

# Understanding World or Predicting Future? A Comprehensive Survey of World Models

JINGTAO DING\*, YUNKE ZHANG\*, YU SHANG<sup>†</sup>, YUHENG ZHANG<sup>†</sup>, ZEFANG ZONG<sup>†</sup>, JIE FENG<sup>†</sup>, YUAN YUAN<sup>†</sup>, HONGYUAN SU<sup>†</sup>, NIAN LI<sup>†</sup>, NICHOLAS SUKIENNIK, FENGLI XU, YONG LI, Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, China

The concept of world models has garnered significant attention due to advancements in multimodal large language models such as GPT-4 and video generation models such as Sora, which are central to the pursuit of artificial general intelligence. This survey offers a comprehensive review of the literature on world models. Generally, world models are regarded as tools for either understanding the present state of the world or predicting its future dynamics. This review presents a systematic categorization of world models, emphasizing two primary functions: (1) constructing internal representations to understand the mechanisms of the world, and (2) predicting future states to simulate and guide decision-making. Initially, we examine the current progress in these two categories. We then explore the application of world models in key domains, including autonomous driving, robotics, and social simulacra, with a focus on how each domain utilizes these aspects. Finally, we outline key challenges and provide insights into potential future research directions. We summarize the representative papers along with their code repositories in <https://github.com/tsinghua-fib-lab/World-Model>.

**CCS Concepts:** • Computing methodologies → Machine learning; Artificial intelligence; Modeling and simulation.

**Additional Key Words and Phrases:** World model, model-based RL, video generation, embodied environment, autonomous driving, robots, social simulacra

## 1 INTRODUCTION

The scientific community has long aspired to develop a unified model that can replicate its fundamental dynamics of the world in pursuit of Artificial General Intelligence (AGI) [109]. In 2024, the emergence of multimodal large language models (LLMs) and video generation models like Sora [146] has intensified discussions surrounding such **World Models**. While these models demonstrate an emerging capacity to capture aspects of world knowledge—such as Sora’s generated videos, which appear to perfectly adhere to physical laws—questions persist regarding whether they truly qualify as comprehensive world models. Therefore, a systematic review of recent advancements, applications, and future directions in world model research is both timely and essential as we look toward new breakthroughs in the era of artificial intelligence (AI).

The definition of a world model remains a subject of ongoing debate, generally divided into two primary perspectives: *understanding the world* and *predicting the future*. As depicted in Figure 1, early work by Ha and Schmidhuber [66] focused on abstracting the external world to gain a deep understanding of its underlying mechanisms. In contrast, LeCun [109] argued that a world model should not only perceive and model the real world but also possess the capacity to envision possible future states to inform decision-making. Video generation models such as Sora represent an approach that concentrates on simulating future world evolution and thus align more closely

---

\*These two authors contributed equally.

†These authors contributed equally.

Author’s address: Jingtao Ding\*, Yunke Zhang\*, Yu Shang<sup>†</sup>, Yuheng Zhang<sup>†</sup>, Zefang Zong<sup>†</sup>, Jie Feng<sup>†</sup>, Yuan Yuan<sup>†</sup>, Hongyuan Su<sup>†</sup>, Nian Li<sup>†</sup>, Nicholas Sukiennik, Fengli Xu, Yong Li, ding15@tsinghua.org.cn, liyong07@tsinghua.edu.cn, Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, China.

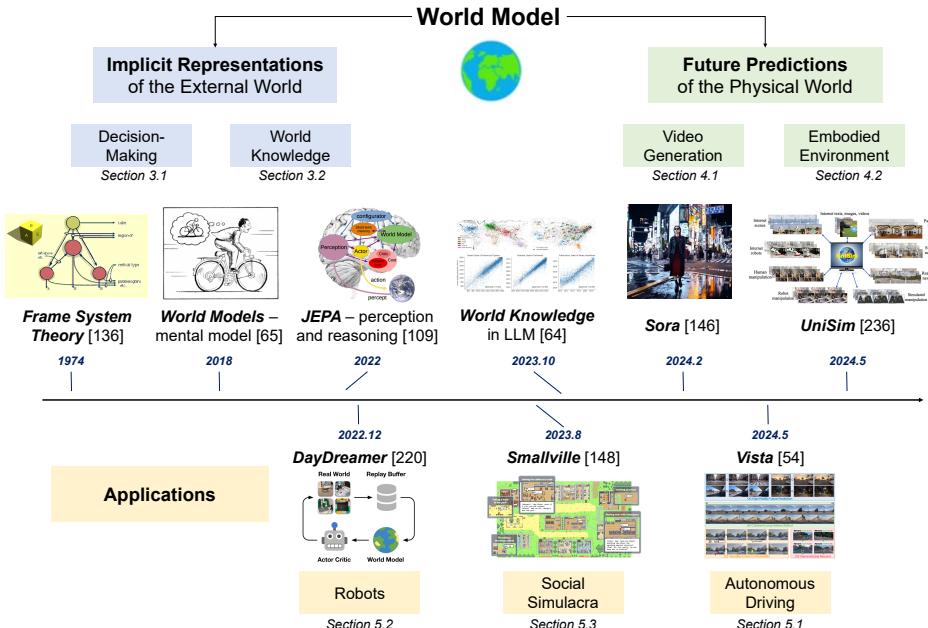


Fig. 1. The overall framework of this survey. We systematically define the essential purpose of a world model as understanding the dynamics of the external world and predicting future scenarios. The timeline illustrates the development of key definitions and applications.

with the predictive aspect of world models. This raises the question of whether a world model should prioritize understanding the present or forecasting future states. In this paper, we provide a comprehensive review of the literature from both perspectives, highlighting key approaches and challenges.

The potential applications of world models span a wide array of fields, each with distinct requirements for understanding and predictive capabilities. In autonomous driving, for example, world models need to perceive road conditions in real-time [198, 216] and accurately predict their evolution [143, 187, 264], with a particular focus on immediate environmental awareness and forecasting of complex trends. For robotics, world models are essential for tasks such as navigation [179], object detection [204], and task planning [69], requiring a precise understanding of external dynamics [51] and the ability to generate interactive and embodied environments [148]. In the realm of simulation of virtual social systems, world models must capture and predict more abstract behavioral dynamics, such as social interactions and human decision-making processes. Thus, a comprehensive review of advancements in these capabilities, alongside an exploration of future research directions and trends, is both timely and essential.

Existing surveys on world models can generally be classified into two categories, as shown in Table S1. The first category primarily focuses on describing the application of world models in specific fields such as video processing and generation [24, 266], autonomous driving [62, 112, 231], and agent-based applications [266]. The second category [130] concentrates on the technological transitions from multi-modal models, which are capable of processing data across various modalities, to world models. However, these papers often lack a systematic examination of what precisely constitutes a world model and what different real-world applications require from these models.

In this article, we aim to formally define and categorize world models, review recent technical progress, and explore their extensive applications.

The main contributions of this survey can be summarized as follows: (1) We present a novel categorization system for world models structured around two primary functions: *constructing implicit representations to understand the mechanism of the external world* and *predicting future states of the external world*. The first category focuses on the development of models that learn and internalize world knowledge to support subsequent decision-making, while the latter emphasizes enhancing predictive and simulative capabilities in the physical world from visual perceptions. (2) Based on this categorization, we classify how various key application areas, including autonomous driving, robots, and social simulacra, emphasize different aspects of world models. (3) We highlight future research directions and trends of world models that can adapt to a broader spectrum of practical applications.

The remainder of this paper is organized as follows. In Section 2, we introduce the background of the world model and propose our categorization system. Section 3 and Section 4 elaborate on the details of current research progress on two categories of world models, respectively. Section 5 covers applications of the world model in three key research fields. Section 6 outlines open problems and future directions of world models.

## 2 BACKGROUND AND CATEGORIZATION

In this section, we explore the evolving concepts of world models in the literature and categorize efforts to construct world models into two distinct branches: internal representation and future prediction.

The concept of building an internal model of the world has a long history in AI, dating back to foundational work such as Marvin Minsky's frame representation in the 1960s [136], designed to systematically capture structured knowledge about the world. Ha *et al.* [65, 66] significantly revived and popularized the term "world model" in 2018 by proposing neural-network-based implicit models for learning latent representations. This line of research aligns with the psychological theory of "mental models" [94]<sup>1</sup>, which holds that humans perceive the external world by abstracting it into simplified elements and relationships—an underlying philosophical root reflected alike in both frames and world models. This principle suggests that our descriptions of the world, when viewed from a deep, internal perspective, typically involve constructing an abstract representation that suffices without requiring detailed depiction. Building upon this conceptual framework, the authors introduce an agent model inspired by the human cognitive system, as illustrated in Figure 1. In this pioneering model, the agent receives feedback from the real-world environment, which is then transformed into a series of inputs that train the model. This model is adept at simulating potential outcomes following specific actions within the external environment. Essentially, it creates a mental simulation of potential future world evolutions, with decisions made based on the predicted outcomes of these states. This methodology closely mirrors the Model-based Reinforcement Learning (MBRL) method, where both strategies involve the model generating internal representations of the external world. These representations facilitate navigation through and resolution of various decision-making tasks in the real world.

In the visionary article on the development of autonomous machine intelligence in 2022 [109], Yann LeCun introduced the Joint Embedding Predictive Architecture (JEPA), a framework mirroring the human brain's structure. As illustrated in Figure 1, JEPA comprises a perception module that processes sensory data, followed by a cognitive module that evaluates this information, effectively embodying the world model. This model allows the brain to assess actions and determine the

---

<sup>1</sup><https://plato.stanford.edu/entries/mental-representation/>

most suitable responses for real-world applications. LeCun's framework is intriguing due to its incorporation of the dual-system concept, mirroring "fast" and "slow" thinking. System 1 involves intuitive, instinctive reactions: quick decisions made without a world model, such as instinctively dodging an oncoming person. In contrast, System 2 employs deliberate, calculated reasoning that considers the future state of the world. It extends beyond immediate sensory input, simulating potential future scenarios, like predicting events in a room over the next ten minutes and adjusting actions accordingly. This level of foresight requires constructing a world model to effectively guide decisions based on the anticipated dynamics and evolution of the environment. In this framework, the world model is essential for understanding and representing the external world. It models the state of the world using latent variables, which capture key information while filtering out redundancies. This approach allows for a highly efficient, minimalistic representation of the world, facilitating optimal decision-making and planning for future scenarios.

The ability of models to capture world knowledge is critical for their effective performance in a wide range of real-world tasks. In the recent wave of works on large language models starting from 2023, several have demonstrated the presence of latent world knowledge. In other words, these models capture intuitive knowledge, including spatial and temporal understanding, which enables them to make predictions about real-world scenarios [64, 133]. Furthermore, LLMs are capable of modeling the external world through cognitive maps, as indicated by recent research revealing the brain-like structures embedded within them [117]. These models can even learn to predict future events based on prior experiences, thereby enhancing their utility and applicability in real-world contexts.

The above world models primarily represent an implicit understanding of the external world. Powered by generative learning like diffusion modeling and model architecture like transformer, recent video generation models (e.g., Sora [146], Keling [104], Gen-2 [29], etc.) take text instructions or real-world visual data as input and output high-quality video frames. Notably, these models demonstrate exceptional modeling capabilities, such as maintaining consistency in 3D video simulations, producing physically plausible outcomes, and simulating digital environments. These capabilities suggest that they not only mimic the appearance of but also model the real-world dynamics within simulation scenarios, focusing on realistically modeling dynamic world changes rather than merely representing static world states.

Whether focusing on learning internal representations of the external world or simulating its operational principles, these concepts coalesce into a shared consensus: the essential purpose of a world model is to understand the dynamics of the world and compute the next state with certainty (or with some guarantee), which empowers the model to extrapolate longer-horizon evolution and to support downstream decision-making and planning. From this perspective, we conduct a thorough examination of recent advancements in world models, analyzing them through the following lenses, as depicted in Figure 1.

