Y})). \tag{6.24}$$

I formulate the problem specific optimization objective $\mathcal{L}_{SLP}$ as the mean balanced information contribution

$$\begin{aligned} \mathcal{L}_{SLP} = -\frac{1}{|Y|} \sum_{i,j \in Y} & [\alpha_{IB}(1 - y_{i,j}) log(1 - \hat{y}_{i.j}) \\ & + (1 - \alpha_{IB}) y_{i,j} log(\hat{y}_{i,j})] \end{aligned} \tag{6.25}$$

where $\hat{y}_{i,j}$ and $y_{i,j}$ is the predicted and observed soft lane probability for the element located at i, j. $|Y|$ denotes the number of traversable elements. The information contribution ratio α_{IB} provides the optimal interpolation between positive and negative traversal observations.

One can view (6.25) as the cross entropy objective with an additional dynamic regularizer between positive and negative observations. Experiments show that the balanced information contribution cross-entropy objective (6.25) performs better than finetuning a static hyperparameter weighting [45], and allows learning probabilistic predictions despite occasional abnormal observations unlike the barrier loss objective [240].

The negative log likelihood NLL_{SLP} of an observed sample y according to a model prediction $\hat{y}$ based on modeling $p(y|\hat{y})$ as a Bernoulli distribution is

$$\text{NLL}_{\text{SLP}} = -\sum_{i,j \in Y} \left[y_{i,j}\, log(\hat{y}_{i,j}) + (1 - y_{i,j})\, log(1 - \hat{y}_{i,j}) \right]. \tag{6.26}$$

6.4.3.2 Directional Probability (DP) Modeling

The likelihood of local traversal directionality at each location (i, j) is modeled by the predicted vector $\hat{w}_{i,j}$ called directional probability (DP). The $\hat{w}_{i,j}$ models a categorical probability distribution representing the direction interval $\theta \in [0, 2\pi)$ by M uniformly

spaced intervals

$$w_{i,j} = (p(\theta \in [0, \frac{2\pi}{M})), \ldots, p(\theta \in \left[\frac{(M-1)2\pi}{M}, 2\pi\right)))^T. \tag{6.27}$$

The learning signal is created by encoding observed trajectories into $w_{i,j}$ as a discrete von Mises distribution. In the case of multiple overlapping trajectories the individual distributions are superimposed and renormalized. Learning to match distributions improve multimodal prediction compared with learning to predict single values by maximum likelihood estimation [45].

The optimization objective $\mathcal{L}_{DP}$ is formulated as learning to predict the directional distribution by minimizing the mean KL divergence between predicted $\hat{w}_{i,j}$ and observed $w_{i,j}$ directionality over all elements $w_{i,j} \in W$

$$\mathcal{L}_{DP} = \frac{1}{|W|} \sum_{i,j \in W} D_{KL}(w_{i,j} || \hat{w}_{i,j}). \tag{6.28}$$

Note that the learning signal used to optimize the DP objective (6.28) lacks false negatives and therefore does not require regularization like the SLP objective (6.25).

The negative log likelihood NLL_{DP} of an observed sample $w_{i,j}$ according to a model prediction $\hat{w}_{i,j}$ based on modeling $p(w|\hat{w})$ as a categorical distribution is

$$\text{NLL}_{\text{DP}} = -\sum_{i,j \in Y} \sum_{m=1}^{M} w_{i,j}^{(m)} \, log(\hat{w}_{i,j}^{(m)}). \tag{6.29}$$

6.4.3.3 Maximum likelihood lane graph inference

Evaluating the goodness of local navigational patterns using the predicted DSLP field is straightforward. To also evaluate the usefulness of the predicted DSLP field for inferring global navigational patterns, I present a sampling-based method to generate a maximum likelihood road lane graph fitted to the predicted DSLP field. The graph generation process is illustrated in Fig. 6.11.

First, I infer entry and exit points at the edges of the predicted DSLP field. A non-maximum suppression (NMS) operation is performed on the SLP field $\hat{Y}$ to find the most likely path centers. Each point is designated as an entry and/or exit point according to the predicted DP field $\hat{W}$. Additional entry and exit points are inferred from directional field regions which are coherent but lack a NMS point.

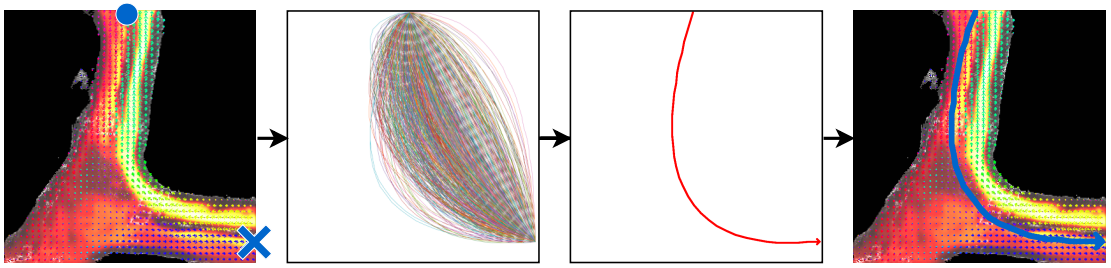

FIGURE 6.11: The maximum likelihood graph is generated by connecting entry (•) and exit (×) points by the most probable of many sampled paths given the predicted DSLP field.

Secondly, I incrementally build a graph by searching for valid connecting paths between all entrance and exit points by a sampling-based approach. A set of second-degree polynomial spline paths is generated between an entry and exit pair by randomly sampling a valid spline control point $(i, j)^*$ from a normal distribution with rejection sampling. The likelihood of each sampled path is evaluated using the location and directionality of M equidistant points along the path given the predicted DSLP field using (6.26) and (6.29). The path with the lowest total NLL is selected as the best path. Repeating this process results in a set of most likely paths representing the maximum likelihood graph. A post-processing operation removes undesired edges between neighboring lanes (i.e. u-turns) using a simple distance threshold heuristic. Representing navigational patterns by splines is a useful inductive bias, as agents tend to navigate structured environments in a continuous and smooth manner.

6.4.3.4 Experiments

I evaluate the model performance on the right-side driving daytime Boston scenes in the nuScenes dataset [561] similar to my baseline methods [313, 314]. The observation accumulation method described in Sec. 6.2.4 generates a partially observed training sample x every 1 m using accumulated observations from six 360° field-of-view RGB cameras and a top-mounted 32 beam lidar and a single pretrained semantic segmentation model [1]. Each x is augmented 20 times. Partitioning the generated training samples into the nonoverlapping regions shown in Fig. 6.12 results in 60,960 (34.7 %), 40,960 (23.3 %), and 73,780 (42.0 %) samples for regions 1 to 3. Evaluation region 4 contains samples generated every 10 m without augmentation. I use a semantic segmentation model pretrained on two different public datasets [1]. I accumulate observations using ground truth pose information to reduce engineering effort, as prior work demonstrates the feasibility of accumulation based on pose estimation [1]. The plausible world state model input representation $\hat{x}$ consists of a five-layered 256×256 grid map encompassing a 51.2×51.2 m region similar to prior work [313].

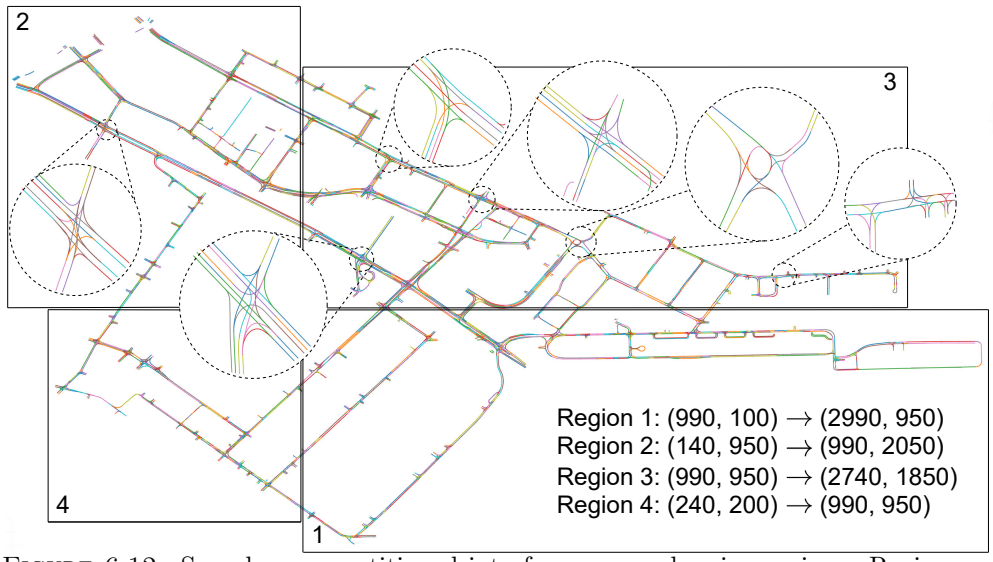

FIGURE 6.12: Samples are partitioned into four nonoverlapping regions. Regions are specified by bottom-left and top-right corners in world coordinates.

I conduct a model hyperparameter study and find that a smaller 1.4 M parameter model generalizes best. The model as depicted in Fig. 6.10 has a common 8-layered CNN encoder with filter count increasing from 16 to 256, and two 8-layered CNN decoders with bilinear upsampling and filter count decreasing from 64 to 8. See the code for further implementation details.

I use the following benchmarks to evaluate my DSLP model. I compare the global navigation pattern inference performance against the two most relevant and recently published SOTA supervised models STSU [314] and LaneGraphNet [313]. Both baselines are trained on nuScenes data [561] to predict lane graphs using complete ground truth graphs as supervision. I compare the local probability field estimation performance against the prior self-supervised SOTA model called DSLA [45].

Local probability field estimation. I evaluate the predicted soft lane $\hat{Y}$ and directional $\hat{W}$ probability fields by computing the summed negative log-likelihood (NLL) of the ground truth lane map using (6.26) and (6.29). Lower NLL means the ground truth lane map is more likely according to the model. Directional accuracy measures the ratio of elements within $\pm 45°$ of the ground truth direction.

Global navigational pattern inference. I evaluate the usefulness of the predicted probability fields for inferring global navigational patterns by computing the intersection over union (IoU) and F1 score between the maximum likelihood graph and ground truth lane map. My method does not consider the spacing of graph nodes as an integral part of navigational patterns and thus does not view node displacement as a relevant performance metric.

TABLE 6.6: Performance of predicted local probability fields

		NLL_{SLP}	NLL_{DP}	NLL	Dir. acc.
DSLA [45]		2.499	12.596	15.095	0.864
DSLP	const α	0.423	12.241	12.663	0.855
	mean α_{IB}	0.444	12.038	12.482	0.881
	α_{IB}	0.556	11.769	12.325	0.892
	full obs.	**0.539**	**11.666**	**12.205**	**0.900**

Ablation studies. I evaluate the advantage of my proposed predictive world modeling approach [1] for learning navigational patterns from sampled plausible completed worlds $\hat{x}$ instead of partially observed worlds x. I conduct an experiment using unaugmented samples to quantify the performance contribution of my geometric data augmentation method [45] on real-world data. I conduct experiments on dataset splits including a different number of regions to estimate how performance increases with additional data.

6.4.3.5 Results

Local probability field estimation. Table 6.6 presents evaluation results for the predicted probability fields. My proposed DSLP model optimized with the information balance regularizer α_{IB} (6.22) predicts the least biased probability field among all models trained and evaluated on accumulated past observation inputs. I conclude that the probabilistic objective (6.25) substantially reduces bias compared with the non-probabilistic DSLA affordance objective [45]. Training and evaluating on accumulated past and future observation inputs in an offline map creation manner (i.e. full obs.) reduces bias, demonstrating that more comprehensively observed environments result in better performance. I performed experiments with different constant α values to demonstrate the merit of the proposed hyperparameter-free regularizer α_{IB} (6.22). The best constant weight α value 0.1, found over five hyperparameter experiments, results in worse performance than using α_{IB}. I demonstrate the merit of dynamic, per-sample computed α_{IB} values (6.22) by running an experiment with the constant mean α_{IB} value 0.122 computed over all training samples, which results in worse performance. See Fig. 6.13 for probability field visualizations.

Global navigational pattern inference. Table 6.7 presents results showing that the maximum likelihood graph fitted to the probability field predicted by my self-supervised DSLP and prior DSLA model [45] from partially observed world representations x, outperforms the supervised SOTA baselines STSU [314] and LaneGraphNet [313] trained on ground truth lane graphs. My self-supervised method not only improves upon the

TABLE 6.7: Performance of global navigational pattern inference

		IoU	F1 score
STSU [314]		0.389	0.560
LaneGraphNet [313]		0.420	0.574
DSLA [45]		0.427 (0.128)	0.839 (0.07)
DSLP	constant α	0.418 (0.146)	**0.853** (0.08)
	mean α_{IB}	0.410 (0.147)	0.846 (0.08)
	α_{IB}	0.442 (0.125)	0.834 (0.07)
	full obs.	**0.454 (0.128)**	0.839 (0.08)

TABLE 6.8: Ablation studies

WM	Aug.	$\text{NLL}^*_{\text{SLP}}$	NLL_{DP}	NLL^*	Dir. acc.	IoU
✓	✓	0.189	**11.769**	**11.958**	**0.892**	0.442
✗	✓	0.266	12.785	13.051	0.853	0.223
✓	✗	**0.167**	13.764	13.931	0.848	**0.453**

*Mean over all elements

supervised baseline results while limited to the same training data domain, but is also a scaleable solution for real-world mobile robotics as the model can improve by continual learning from new observational experience. While the baselines do not specify train and evaluation regions for an ideal comparison, my experiments in Table 6.9 show my model surpassing the supervised baseline methods also when training on one region only, demonstrating that the exact train and evaluation region split is not critical for achieving my favorable results. I note that the probabilistic DSLP model outperforms the non-probabilistic DSLA affordance model [45], the proposed regularizer α_{IB} (6.22) outperforms the best constant hyperparameter regularizer α and the mean α_{IB} value, and that more comprehensively observed environments result in better performance. See Fig. 6.13 for visualizations of inferred navigational paths and dense lane maps used for evaluation against baselines.

