A Survey of Embodied World Models

Yu Shang[1] **Yinzhou Tang**[1*] **Xin Zhang**[1*] **Shengyuan Wang**[1*]
Yuwei Yan[2] **Honglin Zhang**[1] **Zhiheng Zheng**[1] **Jie Zhao**[1]
Jie Feng[1] **Chen Gao**[1] **Jiansheng Chen**[3] **Yong Li**[1†]
[1]Tsinghua University
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]University of Science and Technology Beijing
shangy21@mails.tsinghua.edu.cn, liyong07@tsinghua.edu.cn

Abstract

World models have emerged as a pivotal research direction, with recent breakthroughs in generative AI underscoring their potential for advancing artificial general intelligence. For embodied AI, world models are critical for enabling robots to effectively understand, interact with, and make informed decisions in real-world physical environments. This survey systematically reviews recent progress in embodied world models, under a novel technical taxonomy. We hierarchically organize the field by model architectures, training methodologies, application scenarios, and evaluation approaches, thus offering researchers a clear technical roadmap. We first thoroughly discuss vision-based generative world models and latent space world models, along with their corresponding training paradigms. We then explore the multifaceted roles of embodied world models in robotic applications, from functioning as cloud-based simulation environments to on-device agent brains. Additionally, we summarize important evaluation dimensions for benchmarking embodied world models. Finally, we outline key challenges and provide insights into promising future research directions within this crucial domain. We summarize the representative works discussed in this survey at `https://github.com/tsinghua-fib-lab/Awesome-Embodied-World-Model`.

1 Introduction

World models, which involve constructing representations of world states and modeling state transitions, have emerged as a cutting-edge research topic in recent years. The concept of world models can be traced to psychological studies on hypothetical thinking [22], where the mind internally simulates future world states. In 2018, Ha et al. [36] introduced an early AI-era realization of a world model for reinforcement learning, ingeniously integrating spatial compression of high-dimensional perceptual inputs with temporal dynamics modeling. The recent blooming of generative AI has significantly advanced world modeling research, with interactive visual generative models [1, 11, 98, 99] such as Genie [11] and HunyuanWorld [98] demonstrating exceptional capabilities in both visual generation and future state prediction, establishing them as powerful world model implementations. Parallel to these visually explicit approaches, another paradigm represented by JEPA [3] advocates for latent-space world state representations to enable more efficient action planning. While both directions are rapidly evolving, their comparative merits remain an open question for further exploration. To structure the field, we classify existing embodied world models based on their generation modalities, control signals, and generation views, which is depicted in Figure 1.

[*]These authors contributed equally.
[†]Corresponding Author.

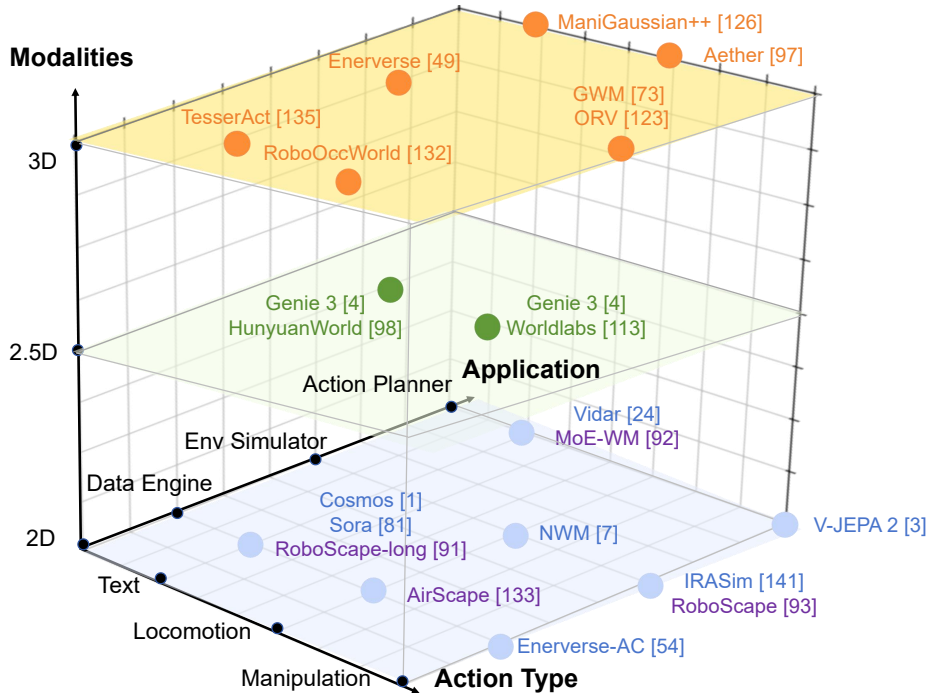


Figure 1: Classification of existing embodied world models according to the generation modalities, action types and applications.