- **Implicit representation of the external world** (Section 3): This research category constructs a model of environmental change to enable more informed decision-making, ultimately aiming to predict the evolution of future states. It fosters an implicit comprehension by transforming external realities into a model that represents these elements as latent variables. Furthermore, with the advent of large language models (LLMs), efforts previously concentrated on traditional decision-making tasks have been significantly enhanced by the detailed descriptive power of these models regarding world knowledge. We further focus on the integration of world knowledge into existing models.
- **Future predictions of the external world** (Section 4): We initially explore generative models that simulate the external world, primarily using visual video data. These works

emphasize the realness of generated videos that mirror future states of the physical world. As recent advancements shift focus toward developing a truly interactive physical world, we further investigate the transition from visual to spatial representations and from video to embodiment. This includes comprehensive coverage of studies related to the generation of embodied environments that mirror the external world.

- **Applications of world models** (Section 5): World models have a wide range of applications across various fields, including autonomous driving, robotics, and social simulacra. We explore how the integration of world models in these domains advances both theoretical research and practical implementations, emphasizing their transformative potential in real-world applications.

### 3 IMPLICIT REPRESENTATION OF THE EXTERNAL WORLD

This section examines how world models enable informed decision-making by representing the environment as latent variables. Section 3.1 focuses on the world models in model-based RL (MBRL), while Section 3.2 explores the integration of world knowledge into advanced AI models, especially LLMs, enhancing real-world task performance.

#### 3.1 World Model in Decision-Making

In decision-making tasks, understanding the environment is the major task in setting a foundation for optimized policy generation. As such, the world model in decision-making should include a comprehensive understanding of the environment. It enables us to take hypothetical actions without affecting the real environment, facilitating a low trial-and-error cost. In literature, research on how to learn and utilize the world model was initially proposed in the field of model-based RL. Furthermore, recent progress on LLM and MLLM also provide comprehensive backbones for world model construction. With language serving as a more general representation, language-based world models can be adapted to more generalized tasks. The two schemes of leveraging world models in decision-making tasks are shown in Figure 2.

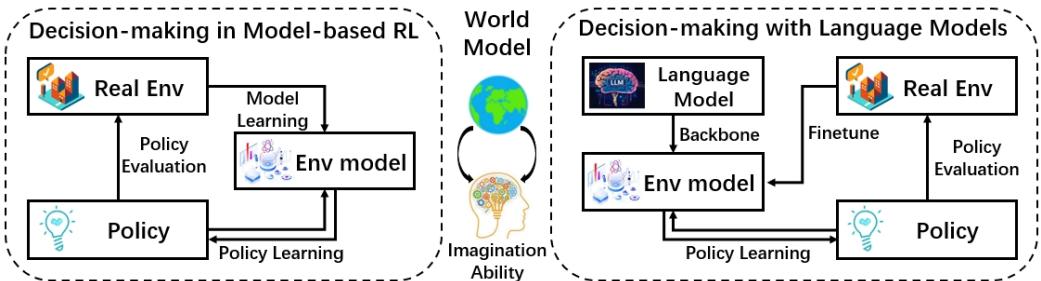


Fig. 2. Two schemes of utilizing world model in decision-making.

**3.1.1 World model in model-based RL.** In decision-making, the concept of the world model largely refers to the environment model in MBRL. A decision-making problem is typically formulated as a Markov Decision Process (MDP), denoted with a tuple  $(S, A, M, R, \gamma)$ , where  $S, A, \gamma$  denotes the state space, action space and the discount factor each. The world model here consists of  $M$ , the state transition dynamics and  $R$ , the reward function. Since the reward function is defined in most cases, the key task of MBRL is to learn and utilize the transition dynamics, which can further support policy optimization.

**World Model Learning.** To learn an accurate world model, the most straightforward approach is to leverage the mean squared prediction error on each one-step transitions [89, 90, 108, 129, 163],

$$\min_{\theta} \mathbb{E}_{s' \sim M^*(\cdot|s,a)} [| | s' - M_{\theta}(s,a) | |_2^2 ], \quad (1)$$

where  $M^*$  is the real transition dynamics used to collect trajectory data and  $M_{\theta}$  is the parameterized transition to learn. Apart from directly utilizing the deterministic transition model, Chua et al.[27] further model the aleatoric uncertainty with the probabilistic transition model. The objective is to minimize the KL divergence between the transition models,

$$\min_{\theta} \mathbb{E}_{s' \sim M^*(\cdot|s,a)} [\log(\frac{M^*(s'|s,a)}{M_{\theta}(s'|s,a)})]. \quad (2)$$

In both settings, the phase of the world model learning task can be transformed into a supervised learning task. The learning labels are the trajectories derived from real interaction environments, also called the simulation data [128].

For high-dimensional environments, representation learning is essential for effective world-model training in MBRL. Early work by Ha and Schmidhuber[65] reconstructs images through an autoencoder-latent-state pipeline, whereas Hafner et al. [68, 70] couple a visual encoder with latent dynamics to master pixel-based control tasks. Their latest iteration, DreamerV3[71], adds robust normalization and balancing techniques, solving over 150 tasks—including diamond collection in Minecraft—without human data or domain-specific tuning. Memory-centric extensions such as Recall-to-Imaging by Samsami et al.[171] further enhance long-horizon reasoning. A complementary trend is unified model learning via next-token prediction with transformer architectures, as shown by Janner et al. [90] and expanded by Schubert et al. [177]. Further, Georgiev et al. [55] train a large off-policy multi-task world model whose smooth latent dynamics enable efficient per-task policy learning with first-order gradients, achieving strong scalability and performance without online planning. Recent work by Jonathan Richens et al.[167] further reinforces the necessity of world models, showing that any agent capable of generalizing to multi-step goal-directed tasks must have learned a predictive model of its environment, with the world model emerging from the agent’s policy. This insight aligns with the ongoing trend of incorporating predictive modeling into reinforcement learning to handle more complex and goal-oriented tasks.

**Policy Generation with World Model.** With an ideally optimized world model, one most straightforward way to generate a corresponding policy is model predictive control (MPC)[102]. MPC plans an optimized sequence of actions given the model as follows:

$$\max_{a_{t:t+\tau}} \mathbb{E}_{s_{t'+1} \sim p(s_{t'+1}|s_t, a_t)} [\sum_{t'=t}^{t+\tau} r(s_{t'}, a_{t'})], \quad (3)$$

where  $\tau$  denotes the planning horizon. Nagabandi et al.[141] adopt a simple Monte Carlo method to sample action sequences. Rather than sampling actions uniformly, Chua et al.[27] propose a new probabilistic algorithm that ensembles with trajectory sampling. Further literature also improves the optimization efficiency by leveraging the world model usage [68, 79, 208, 245]. Hansen et al.[72] introduced an improved model-based RL algorithm called TD-MPC2 that integrates trajectory optimization within the latent space of a learned implicit world model. It achieves strong performance across diverse continuous control tasks and demonstrates scalability by training large agents with hundreds of millions of parameters across multiple domains.

Another popular approach to generating world model policies is the Monte Carlo Tree Search (MCTS). By maintaining a search tree where each node refers to a state evaluated by a predefined value function, actions will be chosen such that the agent can be processed to a state with a higher

value. AlphaGo and AlphaGo Zero are two significant applications using MCTS in discrete action space [189, 190]. Moerland et al. [138] extended MCTS to solve decision problems in continuous action space. Oh et al. [144] proposed a value prediction network that applies MCTS to the learned model to search for actions based on value and reward predictions.

**3.1.2 World model with language backbone.** The rapid growth of language models, especially LLM and MLLM, benefits development in many related applications. With language serving as a universal representation backbone, language-based world models have shown their potential in many decision-making tasks.

**Direct Action Generation via LLM World Models.** LLM is capable of directly generating actions in decision-making tasks based on corresponding constructed world models. For example, in the navigation scenarios, Yang et al. [234] transfer pre-trained text-to-video models to domain-specific tasks for robot control, successfully annotating robot manipulation with text instructions as LLM outputs. Zhou et al. [263] further learn a compositional world model by factorizing the video generation process. Such a method enables a strong few-shot transfer ability to unseen tasks.

Besides training or fine-tuning specialized language-based world models, LLMs and MLLMs can be directly deployed to understand the world environment in decision-making tasks. For example, Long et al. [126] propose a multi-expert scheme to handle visual language navigation tasks. They construct a standardized discussion process where eight LLM-based experts participate to generate the final movement decision. An abstract world model is constructed from the discussion and further imagination (of future states) of the experts to support action generation. Zhao et al. [255] further combine LLMs and open-vocabulary detection to construct the relationship between multi-modal signals and key information in navigation. They propose an omni-graph to capture the structure of the local space as the world model for the navigation task. Meanwhile, Yang et al. [239] utilize an LLM-based imaginative assistant to infer the global semantic graph as the world model based on the environment perception, and another reflective planner to directly generate actions.

**Modular Usage of LLM World Models.** Although taking LLM outputs as actions directly is straightforward in application and deployment, the decision quality in such a scheme heavily relies on the reasoning ability of the LLM itself. Although this year has witnessed the large potential of LLM reasoning capability [225], it can be further improved by integrating LLM-based world models as modules with external model-based verifiers or other effective planning algorithms [96].

Guan et al.[61] extract explicit world models by prompting GPT-4 to generate and iteratively refine PDDL domain descriptions, then pair these models with off-the-shelf planners, yielding good planning performance with much less human intervention. Xiang et al.[224] deploys an embodied agent in a world model, the simulator of VirtualHome [156], where the corresponding embodied knowledge is injected into LLMs. To better plan and complete specific goals, they propose a goal-conditioned planning schema where Monte Carlo Tree Search (MCTS) is utilized to search for the true embodied task goal. Lin et al.[119] introduce an agent, Dynalang, which learns a multimodal world model to predict future text and image representations, and which learns to act from imagined model rollouts. The policy learning stage utilizes an actor-critic algorithm purely based on the previously generated multimodal representations. Liu et al. [125] further cast reasoning in LLMs as learning and planning in Bayesian adaptive Markov decision processes (MDPs). LLMs, like the world model, perform in an in-context manner within the actor-critic updates of MDPs. The proposed RAFA framework shows significantly increased performance in multiple complex reasoning tasks and environments, such as ALFWORLD [188].

Table 1. Overview of recent works in world knowledge learned by models.

Category	Methods/Model	Year&Venue	Modality	Content
Common Sense & General Knowledge	KoLA [242]	2024 ICLR	Language	Benchmark
	EWOK [86]	2024 arxiv	Language	Benchmark
	Geometry of Concepts [117]	2024 arxiv	Language	Analysis
Knowledge of Global Physical World	Space&Time [64]	2024 ICLR	Language	Analysis
	GeoLLM [133]	2024 ICLR	Language	Learning
	GeoLLM-Bias [132]	2024 ICML	Language	Learning
	GPT4GEO [169]	2023 NeurIPS(FMDM)	Language	Benchmark
	CityGPT [42]	2024 arxiv	Language	Learning
Knowledge of Local Physical World	CityBench [44]	2024 arxiv	Language&Vision	Benchmark
	Predictive [59]	2024 NMI	Vision	Learning
	Emergent [93]	2024 ICML	Language	Learning
	E2WM [224]	2023 NeurIPS	Language	Learning
Knowledge of Human Society	Dynalang [119]	2024 ICML	Language&Vision	Learning
	Testing ToM [194]	2024 NHB	Language	Benchmark
	High-order ToM [195]	2024 arxiv	Language	Benchmark
	COKE [219]	2024 ACL	Language	Learning
	MuMA-ToM [185]	2024 ACL	Language&Vision	Benchmark
SimToM [215]	SimToM [215]	2024 ACL	Language	Learning

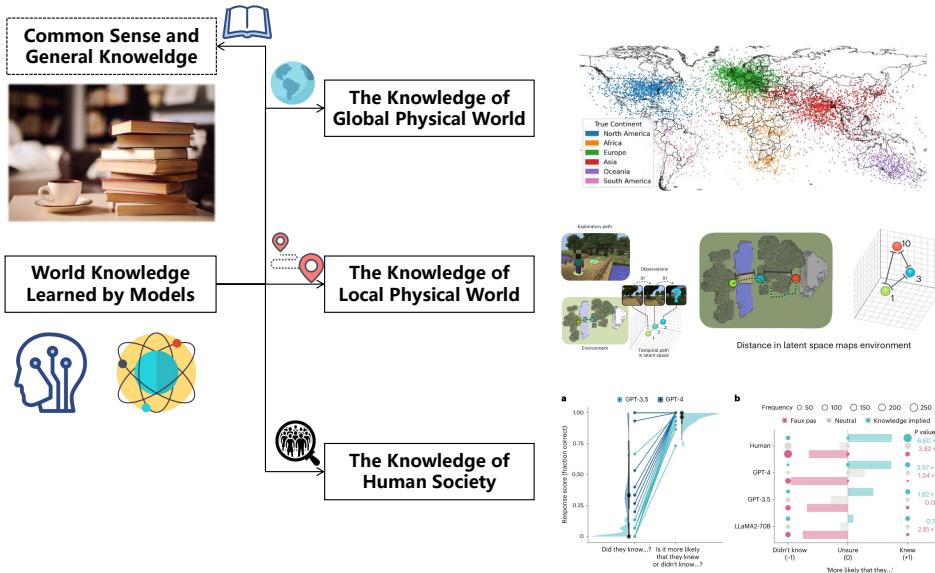


Fig. 3. World knowledge in large language models for world model.