Ablation studies. Table 6.8 shows that leveraging the predictive world model (WM) [1] and proposed data augmentation (Aug.) [45] method reduces bias in the predicted probabilistic fields. I note that the unaugmented experiment generates output biased towards ego-agent trajectories, resulting in worse overall NLL while the maximum likelihood graph remains accurate. I believe this indicates the potential to further improve the graph generation algorithm to better leverage the more accurate probability field prediction. I refer to prior work [1] for world model performance evaluation.

Table 6.9 shows that increased observational experience reduces bias in the predicted probability field, providing evidence that the model can be trained to infer an unbiased probability prediction in the limit of infinite data

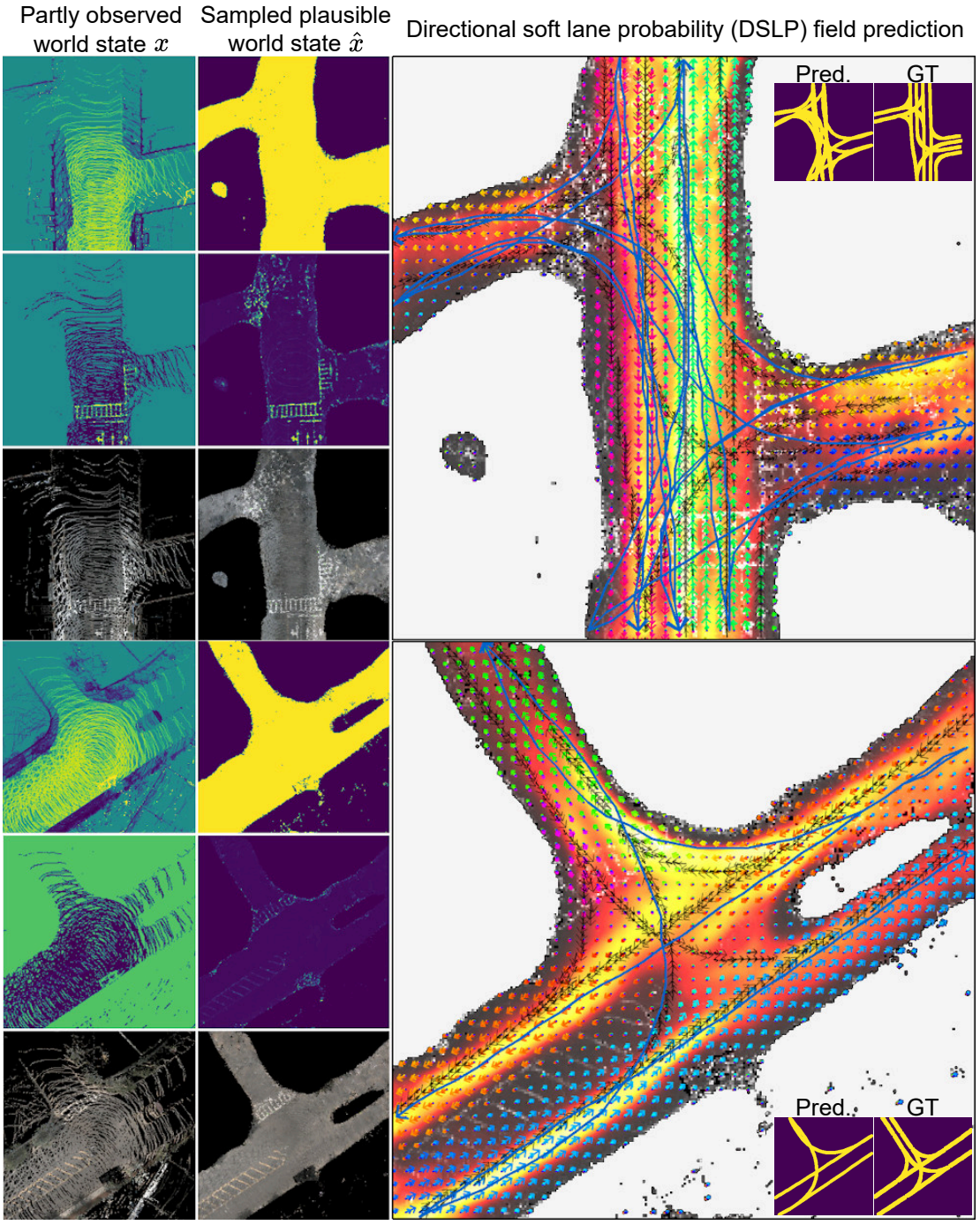

FIGURE 6.13: Model output visualizations. The left column shows accumulated partial observations x. The middle column shows plausible world states $\hat{x}$ sampled from x. The right column visualizes the predicted probability fields $\hat{Y}$ and $\hat{W}$, the maximum likelihood graph, and dense lane maps for evaluation.

TABLE 6.9: Performance with varying data amounts

# Regions	NLL_{SLP}	NLL_{DP}	NLL	Dir. acc.	IoU
{1}	**0.478**	12.696	13.174	0.861	0.423
{1, 2}	0.544	12.013	12.557	0.874	**0.444**
{1, 2, 3}	0.556	**11.769**	**12.325**	**0.892**	0.442

Inference time. I analyze the time taken for one iteration of my proposed system as follows. The mean inference time for the predictive world model and DSLP model is 0.175 sec and 0.017 sec, resulting in a total mean time of 0.192 sec per iteration or 5.21 Hz on an RTX 4090 GPU. I conclude that my method is feasible to run in real-time as it introduces a 0.192 sec overhead with a real-time SLAM implementation [557] operating faster than sensor frame rates.

6.5 Discussion and Limitations

The experiments involving the predictive environment state representation shows promising results in terms of spatial accuracy, semantic inferrability, and diversity of posterior and unconditional sampling of plausible states. However, the current observation accumulation implementation lacks a robust point cloud odometry estimation and loop closure. The proof-of-concept software implementation is does not run in real-time and thus not applicable for real-world usage. Refactoring the proposed framework by an existing optimized SLAM framework [286, 376, 555] would solve this limitation. Additionally, a more robust observation accumulation implementation would reduce the amount of degenerate samples with failed scan matchings that hampers the predictive world model learning performance by random noise and nonsensical environment structures.

The current predictive state representation as top-down 2D grid representations are adequate for mobile robot navigation operating in planar environments like autonomous vehicle. However, in the general case, general-purpose mobile robots must be able to leverage a full 3D state representation of the environment. This thesis propose extending the predictive states to 3D representations using voxel grids [562] or neural radiance fields [197, 294] in order to enable spatial reasoning in fully general complex 3D structures.

The presented implementation of the predictive world model only models elements of the static environment as predictive state representations. Incorporating temporal dynamics into the predictive world model would enable the model to incorporate representation of dynamic objects and agents, as well as predicting agent behavior and outcomes of actions altering the physical environment.

Next I identify limitations and directions for future work for learning navigational pattern based on the predictive state representation framework. The representation of spatially small but semantically important environmental cues, such as road markings, is inefficiently represented by uniform grid maps. Traffic information on signs is not represented at all. I propose to instead detect and semantically draw road markings and signs in the input representation. Graph generation can be improved by inferring start and end points within the BEV, sampling higher-order splines, and decomposing splines into a sparse graph [240]. Understanding navigational patterns may require a temporal memory of past observations to resolve ambiguity. I propose an additional module that maintains a latent environment encoding by learning from sequences instead of i.i.d. data.

6.6 Summary

This chapter presents an implementation and experimental evaluation of the proposed predictive state representation based on latent compositional semantics in the autonomous driving domain. The presented VLM trained to predicit dense embedding maps of latent compositional semantics is evalauted and compared with previous SOTA baseline methods. Results show that the proposed approach outperforms existing methods in terms of both accuracy and scalability, demonstrating its potential for real-world applications.

Next, the open-vocabulary predictive world model (OV-PWM) is experimentally validated on autonomous driving data generated by the CARLA simulator. The experiments show OV-PWM can learn a compact latent representations and generate diverse and accurate worlds with fine detail like road markings, achieving 69 mIoU over six query semantics on an urban evaluation sequence. The results supports using OV-PWM as a versatile continual learning paradigm for providing spatiosemantic memory and learned internal simulation capabilities to future general-purpose mobile robots.

The remaining section presents a self-supervised method for inferring global navigation patterns from partially observed environments using probabilistic modeling leveraging a novel regularizer based on information balance. Experimental results show that the presented self-supervised approach outperforms fully supervised SOTA baselines even when trained on the same amount of data, and that the model can infer unbiased probability predictions with infinite data. The timing experiments show the proposed method can run in real-time provided a faster than real-time SLAM implementation.

Chapter 7

Conclusions

7.1 Summary of Thesis

The human brain continuously generates predictions about sensory inputs based on prior knowledge and experience, which is known as predictive coding. The hippocampus plays a crucial role in this process by learning sequences, forming memory-based predictions, processing prediction errors, and integrating multi-modal information. This thesis proposes an artificial hippocampus based on open-vocabulary predictive states representations generated by a predictive world model for future general-purpose mobile reasoning agents. The predictive states represent the environment by a dual latent and explicit representation. The compact latent representation can represent environments with high spatial accuracy and rich semantics suitable for state-transition modeling. The explicit state representation represent the environment by spatially grounded open-vocabulary semantic embeddings readable by multimodal large language models (LLMs). This thesis propose the predictive state representation as a new direction to allow multimodal LLMs to achieve spatial comprehension and perform spatio-semantic reasoning.

The thesis presents latent compositional semantics as a mathematical model of the unconditional open-vocabulary semantic embeddings underlying the predictive state representation. The sufficient similarity semantic inference method is presented as a means to query overlapping semantics. Experimental results show that VLM can discover latent compositional semantics representing sets of semantics by examples of independent visual examples of member semantics.

The proposed open-vocabulary predictive world model (OV-PWM) is learns hierarchical distributions of compact latent representation directly from raw observations using a dual-encoder hierarchical variational autoencoder (HVAE) with posterior matching

optimization. The predictive world model can generate diverse complete environment states for unobserved regions by iteratively sampling from the learned hierarchical posterior distribution. Experimental results show OV-PWM can learn to predictite a diverse set of spatially and semantically accurate predictive environment states conditioned on partially observed environments.

The predictive state representation is used to learn navigational patterns from observation only using a self-supervised method that outperforms fully supervised SOTA baseline models even when trained on the same amount of data.

In short, the thesis presents how the concept of an artificial hippocampus and predictive coding can be implemented in artificial systems and applied in autonomous driving domains to generate spatio-semantic representations for future sensory inputs based on past observational experience. The proposed OV-PWM model provides a versatile continual learning paradigm that can provide spatial memory and learned internal simulation capabilities to general-purpose mobile robots, making it a promising approach for developing future intelligent autonomous systems akin to general-purpose mobile reasoning agents.

7.2 Limitations and Future Work

While the proposed predictive state representation as a conceptual implementation of artificial hippocampus and predictive coding shows promising proof of concept results, there are several limitations and areas for future work.

The thesis explores the conceptual analogies between the predictive environment representation and biological neural networks like the hippocampus. This investigation is not exhaustive and future work could focus on integrating additional insights from neuroscience to improve the proposed method's efficiency, performance, and generality, in addition to deeper theoretical grounding.

The current top-down 2D grid map representation of the explicit spatio-semantic memory state has limitations in capturing spatio-semantics in the general case. In particular, the 2D representation cannot principally represent complex 3D environments and dynamic objects. Future work could explore using alternative 3D representations like voxel grids or neural radiance fields adapted in order to better represent spatio-semantic memories of general 3D environments.

The current implementation of the predictive world model only accounts for static elements in the environment and does not consider temporal dynamics, such as dynamic

objects and agent behaviors. Incorporating temporal information into the predictive world model could enable more accurate representation and prediction of agent behavior and outcomes of actions altering the physical environment.

The current spatial representation as top-down 2D grid representations is adequate for mobile robot navigation operating in planar environments like autonomous vehicles but lacks sufficient granularity when representing semantically important environmental cues, such as road markings, and representability for 3D information like traffic signs. Future work could focus on incorporating temporal context into a "latent memory" module to maintain an temporal environment encoding that learns from sequences instead of i.i.d. data.

Appendix A

Deriving Cosine Distance from Negative Log Likelihood Minimization

Here we show that minimizing the negative log likelihood $p(x|Z)$ in (5.22) is equivalent to minimizing the cosine distance for normalized OV semantic embeddings modeled by the OV-PWM model. Proposing that the output variable distribution is a Normal distribution and presuming the stochastic process variance σ^2 is constant and thus does not affect the minimization objective

$$\begin{aligned}
\min -\log(p(x|Z) &= \min -\log \mathcal{N}(x|\mu(Z), 2\sigma^2\mathcal{I}) && \text{(A.1)}\\
&= \min -\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu(Z))^2}{\sigma^2}\right)\right] && \text{(A.2)}\\
&= \min -\left[\log(2\pi\sigma^2)^{-\frac{1}{2}} - \frac{1}{2}\frac{(x-\mu(Z))^2}{\sigma^2}\right] && \text{(A.3)}\\
&= \min \frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{(x-\mu(Z))^2}{\sigma^2}\right] && \text{(A.4)}\\
&\propto \min \frac{1}{2}(x-\mu(Z))^2 && \text{(A.5)}\\
&= \min \frac{1}{2}(x-\mu(Z))^T(x-\mu(Z)) && \text{(A.6)}\\
&= \min \frac{1}{2}\left(x^Tx - 2x^T\mu(Z) + \mu(Z)^T\mu(Z)\right) && \text{(A.7)}\\
&= \min \frac{1}{2}\left(1 - 2x^T\mu(Z) + 1\right) && \text{(A.8)}\\
&= \min \frac{1}{2}(2 - 2x^T\mu(Z)) && \text{(A.9)}\\
&= \min(1 - x^T\mu(Z)). && \text{(A.10)}
\end{aligned}$$

Noting that the predicted OV semantic embeddings $\hat{x}^*$ correspond to $\mu(Z)$ shows that (A.10) is is the cosine distance (5.25) and thus completes the derivation.

Appendix B

Mathematical proofs

This Appendix provides full mathematical proofs for all theoreoms, propositions, and lemmas.