Current representative world models are largely confined to virtual game environments [11, 35, 131, 113] and autonomous driving [31, 108, 111]. World modeling for embodied scenarios is an emerging field, and constructing powerful physical world models represents a critical technological path toward enabling embodied agents to achieve Artificial General Intelligence (AGI). Such embodied world models are essential for robots to effectively understand and interact with their environments. By constructing accurate representations of the world, these agents can learn to reason, make informed decisions, adapt to new scenarios, and produce effective actions. The key advantages of world models for embodied agents manifest in two key dimensions. First, by encoding representations of physical environments and their transition dynamics, embodied world models facilitate long-horizon planning in complex, real-world scenarios [17], which is a capability that surpasses the limitations of basic imitation learning approaches. Second, through the distillation of universal world priors, they significantly enhance out-of-distribution (OOD) generalization [130], enabling robust adaptation to previously unseen tasks and scenarios.

From a technical perspective, current embodied world modeling approaches primarily follow three architectural paradigms: video generation-based models that operate in observable visual space [112, 139, 24], 3D reconstruction-enhanced models that incorporate geometric representations [49, 71, 123], and latent world models that maintain compressed state representations for efficient robotic action planning [117, 3]. The first two approaches extend established computer vision and multimodal generative techniques to model world states in pixel space, while the third paradigm focuses on learning compact latent representations of the world. In terms of training methodologies, embodied world models typically employ conditional generation objectives that incorporate external instructions to predict state transitions, including text-conditioned generation [139, 52, 65, 63] and action-trajectory [141, 115, 34] controlled future prediction. Recent advances have introduced visual-action joint prediction frameworks to enhance action understanding [107, 61, 140], as well as physics-constrained learning paradigms [135, 93] to ensure generated content adheres to physical laws. Embodied world models serve three primary roles in robotic applications. First, they function as cloud-based data synthesis engines [52, 53, 132], generating high-quality synthetic training data essential for training advanced robotic policy models like VLA (Visual Language Action) and VLN (Visual Language Navigation) models. Second, they act as environment proxies [62, 85, 73], supporting the evaluation of embodied agents in simulated settings without the need for a real environment. Finally, they can operate as the on-device ”brain” for robots [3, 137, 126], performing

Table 1: Categorization of existing related surveys.

Survey	Year	Main Topic	Limitations
Zhu et al. [142]	2024	General world model	Limited to general applications
Ding et al. [20]	2024	General world model	Limited to concepts and applications
Xie et al. [120]	2025	General world model	Limited to 3D cognition ability
Guan et al. [33], Feng et al. [23], Tu et al. [101]	2024, 2025	Driving world model	Different research domain
Liu et al. [69]	2025	Embodied AI	Comprehensive introduction of embodied robots and simulators, with limited techniques of world models
Long et al. [72]	2025	Embodied AI	Overview of physical simulators and world models, only discussing world model architectures without more technical details
Liang et al. [64]	2025	Embodied AI	Comprehensive overview of embodied learning, only briefly discussing world model concepts and architectures

real-time future state inference to guide robotic action planning. For clarity, we organize existing representative embodied world models into three axes: modalities, action types and applications, as illustrated in Figure 1.

Existing surveys on related topics primarily fall into three categories, as summarized in Table 1. The first category is about general world models [142, 20, 120], which extensively discuss the evolution of concepts, fundamental functionalities, and applications across various fields, offering broad coverage but lacking in-depth analysis of underlying technologies. The second category focuses on world models in autonomous driving [33, 29, 23, 101], providing a comprehensive overview of the latest progress in that specific domain. The third category encompasses embodied AI-related surveys [69, 72, 64]. While these works discuss general techniques for developing embodied agents, including world models, they do not provide a systematic technical review specifically dedicated to embodied world models. Differently, our survey is the first to focus on embodied world models, comprehensively discussing the full-stack technology from model architectures and training paradigms to applications and evaluations, organized meticulously along technical routes.

The main contributions of this survey can be summarized as follows:

- We provide a systematic and up-to-date review of the rapidly developing research on embodied world models, summarizing the significant value of world models for embodied agents.
- We propose a novel technical taxonomy that hierarchically organizes the field into model architectures, training methodologies, application scenarios and evaluation approaches, providing researchers with a clear technical roadmap.
- We highlight future research directions and trends of embodied world models, along with promising future research questions, to further inspire subsequent research in the community.

The following content of this survey is structured as follows: Section 2 introduces fundamental concepts and background related to embodied world models. Section 3 delves into the current model architectures for embodied world models. Section 4 discusses training paradigms for embodied world models. Section 5 explores the functionalities and applications of embodied world models. Section 6 presents diverse evaluation perspectives for embodied world models. Finally, Section 7 discusses the current challenges and outlines future research directions in this field. The framework of the survey is illustrated in Figure 2.