### 3.2 World Knowledge Learned by Models

After pretraining on large-scale web text and books [145, 201], large language models attain extensive knowledge about the real world and common sense relevant to daily life. This embedded knowledge is considered crucial for their remarkable ability to generalize and perform effectively in real-world tasks. For instance, researchers leverage the common sense of large language models for task planning [257], robot control [82], and image understanding [123]. Furthermore, Li et al.[117] discover brain-like structures of world knowledge embedded in the high-dimensional vectors that represent the universe of concepts in large language models. Also, Li et al.[111] demonstrate that language models partially converge towards representations isomorphic to those of vision models.

Unlike common sense and general knowledge, we focus on world knowledge within large language models from the perspective of a world model. As shown in Figure 3, based on objects and spatial scope, the world knowledge in the large language models can be categorized into three parts: 1) knowledge of the global physical world; 2) knowledge of the local physical world; and 3) knowledge of human society. We summarize recent works in Table 1.

**3.2.1 Knowledge of the Global Physical World.** We first introduce research focused on analyzing and understanding the knowledge of the global physical world. Gurnee et al. [64] present the first evidence that large language models genuinely acquire spatial and temporal knowledge of the world, rather than merely collecting superficial statistics. They identify distinct "spatial neurons" and "temporal neurons" in LLaMA2 [201], suggesting that the model learns linear representations of space and time across multiple scales. Distinct from previous observations focused on embedding space, Manvi et al. [132, 133] develop effective prompts about textual address to extract intuitive real-world knowledge about the geospatial space and successfully improve the performance of the model in various downstream geospatial prediction tasks.

While large language models do acquire some implicit knowledge of the real world [64, 117], the quality of this knowledge remains questionable [42, 169]. For example, Feng et al. [42] find that the urban knowledge embedded in large language models is often coarse and inaccurate. To address this, they propose an effective framework to improve the acquisition of urban knowledge of specific cities in large language models. From these works, we can see that although large language models have demonstrated the ability to capture certain aspects of real-world knowledge [64, 117, 169], it is clear that further efforts are needed to enhance this knowledge to enable broader and more reliable real-world applications [43].

**3.2.2 Knowledge of the Local Physical World.** Unlike the knowledge of the global physical world, the local physical world represents the primary environment for human daily life and most real-world tasks. Therefore, understanding and modeling the local physical world is a more critical topic for building a comprehensive world model. We first introduce the concept of the cognitive map [200], which refers to the mental representation that humans form to navigate and understand their environment, including spatial relationships and landmarks. Although initially developed to explain human learning processes, researchers have discovered similar structures in large language models [117] and have leveraged these insights to enhance the efficiency and performance of artificial models in learning and understanding the physical world.

Recent studies explore actively encouraging models to learn abstract knowledge through cognitive map-like processes across various environments. For example, Cornet et al. [59] show that in a simplified Minecraft world, visual predictive coding lets an agent build a spatial cognitive map purely from pixels. Once trained, the latent map encodes its metric distance to any target, enabling accurate rollout of future observations. Lin et al. [119] investigate teaching models to understand the game environments through a world model learning procedure, specifically by predicting the subsequent frame of the environment. In this way, the model can generate better actions in dynamic environments. Moreover, Jin et al. [93] find that language models can learn the emergent representations of program semantics by predicting the next token.

**3.2.3 Knowledge of the Human Society.** Beyond the physical world, understanding human society is another crucial aspect of world models. David Premack and Guy Woodruff proposed the Theory of Mind [155], which was later developed to explain how individuals infer the mental states of others around them. Recent works have extensively explored how large language models develop and demonstrate this social world model [173, 194]. Sap et al. [173] conduct an investigation focusing on evaluating the performance of large language models across various Theory of Mind tasks

to determine whether their human-like behaviors reflect genuine comprehension of social rules and implicit knowledge. Strachan et al. [194] conduct a comparative analysis between human and LLM performance on diverse Theory of Mind abilities, such as understanding false beliefs and recognizing irony. While their findings demonstrate the potential of GPT-4 in these tasks, they also identify its limitations, particularly in detecting faux pas.

To address these limitations, researchers propose innovative methods to enhance the abilities of large language models in Theory of Mind for complex real-world applications. Wu et al. [219] introduce COKE, which constructs a knowledge graph to help large language models explicitly use theory in mind through cognitive chains. Additionally, Alex et al. [215] develop SimTom, a two-stage prompting framework, to enhance the performance of large language models in theory of mind tasks.

## 4 FUTURE PREDICTION OF THE PHYSICAL WORLD

### 4.1 World Model as Video Generation

The integration of video generation into world models marks a significant leap forward in the field of environment modeling [146]. Traditional world models primarily focused on predicting discrete or static future states [66, 109]. However, by generating video-like simulations that capture continuous spatial and temporal dynamics, world models [146, 233] have evolved to address more complex, dynamic environments. This breakthrough in video generation has pushed the capabilities of world models to a new level.

**4.1.1 Towards Video World Models.** A video world model is a computational framework designed to simulate and predict the future state of the world by processing past observations and potential actions within a visual context [146]. This concept builds on the broader idea of world models, which strive to capture the dynamics of an environment and enable machines to predict how the world will evolve over time. In the case of a video world model, the focus is on generating sequences of visual frames that represent these evolving states.

**Sora as a World Simulator.** Sora [146] is a large-scale video generation model, which is designed to generate high-quality, temporally consistent video sequences, up to one minute long, based on various input modalities such as text, images, and videos. Sora leverages a combination of powerful neural network architectures, including encoder-decoder frameworks and transformers, to process multimodal inputs and generate visually coherent simulations. Sora's core capabilities lie in its ability to generate videos that align with real-world physical principles, such as the reflection of light on surfaces or the melting of candles. These properties suggest that Sora has the potential to act as a world simulator, predicting future states of the world based on its understanding of the initial conditions and simulation parameters.

**Sora's Limitations.** However, despite its impressive video generation abilities, Sora has several limitations in terms of understanding and simulating the external world. One key limitation concerns causal reasoning [24, 266], wherein the model is limited in simulating dynamic interactions within the environment. Thus, Sora can only passively generate video sequences based on an observed initial state, but cannot actively intervene or predict how changes in actions might alter the course of events. Another limitation is that it still fails to reproduce correct physical laws consistently [97]. While Sora can generate visually realistic scenes, it struggles with accurately simulating real-world physics, such as the behavior of objects under different forces, fluid dynamics, or the accurate depiction of light and shadow interactions.

**Other Video World Models.** Sora has undoubtedly catalyzed a significant wave of research into video world models, inspiring a surge of advancements in this field. Following Sora's success in generating high-quality video sequences, numerous subsequent models have been developed,

each aiming to push the boundaries of what video world models can achieve. For example, some approaches have extended video lengths to enable long-form video simulation [77, 121, 241]. In addition to conventional language-guided video generation, more modalities are being integrated, such as images and actions [223, 258]. Researchers are also shifting their focus from basic video generation, which lacks user control, to interactive simulations that aim to replicate the decision space of the real world and facilitate decision-making [87, 218, 223, 235, 237, 249]. Several studies have worked to enhance the smoothness of action transitions, improve the accuracy of physical laws, and maintain temporal consistency [17, 166, 229, 233]. Meanwhile, the concept of world models has evolved beyond imagination and is being applied in various scenario-specific simulations, including natural environments, games, and autonomous driving [12, 16, 77, 121, 134, 135, 209, 211, 261]. Table 2 summarizes the categorization of improvements in video world models across different aspects.

**4.1.2 Capabilities of Video World Models.** Despite the ongoing debate about whether models like Sora can be considered full-fledged world models, there is no doubt that video world models hold tremendous potential for advancing environment simulation and prediction [24, 97, 266]. These models can offer a powerful approach to understanding and interacting with complex environments by generating realistic, dynamic video sequences. To achieve this level of sophistication, this section outlines the key capabilities that video world models must possess to set them apart from traditional video generation models.

**Long-Term Predictive Ability.** A robust video world model should be capable of making long-term predictions that adhere to the dynamic rules of the environment over an extended period. This capability allows the model to simulate how a scenario evolves, ensuring that the generated video sequences remain consistent with the temporal progression of the real world. Although Sora has achieved the generation of minute-long video sequences with high-quality temporal coherence, it is still far from being able to simulate complex, long-term dynamics found in real-world environments. Recent efforts have explored extending video lengths to capture longer-term dependencies and improve temporal consistency [77, 121, 241].

**Multi-Modal Integration.** In addition to language-guided video generation, video world models are increasingly integrating other modalities, such as images and actions, to enhance realism and interactivity [223, 258]. The integration of multiple modalities allows for richer simulations that better capture the complexity of real-world environments, improving both the accuracy and diversity of generated scenarios.

**Interactivity.** Another critical capability of video world models is their potential for controllability and interactivity. An ideal model should not only generate realistic simulations but also allow for interaction with the environment. This interactivity involves simulating the consequences of different actions and providing feedback, enabling the model to be used in applications requiring dynamic decision-making. Recent work is focusing on enhancing control over the simulations, allowing for more user-guided exploration of scenarios [218, 237].

**Diverse Environments.** Finally, video world models are being adapted to a variety of scenario-specific simulations, including natural environments, autonomous driving, and gaming. These models are evolving beyond basic video generation to replicate real-world dynamics and support a wide range of applications [16, 121, 211].

## 4.2 World Model as Embodied Environment

The development of world models for embodied environments is crucial for simulating and predicting how agents interact with and adapt to the external world. Initially, generative models focused on simulating visual aspects of the world, using video data to capture dynamic changes in the

Table 2. Overview of recent models in video generation across various categories, which summarizes key models in long-term video generation, multi-modal learning, interactive video generation, temporal consistency, and diverse environment modeling.

Category	Model	Description	Technique
Long-term	NUWA-XL [241]	“Coarse-to-fine” Diffusion over Diffusion architecture for long video generation.	Diffusion
	LWM [121]	Training large transformers on long video and language sequences.	Transformer
	GAIA-1 [77]	Generative world model predicting driving scenarios for autonomous driving.	Transformer, Diffusion
Multimodal	3D-VLA [258]	Integrates 3D perception, reasoning, and action in a world model for embodied AI.	Diffusion
	Pandora [223]	World-state simulation and real-time control with free-text actions.	LLM
	Genie [16]	Generative model from text, images, and sketches.	Transformer
Interactive	UniSim [235]	Simulates real-world interactions for vision-language and RL training.	Diffusion, RL
	VideoDecision[237]	Extends video models to real-world tasks like planning and RL.	Transformer, Diffusion
	iVideoGPT [218]	Combines visual, action, and reward signals for interactive world modeling.	Transformer
	PhysDreamer [249]	Simulates 3D object dynamics to generate responses to novel interactions.	Diffusion
	PEEKABOO [87]	Enhances interactivity with spatiotemporal control without extra training.	Diffusion Transformer
Consistency	WorldGPT [233]	Improves temporal consistency and action smoothness with multimodal learning and refined key frame generation.	Diffusion
	DiffDreamer [17]	Long-range scene extrapolation with improved consistency.	Diffusion
	ConsistI2V [166]	Enhances visual consistency in image-to-video generation.	Diffusion
Diverse environments	WorldDreamer [211]	World model capturing dynamic elements across diverse scenarios.	Transformer
	Genie [16]	Unsupervised generative model for action-controllable virtual environments.	Transformer
	MUVO [12]	Multimodal world model using camera and lidar data.	Transformer
	UniWorld [135]	3D detection and motion prediction in autonomous driving.	Transformer

environment. More recently, the focus has shifted towards creating fully interactive and embodied simulations. These models not only represent the visual elements of the world but also incorporate spatial and physical interactions that more accurately reflect real-world dynamics. By integrating spatial representations and transitioning from video-based simulations to immersive, embodied environments, world models can now provide a more comprehensive platform for developing agents capable of interacting with complex real-world environments.