B.1 Proof for Lemma 4.3

Proof. All normalized semantic embeddings z are vectors in the set of vectors constituting the unit hypersphere

$$z \in S^{D-1} = \{z \in \mathbb{R}^D : \|z\| = 1\}. \tag{B.1}$$

The distribution of uniformly sampled random vectors $Z \sim U(S^{D-1})$ is isotropic (i.e. properties rotationally invariant). The covariance matrix Σ of isotropic distributions equals the diagonal matrix I_D:

$$\Sigma(Z) = \mathbb{E}\, ZZ^T = I_D. \tag{B.2}$$

For the expected inner product of two independent random vectors $Z^{(i)}$, $Z^{(j)}$ sampled from an isotropic distribution it follows

$$\mathbb{E}\langle Z^{(i)}, Z^{(j)}\rangle^2 = \mathop{\mathbb{E}}_{Z^{(j)}} \mathop{\mathbb{E}}_{Z^{(i)}} \left[\langle Z^{(i)}, Z^{(j)}\rangle^2 | Z^{(j)}.\right] \tag{B.3}$$

Assuming a particular but arbitrary vector $z^{(j)}$ and substituting (B.2) the inner expectation becomes

$$\begin{aligned} \underset{Z^{(i)}}{\mathbb{E}} \langle Z^{(i)}, z^{(j)} \rangle^2 &= z^{(j)T} \, \mathbb{E} \left[Z^{(i)} Z^{(i)T} \right] z^{(j)} \\ &= z^{(j)T} I_D z^{(j)} \\ &= z^{(j)T} z^{(j)} \\ &= \| z^{(j)} \|^2 \end{aligned} \tag{B.4}$$

The outer expectation after substituting (B.4) and (B.2) becomes

$$\begin{aligned} \underset{Z^{(i)}}{\mathbb{E}} \langle Z^{(i)}, z^{(j)} \rangle^2 = \underset{Z^{(j)}}{\mathbb{E}} \| z^{(j)} \|^2 &= \mathbb{E} \, Z^{(j)T} Z^{(j)} \\ &= \mathbb{E} \, tr \left[Z^{(j)T} Z^{(j)} \right] \\ &= \mathbb{E} \, tr \left[Z^{(j)} Z^{(j)T} \right] \\ &= tr \left[\mathbb{E} \, Z^{(j)} Z^{(j)T} \right] \\ &= tr \left[I_D \right] = D. \end{aligned} \tag{B.5}$$

Expanding the inner product of two normalized random Euclidean vectors $\hat{Z}^{(i)}$, $\hat{Z}^{(j)}$ sampled from an isotropic distribution

$$\begin{aligned} \mathbb{E} \langle \hat{Z}^{(i)}, \hat{Z}^{(j)} \rangle &= \mathbb{E} \, \hat{Z}^{(i)} \cdot \hat{Z}^{(j)} \\ &= \mathbb{E} \, \tfrac{Z^{(i)}}{\| Z^{(i)} \|} \cdot \tfrac{Z^{(j)}}{\| Z^{(j)} \|} \\ &= \mathbb{E} \, \tfrac{1}{\| Z^{(i)} \| \| Z^{(j)} \|} \langle Z^{(i)}, Z^{(j)} \rangle \\ &= \tfrac{\sqrt{D}}{\sqrt{D}\sqrt{D}} = \tfrac{1}{\sqrt{D}}. \end{aligned} \tag{B.6}$$

Taking the limit shows that any two random vectors are orthogonal in high-dimensional isotropic vector spaces

$$\lim_{D \to \infty} \mathbb{E} \langle \hat{Z}^{(i)}, \hat{Z}^{(j)} \rangle = 0. \tag{B.7}$$

As orthogonality is invariant to vector length

$$\mathbb{E} \langle \hat{Z}^{(i)}, \hat{Z}^{(j)} \rangle = \mathbb{E} \langle Z^{(i)}, Z^{(j)} \rangle = \tfrac{1}{\sqrt{D}}. \tag{B.8}$$

Noting that inner product $\langle Z^{(i)}, Z^{(j)} \rangle$ equals cosine distance similarity $sim(Z^{(i)}, Z^{(j)})$ for Euclidean spaces completes the proof. □

B.2 Proof for Lemma 4.5

Proof. Supposing the optimal compositional semantic embedding z^* is found given a set of K sub-semantic embeddings $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(K)}\}$ such that

$$z^* = \arg\max \frac{1}{K} \sum_{i=1}^{K} sim(z^*, z^{(k)}) - \mathbb{E}\, sim(z^*, z') \tag{B.9}$$

where z' is a semantic embedding of any unrelated object description.

Note that the sub-semantics $\mathcal{Z}$ can be ordered by similarity with z^*, and that the least similar sub-semantic z_{min} and its similarity value ϵ is known

$$z_{min} = \arg\min sim(z^*, z) \forall z \in \mathcal{Z}. \tag{B.10}$$

$$\epsilon = sim(z^*, z_{min}). \tag{B.11}$$

A hyperspherical cap S_{cap}^{D-1} is defined by z^* as the normal center vector and the angle θ_{min} between z^* and z_{min}

$$S_{cap}^{D-1} = \{z \in \mathbb{R}^D : \|z\| = 1, \theta_z \leq \theta_{min}\} \tag{B.12}$$

where the angles θ are related to similarities by

$$\theta_z = \arccos(sim(z^*, z)) \tag{B.13}$$

$$\theta_{min} = \arccos(sim(z^*, z_{min})). \tag{B.14}$$

Since

$$sim(z^*, z) \geq sim(z^*, z_{min}) \Leftrightarrow \theta_z \leq \theta_{min} \;\; \forall z \in \mathcal{Z} \tag{B.15}$$

$$sim(z^*, z^*) = 1 \Leftrightarrow \theta_{z^*} = 0 < \theta_{min} \tag{B.16}$$

all $z \in \mathcal{Z}$ and z^* are in S_{cap}^{D-1}.

□

B.3 Proof for Theorem 4.2

Proof. The optimal compositional semantic embedding $z^* \in \mathbb{R}^D$ representing a set of K sub-semantics $z \in \mathcal{Z}$ in a uniform distribution over the unit hypersphere $U(S^{D-1})$ is

$$z^* = \arg\max \sum_{i=1}^{K} sim(z^*, z^{(i)}) = \arg\max \sum_{i=1}^{K} (z^*)^T z^{(i)}. \tag{B.17}$$

Maximizing cosine distance similarity $sim(z^*, z)$ is equivalent to minimizing squared distance $||z^* - z||^2$ on the unit hypersphere as

$$\begin{aligned}
\min \sum_{i=1}^{K} ||z^* - z^{(i)}||^2 &= \sum_{i=1}^{K} (z^* - z^{(i)})^T (z^* - z^{(i)}) \\
&= \min \sum_{i=1}^{K} \left[||z^*|| - 2(z^*)^T z^{(i)} + ||z^{(i)}|| \right] \\
&= \min \sum_{i=1}^{K} \left[2 - 2(z^*)^T z^{(i)} \right] \\
&= \min \left[2K - 2 \sum_{i=1}^{K} (z^*)^T z^{(i)} \right] \\
&\propto \min \left[- \sum_{i=1}^{K} (z^*)^T z^{(i)} \right] \\
&= \max \sum_{i=1}^{K} (z^*)^T z^{(i)}
\end{aligned} \tag{B.18}$$

The vector z^* maximizing (B.17) can thus be found from the derivative with respect to the vector z^*

$$\frac{d}{dz^*} \sum_{i=1}^{K} ||z^* - z^{(i)}||^2 = 0. \tag{B.19}$$

To apply the general chain rule [563], we rewrite (B.19) with variable substitution so that each operation in the function is factored into single variable components for easily finding partial differentials:

$$\begin{aligned}
\sum_{i=1}^{K} ||z^* - z^{(i)}||^2 = g = \sum_{i=1}^{K} ||f||^2 \\
f = z^* - z^{(i)}.
\end{aligned} \tag{B.20}$$

Applying the chain rule and noting that $||f||^2 = f^T f$ gives

$$\begin{aligned}
\frac{\partial g}{\partial z^*} &= \frac{\partial g}{\partial f}\frac{\partial f}{\partial z^*} \\
&= \sum_{i=1}^{K} 2f^T \frac{\partial}{\partial z^*}\left(z^* - z^{(i)}\right) \\
&= 2\sum_{i=1}^{K}(z^* - z^{(i)})^T \left[\tfrac{\partial}{\partial z_1^*}(z^* - z^{(i)}), \ldots, \tfrac{\partial}{\partial z_D^*}(z^* - z^{(i)})^T\right] \\
&= 2\sum_{i=1}^{K}(z^* - z^{(i)})^T \left[e_1, \ldots, e_D\right] \\
&= 2\left[\sum_{i=1}^{K}(z_1^* - z_1^{(i)}), \ldots, \sum_{i=1}^{K}(z_D^* - z_D^{(i)})\right]^T = 0
\end{aligned} \tag{B.21}$$

where e_d is the one-hot vector with the d^{th} element set to 1. Equation (B.21) is an element-wise system of equations stating that for every d^{th} element

$$\sum_{i=1}^{K}(z_d^* - z_d^{(i)}) = 0 \tag{B.22}$$

meaning the optimal z^* maximizing (B.17) equals the centroid of the sub-semantics $z^{(i)} \in \mathcal{Z}$

$$z^* = \frac{1}{K}\sum_{i=1}^{K} z^{(i)}. \tag{B.23}$$

To prove z^* specified by (B.23) satisfies Definition 4.1 we write

$$\mathbb{E}\, sim(z^*, z) = \mathbb{E}\left[\left(\tfrac{1}{K}\sum_{i=1}^{K} z^{(i)}\right) \cdot z\right] = \tfrac{1}{K}\sum_{i=1}^{K} \mathbb{E}\, z^{(i)} \cdot z. \tag{B.24}$$

As z equals one of the $z^{(i)} \in \mathcal{Z}$ we can assume $z = z^{(k)}$ without loss of generality and expand the sum in (B.24) as

$$\begin{aligned}
\mathbb{E}\, sim(z^*, z) = \tfrac{1}{K}\Big(&\mathbb{E}[z^{(1)} \cdot z^{(k)}] + \ldots \\
&+ \mathbb{E}[z^{(k)} \cdot z^{(k)}] + \ldots + \mathbb{E}[z^{(K)} \cdot z^{(k)}]\Big)
\end{aligned} \tag{B.25}$$

We find a lower bound for (B.25) by applying Lemma 4.3 and noting that the expected similarities $sim(z^{(i)}, z^{(j)}) \forall z^{(i)}, z^{(j)} \in \mathcal{Z}$ must be higher or equal to random vectors, and

that $z^{(k)} \cdot z^{(k)} = 1$

$$\begin{aligned} \mathbb{E}\, sim(z^*, z) &\geq \tfrac{1}{K}\left(D^{-\frac{1}{2}} + \ldots + 1 + \ldots + D^{-\frac{1}{2}}\right) \\ &= \tfrac{1}{K}\left((K-1)D^{-\frac{1}{2}} + 1\right). \end{aligned} \tag{B.26}$$

Substituting the bound (B.26) into Definition 4.1 and applying Lemma 4.3 on the RHS

$$\mathbb{E}\, sim(z^*, z) \geq \tfrac{1}{K}\left((K-1)D^{-\frac{1}{2}} + 1\right) > D^{-\frac{1}{2}}. \tag{B.27}$$

Rearranging the two leftmost inequalities in (B.27)

$$(K-1)D^{-\frac{1}{2}} + 1 > KD^{-\frac{1}{2}} \tag{B.28}$$

$$KD^{-\frac{1}{2}} - D^{-\frac{1}{2}} + 1 - KD^{-\frac{1}{2}} > 0 \tag{B.29}$$

$$-D^{-\frac{1}{2}} > -1 \tag{B.30}$$

$$D^{-\frac{1}{2}} < 1 \tag{B.31}$$

$$\sqrt{D} > 1 \tag{B.32}$$

which is true for $D > 1$ and thus proves Theorem 4.2.

□

B.4 Proof for Theorem 4.4

Proof. A random vector z' sampled from the uniform distribution over the unit hypersphere $U(S^{D-1})$ is equally likely to be a point anywhere on S^{D-1}. The probability z' is sampled in a particular surface region $A_{D,r}$ is

$$P(z' \in A_{D,r}) = \tfrac{A_{D,r}}{A_D} \tag{B.33}$$

where A_D is the total surface region.

The probability z' is sampled into the surface region defined by the hyperspherical cap S_{cap}^{D-1} with surface area A_{cap} given in Lemma 4.5 is therefore

$$P(z' \in S_{cap}^{D-1}) = \tfrac{A_{cap}}{A_D}. \tag{B.34}$$

The surface area ratio of a hyperspherical cap [540] is

$$A_{D,r} = \tfrac{1}{2} A_D I_{\sin^2(\theta)}(\tfrac{D-1}{2}, \tfrac{1}{2}) \tag{B.35}$$

where $I_x(a, b)$ is the regularized incomplete beta function.

Substituting (B.35) into (B.33) gives

$$P(z' \in S_{cap}^{D-1}) = \tfrac{1}{2} I_{\sin^2(\theta)}(\tfrac{D-1}{2}, \tfrac{1}{2}). \tag{B.36}$$

The probability that z' is not sampled in S_{cap}^{D-1} is

$$\begin{aligned} P(z' \notin S_{cap}^{D-1}) &= 1 - P(z' \in S_{cap}^{D-1}) \\ &= 1 - \tfrac{1}{2} I_{\sin^2(\theta)}(\tfrac{D-1}{2}, \tfrac{1}{2}). \end{aligned} \tag{B.37}$$

By Lemma 4.5 and (B.14) we know

$$\forall z' \;\; sim(z^*, z_{min}) \geq sim(z^*, z') \Leftrightarrow z' \notin S_{cap}^{D-1}. \tag{B.38}$$

Substituting the bound $sim(z^*, z_{min})$ by (B.15) gives

$$\forall z', z \in \mathcal{Z} \;\; sim(z^*, z) \geq sim(z^*, z') \Leftrightarrow z' \notin S_{cap}^{D-1}. \tag{B.39}$$

Substituting the LHS of (B.39) into (B.37) and recollecting (B.14) proves Theorem 4.4.