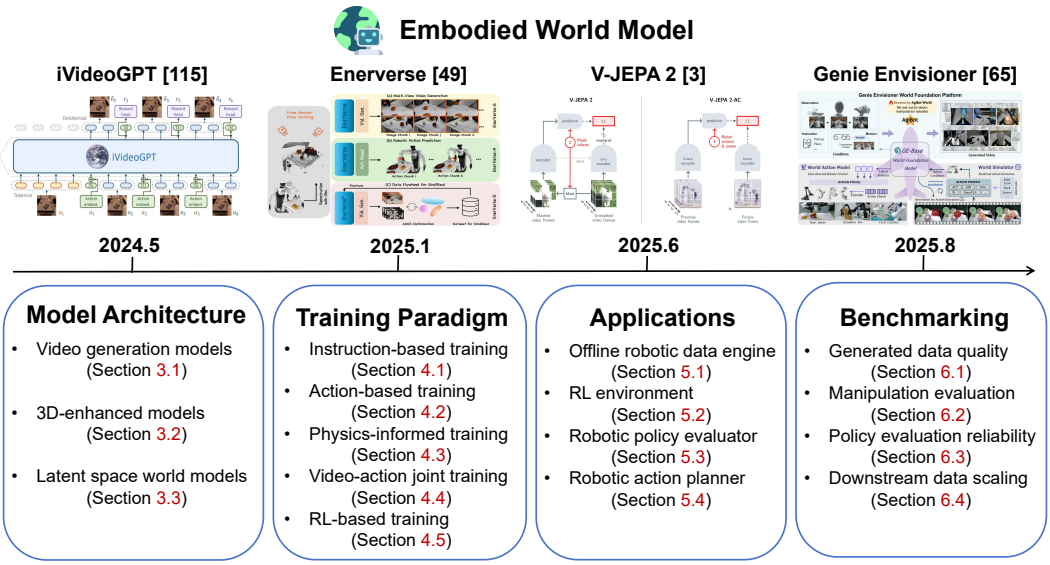


Figure 2: The framework of the survey. We systematically review the model architectures, training paradigm, application scenarios and benchmarks of embodied world models.

2 Background

The concept of world models originates in cognitive science. Craik's theory of mental models [56] proposed that humans perceive and reason about the external world by constructing abstract internal representations. System dynamics research further emphasized the role of internal models in simulating and predicting the behavior of complex systems [26], thereby laying the conceptual foundation for reinforcement learning and robotic control. In the 1990s, Sutton's Dyna architecture [95] was the first to unify learning, planning, and reacting, demonstrating how an agent could accelerate policy improvement by both learning and leveraging an internal environment model. In parallel, Schmidhuber introduced a neural system that, for the first time, explicitly separated a controller from a world model [89, 90, 88]. In this design, the controller selected actions to maximize cumulative reward, while the world model learned to predict environmental dynamics, enabling multi-step forward planning. This crystallized the core principle of world models—decoupling policy from environment simulation and using the model for internal rollouts. Nearly three decades later, this logic was revived in the deep learning era. Ha and Schmidhuber [36] reintroduced the term "world model" in machine learning and proposed a hierarchical architecture with three components: a controller (C) for decision-making, a vision model (V) for compressing high-dimensional perceptual inputs, and a memory model (M) serving as the world model to capture temporal dynamics. Conceptually, this mirrors the 1990 controller–world model separation, but it is implemented with modern deep neural networks that provide stronger representation, compression, and planning capabilities. By simulating the environment in latent space and predicting hypothetical outcomes, this framework significantly improved generalization and transfer across tasks, though early applications remained constrained by limited computational resources and immature algorithms. Within this evolving landscape, representative world models emerged, including Genie [11], exemplifying the autoregressive paradigm, and Sora [142], showcasing the potential of diffusion–Transformer architectures for high-fidelity dynamic world generation.

The concept of embodied intelligence can be traced back to Turing's proposal in the 1950s, which emphasized that machine intelligence should not be confined to symbolic reasoning but must also possess the ability to perceive and act in order to interact with the environment [102]. In the 1980s, Brooks and colleagues advanced this idea by criticizing traditional AI's overreliance on symbolic representations and proposing behavior-based robotics, highlighting that intelligence emerges through embodiment and interaction with the environment [10]. This work established the basic paradigm of the perception–planning–control (PPC) loop. As shown in the Fig 3, within this framework, traditional methods primarily adopted layered architectures: perception relied on vision models, high-level planning was based on logical rules [27], and low-level execution depended on