World models as embodied environments can be divided into three categories: indoor, outdoor, and dynamic environments, as shown in Figure 4, and the relevant works are summarized in Table 3. It can be summarized that most current works focus on developing static, existing indoor and outdoor embodied environments. An emerging trend is to predict the dynamic, future world

Table 3. Comparison of existing works on world models as embodied environments, including indoor, outdoor, and dynamic environments. In the ‘Modality’ column, ‘V’ refers to vision, ‘L’ refers to lidar, ‘T’ refers to text, and ‘A’ refers to audio. In the ‘Num of Scenes’ column, ‘-’ means no reported data, and ‘Arbitrary’ means the method can support generating any number of scenes.

Type	Name	Environment	Year	Num of Scenes	Modality	Physics	3D Assets
Indoor	AI2-THOR [101]	Home	2017	120	V	✓	✓
Indoor	Matterport 3D [18]	Home	2018	90	V	✗	✗
Indoor	Virtual Home [156]	Home	2018	50	V	✓	✓
Indoor	Habitat [174]	Home	2019	-	V	✓	✓
Indoor	SAPIEN [222]	Home	2020	46	V	✓	✓
Indoor	iGibson [183]	Home	2021	15	V, L	✓	✓
Indoor	AVLEN [150]	Home	2022	85	V, T, A	✓	✓
Indoor	ProcTHOR [31]	Home	2022	Arbitrary	V	✓	✓
Indoor	Holodeck [238]	Home	2024	Arbitrary	V	✓	✓
Indoor	AnyHome [48]	Home	2024	Arbitrary	V	✓	✓
Indoor	LEGENT [21]	Home	2024	Arbitrary	V, T	✓	✓
Indoor	TDW [49]	Home	2021	-	V, A	✓	✓
In & Outdoor	GRUtopia [205]	Home, City	2024	100k	V, T	✓	✓
Outdoor	MineDOJO [41]	Game	2022	-	V	✗	✗
Outdoor	MetaUrban [221]	City	2024	13800	V, L	✓	✓
Outdoor	UrbanWorld [180]	City Building	2024	Arbitrary	V	✓	✓
Outdoor	EmbodiedCity [52]	City	2024	87.1k	V, T	✓	✓
Dynamic	UniSim [236]	Home, City, Simulation	2023	Arbitrary	V, T	✗	✗
Dynamic	StreetScapes [32]	Street View	2024	Arbitrary	V, T	✗	✗
Dynamic	AVID [168]	Home, Game	2024	Arbitrary	V, T	✗	✗
Dynamic	EVA [23]	Home, Simulation	2024	Arbitrary	V, T	✗	✗
Dynamic	Pandora [223]	Home, Game, Simulation, Street View	2024	Arbitrary	V, T	✗	✗

through generative models producing first-person, dynamic video-based simulation environments. Such environments can offer flexible and realistic feedback for training embodied agents, enabling them to interact with ever-changing environments and improve their generalization ability.

**4.2.1 Indoor Environments.** Indoor environments offer controlled, structured scenarios where agents can perform detailed, task-specific actions such as object manipulation, navigation, and real-time interaction with users [18, 53, 101, 150, 156, 174, 183, 222]. Early works on establishing indoor environments like AI2-THOR [101] and Matterport 3D [18] focus on providing only visual information. These works build indoor environments by providing photorealistic settings where agents can practice visual navigation and engage in interactive tasks that mimic real-life home activities. These environments emphasize the importance of using visual-based reinforcement learning techniques that allow agents to optimize their decision-making based on environmental cues. By simulating real-world tasks like cooking or cleaning, these platforms assess an agent’s capacity to generalize learned behaviors across different types of spaces and objects. A line of further works contributes toward expanding the data modalities of the provided environments. Among these, iGibson [183] introduces Lidar observation as additional signal feedback, contributing to more accurate environment perception of agents. AVLEN [150] further supplements audio signals allowing agents to execute tasks such as object manipulation and navigation in household-like settings. The challenge here lies in enabling agents to understand and act on multimodal input including vision, language, and sound within a constrained space. Adding a social dimension, environments like GRUtopia [205] introduce agents to spaces where they must navigate and interact with both objects and NPCs. Here, agents need to understand social dynamics, such as positioning and task sharing, which requires more advanced forms of interaction modeling. The inclusion of social interaction modules in these settings demonstrates how agents can be trained to balance human-like social behaviors with task performance. More recently, with the development

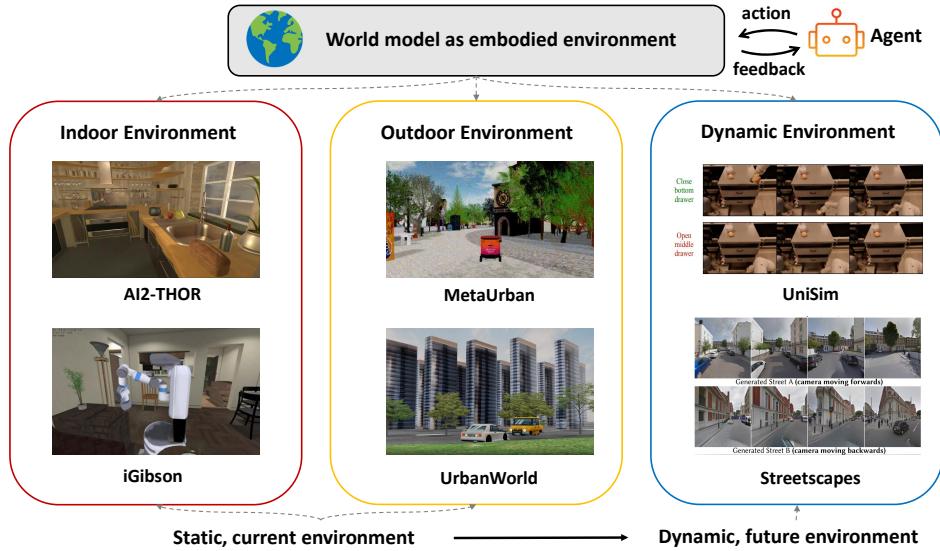


Fig. 4. Classification of world models as interactive embodied environments, including indoor, outdoor and dynamic environments. The modeling of the outside world is evolving from constructing static, current environments to predicting dynamic, future environments.

of LLMs, some works [21, 48, 238] seek to provide a flexible environment generation pipeline, supporting the generation of arbitrary indoor environments with language instructions.

**4.2.2 Outdoor Environments.** In contrast to indoor environments, creating outdoor environments [41, 49, 180, 205, 221] faces greater challenges due to their larger scale and increased variability. Some existing works focus on urban environments, such as MetaUrban [221], where agents are deployed to navigate in large-scale urban environments, where they encounter challenges like dynamically changing traffic, varied building structures, and social interactions with other entities. These tasks often require the use of context-aware navigation algorithms that allow agents to adjust their trajectories and behaviors based on the layout and conditions of the environment. However, the environments in MetaUrban are created by retrieving and organizing 3D assets from existing libraries. Recently, utilizing advanced generative techniques, UrbanWorld [180] significantly enhances the scope of outdoor environments, using 3D generative models to create complex, customizable urban spaces that allow for more diverse urban scenarios. This shift from static asset-based environments to generative ones ensures that agents are exposed to a wider variety of tasks, from navigating unfamiliar street layouts to interacting with new types of objects or structures. In addition to the above real open-world generation works, there are also some virtual open-world platforms like MineDOJO [41] that extend these challenges even further by simulating procedurally generated, sandbox-like environments. These platforms, inspired by the open-ended world of Minecraft, push agents to engage in tasks like resource collection, construction, and survival, demanding continuous exploration and adaptive learning. In such environments, agents are motivated to seek out new information and adapt their behavior to finish given tasks. Training in such environments can help agents learn knowledge across a broad range of tasks and terrains, enabling them to operate effectively in various outdoor environments.

**4.2.3 Dynamic Environments.** Dynamic environments mark a significant evolution from traditional, static simulators by utilizing generative models to create flexible, real-time simulations. Unlike predefined environments that require manual adjustments, these models allow for the dynamic creation of a wide variety of scenarios, enabling agents to experience diverse, first-person perspectives. This shift provides agents with richer, more varied training experiences, improving their adaptability and generalization in complex, unpredictable real-world situations. A representative work is UniSim [236], which dynamically generates robot manipulation video sequences based on input conditions like spatial movements, textual commands, and camera parameters. Leveraging multimodal data from 3D simulations, real-world robot actions, and internet media, this system generates varied, realistic environments where agents can practice tasks like object manipulation and navigation. The key advantage of this approach is its flexibility, allowing agents to adapt to various scenarios without the limitations of static physical environments. Pandora [223] expands the dynamic environment generation from robot actions in Unisim to wider domains including human and robot actions in both indoor and outdoor scenes. Another subsequent work, AVID [168] builds on UniSim by conditioning on actions and modifying noise predictions from a pre-trained diffusion model to generate action-driven visual sequences for dynamic environment generation. Beyond the video diffusion-based framework of Unisim, EVA [23] introduces an additional vision-language model for embodied video anticipation, producing more consistent embodied video predictions. As for the generation of open-world dynamic environments, Streetscapes [32] employs autoregressive video diffusion models to simulate urban environments where agents must navigate dynamic challenges like changing weather and traffic. These environments offer consistently coherent, yet flexible, urban settings, exposing agents to real-world-like variability. The core trend in dynamic environments is the use of generative world models that provide scalable, adaptable simulations. This approach significantly reduces the manual effort required for environment setup, allowing agents to train across a diverse range of scenarios quickly. Moreover, the focus on first-person training closely mimics real-world decision-making, enhancing the agents' ability to adapt to evolving situations. These advances are key in developing embodied environments supporting agent learning in complex, dynamic scenarios.

Given the above developments, it is evident that world models as embodied environments have made significant advances in simulating and predicting how agents interact with dynamic, real-world scenarios. Current research predominantly focuses on developing indoor, static environments, with notable efforts expanding to large-scale outdoor environments and dynamic simulation environments. A promising direction is to construct dynamic environments, which can provide first-person, action-conditioned future world prediction, enabling agents to better adapt to unseen conditions. These methods are promising to offer flexible, scalable environments for training embodied agents, enhancing their generalization capabilities for real-world tasks.

## 5 APPLICATION

### 5.1 Autonomous Driving

In recent years, with the rapid advancement of vision-based generative models [14, 75, 193] and multimodal large language models [1, 122], world models have attracted growing interest in the field of autonomous driving. The modern autonomous driving pipeline is typically divided into four key components: *perception*, *prediction*, *planning*, and *control*. Among these, the perception and prediction stages correspond to driving scene understanding—i.e., learning an implicit representation of the vehicle's external environment. In parallel, recent surveys [62] highlight the emergence of end-to-end world simulators that learn to simulate realistic driving environments based on multimodal inputs—such as images, point clouds, trajectories, and language—and then

generate future states to support downstream tasks like planning and decision-making. These two perspectives align well with our earlier categorization of world models, and in the following, we detail their applications and developments within the autonomous driving domain accordingly.