□

B.5 Proof for Proposition 4.6

Proof. Non-uniformity means the distribution of vectors is not maximally dispersed over the hypersphere [549]. Recalling Lemma 4.3 for uniform distributions, the expected similarity of two non-uniformly distributed independent vectors $Z^{(i)}, Z^{(j)} \sim p(Z)$ is therefore

$$\mathbb{E}\, sim(Z^{(i)}, Z^{(j)}) = C \geq \tfrac{1}{\sqrt{D}}. \tag{B.40}$$

By substituting (B.40) in Definition 4.1 gives

$$\mathbb{E}\, sim(z^*, z) > \mathbb{E}\, sim(z^*, z') = C. \tag{B.41}$$

Expanding the LHS of (B.41) using the same idea as in (B.24) and (B.25)

$$\mathbb{E}\, sim(z^*, z) \geq \tfrac{1}{K} \left[(K-1)C + 1\right]. \tag{B.42}$$

Substituting (B.42) into (B.41)

$$\frac{1}{K}\left[(K-1)C+1\right] > C \tag{B.43}$$

$$KC - C + 1 > KC \tag{B.44}$$

$$-C > -1 \tag{B.45}$$

$$C < 1. \tag{B.46}$$

Since $sim(z^{(i)}, z^{(j)}) \in [-1, 1[\;\; s.t. z^{(i)} \neq z^{(j)}$ the inequality (B.41) is true for all distributions $p(z)$ except the singular distribution and thus proves Proposition 4.6.

□

B.6 Proof for Proposition 4.7

Proof. The global convergence guarantee for convex optimization problems [564] proves that for any convex function f is guaranteed that the value $z^{*(t)}$ converges to the optimal value z^*

$$\lim_{t\to\infty} f(z^{*(t)}) = f(z^*) \tag{B.47}$$

given a sufficiently small learning rate λ.

We prove that the cosine similarity optimization objective (B.17) is a convex problem by noting that the set (B.1) is convex and show that the Hessian matrix

$$H\left(f(z^*)\right) = \nabla^2_{z^*} f(z^*) = \left[\frac{\partial}{\partial z_i^* \partial z_j^*} f(z^*)\right] \tag{B.48}$$

is a positive semidefinite matrix [563]. Note that $f(z^*)$ substitutes $\sum_{k=1}^{K} sim(z^*, z^{(k)})$. Recalling the form of the first partial derivatives (B.21) and taking another partial derivative for an arbitrary element

$$\frac{\partial}{\partial z_i^* \partial z_j^*} f(z^*) = \frac{\partial}{\partial z_i^*}\left[K z_j^* - \sum_{k=1}^{K} z_j^{(k)}\right] = K\mathbf{1}_{i=j}. \tag{B.49}$$

The Hessian matrix is thus the scaled identity matrix

$$H\left(f(z^*)\right) = \left[Ke_i\right] = KI_D \tag{B.50}$$

meaning $f(z^*)$ is a convex function.

Finally noting that

$$\mathbb{E}\left[\frac{1}{L}\sum_{i=1}^{L} z \in \mathcal{Z}^{(t)}\right] = \frac{1}{K}\sum_{i=1}^{K} z^{(i)} \in \mathcal{Z},\ \mathcal{Z}^{(t)} \subseteq \mathcal{Z} \tag{B.51}$$

shows that the optimal convergence value obtained by optimizing (B.17) by gradient descent results in the optimal compositional semantic embedding z^* obtained by (B.23).

□

Appendix C

Aligning Predictive States and LLMs by Computational Geometry

Large language models (LLMs) can serve as general-purpose agents in robotics and autonomous vehicle applications. Agents are called general-purpose due to their capability and versatility to understand and follow natural language instructions in the context of the perceived environment and agent state. The capability to understand instructions grounded in the environment enables spatial reasoning. Spatial reasoning is the ability to complete tasks involving question-answering, action or manipulation, and navigating environments involving semantic objects and leveraging external knowledge. Spatial reasoning for autonomous driving agent systems include making rational decisions for navigating dynamic environments. Examples include whether to panic break for a plastic bag on the highway, or how to traverse complicated rule-constrained environments like intersections. See Sec. 1.2 for further details.

Existing LLM-based agents capable of spatial reasoning utilize various spatio-semantic representations of the environment. Parametric object lists [82, 84, 90, 384, 386, 412, 565–568] represents the environment in terms of a set of discrete semantic objects. Object list representations are intuitive and simple to implement for proof of concept works. However, enumerating all possible useful objects and their properties as explicit textual entries in each reasoning step is not a practical nor computationally efficient approach due to context window length considerations [95, 155, 569]. Transforming images into visual tokens [73, 411, 570] as input for multimodal LLMs is also intuitive and simple to implement to learn to represent the environment by finetuning a vision encoder or jointly optimizing the encoder and LLM on multimodal data. Training multimodal LLMs on

sequential visual input can provide a form of latent spatio-semantic memory [571–574]. Such spatio-semantic memories lack the spatial grounding of object semantics associated with explicit maps, as well as human-machine interpretablity and the ability to input spatio-semantic information to the machine. Other notable environment representations include 3D reconstruction [399], object-centric, topological maps [400], scene graphs [91, 401], and top-down metric grid maps [85].

This thesis propose predictive environment states x^* aligned to the embedding space of LLMs as a principled spatio-semantic memory representation for LLMs. The advantage of LLM-readable x^* are explicit and human communicable map representations of the environment. The latent compositional semantic embeddings of x^* is also an efficient representation where an ideal single embedding can represent a set of hundreds of random object semantics as demonstrated in Chap. 4. The encoder used to transform predictive states x^* into latent state tokens Z in the LLM embedding space $\mathcal{Z}$ is technically similar to the visual encoder using in multimodal LLMs. The primary challenge is how to design a self-supervised optimization objective to enable continual learning from observation experience [330].

The optimization objective is based on computational geometry as the bridge between geometric and textual representations. A self-supervised ground truth representation for the 2D open vocabulary predictive states x^* presented in this thesis is be derived pragmatically as follows. First, the state $x^* \in \mathbb{R}^{H\times W\times D}$ is queried element-wise by a known object semantic $x_q \in \mathbb{R}^D$ using sufficient similarity inference (4.21) presented in Chap. 4:

$$sim(x^*_{i,j}, x_q) > \tau_q \Rightarrow MemberOf(x^*_{i,j}, \text{query semantic}). \tag{C.1}$$

The query results in a set of boolean masks as dense representation of geometries $m \in \mathbb{B}^{H\times W}$. Secondly, the masks m can be converted into textual descriptions such as a sequence of polygon vertices $[(i,j)_1, \ldots, (i,j)_M]$ [575] using a computational geometry library like CGAL [576]. Finally, the encoder f_θ is optimized so the transformed spatio-semantic information in $Z = (z^{(1)}, \ldots, z^{(K)})$ allows the LLM to predict the same geometric textual description based on a semantic query and in-context example of the computational geometry file structure. The loss is the negative log likehood of the tokenized sequences y and y' of the programatic txt and predicted txt' text descriptions

$$\mathcal{L}_{NLL}(txt, txt') = -\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{n}\log p(y_i = y'_i|y_1, \ldots, y_{i-1})\right] \tag{C.2}$$

See Fig. C.1 for a visual representation of the alignment framework. A pictorial demonstration of learning to align "road" object semantics is shown in Fig. C.2.

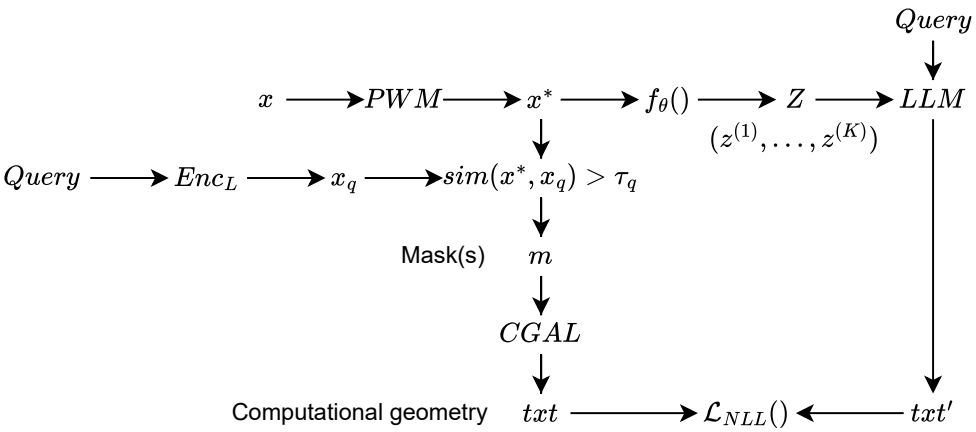

FIGURE C.1: The alignment function f_θ transforms predictive state representations x^* to latent environment tokens Z in an LLM embedding space. f_θ is optimized using computational geometry as a bridge between geometric and textual representations.

Prior work leveraging LLMs in the semantic segmentation task provide empirical proof that LLMs are adept at predicting coordinate sequences based on image data, natural language instruction, and few-shot learning examples. LLaFS [577] use world knowledge in LLMs to guide dense segmentation of queried object semantics. The LLM takes visual tokens generated by an image encoder and learns to predict fixed-length polygons representing object masks. The LLM-based mask predictor improves SOTA few-shot semantic segmentation. PolyFormer [578] and SeqTR [579] trains a multimodal transformer model to autoregressively predicit object masks based on an image and natural language referring expression query text encoded into visual and text tokens, respectively. The work demonstrates that polygon outputs perform better than dense pixel-wise output in the semantic semgmentation task. BoundaryFormer [580] presents a differentiable rasterizer based on the signed distance function for end-to-end trainable autoregressive polygon prediction.

Future work include experimentally prove that aligning the predictive state representation x^* by computational geometry objectives, like autoregressive polygon prediction [577, 578, 580], is a principled way to ground multimodal LLMs in a memory-enabbled spatio-semantic state representation. The proposed alignment method for f_θ supports continual learning on observational data as the learning signal is based on a self-supervised target.

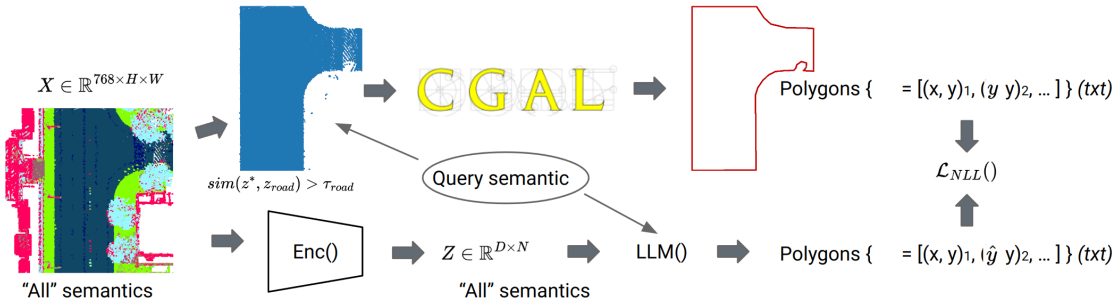

FIGURE C.2: The alignment method demonstrated by a visual example of how a "road" semantic object. The resulting polygon is programatically inferred and used as a self-supervised learning signal.

Appendix D

ViCE Pseudocodes

Algorithm 1 explains the generation of M views for a batch of N images. The algorithm samples an image $X^{(n)}$ and computes a superpixel index map $A^{(n)}$. M views are generated from the sampled image and superpixel index map. Each of these views are randomly masked before being resized to the same pixel dimension. Only mutual regions existing in all views are kept. All views are geometrically augmented by random horizontal flipping, and appearance augmented by color distortion and randomly blurred. All generated views are gathered and converted into a 4D tensor.

Algorithm 1 View generation

$\tilde{X} := \{\}$ ▷ Empty sets
$\tilde{A} := \{\}$
for $n \in \{1, \ldots, N\}$ **do**
  $X^{(n)} \sim$ dataloader ▷ Sample an image
  $A^{(n)} := \text{superpixels}(X^{(n)})$

  $\tilde{X}^{(n)}, \tilde{A}^{(n)} := \text{gen\_views}(X^{(n)}, A^{(n)})$
  $\#\ \tilde{X}^{(n)} = \{\tilde{X}^{(1,n)}, \ldots, \tilde{X}^{(M,n)}\}$
  $\#\ \tilde{A}^{(n)} = \{\tilde{A}^{(1,n)}, \ldots, \tilde{A}^{(M,n)}\}$

  $\tilde{X}^{(n)}, \tilde{A}^{(n)} := \text{mask\_views}(\tilde{X}^{(n)}, \tilde{A}^{(n)})$
  $\tilde{X}^{(n)}, \tilde{A}^{(n)} := \text{resize\_views}(\tilde{X}^{(n)}, \tilde{A}^{(n)})$
  $\tilde{X}^{(n)}, \tilde{A}^{(n)} := \text{mutual\_regions}(\tilde{X}^{(n)}, \tilde{A}^{(n)})$

  $\tilde{X}^{(n)}, \tilde{A}^{(n)} := \text{geometric\_aug}(\tilde{X}^{(n)}, \tilde{A}^{(n)})$
  $\tilde{X}^{(n)} := \text{appearance\_aug}(\tilde{X}^{(n)})$

  $\tilde{X} := \tilde{X} + \tilde{X}^{(n)}$ ▷ Add new views to set
  $\tilde{A} := \tilde{A} + \tilde{A}^{(n)}$
end for
$\tilde{X} := \text{to\_tensor}(\tilde{X})$ ▷ $\tilde{X} \in \mathbb{R}^{B \times 3 \times h \times w}$
$\tilde{A} := \text{to\_tensor}(\tilde{A})$ ▷ $\tilde{A} \in \mathbb{R}^{B \times 1 \times h \times w}$

Algorithm 2 explains the learning algorithm. The model f_θ generates an embedding map $\hat{Z}$ from the image view tensor $\tilde{X}$. The single tensor $\hat{Z}$ is decomposed into B tensors $\hat{Z}^{(b)}$ each corresponding to a single view. Next, four trees are created to contain the latent visual embeddings z for all elements in each mutual region i. A mean vectors z^* is computed to represent regions. Each mean vector gets computed a concept compatibility score s^* as distance to each cluster $C = (c^{(1)}, \ldots, c^{(K)})$. The swapped prediction objective is computed using the score vectors s^* stored in the tree T_{S^*}. The model parameters θ and set of visual concept vectors C are optimized to reduce the loss $\mathcal{L}$.