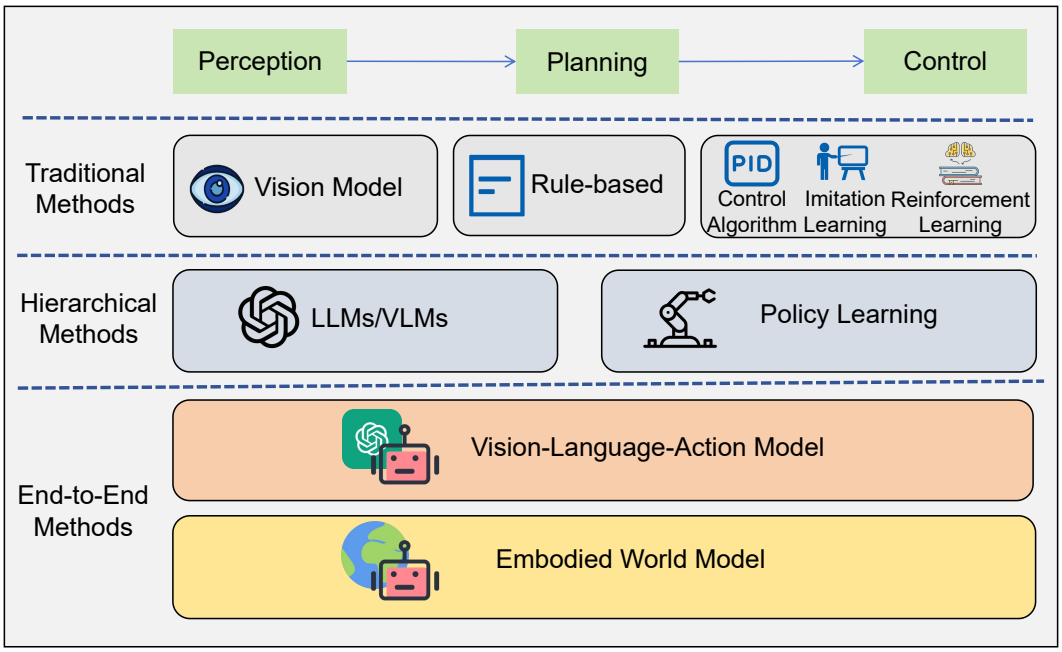


Figure 3: Three types of robotic learning frameworks. Traditional methods split tasks into visual perception, rule-based planning, and control stages. Hierarchical methods use LLMs/VLMs for perception and planning, and use separate policy learning models. End-to-end approaches integrate perception, planning, and control within VLA and world models.

classical control algorithms [28], imitation learning, and reinforcement learning. However, these approaches relied heavily on manual design, and the limited flow of information across layers hindered adaptation to complex environments.

In the current large model era, the landscape of embodied learning and decision-making is defined by two prominent, parallel-evolving paradigms: **Vision-Language-Action (VLA) models** and **World Models**. Here, we provide a detailed comparative analysis of them to reveal their fundamental differences. First, in terms of input and output information, both VLA and world models typically take a textual instruction and the current observation as input, with some VLA architectures also incorporating the robot's state. As for model output, VLA models directly produce a predicted action for the robot to execute, essentially functioning as a direct mapping from perception and language to action. In contrast, world models output a sequence of future observations, which can be either explicit visual frames or latent representations. The generation of a final action from this prediction typically requires a subsequent step, such as introducing an action decoder or employing rejection sampling for optimal action planning. We illustrate the workflow and IO comparison of VLA and world models for embodied decision-making in Figure 4. Second, in terms of base model architecture, VLA models are fundamentally built upon an autoregressive Large Language Model (LLM) backbone. This approach tokenizes and aligns visual and action information with text, essentially recasting the embodied decision-making as a token prediction task in the language space. Conversely, World Models generally utilize vision-centric generative models, such as DiT, as their backbone. Their core function is to align textual and action information with spatiotemporal visual tokens, enabling the generation of future visual states. We present an illustrative comparison of their token learning paradigm in Figure 5.

Each of the two approaches possesses respective advantages and disadvantages. VLA models, which still follow a traditional imitation learning path, learn a direct mapping between visual perception, language understanding, and action execution from human teleoperation data. The key advantages of VLA are the effectiveness of policy warm-up via utilizing high-quality teleoperation data, and superior complex instruction understanding and reasoning capabilities inherited from LLMs. However, they suffer from three major drawbacks: (1) poor data scalability, as teleoperation data is costly to collect; (2) limited generalization, as they are often trained in idealized environments and struggle with out-of-distribution scenarios; and (3) a lack of common sense, as imitation learning alone