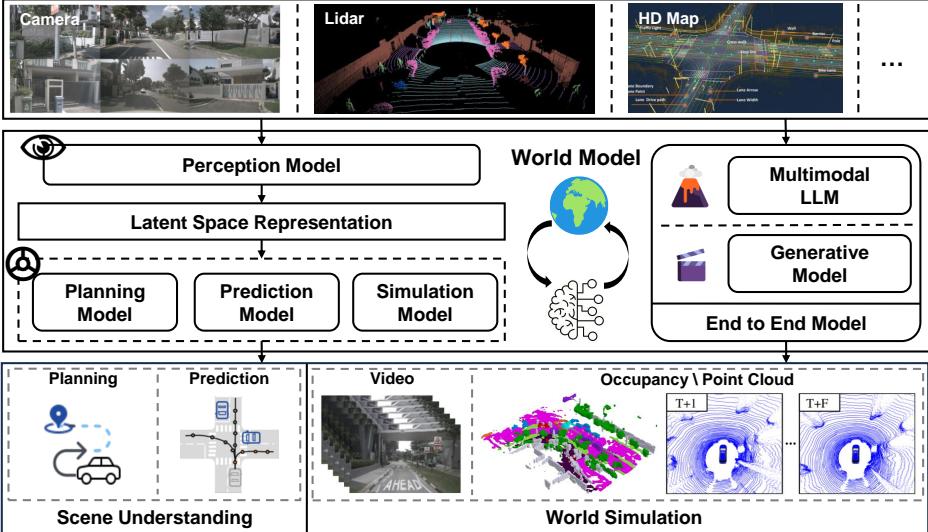


Fig. 5. Application of world model in autonomous driving.

**5.1.1 Learning Implicit Representations.** Autonomous vehicles typically utilize cameras, radar, and lidar to perceive the real world, gathering information through images, video data, and point cloud data. In the initial decision-making paradigm [20, 175], models often take perceptual data as input and directly output motion planning results for the autonomous vehicle. Conversely, when humans operate vehicles, they typically observe and predict the current and future states of other traffic participants to determine their own driving strategies [83]. Thus, learning the implicit representation of the world through perceptual data and predicting the future states of the surrounding environment is a crucial step in enhancing the decision-making reliability of autonomous vehicles. We consider this process as it manifests in how autonomous vehicles learn a world model in latent space.

As shown in the left half of Figure 5, before the advent of multimodal large models and end-to-end autonomous driving technologies [80], the perception and prediction tasks of autonomous vehicles were typically assigned to distinct modules, each trained on their respective tasks and datasets. The perception module processed data from images, point clouds, and other sources to accomplish tasks such as object detection and map segmentation, projecting the perceived world into an abstract geometric space. Furthermore, the prediction module would typically operate within these geometric spaces to forecast the future states of the surrounding environment, including the trajectories and motions of traffic participants.

The processing of perceptual data is closely tied to the evolution of deep learning technologies, as shown in Table 4. Pointnet [158], introduced in 2017, was the first to employ deep learning methods for processing point cloud data. As convolutional neural networks advanced, perception techniques based on image data, exemplified by YOLOP [216] and MultiNet [198], emerged and excelled in driving scene understanding tasks [74, 107, 203, 262]. In recent years, the transformer architecture has gained prominence in natural language processing, and this technology has also been applied to image data understanding. BEVFormer [118] utilizes the attention mechanism to

Table 4. Comparison of existing works in scene understanding and world simulation.

	Task	Work	Year	Data Modality	Technique	Task Description
Driving Scene Understanding	Perception	Faster r-cnn [165]	2015	Camera	CNN	Object Detection
		Pointnet [158]	2017	Lidar	MLP	3D Classification
		MultiNet [198]	2018	Camera	CNN	Semantic Segmentation
		OmniDet [107]	2021	Camera	CNN & Attention	Multi-task Visual Perception
		YOLOP [216]	2022	Camera	CNN	Object Detection
		BEVFormer [118]	2022	Camera	Attention	3D Visual Perception
		Transfusion [8]	2022	Camera & Lidar	Transformer	3D Object Detection
	Prediction	Wayformer [142]	2022	Geometric Space	Attention	Trajectory Prediction
		MTR [187]	2022	Geometric Space	Transformer	Trajectory Prediction
		QCNet [264]	2023	Geometric Space	Transformer	Trajectory Prediction
	End to End Scene Understanding	HPTR [253]	2023	Geometric Space	Transformer	Trajectory Prediction
		Jiang <i>et al.</i> [92]	2023	Geometric Space	Diffusion	Trajectory Prediction
Driving World Simulation	Motion Simulation	UniAD [80]	2023	Camera	Transformer	Motion Planning
		TOKEN [199]	2024	Camera	MLLM	Motion Planning
		OmniDrive [107]	2024	Camera	MLLM	Motion Planning
		SUMO [127]	2000	Geometric Space	Rule-based	Traffic Simulation
	End to End Sensor Simulation	Metadrive [116]	2022	Geometric Space	Data-driven	Traffic Simulation
		Trafficbots [252]	2023	Geometric Space	Transformer	Traffic Simulation
	End to End Sensor Simulation	Waymax [63]	2024	Geometric Space	Data-driven	Traffic Simulation
		GAIA-1 [78]	2023	Camera	Transformer	Video Generation
		DriveDreamer [210]	2023	Camera	Diffusion	Video Generation
		Drive-WM [213]	2023	Camera	Diffusion	Video Generation

integrate images from multiple camera angles, constructing an abstract geometric space from a bird's-eye view, and achieving state-of-the-art results in various tasks, including object detection. Additionally, Transfusion [8] enhances perceptual accuracy by fusing lidar and camera data through a cross-attention approach. Building on the perceptual results, a series of techniques such as RNNs [6, 98, 267], CNNs [25, 28, 152], and Transformers [84, 143, 187, 264] have been employed to encode historical scene information and predict the future behaviors of traffic participants.

With the emergence and rapid development of multimodal large language models in recent years, many efforts have sought to apply the general scene understanding capabilities of these models to the field of autonomous driving. TOKEN [199] tokenizes the whole traffic scene into object-level knowledge, using the reasoning ability of the language model to handle the long-tail prediction and planning problems. OmniDrive [107] sets up llm-based agents and covers multiple tasks including scene description, counterfactual reasoning and decision making through visual question-answering.

**5.1.2 World Simulators.** As shown in Table 4, before the emergence of multimodal large models and vision-based generative models, traffic scenario simulations are often conducted in geometric spaces. The scene data on which these simulations rely is typically collected by the perception modules of autonomous vehicles or constructed manually. These simulations represent future states of the scenario in the form of geometric trajectories [63, 116, 127, 252], which require further modeling and rendering to produce outputs suitable for vehicle perception. The cascading of multiple modules often results in information loss and increases the complexity of simulations, making scenario control more challenging. Furthermore, realistic scene rendering typically requires substantial computational resources, which limits the efficiency of virtual traffic scenario generation.

Using diffusion-based video generation models as a world model partially addresses the aforementioned issues. By training on large-scale traffic scenario datasets, diffusion models can directly generate camera perception data that closely resembles reality. Additionally, the inherent controllability of diffusion models, combined with text-image alignment methods like CLIP [161], enables

users to exert control over scenario generation in a straightforward manner. The GAIA-1 [78] and DriveDreamer series [210, 256] are among the first to employ this method for constructing world models. Building on this foundation, Drive-WM [213] introduces closed-loop control for planning tasks, while Vista [54] focuses on improving the resolution of generated results and extending prediction duration. In addition to methods that predict future states in video space, many other works have explored different forms of vehicle perception data. OccWorld [260] and OccSora [206] predict the future state of the world by forecasting 3D occupancy grids, whereas Copilot4D [248] constructs a world model by predicting changes in radar point cloud data. Compared to video data, these types of features better reflect the spatial characteristics of traffic scenarios.

## 5.2 Robots

World models have emerged as a transformative paradigm in robotics, empowering robots to perceive, predict, and act effectively in complex environments. This progress is driven in part by advances in neural architectures [75, 202] and learning algorithms [162, 178], which enable robots to build implicit representations that capture key aspects of the external world. Complementarily, prediction models [46, 47] offer the ability to forecast future environmental states, moving beyond static abstractions to support anticipatory and adaptive behavior. Together, these capabilities make it increasingly feasible for robots to learn directly from real-world interactions. In Table 5, we summarize the core learning tasks involved in constructing world models for robotics, categorized according to the three major perspectives outlined above (typical examples shown in Figure S1).

**5.2.1 Learning Implicit Representation.** Traditional robotic tasks (e.g., object grasping) are typically performed in highly structured environments where the critical components are explicitly modeled [38, 100], eliminating the need for the robot to independently learn or adapt its understanding of the world. However, when the robot is deployed in unfamiliar environments, especially those in which key features or dynamics have not been explicitly modeled, tasks that were previously successful may fail as the robot struggles to generalize to these unknown features [95, 137]. Thus, enabling a robot to learn an implicit representation of its environment is a crucial first step toward achieving intelligence.

To help a robot understand the objects in the world, visual models such as convolutional neural networks (CNNs) [57, 103, 110] and vision transformers (ViT) [34, 204] integrate visual characteristics of entities into representations, making it possible for robots to recognize critical objects for tasks. RoboCraft [184] transfers visual observation into particles and captures the structure of the underlying system through a graph neural network. Moreover, other attempts are made for the sensing of physical 3D space. PointNet [157, 159] first encodes the unstructured 3D point clouds with asymmetrical functions, capturing the spatial characteristics of the environment. A recent work [59] assembles observations acquired along local exploratory paths into a global representation of the physical space within its latent space, enabling robots to tail and approach specific targets. SpatialLM [197] further advances this direction by processing raw 3D point clouds into structured 3D scene representations with semantic labels, enhancing spatial reasoning for complex tasks in robotics and autonomous driving. With the advancement of language comprehension in LLMs [15, 35, 201], a novel paradigm for enabling robots to capture task intentions involves describing the task in textual form and then obtaining a textual representation through LLMs [56, 81, 140, 207]. BC-Z [88] utilizes language representations as task representations, enhancing the multi-task performance of robots. Text2Motion [120] splits the natural language instruction into task-level and motion-level plans with LLM to handle complex sequential manipulation tasks.

**5.2.2 Predicting Future States of the Environment.** Robotic tasks are inherently sequential and long-term, where early decisions significantly influence future outcomes [191]. Effective robotic planning

Table 5. Core learning tasks involved in constructing world models for robotics.

	Task	Model	Year	Input	Backbone
Learning Inner Representation	Visual Representation	CNN [110] ViT [34] RoboCraft [184]	1998 2020 2024	Image Image Image	CNN Transformer GNN
	3D Representation	PointNet [157] Predictive Coding [59] SpatialLM [197]	2017 2024 2025	3D point clouds Image 3D point clouds	MLP ResNet MLLM
	Task Representation	BC-Z [88] Text2Motion [120] Gensim [207]	2022 2023 2023	Text& Video Text Text	LLM& ResNet LLM LLM
Predicting Future Environment	Video Prediction	UniPi [36]	2024	Video	Diffusion
		VIPER [39]	2024	Video	Transformer
Real-world Planning	Real-World Adaptation	GR-2 [19]	2024	Text & Video	Transformer
		IRASim [265]	2024	Trajectory	Diffusion
	Evaluation	DayDreamer [220] SWIM [134] CoSTAR [2]	2023 2023 2021	Video Video Multimodal	RSSM Transfer Learning Belief Space
		OpenEQA [131]	2024	Image& Text	LLM

thus relies heavily on accurately predicting future environmental states, enabling proactive decision-making and reducing costly errors. Classic closed-loop algorithms [10, 99], which select actions solely based on current observations, are typically short-sighted, risking irreversible mistakes. Additionally, approaches depending on explicit dynamic models crafted from expert knowledge tend to be limited in flexibility and robustness.

A key recent insight is the use of generative video models—particularly those leveraging diffusion [11, 22, 40, 73] and transformer architectures [230, 243]—to implicitly learn environmental dynamics directly from visual data. For instance, UniPi [36] frames action prediction explicitly as a video generation problem, conditioning a constrained diffusion model on the current state to visualize future scenarios. Similarly, VIPER [39] employs a pretrained autoregressive transformer to guide robotic actions, effectively leveraging rich representations learned from expert demonstration videos. IRASim [265] leverages diffusion models for trajectory-to-video generation tasks, starting from an initial given frame. Moreover, models like GR-2 [19, 217] benefit from the vast scale of internet videos to establish robust priors, subsequently fine-tuning on specific robotics tasks to generate accurate image predictions and action trajectories.

Collectively, these methods demonstrate the promise of generative, vision-centric modeling as a foundation for anticipatory robotic control, significantly enhancing robots’ capabilities to reason (to some extent) about future states and improve long-term task performance.