Algorithm 2 Learning algorithm

```
# Generate embedding maps
Ẑ := f_θ(X̃)                                   ▷ Ẑ ∈ ℝ^{B×D×h×w}
{Ẑ^(1), ..., Ẑ^(B)} := decompose(Ẑ)

# Create embedding and score trees
T_Z(n, m, i) := {}                             ▷ Empty depth-3 trees
T_Z*(n, m, i) := {}
T_S*(n, m, i) := {}
for b ∈ {1, ..., B} do
    Z̃^(b) := unroll(Ẑ^(b))                     ▷ Z̃^(b) ∈ ℝ^{hw×D}
    Ã^(b) := unroll(Ã^(b))                     ▷ Ã^(b) ∈ ℝ^{hw}
    n, m := img_view_index(b)
    I := num_regions(Ã^(b))
    for i ∈ {1, ..., I} do
        # Compute mean vectors for region
        {ẑ^(j)} := extract_region(Z̃^(b), Ã^(b), i)
        T_Z(n, m, i) := {ẑ^(j)}
        z^(i)* := mean(T_Z(n, m, i))
        T_Z*(n, m, i) := z^(i)*

        # Compute score vectors for region
        s^(i)* = (T_Z*(n, m, i))^T C
        T_S* := s^(i)*
    end for
end for

L = swapped_prediction(T_S*)

optimize(θ, C, L)
```

The swapped prediction objective is explained in Algorithm 3. First, we compute an optimal assignment of visual concepts Q based on the scores in the first view $m = 1$. The loss is minimized when predicted visual embeddings in secondary views $m \geq 1$ are closer to the optimally assigned visual concept vectors for each region i in all views m of all images n. This results in a cross-entropy optimization objective when both assignments $q^{(i)}$ and compatibility scores $s^{(i)*}$ are normalized.

Algorithm 3 Swapped prediction objective

$\mathcal{L} := 0$
$Q :=$ optimal_assignment(T_{S^*})
for $n \in \{1, \ldots, N\}$ **do**
  for $m \in \{2, \ldots, M\}$ **do**
    for $i \in \{1, \ldots, I\}$ **do**
      $q^{(i)} := Q(n, i)$
      $s^{(i)*} := T_{S^*}(n, m, i)$
      $p^{(i)} := \sigma\left(\frac{1}{\tau} s^{(i)*}\right)$
      $\mathcal{L} \mathrel{-}= q^{(i)} log\, p^{(i)}$
    end for
    $\mathcal{L} := \mathcal{L}/I$
  end for
end for
$\mathcal{L} := \mathcal{L}/(N(M-1))$

Bibliography

[1] Robin Karlsson, Alexander Carballo, Keisuke Fujii, Kento Ohtani, and Kazuya Takeda. Predictive world models from real-world partial observations. In *IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*, pages 152–166, 2023.

[2] Robin Karlsson, Ruslan Asfandiyarov, Alexander Carballo, Keisuke Fujii, Kento Ohtani, and Kazuya Takeda. Open-vocabulary predictive world models from sensor observations. *Sensors*, 24(14), 2024. ISSN 1424-8220. doi: 10.3390/s24144735. URL https://www.mdpi.com/1424-8220/24/14/4735.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, volume 33, 2020.

[4] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, PAMI-8(6):679–698, 1986.

[5] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021.

[6] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022.

[7] D. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, (ICLR)*, 2014.

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[9] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785.

[10] Immanuel Kant. *Critique of Practical Reason*. 1788.

[11] Plato. *The Republic of Plato*. Basic Books, 1991. Originally written ca. 375 BCE.

[12] Aristotle. *Politics*. Penguin Classics, 1981. Originally written ca. 350 BCE.

[13] J. Rousseau. *Of the Social Contract and Other Political Writings*. Cambridge University Press, 1997. Originally written 1762.

[14] J. Rousseau. *The Division of Labor in Society*. Free Press, 2014. Originally written 1893.

[15] Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. The MIT Press, 2018. ISBN 9780262039246.

[16] A. Smith. *The Wealth of Nations*. W. Strahan and T. Cadell, London, 1776. Originally written 1776.

[17] K. Marx. *Das Kapital*. Verlag von Otto Meisner, 2011. Originally written 1867–1894.

[18] T. Piketty. *Capital in the Twenty-First Century*. Belknap Press, 2017. Originally written 2013.

[19] Joseph Stiglitz and Bruce Greenwald. *Towards a New Paradigm in Monetary Economics*. Cambridge University Press, 2003.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, volume 25, pages 1097–1105, 2012.

[21] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

[22] Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. Dendral: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209–261, 1993. ISSN 0004-3702.

[23] Edward Shortliffe. Mycin: A knowledge-based computer program applied to infectious diseases. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 10 1977.

[24] J. Pearl. *Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20 (3):273–297, sep 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411.

[26] Kurt Hornik, Maxwell B. Stinchcombe, and Halbert L. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013.

[28] Stuart J. Russel and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 4th ed. edition, 2020. ISBN 0-13-461099-7.

[29] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, page 26–33, USA, 2001. doi: 10.3115/1073012.1073017.

[30] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[31] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103535.

[32] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[33] M. Bojarski et al. End to end learning for self-driving cars. *arXiv preprint*, 2016.

[34] Jianyu Chen, Zhuo Xu, and Masayoshi Tomizuka. End-to-end autonomous driving perception with sequential latent representation learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1999–2006, 2020.

[35] Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books, USA, 2019. ISBN 1524748250.

[36] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62:54–60, 2019.

[37] Jürgen Schmidhuber. Prof. schmidhuber's highlights of robot car history. https://people.idsia.ch/~juergen/robotcars.html, 2009. Retrieved 29 June 2024.

[38] Chris Urmson, Joshua Anhalt, J. Andrew Bagnell, Christopher R. Baker, Robert Bittner, M. N. Clark, et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25, 2008.

[39] Julia Nilsson, Jonatan Silvlin, Mattias Brannstrom, Erik Coelingh, and Jonas Fredriksson. If, when, and how to perform lane change maneuvers on highways. *IEEE Intelligent Transportation Systems Magazine*, 8(4):68–78, 2016. doi: 10.1109/MITS.2016.2565718.

[40] Heiko G. Seif and Xiaolong Hu. Autonomous driving in the icity—hd maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016. ISSN 2095-8099. doi: https://doi.org/10.1016/J.ENG.2016.02.010.

[41] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J. Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, Michael Sokolsky, Ganymed Stanek, David Stavens, Alex Teichman, Moritz Werling, and Sebastian Thrun. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168, 2011. doi: 10.1109/IVS.2011.5940562.

[42] B. Paden, M. Cap, S. Yong, D. Yershov, and E. Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 2016.

[43] R. Kala and K. Warwick. Multi-level planning for semi-autonomous vehicles in traffic scenarios based on separation maximization. *Journal of Intelligent & Robotic Systems*, 72:559–590, 2013.

[44] Brett M. Leedy, Joseph S. Putney, Cheryl Bauman, Stephen Cacciola, J. Michael Webster, and Charles F. Reinholtz. Virginia tech's twin contenders: A comparative study of reactive and deliberative navigation. *Journal of Field Robotics*, 23(9):709–727, 2006. doi: https://doi.org/10.1002/rob.20143.

[45] R. Karlsson and E. Sjoberg. Learning a directional soft lane affordance model for road scenes using self-supervision. In *IV*, 2021.

[46] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, may 2019. ISSN 0001-0782. doi: 10.1145/3282486.

[47] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. volume abs/1708.06374, 2017.

[48] Léon Bottou. From machine learning to machine reasoning. *Mach. Learn.*, 94(2): 133–149, feb 2014. ISSN 0885-6125. doi: 10.1007/s10994-013-5335-x.

[49] Josh Tenenbaum. Building machines that learn and think like people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 5, 2018.

[50] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, aug 2015. ISSN 0001-0782.

[51] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *Robotics: Science and Systems*, June 2019.

[52] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2084–2093, 2020. doi: 10.1109/WACV45572.2020.9093332.

[53] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 568–575, 2018. doi: 10.1109/IROS.2018.8594386.

[54] A. Kendall et al. Learning to drive in a day. In *ICRA*, 2019.

[55] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[56] Richard Sutton and Anderew Barto. *Reinforcement Learning*. The MIT Press, 2018. ISBN 9780262039246.

[57] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[58] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in

a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254, 2019. doi: 10.1109/ICRA.2019.8793742.

[59] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022. doi: 10.1109/TITS.2021.3054625.

[60] Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2641–2646, 2015. doi: 10.1109/ICRA.2015.7139555.

[61] Marwa Ahmed, Chee Peng Lim, and Saeid Nahavandi. A deep q-network reinforcement learning-based model for autonomous driving. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 739–744, 2021.

[62] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *ArXiv*, abs/1610.03295, 2016.

[63] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[64] Sangeet S. Khemlani, Aron K. Barbey, and Philip N. Johnson-Laird. Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00849. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00849.

[65] Dario Amodei, Christopher Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. Concrete problems in ai safety. *ArXiv*, abs/1606.06565, 2016.

[66] Akifumi Wachi. Failure-scenario maker for rule-based agent using multi-agent adversarial reinforcement learning and its application to autonomous driving. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6006–6012, 7 2019. doi: 10.24963/ijcai.2019/832.

[67] Panagiotis Tigas, Angelos Filos, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Robust imitative planning: Planning from demonstrations under uncertainty. In *NeurIPS2019 Workshop on Machine Learning for Autonomous Driving*, 2019.

[68] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 2891–2897. AAAI Press, 2017.

[69] Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 661–668. PMLR, 13–15 May 2010.

[70] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020.

[71] Shiyi Wang, Yuxuan Zhu, Zhiheng Li, Yutong Wang, Li Li, and Zhengbing He. Chatgpt as your vehicle co-pilot: An initial attempt. *IEEE Transactions on Intelligent Vehicles*, 8(12):4706–4721, 2023. doi: 10.1109/TIV.2023.3325300.

[72] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7020–7030, 2023.

[73] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[74] Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, and Dieter Fox. Conditional driving from natural language instructions. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 540–551. PMLR, 30 Oct–01 Nov 2020.

[75] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.

[76] Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[78] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *arXiv preprint arxiv:2203.03897*, 2023.

[79] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *arXiv preprint arXiv:2304.08485*, 2023.

[80] Feng Liang, Bichen Wu, Xiaoliang Dai, et al. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[81] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open vocabulary semantic segmentation with pre-trained vision-language model. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2022.

[82] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

[83] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR, 17–23 Jul 2022.

[84] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*, 2023.

[85] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. LM-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[86] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2980–2987, 2023. doi: 10.1109/ICRA48891.2023.10160609.

[87] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8545–8556, 2023. doi: 10.1109/ICCV51070.2023.00788.

[88] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14633–14642, 2023.

[89] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 910–919, 2024. doi: 10.1109/WACVW60836.2024.00102.

[90] Michael Ahn, Anthony Brohan, Noah Brown, et al. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[91] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[92] Anthony Brohan, Noah Brown, Justice Carbajal, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[93] Christopher Summerfield. *Natural General Intelligence: How understanding the brain can help us build AI*. Oxford University Press, 12 2022. ISBN 9780192843883. doi: 10.1093/oso/9780192843883.001.0001.

[94] J. Marino. Predictive coding and variational autoencoders and biological connections. *Neural Computation*, 34:1–44, 2019.

[95] Robin Karlsson, Francisco Lepe-Salazar, and Kazuya Takeda. Compositional semantics for open vocabulary spatio-semantic representations. In *arxiv:2310.04981*, 2023.

[96] Gary F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. The MIT Press, 04 2001. ISBN 9780262279086. doi: 10.7551/mitpress/1187.001.0001. URL https://doi.org/10.7551/mitpress/1187.001.0001.

[97] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman and Hall/CRC, 2020. ISBN 9780429029608. doi: 10.7551/mitpress/1187.001.0001. URL https://doi.org/10.1201/9780429029608.

[98] François Chollet. On the measure of intelligence. *ArXiv*, abs/1911.01547, 2019.

[99] E. Spelke and K. Kinzler. Core knowledge. *Developmental science*, 10:89–96, 2007.

[100] Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 14(1):29–56, 1990. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(90)90025-R.

[101] Nancy N. Soja, Susan Carey, and Elizabeth S. Spelke. Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, 38(2):179–211, 1991. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(91)90051-5. URL https://www.sciencedirect.com/science/article/pii/0010027791900515.

[102] Robin Karlsson, Tomoki Hayashi, Keisuke Fujii, Alexander Carballo, Kento Ohtani, and Kazuya Takeda. Vice: Improving dense representation learning by superpixelization and contrasting cluster assignment. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[103] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[104] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[105] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[106] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. doi: 10.1109/CVPR.2016.278.

[107] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[108] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[109] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[110] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

[111] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *ArXiv*, abs/2103.01988, 2021.

[112] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[113] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[114] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[115] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[116] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

[117] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320, 2021.

[118] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020.

[119] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[120] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, Apr. 2020.

[121] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV*, pages 2959–2968, 2019. doi: 10.1109/ICCV.2019.00305.

[122] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Loy. Online deep clustering for unsupervised representation learning. In *CVPR*, pages 6688–6697, 2020.

[123] Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. Deep clustering by Gaussian mixture variational autoencoders with graph embedding. In *ICCV*, pages 6439–6448, 2019. doi: 10.1109/ICCV.2019.00654.

[124] Yunfan Li, Hu Peng, Liu Zitao Peng Dezhong, Tianyi Zhou, and Peng Xi. Unsupervised semantic segmentation by contrasting object mask proposals. In *AAAI*, 2021.

[125] Bruno Lévy and Erica L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[126] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, volume 26, pages 2292–2300, 2013.

[127] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. *ICCV*, pages 7526–7534, 2021.

[128] Xiaoni Li, Y. Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[129] Olivier J. H'enaff, Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *ICCV*, pages 10066–10076, 2021.

[130] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *NeurIPS*, 2020.

[131] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

[132] Xiaoni Li, Y. Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *ACM MM*, 2021.

[133] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *CVPR*, 2021.

[134] Pedro H. O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020.