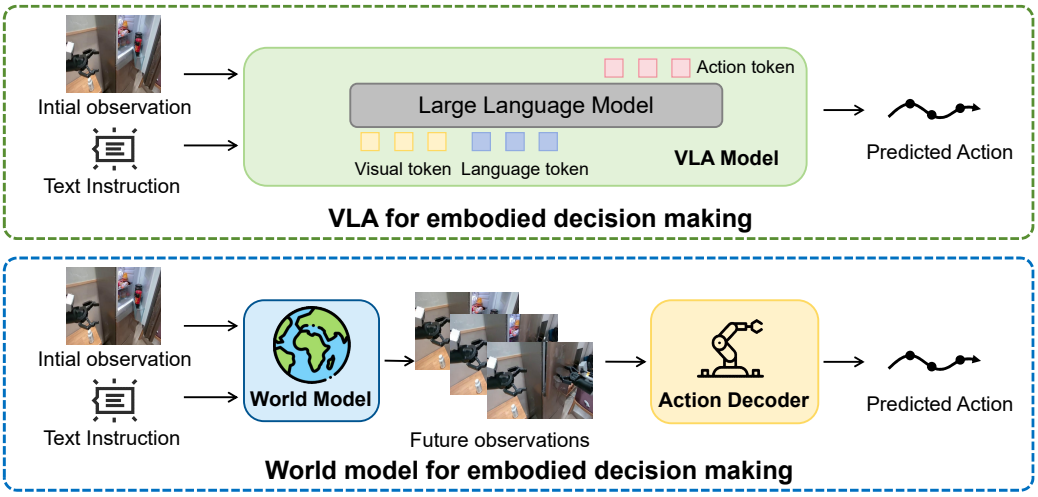


Figure 4: Workflow and input/output comparison of the world model and VLA model for embodied decision-making.

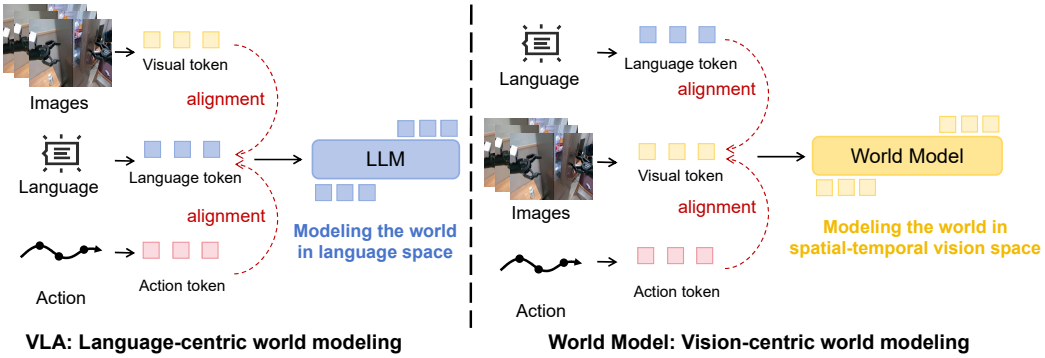
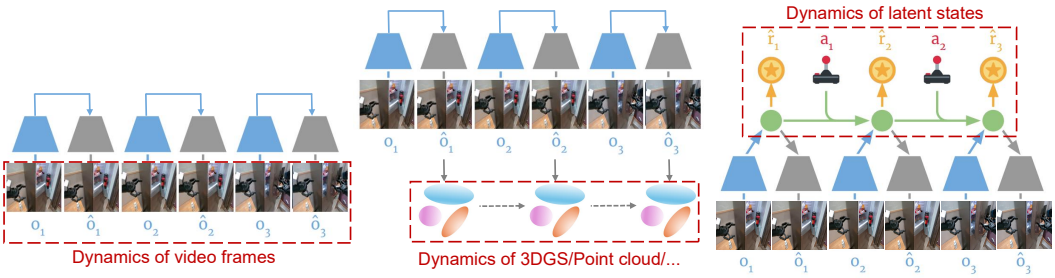


Figure 5: A comparative overview of VLA and world models, illustrating their distinct token space and modality alignment. VLA models process the world in a language token space, whereas world models are grounded in a spatial-temporal vision token space.

cannot model the deep causal principles of the physical world. Differently, world models aim to build an internal, predictive understanding of the physical world by learning state dynamics from large-scale unsupervised data. The primary strength is data efficiency and scalability, because world models can leverage vast amounts of unlabeled video data for training. They also offer better generalization by learning universal world dynamics and concept combination. The main disadvantage of world models is their current technical immaturity, as their capacity for deep action and instruction understanding still needs significant improvement.

Given that the physical world is fundamentally composed of spatiotemporal visual dynamics, which are more accurately captured by visual tokens than the high-level semantic abstractions of language tokens, we argue that world models are a more promising and better-suited paradigm for embodied learning over a long period. Current frontier research of world models broadly follows three directions. One approach is video-based, which conceptualizes the world as a continuous sequence of images and learns environmental dynamics by predicting future frames, thereby implicitly capturing physical laws and object interactions. The advantage of this path lies in leveraging the virtually unlimited supply of 2D video data on the internet to learn highly complex and high-fidelity visual dynamics. Another approach is 3D-enhanced world models, which prioritize geometric accuracy and physical consistency, making it indispensable for safety-critical applications such as robotics. In addition, a third direction, exemplified by JEPA [3], emphasizes learning compact and abstract world state representations in latent space. By modeling environmental dynamics at the level of



(a) Modeling the world in 2D space (b) Modeling the world in 3D space (c) Modeling the world in latent space

Figure 6: Three representative approaches to build embodied world models: (a) modeling the world in 2D space with video generation models; (b) modeling the world in 3D space with 3D reconstruction models; and (c) modeling the world in latent space with latent world models.

implicit representations, such models can support action prediction and planning more efficiently, while also providing stronger abstraction and generalization capabilities.