**5.2.3 From Simulation to Real World.** Deep reinforcement learning has demonstrated remarkable capabilities in robotics, enabling autonomous performance in complex tasks such as stable locomotion [106, 192], precise object manipulation [33, 244], and intricate activities like tying shoelaces [5]. However, its practicality remains significantly limited by low sample efficiency. For instance, training a robot to solve a Rubik’s Cube in the real world can require tens of thousands of simulated years [3]. Consequently, most robot training is conducted within simulation environments, leveraging distributed training techniques to enhance efficiency [67, 170]. Unfortunately, due to discrepancies between simulations and real-world conditions, policies trained in simulation often fail when directly transferred to physical robots, particularly in complex or novel environments.

A pivotal recent insight is that world models can effectively bridge this simulation-to-reality gap by learning generalized representations of real-world dynamics. For example, NeBula [2] constructs a structured belief space that enables reasoning and rapid adaptation across diverse

Table 6. Representative works of social simulacra.

	Advance	What to simulate	Effects of world model
World Model as Social Simulacra	AI Town [148]	Machine society	Stylized facts
	S3 [51]	Social network	Predictions
	Papachristou <i>et al.</i> [147]	Social network	Stylized facts & Predictions
	Xu <i>et al.</i> [228]	Games	Stylized facts
	EconAgent [114]	Macroeconomics	Stylized facts
	SRAP-Agent [91]	Resource allocation	Stylized facts
	Project Sid [4]	Collective rules (tax)	Stylized facts
World Model in Social Siumlacra	AgentSociety [153]	Social life	Stylized facts
	Agent-Pro [250]	Games	Belief
	Zhang <i>et al.</i> [247]	Machine society	Reflection & Debate
	GovSim [154]	Resource sharing	Cognition
	AgentGroupChat [60]	Group chat	Belief & Memory

robot morphologies and unstructured environments. DayDreamer [220] further demonstrates the capability of generalized world models, allowing robots to directly learn locomotion in real-world environments within hours, significantly reducing the reliance on extensive simulations. Additionally, SWIM [134] highlights the power of human-video-based learning combined with minimal real-world fine-tuning, enabling task generalization with less than 30 minutes of interaction. These examples illustrate that by building robust, real-world-oriented internal representations, world models substantially narrow the gap between simulation and reality, facilitating rapid adaptation and generalization in robotics.

### 5.3 Social Simulacra

The concept of “social simulacra” was originally introduced as a prototyping technique in [149], aimed at helping designers create a virtual social computing system encompassing many diverse agents. The traditional methods of constructing agents based on expert-defined rules [13, 176] or reinforcement learning [259] face issues such as overly simplistic behaviors or a lack of interpretability. However, the emergence of LLMs provides a transformative tool for building more realistic social simulacra, achieving more convincing stylized facts [114] or accurate predictions. Social simulacra can be seen as a form of world model that mirrors real-world social computing systems. From another perspective, the agents within social simulacra also develop implicit representations of the external system; that is, they build an implicit world model that supports the generation of their social behaviors. The relationship between the world model and social simulacra is shown in Figure S2, and the summary of representative works is shown in Table 6.

**5.3.1 Building Social Simulacra Mirroring Real-world Society.** In the era of the rapid rise of LLM agents, building realistic social simulation systems becomes more practical. One of the most famous examples of social simulacra is AI Town [148], a world model composed of 25 generative agents, essentially forming a sandbox social environment. In this virtual community, the agents exhibit believable individual behaviors, and at the group level, emergent social behaviors similar to those that might appear in the real world. Along these lines, more and more attempts are being made to replace humans in various social scenarios with LLM agents, in effect forming their own scenario-specific social simulacra. These works have used the simulacra paradigm in such scenarios as social networks and cooperative or competitive games, among others [50].

Recent studies have extended social simulacra by demonstrating how LLM agents can reflect realistic social interactions and collective behaviors, effectively mirroring human societies. A critical insight is that LLM-driven agents, by modeling aspects such as emotions, attitudes, and decision-making patterns, can reproduce intricate social phenomena across multiple contexts. For example, S3 [51] leverages these human-like features to simulate realistic message propagation

on social networks, successfully capturing the dynamics of real-world public events. Related studies [147] deepen this understanding by analyzing how network structures emerge organically among LLM agents, comparing these structures directly to those formed in human social networks. Beyond networks, the capacity of LLM agents to embody sophisticated strategic reasoning is highlighted in simulations of social games such as Werewolf [228], where the agent outputs exhibit patterns resembling human strategic behaviors such as deception and confrontation. Furthermore, in economic contexts, works such as EconAgent [114] reveal that collective behaviors emerging from individual economic reasoning by LLM agents can convincingly reproduce macroeconomic trends and regularities. Recently, AgentSociety [153] has further expanded this paradigm by creating a large-scale, LLM-powered societal simulation capable of modeling diverse social phenomena such as polarization and response to public policies, thus providing a versatile platform for computational social science research. These studies collectively illustrate the broad potential of social simulacra, demonstrating that agent-based world models powered by LLMs can capture diverse and realistic social, strategic, and economic interactions.

**5.3.2 Agent’s Understanding of External World in Social Simulacra.** LLM agents build their memory by storing observations obtained through interactions with the external environment [251], thereby forming implicit representations and basic cognition of the external world, especially in the context of simulating social scenarios. This cognition is stored in a memory bank in textual form for LLM agents to retrieve and use, enabling them to access useful information and fully leverage experiential knowledge from past interactions with the environment when making decisions.

Agent-Pro [250] transforms the memory of its interactions with the external environment (specifically with other agents in interactive tasks) into what are called ‘beliefs’. Based on these beliefs, it makes the next decision and updates its behavior strategy. These beliefs represent the agent’s social understanding of the environment and other agents within it, relating to Theory of Mind mentioned in Section 3.2. Other works on LLM agents have also adopted similar designs. For example, Zhang *et al.* [247] introduces mechanisms of reflection and debate from a social psychology view for modeling multi-agent collaboration tasks. A more advanced study is GovSim [154], which explores whether cooperative behaviors aimed at sustainable resource development can emerge within a society composed of LLM agents. In this setup, each agent gathers information about the external world and other agents’ behavioral strategies through multi-agent conversations and subsequently forms its own high-level insights, essentially creating an implicit representation of the world model. Another similar application scenario is Interactive Group Chat [60], where human-like behaviors and strategies emerge across four narrative scenarios, including Inheritance Disputes, Law Court Debates, etc.

## 6 OPEN PROBLEMS AND FUTURE DIRECTIONS

The recent advance of hyper-realistic generative AI has brought a lot of attention to development of the world model, with particular focus on the multi-modal big models like Sora [146]. Despite the rapid innovation, there are also a lot of important open problems that remain to be solved.

### 6.1 Physical Rules and Counterfactual Simulation

A key objective of world models is to capture the causal structure of their environments—especially the underlying physical rules—so they can reason about counterfactuals beyond the data distribution [151]. This capacity is crucial for handling rare, mission-critical events (e.g., autonomous-driving corner cases [45]) and for narrowing sim-to-real gaps. Recent progress raises the question of whether large-scale, purely data-driven generative models can acquire such rules from raw visual data alone. While transformer- and diffusion-based video generators such as Sora [146] produce

strikingly realistic sequences, studies reveal persistent physical-law failures—e.g., inaccurate gravity, fluid, or thermal dynamics [212].

Hybrid approaches that explicitly embed physics are emerging as promising alternatives. Genesis[7] illustrates this direction by unifying fast, photo-realistic rendering with a re-engineered universal physics core, allowing language-conditioned data generation grounded in first-principles simulation. PhysGen[124] takes a similar stance at the image-to-video level: it couples a rigid-body simulator with a diffusion refiner, enabling controllable, physically plausible motion from a single image. Complementary diagnostic work underscores why such hybrids are needed. Kang et al.[97] show that scaling diffusion video models yields perfect in-distribution fidelity yet breaks down on out-of-distribution or combinatorial tests, indicating “case-based” rather than rule-based generalization. Motamed et al.[139] reach a similar conclusion with the Physics-IQ benchmark: current video generators achieve visual realism but largely fail on tasks requiring understanding of optics, fluid dynamics, or magnetism.

Taken together, the evidence suggests that data-driven scaling alone is insufficient to recover robust physical laws. Integrating explicit simulators—or otherwise enforcing physical priors [186]—remains a promising path toward world models that generalize to unseen counterfactual scenarios while retaining interpretability and transparency.

## 6.2 Enriching the Social Dimension

Simulating the physical elements alone is not sufficient for an advanced world model, since human behavior and social interaction also play a crucial role in many important scenarios [50, 58, 246]. For example, the behavior of urban dwellers is particularly important for building world models of the urban environment [9, 226]. Previous work shows that the human-like commonsense reasoning capabilities of LLMs provide a unique opportunity to simulate realistic human behavior with generative agents [148]. However, designing autonomous agents that can simulate realistic and comprehensive human behavior and social interactions remains an open problem. Recent studies suggest that theories of human behavior patterns and cognitive processes can inform the design of agentic workflows, which in turn enhance the human behavior simulation capabilities of LLMs [148, 182], representing an important direction for future research. In addition, the evaluation of the realism of generated human behavior still largely relies on subjective human assessment, which is challenging to scale up to a large-scale world model. Developing a reliable and scalable evaluation scheme will be another future research direction that can enrich the social dimension of the world model.

## 6.3 Bridging Simulation and Reality with Embodied Intelligence

The world model has long been envisioned as a critical step towards developing embodied intelligence [174]. It can serve as a powerful simulator that creates comprehensive elements of the environment and models realistic relationships between them. Such an environment can facilitate embodied agents to learn through interaction with a simulated environment, reducing the need for supervision data. To achieve this goal, improving the multi-modal, multi-task, and 3D capacities of generative AI models has become an important research problem for developing general world models for embodied agents. Moreover, closing the simulation-to-reality gaps [76] has been a long-standing research problem for embodied environment simulators, and it is therefore important to transfer the trained embodied intelligence from the simulation environment to the physical world. Collecting more fine-grained sensory data is also a critical step toward this goal, which can be facilitated through the interface of embodied agents. Therefore, an interesting future research direction is to create self-reinforcing loops to harness the synergy power of generative world models and embodied agents.

## 6.4 Simulation Efficiency

Ensuring high simulation efficiency of world models is important for many applications. For example, number of frames per second is a key metric for high quality for learning sophisticated drone manipulating AIs. The popular transformer architecture of most big generative AIs poses a huge challenge for high-speed simulation because its autoregressive nature can only generate one token at a time. Several strategies are proposed to accelerate the inference of large generative models, such as incorporating big and small generative models [181] and distilling big models [182]. More holistic solutions include building a simulation platform that optimally schedule LLM requests [232]. High computation cost is also a problem for classic physics simulators when they are tasked to simulate large and complex systems. Previous research finds deep learning models like graph neural networks can be used to efficiently approximate physical systems [172]. Therefore, an important research direction will be to explore the synergy between smaller deep learning models and big generative AI models. Additionally, the overall improvement from underlying hardware to programming platform and AI models is also needed to achieve substantial speedup.

## 6.5 Ethical and Safety Concerns

**Data Privacy.** The recent trend of building world models with big generative AIs raises significant concerns of privacy risk, largely due to the massive and often opaque training data [240]. Extensive research effort is devoted to assessing the risk of inferring private information with big generative AIs like LLM [115], which could be especially sensitive in the context of video generation models. To be compliant with privacy regulations like GDPR [196], it is important to improve the transparency of the life cycle of generative AIs, helping the public understand how data is collected, stored, and used in these AI models.

**Simulating Unsafe Scenario.** The incredibly intelligent power of generative AIs makes safeguarding their access a paramount task. Previous studies on LLMs found they can be misled to generate unsafe content with adversarial prompting [85, 105]. The risk of unsafe use of world models can be even larger. Adversarial users might leverage such techniques to simulate harmful scenarios, reducing the cost of planning illegal and unethical activities. Therefore, an important future research direction is to safeguard the usage of world models.

**Accountability.** The ability to generate hyper-realistic text, images, and videos has caused severe social problems of spreading misinformation and disinformation. For example, the emergence of deepfake technology gives rise to large-scale misuses that have widespread negative effects on social, economic, and political systems [214]. Thus, detecting AI-generated content has been a key research problem in addressing these risks [164]. However, this problem is becoming increasingly challenging due to the advance of generative AIs, and it will be even more difficult with the arrival of a world model that can generate consistent, multi-dimensional output. Technology like watermarking could help improve the accountability of world model usage [30]. More research attention, as well as legal solutions, are needed to improve the accountability of world model usage.