[135] Amir Bar, Xin Wang, Vadim Kantorov, Colorado Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised pretraining with region priors for object detection. In *CVPR*, 2022.

[136] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In *NeurIPS*, volume 34, 2021.

[137] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Ching-Feng Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, pages 3987–3996, 2021.

[138] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610, 2021.

[139] Xinlong Wang, Zhang Rufeng, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021.

[140] Tete Xiao, Colorado J. Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *ICCV*, pages 10539–10548, 2021.

[141] Jasper R.R. Uijlings, Koen E.A. Van de Sande, Theo Gevers, and Arnold W.M. Smeulders. Selective search for object recognition. *IJCV*, 104:154–171, 2013. doi: 10.1007/s11263-013-0620-5.

[142] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022.

[143] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *CVPR*, 2022.

[144] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Int. Conf. on Learning Representations*, 2022.

[145] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv*, 2022.

[146] Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019.

[147] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *ECCV*, 2020.

[148] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *NeurIPS*, volume 32, pages 12705–12716, 2019.

[149] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *NeurIPS*, volume 32, 2019.

[150] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, pages 11130–11140, 2021.

[151] Tuan-Hung Vu, Himalaya Jain, Max Bucher, Matthieu Cord, and Patrick Pérez. DADA: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, pages 7364–7373, 2019.

[152] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[153] Juan Luis, Gonzalez Bello, and Munchurl Kim. Forget about the LiDAR: Self-supervised depth estimators with MED probability volumes. In *NeurIPS*, volume 33, pages 12626–12637, 2020.

[154] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[155] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, volume 30, page 6000–6010, 2017.

[156] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondances. In *ICLR*, Apr. 2022.

[157] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[158] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[159] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[160] Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–394, 2010.

[161] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[162] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020.

[163] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, volume 32, 2019.

[164] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[165] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520.

[166] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research (JMLR)*, 9:297–304, Jan. 2010.

[167] Yoav Goldberg and Omer Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv*, abs/1402.3722, 2014.

[168] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[169] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pages 524–531, 2005.

[170] Josef Sivic, Bryan Russell, Alexei A. Efros, Andrew Zisserman, and William Freeman. Discovering object categories in image collections. In *ICCV*, pages 370–377, 2005.

[171] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157, 1999.

[172] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, pages 73–86. Springer, 2012.

[173] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, pages 494–502, 2013.

[174] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *CVPR*, 1:886–893, 2005.

[175] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, pages 433–442, 2019.

[176] Songyang Zhang, Shipeng Yan, and Xuming He. LatentGNN: Learning efficient non-local relations for visual recognition. In *ICML*, pages 7374–7383, 2019.

[177] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P. Xing. Symbolic graph reasoning meets convolutions. In *NeurIPS*, volume 31, pages 1853–1863, 2018.

[178] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Péter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *ArXiv*, abs/2006.03677, 2020.

[179] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[180] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[181] S. Harris. Distributional structure. *WORD*, 10:146–162, 1954.

[182] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2227–2237, 2018.

[183] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1188–1196. PMLR, 2014.

[184] Adrew Maas and Andrew Ng. A probabilistic model for semantic word vectors. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[185] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[186] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. In *Adv. in Neural Information Processing Systems*, 2010.

[187] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[188] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2022.

[189] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[190] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022.

[191] Zhuowen Tu Zheng Ding, Jieke Wang. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning (ICLR)*, 2023.

[192] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[193] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, et al. Generalized decoding for pixel, image, and language. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15116–15127, 2023.

[194] Bowen Cheng, Ishan Misra, Alexander Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[195] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, et al. Regionclip: Region-based language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[196] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

[197] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[198] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

[199] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[200] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, June 2022.

[201] John McCarthy. Programs with common sense. In *Proc. Symposium on Mechanisation of Thought Processes*, volume 1, pages 77–84, 1958.

[202] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[203] Saul A. Kripke. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1 – 14, 1959.

[204] M. Ross Quillian. A design for an understanding machine. Paper presented at a colloquium: Semantic Problems in Natural Language, King's College, Cambridge, England, Sep 1961.

[205] Marvin L. Minsky. A framework for representing knowledge. In *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, 1975.

[206] M.W. Eysenck and M.T. Keane. *Cognitive Psychology: A Student's Handbook - 8th ed.* Psychology Press, 2020. ISBN 9781351058506.

[207] Jeffrey R. Binder and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11):527–536, 2011. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2011.10.001.

[208] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(76)90013-X.

[209] Ling ling Wu and Lawrence W. Barsalou. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132(2):173–189, 2009. ISSN 0001-6918. doi: https://doi.org/10.1016/j.actpsy.2009.02.002.

[210] Lawrence W. Barsalou. *The Human Conceptual System*, page 239–258. Cambridge Handbooks in Psychology. Cambridge University Press, 2012.

[211] Willard Van Orman Quine. *From a Logical Point of View.* Harvard University Press, 1953. ISBN 9780674323513.

[212] Ludwig Wittgenstein. *Philosophical Investigations.* Blackwell Publishing, Inc., 1953. ISBN 9780631231271.

[213] George Lakoff. *Women, Fire, and Dangerous Things.* University of Chicago Press, 1987. ISBN 0-226-46803-8.

[214] Ithaca Schwartz. *Naming, Necessity, and Natural Kinds.* Cornell University Press, 1977. ISBN 9780801410499.

[215] O. Ivanov, M. Figurnov, and D. Vetrov. Variational autoencoder with arbitrary conditioning. In *ICLR*, 2019.

[216] Y. Li, S. Akbar, and J. Oliva. Acflow: Flow models for arbitrary conditional likelihoods. In *PMLR*, 2020.

[217] R. Strauss and J. Oliva. Arbitrary conditional distributions with energy. In *NeurIPS*, 2021.

[218] D. Ballard. Modular learning in neural networks. In *AAAI*, 1987.

[219] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[220] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 2017.

[221] R. Yeh, C. Chen, T. Lim, A. Schwing, M. Hasegawa-Johnson, and M. Do. Semantic image inpainting with deep generative models. In *CVPR*, 2017.

[222] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

[223] G. Liu, F. Reda, K. Shih, T. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018.

[224] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019.

[225] W. Cai and Z. Wei. Piigan: Generative adversarial networks for pluralistic image inpainting. *IEEE Access*, 8:48451–48463, 2019.

[226] Y. Liu, Z Wang, Y. Zeng, H. Zeng, and D. Zhao. Pd-gan: Perceptual-details gan for extremely noisy low light image enhancement. In *ICASSP*, 2021.

[227] D. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, 2013.

[228] C. Zheng, T. Cham, and J. Cai. Pluralistic image completion. In *CVPR*, 2019.

[229] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, 2020.

[230] J. Peng, D. Liu, S. Xu, and H. Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *CVPR*, 2021.

[231] R. Strauss and J. Oliva. Posterior matching for arbitrary conditioning. In *NeurIPS*, 2022.

[232] A. Nazabal, P. Olmos, Z. Ghahramani, and I. Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107, 2018.

[233] C. Ma, S. Tschiatschek, K. Palla, J. Hernández-Lobato, S. Nowozin, and C. Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *ICML*, 2019.

[234] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 2017.

[235] Chao Ma, Sebastian Tschiatschek, José Miguel Hernández-Lobato, Richard E. Turner, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. *NeurIPS*, 2020.

[236] Ignacio Peis, Chao Ma, and José Miguel Hernández-Lobato. Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo. *NeurIPS*, 2022.

[237] Mark Collier, Alfredo Nazabal, and Christopher K.I. Williams. Vaes in the presence of missing data. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*, 2020.

[238] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. *ICLR*, 2018.

[239] E Denton and R Fergus. Stochastic video generation with a learned prior. *ICML*, 2018.

[240] R. Karlsson, D. Wong, S. Thompson, and K. Takeda. Learning a model for inferring a spatial road lane network graph using self-supervision. In *ITSC*, 2021.

[241] Hanspeter A. Mallot, Heinrich H. Bülthoff, J. Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biol Cybern.*, 64(3), 1991.

[242] Massimo Bertozzi, Alberto Broggi, and Alessandra Fascioli. An extension to the inverse perspective mapping to handle non-flat roads. In *IV*, 1998.

[243] Massimo Bertozzi, Alberto Broggi, and Alessandra Fascioli. Stereo inverse perspective mapping: theory and applications. *Image Vis. Comput.*, 16:585–590, 1998.

[244] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. *ITSC*, 2020.

[245] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019.

[246] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. *CVPR*, 2020.

[247] Yurong You, Wei-Lun Chao Yan Wang, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *ICLR*, 2020.

[248] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *ICLR*, 2020.

[249] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. *CVPR*, 2020.

[250] Victor Guizilini, Razvan Ambruş, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. *CVPR*, 2021.

[251] Sebastian Schulter, Manmohan Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to look around objects for top-view representations of outdoor scenes. *ECCV*, 2018.

[252] Kaustubh Mani, Swapnil Daga, Shubhika Garg, N. Sai Shankar, J. Krishna Murthy, and K. Madhava Krishna. Mono lay out: Amodal scene layout from a single image. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1678–1686, 2020.

[253] Julian Philion and Sarah Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. *ECCV*, 2020.

[254] Christopher Reading, Alexander Harakeh, Jungjin Chae, and Steven Waslander. Categorical depth distribution network for monocular 3d object detection. *CVPR*, 2021.

[255] Anqi Hu, Ziyu Murez, Nikhil Mohan, Sebastian Dudas, Jonathan Hawke, Vignesh Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. *ICCV*, 2021.

[256] Cheng Lu, Michiel van de Molengraft, and Gerke Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks. *IEEE Robotics and Automation Letters*, 4:445–452, 2018.

[257] Timothy Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *BMVC*, 2018.

[258] Timothy Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. *CVPR*, 2020.

[259] Nicholas Hendy, Christopher Sloan, Feng Tian, Pengcheng Duan, Nadav Charchut, Ye Yuan, Xiaohui Wang, and Jimmy Philbin. Fishing net: Future inference of semantic heatmaps in grids. *CVPR*, 2020.

[260] Wei Yang, Qi Li, Wen Liu, Yuan Yu, Shuang Liu, He He, and Jun Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. *CVPR*, 2021.

[261] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Malik. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. *CoRL*, 2021.

[262] K. Chitta, A. Prakash, and A. Geiger. Neat: Neural attention fields for end-to-end autonomous driving. *ICCV*, 2021.

[263] S. Casas, A. Sadat, and R. Urtasun. Mp3: A unified model to map, perceive, predict and plan. *CVPR*, 2021.

[264] Q. Li, Y. Wang, Y. Wang, and H. Zhao. Hdmapnet: An online hd map construction and evaluation framework. *ICRA*, 2022.

[265] J. Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Forschungsberichte Kunstliche Intelligenz*, 126, 1990.

[266] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 1991.

[267] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation. *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[268] D. Corneil, W. Gerstner, and J. Brea. Efficient model-based deep reinforcement learning with variational state tabulation. In *ICML*, 2018.

[269] D. Ha and J. Schmidhuber. World models. *arXiv*, 2018.

[270] D. Corneil, W. Gerstner, and J. Brea. Efficient model-based deep reinforcement learning with variational state tabulation. In *ICML*, 2018.

[271] T. Kurutach, A. Tamar, G. Yang, S. Russell, and P. Abbeel. Learning plannable representations with causal infogan. In *NeurIPS*, 2018.

[272] A. Wang, T. Kurutach, K. Liu, P. Abbeel, and A. Tamar. Learning robotic manipulation through visual planning and acting. In *Robotics: Science and Systems (RSS)*, 2019.

[273] Y. LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

[274] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *NeurIPS*, 2017.

[275] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *PMLR*, volume 97, pages 2555–2565, 2019.

[276] A. Laversanne-Finot, A. Pere, and P. Oudeyer. Curiosity driven exploration of learned disentangled goal spaces. In *CoRL*, 2018.

[277] C. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv*, 2019.

[278] T. Kipf, E. van der Pol, and M. Welling. Contrastive learning of structured world models. In *ICLR*, 2020.

[279] N. Watters, L. Matthey, M. Bosnjak, C. Burgess, and A. Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. In *ArXiv*, 2019.

[280] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

[281] W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. In *ICML*, 2018.

[282] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[283] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, et al. Stanley: The robot that won the darpa grand challenge. *Journal of Field Robotics*, 23(9):661–692, 2006.

[284] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[285] Randall Smith and Peter Cheeseman. On the representation and estimation of spatial uncertainty. *The Int. J. of Robotics Research*, 5(4), 1986.

[286] Krishna Murthy Jatavallabhula, Ganesh Iyer, and Liam Paull. ∇slam: Dense slam meets automatic differentiation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2130–2137, 2020.

[287] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Stachniss Cyrill. KISS-ICP: In Defense of Point-to-Point ICP Simple,

Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023.

[288] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. SemanticFusion: Dense 3d semantic mapping with convolutional neural networks. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 4628–4635, 2017.

[289] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to look around objects for top-view representations of outdoor scenes. In *European Conference on Computer Vision (ECCV)*, pages 815–831, 2018.

[290] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, et al. Conceptfusion: Open-set multimodal 3d mapping. In *Proceedings of the Robotics: Science and System (RSS)*, 2023.

[291] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[292] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *Proceedings of the 2022 Conference on Robot Learning (CoRL)*, 2022.

[293] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[294] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2856–2865, June 2021.

[295] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[296] Nur Muhammad, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *Proceedings of the Robotics: Science and System (RSS)*, 2023.

[297] Renjie Pi, Jiahui Gao, Shizhe Diao, et al. Detgpt: Detect what you need via reasoning. In *arXiv preprint arXiv:2305.14167*, 2023.

[298] James Jerome Gibson. The ecological approach to visual perception. In *Houghton, Mifflin and Company*, 1979.

[299] A.D. Milner and M.A. Goodale. Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785, 2008. ISSN 0028-3932. Consciousness and Perception: Insights and Hindsights.

[300] Zhixian Han and Anne Sereno. Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Computation*, 34(1):138–171, 2022.

[301] RobinKarlsson, Alexander Carballo, Francisco Lepe-Salazar, Keisuke Fujii, Kento Ohtani, and Kazuya Takeda. Learning to predict navigational patterns from partial observations. *IEEE Robotics and Automation Letters*, 8(9):5592–5599, 2023. doi: 10.1109/LRA.2023.3291924.