Alongside the evolution of model architectures, the training paradigms of embodied world models have also diversified, largely reflecting the functional role the model assumes. When a world model is designed as a data generation engine, its primary objective is to synthesize high-fidelity observational data to facilitate downstream policy learning. In this setting, instruction-conditioned training, action-conditioned training, style transfer-based methods, and physics-informed approaches are commonly adopted to enhance the realism and diversity of the generated data. By contrast, when the world model serves as an edge-side policy optimizer, the focus shifts toward directly guiding agents in policy generation. In this case, video–action joint training is emerging as the predominant paradigm. Furthermore, due to the persistent limitations of current world models in terms of generalization and controllability, reinforcement learning-based training has been increasingly employed to improve robustness and adaptability in complex, dynamic environments.

World models ultimately serve multiple roles as simulation and prediction engines in embodied intelligence. First, they act as cloud-based data synthesis tools [52, 49], generating high-quality observational data to support the training of advanced policy models such as Visual-Language-Action and Visual-Language-Navigation. Second, they function as environment proxies [62, 85], enabling the evaluation of agents without requiring real-world environments. Finally, world models can operate as on-device "brains" for robots [3, 137], performing real-time future state inference to guide action planning. In this way, world models play a central role in data generation, environment evaluation, and online decision-making, forming a key foundation for the advancement of embodied intelligence.

3 Architectures of Embodied World Models

Current embodied world models are predominantly categorized into three architectural paradigms, distinguished by their world state representation spaces and tailored to address distinct core needs of embodied intelligence: video generation-based models (Fig 6a) operate in observable visual space, primarily synthesizing high-quality robotic video data while modeling the world in 2D pixel space; 3D reconstruction-enhanced models (Fig 6b) integrate explicit 3D geometric representations to encode depth, volumetric structure, and object poses, enabling physically consistent world modeling in 3D space; latent world models (Fig 6c) maintain compact compressed latent states, modeling the world in low-dimensional latent space by distilling task-relevant information from redundant sensory data to boost efficiency, thus facilitating efficient state inference and action planning.

3.1 Video Generation-based Models

Video generation models have recently become the backbone for many embodied world models, as they provide a scalable foundation for simulating high-fidelity and temporally coherent visual dynamics. The key challenge lies in converting these models from non-interactive video generators into interactive world models, where action-conditioning and causal structure are essential for

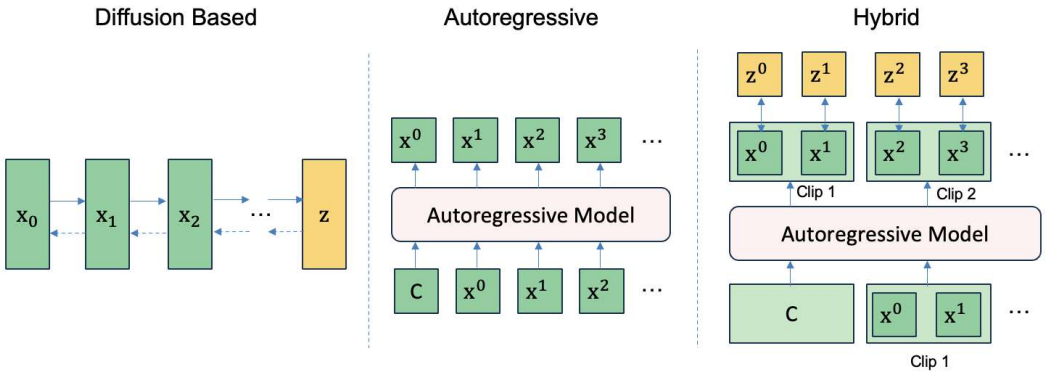


Figure 7: Architectural paradigms of video generation models: diffusion-based, autoregressive, and hybrid approaches.

embodied intelligence. Broadly, three architectural paradigms have emerged in this line: diffusion-based models, autoregressive models, and hybrid approaches that combine both, as illustrated in Figure 7.