## 6.6 Benchmark

Benchmarking world models is both necessary and challenging. Because the community pursues divergent goals—*learning internal representations vs. predicting future worlds*—with heterogeneous technical approaches (e.g., LLM agents, video diffusion) and wide-ranging application domains (autonomous driving, robotics, social simulation), there is no single canonical task or metric. Nevertheless, several recent efforts illustrate how carefully designed testbeds can expose the specific gaps that prevent current models from becoming reliable world simulators.

**Video-centric world simulation.** Qin *et al.* propose WorldSimBench[160], a dual evaluation suite combining human-preference video judgments and action-level consistency across three embodied settings (open-ended sandboxes, autonomous driving, and robot manipulation). Complementarily, Duan *et al.* present WorldScore[37], which decomposes “world generation” into controllability, visual quality, and dynamics across 3,000 camera-specified scenes, enabling head-to-head comparisons of 3D, 4D, and video generators.

**Physical and spatial reasoning.** PhysBench [26] (10k video–image–text triplets) and UrbanVideoBench [254] (5.2k drone clips) demonstrate that current VLMs lack basic physics understanding and urban navigation skills. Xu *et al.* translate psychometric testing into AI evaluation by defining five Basic Spatial Abilities and diagnosing geometry and rotation weaknesses across 13 VLMs [227].

**Embodied decision making.** The Embodied Agent Interface (EAI) [113] standardizes four LLM-based modules—goal parsing, sub-goal generation, action sequencing, and transition modeling—and reports fine-grained errors rather than a single success rate.

Despite these advances, benchmarking world models remains an open challenge. Future work should focus on building more diverse and realistic benchmarks to rigorously test generalization capabilities. Additionally, standardizing evaluation protocols will be key to improving comparability and robustness assessments across environments.

## 7 CONCLUSION

Understanding the world and predicting the future have been long-standing objectives for scientists developing artificial generative intelligence, underscoring the significance of constructing world models across various domains. This paper presents the first comprehensive survey of world models that systematically explores their two primary functionalities: *implicit representations* and *future predictions* of the external world. We provide an extensive summary of existing research on these core functions, with particular emphasis on world models in decision-making, world knowledge learned by models, world models as video generation, and world models as embodied environments. Additionally, we review progress in key applications of world models, including autonomous driving, robotics, and social simulation. Finally, recognizing the unresolved challenges in this rapidly evolving field, we highlight open problems and propose promising research directions with the hope of stimulating further investigation in this burgeoning area.

## REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Ali Agha, Kyohei Otsu, Benjamin Morrell, David D Fan, Rohan Thakker, Angel Santamaría-Navarro, Sung-Kyun Kim, Amanda Bouman, Xianmei Lei, Jeffrey Edlund, et al. Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge. *arXiv preprint arXiv:2103.11470*, 2021.
- [3] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik’s cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.
- [4] Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, et al. Project sid: Many-agent simulations toward ai civilization. *arXiv preprint arXiv:2411.00114*, 2024.
- [5] Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sankh Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.
- [6] Florent Altché and Arnaud de La Fortelle. An lstm network for highway trajectory prediction. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pages 353–359. IEEE, 2017.
- [7] Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond, December 2024.

- [8] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.
- [9] Michael Batty. Digital twins in city planning. *Nature Computational Science*, 4(3):192–199, 2024.
- [10] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.
- [11] Kevin Black, Mitsuhiro Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.
- [12] Daniel Bogdoli, Yitian Yang, and J Marius Zöllner. Muvo: A multimodal generative world model for autonomous driving with geometric representations. *arXiv e-prints*, pages arXiv–2311, 2023.
- [13] William A Brock and Cars H Hommes. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic dynamics and Control*, 22(8-9):1235–1274, 1998.
- [14] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.
- [15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [16] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.
- [17] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2139–2150, 2023.
- [18] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.
- [19] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.
- [20] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving, 2019.
- [21] Zhili Cheng, Zhitong Wang, Jinyi Hu, Shengding Hu, An Liu, Yuge Tu, Pengkai Li, Lei Shi, Zhiyuan Liu, and Maosong Sun. Legent: Open platform for embodied agents. *arXiv preprint arXiv:2404.18243*, 2024.
- [22] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [23] Xiaowei Chi, Hengyuan Zhang, Chun-Kai Fan, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-min Chan, Wei Xue, Wenhai Luo, Shanghang Zhang, et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461*, 2024.
- [24] Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.
- [25] Fang-Chieh Chou, Tsung-Han Lin, Henggang Cui, Vladan Radosavljevic, Thi Nguyen, Tzu-Kuo Huang, Matthew Niedoba, Jeff Schneider, and Nemanja Djuric. Predicting motion of vulnerable road users using high-definition maps and efficient convnets. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1655–1662. IEEE, 2020.
- [26] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.
- [27] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.
- [28] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 international conference on robotics and automation (icra)*, pages 2090–2096. IEEE, 2019.
- [29] Yifan Cui, Xinyi Shan, and Jeanhun Chung. A feasibility study on runway gen-2 for generating realistic style images. *International Journal of Internet, Broadcasting and Communication*, 16(1):99–105, 2024.
- [30] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

- [31] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.
- [32] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [33] Mehmet Dogar, Andrew Spielberg, Stuart Baker, and Daniela Rus. Multi-robot grasp planning for sequential assembly operations. *Autonomous Robots*, 43:649–664, 2019.
- [34] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [35] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.
- [36] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [37] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.
- [38] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.
- [39] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [40] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.
- [41] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.
- [42] Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*, 2024.
- [43] Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv preprint arXiv:2504.09848*, 2025.
- [44] Jie Feng, Jun Zhang, Junbo Yan, Xin Zhang, Tianjian Ouyang, Tianhui Liu, Yuwei Du, Siqi Guo, and Yong Li. Citybench: Evaluating the capabilities of large language model as world model. *arXiv preprint arXiv:2406.13945*, 2024.
- [45] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023.
- [46] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.
- [47] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.
- [48] Rao Fu, Zehao Wen, Zichen Liu, and Srinath Sridhar. Anyhome: Open-vocabulary generation of structured and textured 3d homes. In *European Conference on Computer Vision*, pages 52–70. Springer, 2025.
- [49] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.
- [50] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.
- [51] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.
- [52] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, et al. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024.

- [53] Qiaoz Gao, Govind Thattai, Suhaila Shakiah, Xiaofeng Gao, Shreyas Pansare, Vasu Sharma, Gaurav Sukhatme, Hangjie Shi, Bofei Yang, Desheng Zhang, et al. Alexa arena: A user-centric interactive platform for embodied ai. *Advances in Neural Information Processing Systems*, 36, 2024.
- [54] Shenyuan Gao, Jiazh Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.
- [55] Ignat Georgiev, Varun Giridhar, Nicklas Hansen, and Animesh Garg. Pwm: Policy learning with multi-task world models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [56] Elliot Gestrin, Marco Kuhlmann, and Jendrik Seipp. Nl2plan: Robust llm-driven planning from minimal text descriptions. *arXiv preprint arXiv:2405.04215*, 2024.
- [57] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [58] Jiahui Gong, Jingtao Ding, Fanjin Meng, Chen Yang, Hong Chen, Zuojian Wang, Haisheng Lu, and Yong Li. Behavegpt: A foundation model for large-scale user behavior modeling. *arXiv preprint arXiv:2505.17631*, 2025.
- [59] James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6(7):820–833, 2024.
- [60] Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, et al. Agent group chat: An interactive group chat simulacra for better eliciting collective emergent behavior. *arXiv preprint arXiv:2403.13433*, 2024.
- [61] Lin Guan, Karthik Valmecikan, Sarah Sreedharan, and Subbarao Kambhampati. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.
- [62] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.
- [63] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36, 2024.
- [64] Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.
- [65] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.
- [66] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- [67] Sehoon Ha, Peng Xu, Zhenyu Tan, Sergey Levine, and Jie Tan. Learning to walk in the real world with minimal human effort. *arXiv preprint arXiv:2002.08550*, 2020.
- [68] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.
- [69] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.
- [70] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.
- [71] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.
- [72] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024.
- [73] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. Learning an actionable discrete diffusion policy via large-scale actionless video pre-training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [74] Juncai He and Jinchao Xu. Mgnet: A unified framework of multigrid and convolutional neural network. *Science China Mathematics*, 62(7):1331–1354, May 2019.
- [75] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [76] Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Melissa Mozifian, Florian Golemo, Chris Atkeson, Dieter Fox, Ken Goldberg, John Leonard, et al. Sim2real in robotics and automation: Applications and challenges. *IEEE transactions on automation science and engineering*, 18(2):398–400, 2021.

- [77] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [78] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023.
- [79] Yi-Qi Hu, Hong Qian, and Yang Yu. Sequential classification-based optimization for direct policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [80] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Sici Chai, Senyao Du, Tianwei Lin, Wenhui Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving, 2023.
- [81] Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. *arXiv preprint arXiv:2410.03645*, 2024.
- [82] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.
- [83] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022.
- [84] Zhiyu Huang, Xiaoyu Mo, and Chen Lv. Multi-modal motion prediction with transformer-based neural network for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2605–2611. IEEE, 2022.
- [85] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.
- [86] Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*, 2024.
- [87] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024.
- [88] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.
- [89] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.
- [90] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.
- [91] Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Weiran Shen, Qi Qi, and Yankai Lin. Srap-agent: Simulating and optimizing scarce resource allocation policy with llm-based agent. *arXiv preprint arXiv:2410.14152*, 2024.
- [92] Chiyu Max Jiang, Andre Cormann, Cheolho Park, Ben Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion, 2023.
- [93] Charles Jin and Martin Rinard. Emergent representations of program semantics in language models trained on programs. In *Forty-first International Conference on Machine Learning*, 2024.
- [94] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.
- [95] Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv preprint arXiv:1702.01182*, 2017.
- [96] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldty, and Anil B Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.
- [97] Bingyi Kang, Yang Yue, Rui Lu, Zhipie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.
- [98] Atsushi Kawasaki and Akihito Seki. Multimodal trajectory predictions for urban environments using geometric relationships between a vehicle and lanes. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9203–9209. IEEE, 2020.
- [99] Oussama Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal on Robotics and Automation*, 3(1):43–53, 1987.
- [100] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1:239–249, 2020.

- [101] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.
- [102] Basil Kouvaritakis and Mark Cannon. Model predictive control. *Switzerland: Springer International Publishing*, 38:13–56, 2016.
- [103] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [104] Kuaishou. Kling ai: Next-generation ai creative studio. <https://www.klingai.com/global/>, 2024. (Accessed on 06/05/2025).
- [105] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.
- [106] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.
- [107] Varun Ravi Kumar, Senthil Yogamani, Hazem Rashed, Ganesh Sistu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving, 2023.
- [108] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*, 2018.
- [109] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.
- [110] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [111] Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024.
- [112] Liancan Li, Wei Shao, Wei Dong, Yijun Tian, Qiming Zhang, Kaixiang Yang, and Wenjie Zhang. Data-centric evolution in autonomous driving: A comprehensive survey of big data system, data mining, and closed-loop technologies. *arXiv preprint arXiv:2401.12888*, 2024.
- [113] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024.
- [114] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, 2024.
- [115] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. Llm-pbe: Assessing data privacy in large language models. *arXiv preprint arXiv:2408.12787*, 2024.
- [116] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022.
- [117] Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure, 2024.
- [118] Zhiqi Li, Wenhui Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022.
- [119] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language, 2024.
- [120] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365, 2023.
- [121] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.
- [122] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [123] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [124] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.