[302] T. Salzmann, J. Thomas, T. K"uhbeck, J. Sung, S. Wagner, and A. Knoll. Online path generation from sensor data for highly automated driving functions. In *ITSC*, 2019.

[303] U. Baumann, C. Guiser, M. Herman, and J. Zollner. Predicting ego-vehicle paths from environmental observations with a deep neural network. In *ICRA*, 2018.

[304] D. Barnes, W. Maddern, and I. Posner. Find your own way: Weakly supervised segmentation of path proposals for urban autonomy. In *ICRA*, 2017.

[305] T. Ort et al. Maplite: Autonomous intersection navigation without a detailed prior map. *RL-L*, 2020.

[306] N. Prez-Higueras, F. Caballero, and L. Merino. Learning human-aware path planning with fully convolutional networks. In *ICRA*, 2018.

[307] S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *The Int. J. of Robotics Research*, 30(7), 2011.

[308] K. Kitani, B. Ziebart, A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.

[309] N. Ratliff, D. Silver, and J. Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 27(1):25–53, 2009.

[310] N. Homayounfar, W. Ma, S. Lakshmikanth, and R. Urtasun. Hierarchical recurrent attention networks for structured online maps. In *CVPR*, 2018.

[311] N. Homayounfar, J. Liang, W. Ma, J. Fan, X. Wu, and R. Urtasun. Dagmapper: Learning to map by discovering lane topology. In *ICCV*, 2019.

[312] Y. Guo et al. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *ECCV*, 2020.

[313] J. Z"urn, J. Vertens, and W. Burgard. Lane graph estimation for scene understanding in urban driving. *RA-L*, 2021.

[314] Y. Can, A. Liniger, D. Paudel, and L. Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. In *ICCV*, 2021.

[315] L. Zhang et al. Hierarchical road topology learning for urban map-less driving. In *IROS*, 2022.

[316] L. Mi et al. Hdmapgen: A hierarchical graph generative model of high definition maps. In *CVPR*, 2021.

[317] D. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *NIPS*, 1988.

[318] A. Amini, G. Rosman, S. Karaman, and D. Rus. Variational end-to-end navigation and localization. In *ICRA*, 2018.

[319] M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv*, 2018.

[320] M. Henaff, A. Canziani, and Y. LeCun. Model-predictive policy learning with uncertainty regularization for driving in dense traffic. In *ICLR*, 2019.

[321] D. Chen, V. Koltun, and P. Kr"ahenb"uhl. Learning to drive from a world on rails. In *ICCV*, 2021.

[322] T. Krause, A. Pandey, R. Alami, and A. Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[323] David Papineau. *Philosophical Naturalism*. Blackwell Publishers, 1993. ISBN 9780631189039.

[324] Aristotle. *De Anima (On the Soul)*. Penguin Classics, 1987. Originally written ca. 350 BCE.

[325] David Hume. *A Treatise of Human Nature*. 1739. ISBN 0-7607-7172-3.

[326] Charles E. Spearman. *The Abilities of Man: Their Nature and Measurement*. The MacMillan Company, 1927.

[327] D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, 1949. ISBN 978-0805843002.

[328] Richard J. Haier. *The Neuroscience of Intelligence*. Cambridge University Press, 2016. ISBN 9781316105771.

[329] Richard J Herrnstein and Charles A Murray. *The Bell Curve : Intelligence and Class Structure in American Life*. New York: Free Press, 1994. ISBN 0-02-914673-9.

[330] Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2:79–87, 02 1999. doi: 10.1038/4580.

[331] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364:1211–21, 05 2009. doi: 10.1098/rstb.2008.0300.

[332] John OḰeefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978. ISBN 0198572069.

[333] Karl J. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[334] Noam Chomsky. *Language and Mind*. Cambridge University Press, 2018. ISBN 052167493X.

[335] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability*. Springer, 2005. ISBN 3-540-22139-5.

[336] Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence*. Springer, 2007. ISBN 354023733X.

[337] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, page 17–24, 2007.

[338] N. Mackintosh. *IQ and Human Intelligence*. Oxford University Press, 2011.

[339] R.J. Herrnstein and C. Murray. *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, 2010. ISBN 9781439134917.

[340] Marvin Minsky. *Society of Mind*. Simon & Schuster, 1986. ISBN 0-671-60740-5.

[341] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards,

Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=1ikK0kHjvj.

[342] Aristotle. *The Nicomachean Ethics*. Penguin Classics, 2004. Originally written ca. 350 BCE.

[343] Alan Turing. Computing machinery and intelligence. *Mind*, LIX:433–460, 1950.

[344] Herbert A. Simon and Allen Newell. Heuristic problem solving: The next advance in operations research. *Operations Research*, 6:1–10, 1958.

[345] Herbert A. Simon. Experiments with a heuristic compiler. *J. ACM*, 10(4):493–506, oct 1963. ISSN 0004-5411. doi: 10.1145/321186.321192.

[346] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edition, 1989. ISBN 0201517523.

[347] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(90)90087-6.

[348] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250.

[349] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331 (6022):1279–1285, 2011. doi: 10.1126/science.1192788.

[350] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:1–11, 07 2021. doi: 10.1038/s41586-021-03819-2.

[351] N. Medathati, H. Neumann, G. Masson, and P. Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding*, 150:1–30, 2016.

[352] J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin, Boston, MA, 1979.

[353] D. Milner and M. Goodale. Two visual systems re-viewed. *Neuropsychologia*, 46 (3):774–785, 2008.

[354] Z. Han and A. Sereno. Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Computation*, 34(1):138–171, 2022.

[355] D. Milner. Is visual processing in the dorsal stream accessible to consciousness? *Proc Biol Sci*, 279:2289–2298, 2012.

[356] Paul St ower, Christian Schlieker, Achim Schilling, Claus Metzner, Andreas Maier, and Patrick Krauss. Neural network based successor representations to form cognitive maps of space and language. *Scientific Reports*, 12, 07 2022. doi: 10.1038/s41598-022-14916-1.

[357] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101 (1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X.

[358] Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *Int. J. Rob. Res.*, 32(11):1238–1274, sep 2013. ISSN 0278-3649. doi: 10.1177/0278364913495721. URL https://doi.org/10.1177/0278364913495721.

[359] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136.

[360] E. F. Moore. The shortest path through a maze. *Proceedings of an International Symposium on the Theory of Switching, Part II*, 4(2):285–292, 1959.

[361] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *Proceedings of the 5th International Conference on Computers and Games*, CG'06, page 72–83. Springer-Verlag, 2006. ISBN 3540755373.

[362] Richard Bellman. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

[363] Åström, Karl Johan. Optimal Control of Markov Processes with Incomplete State Information I. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. ISSN 0022-247X. doi: {10.1016/0022-247X(65)90154-X}.

[364] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[365] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967. doi: 10.1109/TIT.1967.1054010.

[366] Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, and Wager TD. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn Psychol.*, 41(1):49–100, 1999. doi: 10.1006/cogp.1999.0734.

[367] Conway MA and Pleydell-Pearce CW. The construction of autobiographical memories in the self-memory system. *Psychol Rev.*, 107(2):261–288, Apr 2000. doi: 10.1037/0033-295x.107.2.261.

[368] Karl Szpunar. Episodic future thought: An emerging concept. *Perspectives on Psychological Science*, 5:142–162, 03 2010. doi: 10.1177/1745691610362350.

[369] Daniel L. Schacter and Donna Rose Addis. The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362:773 – 786, 2007.

[370] R.M. Murray, Z. Li, and S.S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1st edition, 1994. ISBN 0201517523. doi: 10.1201/9781315136370.

[371] Jérôme Barraquand and Jean-Claude Latombe. Robot motion planning: A distributed representation approach. *The International Journal of Robotics Research*, 10(6):628–649, 1991. doi: 10.1177/027836499101000604.

[372] G. Franklin, J.D. Powell, and A. Emami-Naeini. *Feedback Control Of Dynamic Systems*. 01 1994.

[373] Wise SP Donoghue JP. The motor cortex of the rat: cytoarchitecture and microstimulation mapping. *J Comp Neurol.*, 212(1):76–88, 1982. doi: 10.1109/TIT.1967.1054010.

[374] Apostolos P. Georgopoulos, Andrew B. Schwartz, and Ronald E. Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986. doi: 10.1126/science.3749885.

[375] Georgopoulos AP, Kalaska JF, Caminiti R, and Massey JT. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci.*, 2(11):1527–1537, Nov 1982. doi: 10.1523/JNEUROSCI.02-11-01527.

[376] S. Macenski and I. Jambrecic. Slam toolbox: Slam for hte dynamic world. *J. Open Source Softw.*, 6(61):2783, 2021.

[377] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv: Arxiv-2310.12931*, 2023.

[378] Yecheng Jason Ma, William Liang, Hungju Wang, Sam Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. Dreureka: Language model guided sim-to-real transfer. In *Robotics: Science and Systems (RSS)*, 2024.

[379] Gabriel Cristobal, Laurent Perrinet, and Matthias Keil. *Biologically Inspired Computer Vision: Fundamentals and Applications*. Aug. 2015. ISBN 9783527680863. doi: 10.1002/9783527680863.

[380] B. Lake, T. Ullman, J. Tenenbaum, and S. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[381] Robin Karlsson, Alexander Carballo, Keisuke Fujii, Kento Ohtani, and Kazuya Takeda. Predictive world models from real-world partial observations. In *IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*, pages 152–166, 2023.

[382] Wenlong Huang, Pieter Abbeel, Deepak Pathak, et al. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.

[383] Andy Zeng, Maria Attarian, Brian Ichter, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations (ICLR)*, 2023.

[384] Wenlong Huang, Fei Xia, Ted Xiao, et al. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, pages 1769–1782, 2023.

[385] Jacky Liang, Wenlong Huang, Fei Xia, et al. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[386] Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. In *Workshop on Reincarnating Reinforcement Learning at ICLR*, 2023.

[387] Ishika Singh, Valts Blukis, Arsalan Mousavian, et al. Progprompt: Generating situated robot task plans using large language models. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 11523–11530, 2023.

[388] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[389] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[390] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip Torr. Learn to pay attention. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=HyzbhfWRW.

[391] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159. Association for Computational Linguistics, July 2017. doi: 10.18653/v1/P17-1106. URL https://aclanthology.org/P17-1106.

[392] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

[393] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282.

[394] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

[395] Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[396] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations (ICLR)*, 2023.

[397] Timothy P. McNamara, James K. Hardy, and Stephen C. Hirtle. Subjective hierarchies in spatial memory. *Journal of experimental psychology: Learning, memory, and cognition*, 15 2:211–27, 1989.

[398] Andrew J. Davison. Futuremapping: The computational structure of spatial ai systems. In *arXiv preprint arXiv:1803.11288*, 2018.

[399] Fei Xia, Amir R. Zamir, Zhiyang He, et al. Gibson env: Real-world perception for embodied agents. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9079, 2018. doi: 10.1109/CVPR.2018.00945.

[400] Kevin Chen, Junshen K. Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11271–11281, 2021.

[401] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[402] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023. URL https://arxiv.org/abs/2305.11176.

[403] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, Hao Tian, Lewei Lu, Xizhou Zhu, Xiaogang Wang, Yu Qiao, and Jifeng Dai. Drivemlm: Aligning multi-modal

large language models with behavioral planning states for autonomous driving, 2023. URL https://arxiv.org/abs/2312.09245.

[404] Shu Ishida, Gianluca Corrado, George Fedoseev, Hudson Yeo, Lloyd Russell, Jamie Shotton, Joao F. Henriques, and Anthony Hu. Langprop: A code optimization framework using large language models applied to driving. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL https://openreview.net/forum?id=JQJJ9PkdYC.

[405] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, Jacob Andreas, Igor Mordatch, Antonio Torralba, and Yuke Zhu. Pre-trained language models for interactive decision-making. In *Advances in Neural Information Processing Systems*, volume 35, pages 31199–31212. Curran Associates, Inc., 2022.

[406] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[407] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022. ISBN 9781713871088.

[408] Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, and Andrew E. Fano. Causal bert: Language models for causality detection between events expressed in text. In Kohei Arai, editor, *Intelligent Computing*, pages 965–980. Springer International Publishing, 2022. ISBN 978-3-030-80119-9.

[409] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.

[410] Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. Large language models are semi-parametric reinforcement learning agents. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ZcJa1R6j3v.

[411] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy

Zeng, Igor Mordatch, and Pete Florence. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[412] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models, 2023.

[413] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. In *OpenAI*, 2018.

[414] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2024.

[415] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

[416] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 46534–46594, 2023.

[417] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[418] W. Förstner and B. Wrobel. *Photogrammetric Computer Vision*. Springer Cham, 2016.

[419] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019.

[420] V. Guizilini, R. Ambrus, W. Burgard, and A. Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

[421] C. Godard, O. Aodha, M. Firman, and G. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.

[422] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[423] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *ECCV*, 2018.

[424] J. Chang and Y. Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[425] H. Xu and J. Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, 2020.

[426] S. Koelsch and W. A. Siebel. Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9:578–584, 2005.

[427] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *In Proc. 34th International Conference on Machine Learning*, volume 70, 2017.

[428] B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, and A. B. Wiltschko. Machine learning for scent: Learning general- izable perceptual representations of small molecules. In *arXiv:1910.10685.*, 2019.

[429] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2876–2885, 2020.

[430] John McCarthy. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London, 1959. Her Majesty's Stationary Office. URL http://jmvidal.cse.sc.edu/library/maccarthy59a.pdf.

[431] Nils J. Nilsson. *Principles of Artificial Intelligence*. Springer Berlin, Heidelberg, 1st edition, 1982. ISBN 978-3-540-11340-9. doi: 10.1016/C2009-0-27546-5.

[432] Cordell Green. Application of theorem proving to problem solving. In *Proceedings of the 1st International Joint Conference on Artificial Intelligence*, IJCAI'69, page 219–239, San Francisco, CA, USA, 1969. Morgan Kaufmann Publishers Inc.

[433] M. H. Van Emden and R. A. Kowalski. The semantics of predicate logic as a programming language. *J. ACM*, 23(4):733–742, oct 1976. ISSN 0004-5411. doi: 10.1145/321978.321991. URL https://doi.org/10.1145/321978.321991.