Diffusion-based Models. Diffusion models, originally designed for unconditional or text-conditioned video synthesis, are increasingly adapted into world simulators by introducing causal and action-conditioned mechanisms. A landmark development was Sora[81], released by OpenAI in 2024, which achieved unprecedented levels of photorealism and long-horizon video generation, marking the first industrial-scale diffusion model capable of producing cinematic-quality content. Building on this, the community introduced Open-Sora[138], an open-source initiative that reimplemented efficient video VAEs and diffusion transformer backbones, making large-scale video synthesis research accessible beyond proprietary systems. Most recently, Wan[104] has further advanced diffusion-based modeling in the open-source ecosystem, offering 1.3B and 14B parameter variants trained with flow-matching and massive curated datasets. Wan integrates a novel spatio-temporal VAE, scalable pretraining strategies, and multi-task extensions including text-to-video, image-to-video, instruction-guided editing, and real-time generation, consistently achieving state-of-the-art performance across benchmarks. Complementing these large-scale foundation models, CogVideoX[124] demonstrated the scalability of diffusion transformers for long-horizon synthesis, while Vid2World [48] proposed "causalization" to embed autoregressive rollout and causal action guidance directly into the diffusion pipeline. Collectively, these advances illustrate how diffusion-based models have rapidly evolved from generic video synthesis tools into versatile platforms for causally grounded, action-conditioned, and temporally scalable world simulation, while still facing challenges of inference efficiency, responsiveness, and real-time interactivity.

Autoregressive Models. Autoregressive models differ from diffusion approaches in that they generate video tokens sequentially, frame by frame or patch by patch, inherently encoding causality and temporal continuity. This makes them naturally suited for embodied rollouts where actions and states must evolve in a consistent temporal order. By conditioning each prediction on past tokens, autoregressive systems support direct action integration, enabling interactive control with low latency. Genie[11] represents one of the earliest large-scale autoregressive world models, showing that high-capacity transformers can capture action-conditioned dynamics and produce responsive short-horizon rollouts. Building on this, Lumos-1[128] demonstrates how scaling autoregressive transformers to billions of parameters improves both controllability and long-term stability, while iVideoGPT [115] introduces efficient video tokenization schemes that significantly reduce the computational burden of sequential prediction. These works collectively highlight the major strengths of autoregressive modeling—real-time responsiveness, direct causal grounding, and smooth temporal evolution—while also revealing limitations: they often struggle to match the visual fidelity and stylistic diversity achieved by diffusion models, and scaling to long horizons can lead to compounding errors that degrade rollouts over time.

Table 2: Representative video generation models toward world modeling across three paradigms.

Paradigm	Model	Key Contributions
Diffusion	Sora [81]	Photorealism; Long-horizon; Industrial-scale training
	Open-Sora [138]	Open-source; Efficient VAE; DiT backbone
	Wan [104]	Spatio-temporal VAE; Flow-matching; Multi-task coverage
	CogVideoX [124]	Large transformer; High fidelity; Long rollout
	Vid2World [48]	Causalization; Action guidance; Interactivity
Autoregressive	Genie 1 [11]	AR world model; Action-conditioned rollouts
	Lumos-1 [128]	Billion-scale AR; Controllability; Stability
	iVideoGPT [115]	Efficient tokenization; AR rollouts
Hybrid	Genie 2 [18]	Diffusion + AR; Interactive; Long-horizon sims
	NOVA [19]	Non-quantized AR; Dual prediction; Diffusion denoising
	RoboScape-long [91]	Adaptive combination of Diffusion and AR generation
	VideoGPT [143]	AR rollout; Diffusion priors; Realism + Causality
	MAGI-1 [100]	Scalable hybrid; Chunk-wise generation; Streaming

Hybrid Models. To combine the strengths of both paradigms, hybrid designs leverage the high-quality synthesis of diffusion with the causal grounding and interactivity of autoregression. Genie 2[18] exemplifies this approach by integrating a diffusion backbone for visual fidelity with an autoregressive rollout mechanism to maintain interactivity, enabling longer and more responsive simulations than its predecessor. NOVA[19] further advances the hybrid paradigm by discarding vector quantization entirely and adopting a continuous-space autoregressive formulation. It introduces a two-level mechanism—temporal frame-by-frame causal prediction and spatial set-by-set masked prediction—augmented with a diffusion-style denoising objective. Despite its compact size (0.6B parameters), NOVA achieves competitive results with state-of-the-art diffusion models on text-to-video benchmarks, while being significantly more efficient in training and inference. Beyond these, Video-GPT[115] incorporates autoregressive rollout while distilling diffusion priors to enhance realism, and MAGI-1[100] scales hybrid autoregressive–diffusion modeling with chunk-wise generation, enabling efficient streaming video synthesis, controllable long-horizon rollouts, and real-time deployment at trillion-token scale. More recently, Roboscape-long [91] introduces an auto-regressive framework that performs varying-length chunk denoising, thereby enabling the generation of long-horizon embodied videos. Collectively, these systems demonstrate that hybrid designs represent a natural evolution of autoregressive and diffusion paradigms, unifying realism, causality, and efficiency, though challenges remain in optimization stability and scaling to ultra-long rollouts.