- [125] Zhihan Liu, Hao Hu, Shenao Zhang, Hongyi Guo, Shuqi Ke, Boyi Liu, and Zhaoran Wang. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency, 2024.
- [126] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE, 2024.
- [127] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.
- [128] Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. *Science China Information Sciences*, 67(2):121101, 2024.
- [129] Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. *arXiv preprint arXiv:1807.03858*, 2018.
- [130] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. From efficient multimodal models to world models: A survey. *arXiv preprint arXiv:2407.00118*, 2024.
- [131] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Venamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024.
- [132] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.
- [133] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023.
- [134] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.
- [135] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Uniworld: Autonomous driving pre-training via world models. *arXiv preprint arXiv:2308.07234*, 2023.
- [136] Marvin Minsky. A framework for representing knowledge, 1974.
- [137] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [138] Thomas M Moerland, Joost Broekens, Aske Plaat, and Catholijn M Jonker. A0c: Alpha zero in continuous action space. *arXiv preprint arXiv:1805.09613*, 2018.
- [139] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.
- [140] Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, Chenxue Wang, Shichao Liu, and Qing Wang. Clarifygpt: Empowering llm-based code generation with intention clarification. *arXiv preprint arXiv:2310.10996*, 2023.
- [141] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7559–7566. IEEE, 2018.
- [142] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks, 2022.
- [143] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified multi-task model for behavior prediction and planning. *arXiv preprint arXiv:2106.08417*, 2(7), 2021.
- [144] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. *Advances in neural information processing systems*, 30, 2017.
- [145] OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022. (Accessed on 06/05/2025).
- [146] OpenAI. Sora: Creating video from text. <https://openai.com/sora>, 2024. (Accessed on 06/05/2025).
- [147] Marios Papachristou and Yuan Yuan. Network formation and dynamics among multi-lmms. *arXiv preprint arXiv:2402.10659*, 2024.
- [148] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.
- [149] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

- [150] Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *Advances in Neural Information Processing Systems*, 35:6236–6249, 2022.
- [151] Judea Pearl. Causal inference in statistics: An overview. 2009.
- [152] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14074–14083, 2020.
- [153] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.
- [154] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainability behaviors in a society of llm agents. *arXiv preprint arXiv:2404.16698*, 2024.
- [155] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.
- [156] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.
- [157] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [158] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017.
- [159] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 35:23192–23204, 2022.
- [160] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.
- [161] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [162] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [163] Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. In *International conference on machine learning*, pages 7953–7963. PMLR, 2020.
- [164] Md Shohel Rana, Mohammad Nur Nobi, Bedduhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.
- [165] Shaogang Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [166] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhui Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.
- [167] Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents need world models. *arXiv preprint arXiv:2506.01622*, 2025.
- [168] Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.
- [169] Jonathan Roberts, Timo Lüdecke, Sowmen Das, Kai Han, and Samuel Albanie. Gpt4geo: How a language model sees the world’s geography. *arXiv preprint arXiv:2306.00020*, 2023.
- [170] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*, pages 91–100. PMLR, 2022.
- [171] Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. *arXiv preprint arXiv:2403.04253*, 2024.
- [172] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.
- [173] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

- [174] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.
- [175] Dhruv Mauria Saxena, Sangjae Bae, Alireza Nakhaei, Kikuo Fujimura, and Maxim Likhachev. Driving in dense traffic with model-free reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.
- [176] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.
- [177] Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtle, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, and Nicolas Heess. A generalist dynamics model for control. *arXiv preprint arXiv:2305.10912*, 2023.
- [178] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [179] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.
- [180] Yu Shang, Jiansheng Chen, Hangyu Fan, Jingtiao Ding, Jie Feng, and Yong Li. Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965*, 2024.
- [181] Yu Shang, Yu Li, Fengli Xu, and Yong Li. Defint: A default-interventionist framework for efficient reasoning with hybrid large language models. *arXiv preprint arXiv:2402.02563*, 2024.
- [182] Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtiao Ding, Yuan Yuan, Meng Wang, and Yong Li. Beyond imitation: Generating human mobility from context-aware reasoning with large language models. *arXiv preprint arXiv:2402.09836*, 2024.
- [183] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021.
- [184] Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks. *The International Journal of Robotics Research*, 43(4):533–549, 2024.
- [185] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Layla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. *arXiv preprint arXiv:2408.12574*, 2024.
- [186] Hongzhi Shi, Jingtiao Ding, Yufan Cao, Li Liu, Yong Li, et al. Learning symbolic models for graph-structured physical mechanism. In *The Eleventh International Conference on Learning Representations*, 2022.
- [187] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.
- [188] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- [189] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [190] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- [191] Reid G Simmons. Structured control for autonomous robots. *IEEE transactions on robotics and automation*, 10(1):34–43, 1994.
- [192] Laura Smith, Ilya Kostrikov, and Sergey Levine. A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. *arXiv preprint arXiv:2208.07860*, 2022.
- [193] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [194] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.
- [195] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*, 2024.
- [196] Damian A Tamburri. Design principles for the general data protection regulation (gdpr): A formal concept analysis and its evaluation. *Information Systems*, 91:101469, 2020.

- [197] ManyCore Research Team. Spatiallm: Large language model for spatial understanding. <https://github.com/manycore-research/SpatialLM>, 2025.
- [198] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving, 2018.
- [199] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving, 2024.
- [200] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.
- [201] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [202] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [203] Dat Vu, Bao Ngo, and Hung Phan. Hybirdnets: End-to-end perception network, 2022.
- [204] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024.
- [205] Hanqing Wang, Jiahui Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng Zhao, Yilun Chen, Sizhe Yang, et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024.
- [206] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024.
- [207] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. *arXiv preprint arXiv:2310.01361*, 2023.
- [208] Tingwu Wang and Jimmy Ba. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*, 2019.
- [209] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.
- [210] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving, 2023.
- [211] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.
- [212] Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? *arXiv preprint arXiv:2310.07018*, 2023.
- [213] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving, 2023.
- [214] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.
- [215] Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities, 2023.
- [216] Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6):550–562, November 2022.
- [217] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.
- [218] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *arXiv preprint arXiv:2405.15223*, 2024.
- [219] Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*, 2024.
- [220] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.
- [221] Wayne Wu, Honglin He, Yiran Wang, Chenda Duan, Jack He, Zhizheng Liu, Quanyi Li, and Bolei Zhou. Metaurban: A simulation platform for embodied ai in urban spaces. *arXiv preprint arXiv:2407.08725*, 2024.
- [222] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.
- [223] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.
- [224] 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. *Advances in neural information processing systems*,

- 36, 2024.
- [225] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.
  - [226] Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*, 2023.
  - [227] Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics. *arXiv preprint arXiv:2502.11859*, 2025.
  - [228] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*, 2023.
  - [229] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *International Conference on Machine Learning*, pages 39062–39098. PMLR, 2023.
  - [230] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.
  - [231] Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.
  - [232] Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. Opencity: A scalable platform to simulate urban activities with massive llm agents. *arXiv preprint arXiv:2410.21286*, 2024.
  - [233] Deshun Yang, Luhui Hu, Yu Tian, Zihao Li, Chris Kelly, Bang Yang, Cindy Yang, and Yuexian Zou. Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*, 2024.
  - [234] Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint arXiv:2306.01872*, 2023.
  - [235] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.
  - [236] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024.
  - [237] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.
  - [238] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16227–16237, 2024.
  - [239] Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K Marks, Jonathan Le Roux, and Chuang Gan. Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16251–16261, 2024.
  - [240] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.
  - [241] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.
  - [242] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023.
  - [243] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023.
  - [244] Sheng Yu, Di-Hua Zhai, Yuanqing Xia, Haoran Wu, and Jun Liao. Se-resunet: A novel robotic grasp detection method. *IEEE Robotics and Automation Letters*, 7(2):5238–5245, 2022.
  - [245] Yang Yu, Hong Qian, and Yi-Qi Hu. Derivative-free optimization via classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
  - [246] Yuan Yuan, Jingtao Ding, Depeng Jin, and Yong Li. Learning the complexity of urban mobility with deep generative network. *PNAS nexus*, 4(5):pgaf081, 2025.
  - [247] Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.
  - [248] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion, 2024.

- [249] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2025.
- [250] Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yuetong Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*, 2024.
- [251] Zeyu Zhang, Xiaohu Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.
- [252] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. Trafficbots: Towards world models for autonomous driving simulation and motion prediction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1522–1529. IEEE, 2023.
- [253] Zhejun Zhang, Alexander Liniger, Christos Sakaridis, Fisher Yu, and Luc Van Gool. Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding, 2023.
- [254] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, et al. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. *arXiv preprint arXiv:2503.06157*, 2025.
- [255] Ganlong Zhao, Guanbin Li, Weikai Chen, and Yizhou Yu. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16296–16306, 2024.
- [256] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation, 2024.
- [257] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [258] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.
- [259] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances*, 8(18):eabk2607, 2022.
- [260] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023.
- [261] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025.
- [262] Hongyu Zhou, Zheng Ge, Zeming Li, and Xiangyu Zhang. Matrixvt: Efficient multi-camera to bev transformation for 3d perception, 2022.
- [263] Siyuan Zhou, Yilun Du, Jiaiben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.
- [264] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023.
- [265] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.
- [266] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.
- [267] Alex Zyner, Stewart Worrall, and Eduardo Nebot. Naturalistic driver intention and path prediction using recurrent neural networks. *IEEE transactions on intelligent transportation systems*, 21(4):1584–1594, 2019.

## A RELATED SURVEY

Table S1. Comparison with existing surveys. This paper focuses on a comprehensive overview of the systematic definition and the capabilities of world models.

Survey	Venue and Year	Main Focus	Deficiency
[266]	Arxiv, 2024	General world model	Limited to discussion on applications
[130]	Arxiv, 2024	Efficient multimodal models	Limited to discussion on techniques
[24]	Arxiv, 2024	Text-to-video generation	Limited scope
[62]	IEEE T-IV, 2024	Autonomous driving	Limited scope
[112]	Arxiv, 2024	Autonomous driving	Limited scope
[231]	Arxiv, 2024	Autonomous driving	Limited scope

## B FIGURES AND TABLES

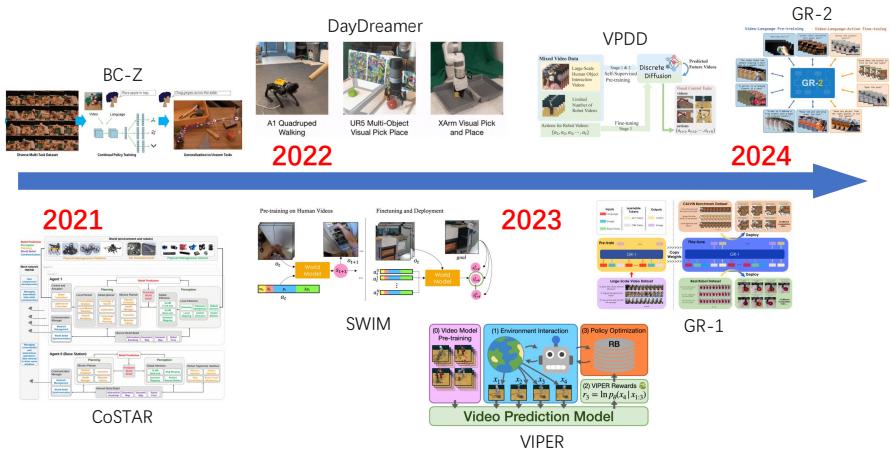


Fig. S1. The development of the robotic world model.

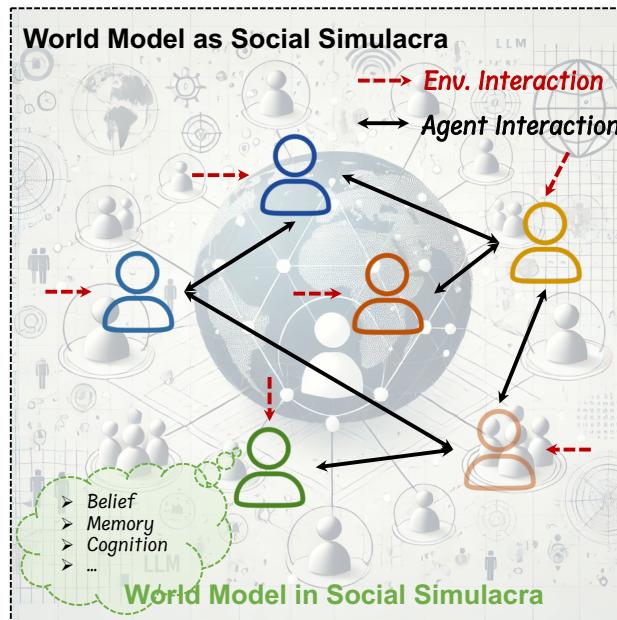


Fig. S2. World model and social simulacra.