[434] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, 2nd edition, 2020.

[435] Michael R. James and Satinder Singh. Learning and discovery of predictive state representations in dynamical systems with reset. *Proceedings of the twenty-first international conference on Machine learning*, 2004.

[436] Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12:1371–1398, 2000.

[437] Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2986–2997, 2022.

[438] Kenneth O. Stanley and Joel Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, 2015.

[439] Drew McDermott. Planning and acting. *Cognitive Sciences*, 2:71–100, 1978.

[440] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1):14–23, 1986. doi: 10.1109/JRA.1986.1087032.

[441] Michael P. Georgeff and Amy L. Lansky. Reactive reasoning and planning. In *AAAI Conference on Artificial Intelligence*, 1987.

[442] Henry Kautz and Bart Selman. Planning as satisfiability. In *Proceedings of the 10th European Conference on Artificial Intelligence*, ECAI '92, page 359–363, USA, 1992. John Wiley & Sons, Inc. ISBN 0471936081.

[443] Allen Newell and Herbert A. Simon. *GPS, a program that simulates human thought*, page 415–428. American Association for Artificial Intelligence, USA, 1995. ISBN 0262621010.

[444] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[445] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 978-0-374-27563-1.

[446] D. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, 2013.

[447] R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *ICML*, 2016.

[448] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020.

[449] R. Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *ICLR*, 2021.

[450] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844.

[451] Judea Pearl. *Heuristics: intelligent search strategies for computer problem solving*. Addison-Wesley Longman Publishing Co., Inc., USA, 1984. ISBN 0201055945.

[452] R udiger Ebendt and Rolf Drechsler. Weighted a search – unifying view and application. *Artificial Intelligence*, 173(14):1310–1342, 2009. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2009.06.004.

[453] Martin Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994. ISBN 9780471619772.

[454] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 11809–11822, 2023.

[455] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[456] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[457] Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. Augmenting autotelic agents with large language models. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 205–226. PMLR, 22–25 Aug 2023.

[458] Toki Migimatsu and Jeannette Bohg. Grounding predicates through actions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3498–3504, 2022. doi: 10.1109/ICRA46639.2022.9812016.

[459] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 893–905. PMLR, 23–24 Jun 2022.

[460] Siwei Chen, Anxing Xiao, and David Hsu. Llm-state: Open world state representation for long-horizon task planning with large language model. *ArXiv*, 2024.

[461] Fu-Jen Chu, Ruinian Xu, and Patricio A. Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics and Automation Letters*, 3(4):3355–3362, 2018. doi: 10.1109/LRA.2018.2852777.

[462] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *ArXiv*, abs/2108.07732, 2021.

[463] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis.

[464] H. Sheif and X. Hu. Autonomous driving in the icity-hd maps as a key challenge of the automotive industry. In *Engineering*, 2016.

[465] B. Sheth and R. Young. Two visual pathways in primates based on sampling of space: Exploitation and exploration of visual information. *Front. Integr. Neurosci.*, 10, 2016.

[466] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, and D. Filliat. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020.

[467] G.D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. doi: 10.1109/PROC.1973.9030.

[468] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744, 2022.

[469] S Amarel. An approach to heuristic problem-solving and theorem proving in the propositional cal-culus. *Systems and Computer Science*, 1967.

[470] Nils J. Nilsson. Problem-solving methods in artificial intelligence. In *McGraw-Hill computer science series*, 1971.

[471] Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

[472] M. Erdmann and M. Mason. An exploration of sensorless manipulation. In *Proceedings. 1986 IEEE International Conference on Robotics and Automation*, volume 3, pages 1569–1574, 1986. doi: 10.1109/ROBOT.1986.1087522.

[473] Blai Bonet and Hector Geffner. An algorithm better than ao*? In *AAAI Conference on Artificial Intelligence*, 2005.

[474] Stuart J. Russell and J. Wolfe. Efficient belief-state and-or search, with application to kriegspiel. In *International Joint Conference on Artificial Intelligence*, 2005.

[475] Sven Koenig and Maxim Likhachev. D*lite. In *AAAI/IAAI*, 2002.

[476] Sven Koenig, Maxim Likhachev, and David Furcy. Lifelong planning a. *Artif. Intell.*, 155:93–146, 2004.

[477] Serena Booth, W. Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i5.25733.

[478] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *NIPS*, pages 6232–6240, 2017.

[479] Yann Le Cunn. Self-supervised learning (keynote talk). AAAI, 2020.

[480] Alexei A. Efros. Self-supervision for learning from the bottom up (invited talk). ICLR, 2021.

[481] Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85:207–238, 1978.

[482] Robert M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115:39–57, 1986.

[483] Robert M. Nosofsky, John K. Kruschke, and Stephen C. McKinley. Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology*, 18:211–233, 1992.

[484] Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973. doi: https://doi.org/10.1016/0010-0285(73)90017-0.

[485] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *NeurIPS*, pages 3859–3869, 2017.

[486] Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2021.

[487] Mohammad Shoeybi, Mostofa Ali Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053, 2019.

[488] Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research (JMLR)*, 6(11):1817–1853, 2005.

[489] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[490] Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence, Mar. 2021.

[491] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

[492] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 55(1):271–304, 2004.

[493] Daniel Kersten and Alan Yuille. Vision as bayesian inference: analysis by synthesis? *Trends Cogn Sci.*, 10(7), 2006.

[494] Moreno Comellas. *Vision as inverse graphics for detailed scene understanding.* PhD thesis, University of Edinburgh, July 1956.

[495] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Loddon Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022.

[496] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, pages 2227–2237, Jun. 2018.

[497] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *Open AI*, 2018.

[498] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40:834–848, 2018.

[499] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *ICCV*, volume 1, pages 10–17, 2003.

[500] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels. In *EPFL Technical Report*, volume 149300, 2010.

[501] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (11):2274–2282, 2012.

[502] Pedro Felzenszwalb and Daniel Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[503] Pablo Arbelaez, Jordi Pont-Tuset, Jon Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[504] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020.

[505] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *ICLR*, 2022.

[506] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *ICLR*, 2021.

[507] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, volume 139, pages 11112–11122, 2021.

[508] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

[509] Tim Kaiser and Nikolas Adaloglou. Understanding SwAV: Self-supervised learning with contrasting cluster assignments. https://theaisummer.com/swav/, 2021.

[510] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL. https://github.com/facebookresearch/vissl, 2021.

[511] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, pages 8026–8037, 2019.

[512] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[513] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[514] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[515] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[516] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, 2018.

[517] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[518] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[519] Kuhn W. Harold. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[520] Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[521] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *ArXiv*, abs/1708.03888, 2017.

[522] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[523] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6706–6716, 2020.

[524] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *ArXiv*, abs/1706.02677, 2017.

[525] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[526] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *ArXiv*, 2020.

[527] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Learning to classify images without labels. *ECCV*, 2020.

[528] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[529] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

[530] R. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. In *The International Journal of Robotics Research*, volume 5, pages 56–68, 1986.

[531] R. Smith and P. Cheeseman. Estimating uncertain spatial relationships in robotics. In *Proceedings of the Second Annual Conference on Uncertainty in Artificial Intelligence*, 1986.

[532] S. Thrun, M. Montemerlo, H. Dahlkamp, D, Stavens, A. Aron, J. Diebel, et al. *Stanley: The Robot That Won the DARPA Grand Challenge*, volume 36 of *Springer Tracts in Advanced Robotics*. Springer, 2007.

[533] Theo M.V. Janssen and Barbara H. Partee. Chapter 7 - compositionality. In *Handbook of Logic and Language*, pages 417–473. North-Holland, 1997. ISBN 978-0-444-81714-3. doi: https://doi.org/10.1016/B978-044481714-3/50011-4.

[534] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

[535] Junhyuk So, Changdae Oh, Yongtaek Lim, Hoyoon Byun, Minchul Shin, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *arxiv:2203.03897*, 2022.

[536] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2021. doi: 10.1109/CVPR46437.2021.00252.

[537] Zhiying Cui, Wu Longshi, and Ruixuan Wang. Open set semantic segmentation with statistical test and adaptive threshold. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[538] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, Nov 1995. doi: 10.1145/219717.219748.

[539] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[540] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4:66–70, 2011. doi: 10.3923/ajms.2011.66.70.

[541] Gerben van den Broeke. What auto-encoders could learn from brains - generation as feedback in deep unsupervised learning and inference. Master's thesis, Aalto University, 2016.

[542] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICML*, 2017.

[543] T. Salimans, A. Karpathy, X. Chen, and D. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *ArXiv*, 2017.

[544] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, 2017.

[545] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, pages 1–16, 2017.

[546] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. *arXiv:2304.02643*, 2023.

[547] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

[548] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017.

[549] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pages 9929–9939, 2020.

[550] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[551] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. In *ArXiv*, 2022.

[552] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[553] Y. Liao, J. Xie, and A. Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. https://arxiv.org/abs/2109.06074, 2021.

[554] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[555] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2013.

[556] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* W.H. Freeman, 1982.

[557] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss. Kiss-icp: In defense of point-to-point icp. *RA-L*, 2022.

[558] R. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *The Int. J. of Robotics Research*, 5(4), 1986.

[559] C. Glasbey and K. Mardia. A review of image warping methods. *Journal of Applied Statistics*, 25:155–171, 1998.

[560] K. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[561] H. Caesar, V. Bankiti, A. Lang, S. Vora, V. Liong, Q. Xu, et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[562] Yusheng Xu, Xiaohua Tong, and Uwe Stilla. Voxel-based representation of 3d point clouds: Methods, applications, and its potential use in the construction industry. *Automation in Construction*, 126:103675, 2021. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2021.103675.

[563] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. doi: 10.1017/9781108679930.

[564] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 0521833787.

[565] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 75392–75412. Curran Associates, Inc., 2023.

[566] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

[567] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian (Shawn) Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34153–34189. Curran Associates, Inc., 2023.

[568] Yue Wu, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Yuanzhi Li, Tom Mitchell, and Shrimai Prabhumoye. Plan, eliminate, and

track – language models are good teachers for embodied agents, 2023. URL https://arxiv.org/abs/2305.02412.

[569] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR, 18–24 Jul 2021.

[570] Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, Demetri Terzopoulos, Ade Famoti, Noboru Kuno, Ashley Llorens, Hoi Vo, Katsu Ikeuchi, Li Fei-Fei, Jianfeng Gao, Naoki Wake, and Qiuyuan Huang. An interactive agent foundation model, 2024. URL https://arxiv.org/abs/2402.05929.

[571] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022.

[572] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. Videollm: Modeling video sequence with large language models, 2023. URL https://arxiv.org/abs/2305.13292.

[573] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16378–16388, 2022. doi: 10.1109/CVPR52688.2022.01591.

[574] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023.

[575] Greg Turk. The ply polygon file format. Web Archive, 2016. URL https://web.archive.org/web/20161204152348/http://www.dcs.ed.ac.

uk/teaching/cs4/www/graphics/Web/ply.html. Accessed via Internet Archive. Accessed 13 July 2024.

[576] The CGAL Project. *CGAL User and Reference Manual*. CGAL Editorial Board, 5.6.1 edition, 2024. URL https://doc.cgal.org/5.6.1/Manual/packages.html.

[577] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3065–3075, June 2024.

[578] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18663, June 2023.

[579] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 598–615. Springer, 2022.

[580] Justin Lazarow, Weijian Xu, and Zhuowen Tu. Instance segmentation with mask-supervised polygonal boundary transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4372–4381, 2022. doi: 10.1109/CVPR52688.2022.00434.

List of Publications

Journal Papers

[1] R. Karlsson, R. Asfandiyarov, A. Carballo, K. Fujii, K. Ohtani, and K. Takeda, "Open-vocabulary Predictive World Models from Sensor Observations," MDPI Sensors, 24(14):4735, https://doi.org/10.3390/s24144735, 2024

[2] R. Karlsson, A. Carballo, F. Lepe-Salazar, K. Fujii, K. Ohtani, and K. Takeda, 'Learning to Predict Navigational Patterns From Partial Observations," IEEE Robotics and Automation Letters (RA-L), 8(9);5592-5599, 2023

Journal Papers (under review)

[3] R. Karlsson, F. Lepe-Salazar, and K. Takeda, "Compositional Semantics for Open Vocabulary Spatio-semantic Representations," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

International Conference Papers

[4] R. Karlsson, A. Carballo, K. Fujii, K. Ohtani, and K. Takeda, "Predictive World Models from Real-World Partial Observations," IEEE International Conference on Mobility, Operations, Services and Technologies (MOST), pp. 152-166, 2023

[5] R. Karlsson, T. Hayashi, K. Fujii, A. Carballo, K. Ohtani, and K. Takeda, "ViCE: Improving Dense Representation Learning by Superpixelization and Contrasting Cluster Assignment," 33rd British Machine Vision Conference 2022, BMVC 2022

[6] R. Karlsson, D. R. Wong, K. Kawabata, S. Thompson, and N. Sakai, "Probabilistic Rainfall Estimation from Automotive Lidar," IEEE Intelligent Vehicles Symposium (IV), 2022

[7] R. Karlsson, D. R. Wong, S. Thompson, and K. Takeda, "Learning a Model for Inferring a Spatial Road Lane Network Graph using Self-Supervision," IEEE International Intelligent Transportation Systems Conference (ITSC), 2021

[8] R. Karlsson and E. Sjoberg, "Learning a Directional Soft Lane Affordance Model for Road Scenes Using Self-Supervision," IEEE Intelligent Vehicles Symposium (IV), 2020

[9] R. Karlsson, "Numerical Simulation of a Freestream MHD Generator System Using the MHD Equations," 30th International Symposium on Space Technology and Science (ISTS), 2015

International Workshops

[10] R. Karlsson, "Learned Reasoning as Framework for AGI in Mobile Robotics," Seventh International Workshop on Symbolic-Neural Learning (SNL), Tokyo, Japan, 2023

List of Awards

1. Best paper award: 2023 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)
2. IEEE ITS Society Nagoya Chapter Young Researcher Award 2023
3. Best paper award: 30th International Symposium on Space Technology and Science (ISTS) 2015
4. Student competition 1st place: All Japan National Indoor Flying Robot Student Contest 2014
5. Japanese language student of the year: Aalto University, Finland, 2011