Overall, video generation-based world models progress along three complementary trajectories: (i) diffusion-based methods enhanced with causal and action-guided extensions, (ii) autoregressive designs that directly support interaction and real-time rollouts, and (iii) hybrid approaches that integrate both paradigms to achieve controllability and realism simultaneously. Together, these advances highlight the transition from generic video generation to embodied simulation, establishing the foundation for world models that are not only visually realistic but also causally grounded, interactive, and adaptive to agent actions.

3.2 3D Reconstruction-enhanced Models

Integrating 3D reconstruction into world models has emerged as a key direction to improve realism and geometric consistency in dynamic world modeling [4, 47]. Reconstruction provides structural priors for generation, while generation in turn enriches reconstruction, forming a bidirectional interplay.

Existing studies generally follow two trajectories: cascade paradigms and unified paradigms. As illustrated in Figure 8, cascade models explicitly reconstruct 3D geometry and render novel views

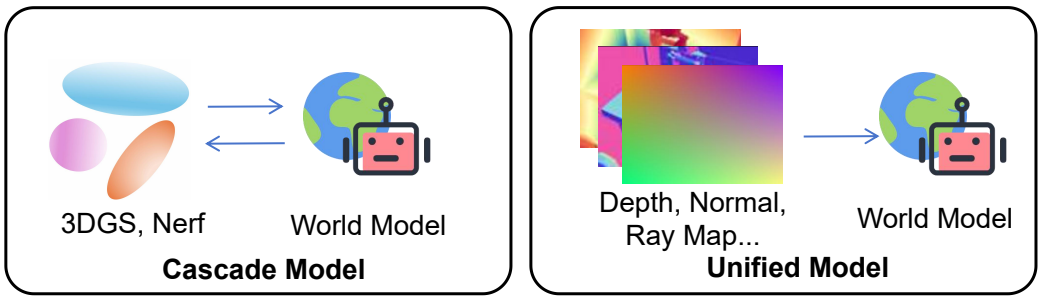


Figure 8: Overview of 3D reconstruction-enhanced world models. The cascade model relies on explicit 3D geometry reconstruction (e.g., 3DGS, NeRF), while the unified model integrates depth, normal, and ray-based projections into a single framework .

(e.g., 3DGS, NeRF), whereas unified models incorporate multi-view priors such as depth, normal, and ray maps into a single framework. In cascade paradigms, reconstruction outputs are incorporated as memory or conditional guidance for generation. Wu et al. [118] introduced a video world model with hierarchical memory modules, including short-term working memory, long-term spatial memory represented by point clouds, and sparse episodic memory. This design enables the model to maintain spatial coherence across long video sequences. Chen et al. [15] proposed a novel world model, which embeds point clouds and camera poses into the video generator. By injecting 3D structural priors, the model achieves high-fidelity, controllable world simulations. Zuo et al. [144] presented GaussianWorld, a framework that models scene evolution in 3D Gaussian space. By explicitly leveraging reconstruction priors and conditioning on RGB observations, the model produces geometrically consistent video predictions. Conversely, generative models can facilitate reconstruction. Min et al. [78] presented a world model which is based 4D pretraining framework and learns compact spatiotemporal bird's-eye view representations by predicting 3D occupancy from past multi-camera images and actions, offering better scalability for downstream tasks. Zhao et al. [134] showed that novel trajectory synthesis and multi-view video generation can serve as enriched supervision signals for 4D reconstruction, thereby improving reconstruction quality. Ni et al. [80] further advanced this idea by proposing ReconDreamer, which introduces progressive trajectory perturbations and online restoration to enhance robustness under large viewpoint shifts. To address cross-view inconsistencies, Ni et al. [79] developed WonderFree, which integrates video diffusion priors with cross-view consistency constraints, enabling efficient and view-aligned video generation. Similarly, HunyuanWorld [98] employs a staged pipeline: Panorama-DiT generates panoramic priors, while depth reconstruction ensures structural alignment. This system supports diverse downstream tasks, including image-to-3D object synthesis and sky generation. In the field of robotics, several recent models are pushing the boundaries of generative AI for manipulation tasks. EnerVerse [49] unifies video generation and 4D Gaussian Splatting into a closed loop. EnerVerse-AC [54] introduces a novel action map to control the video generation for embodied scenarios. Furthermore, ORV [123] additionally provides precise semantic and geometric cues to enhance the temporal consistency and controllability of generation results.

In unified paradigms, geometry and dynamics are jointly represented within a single neural framework to avoid error accumulation inherent to cascaded pipelines. Aether [97] integrates depth and ray map reconstruction into generative modeling, enabling human-like spatial reasoning. Its design allows strong zero-shot generalization to real-world scenes. Chen et al. [16] proposed DeepVerse, a 4D autoregressive world model that unifies visual, depth, and pose information in a shared state space. By incorporating geometry-aware memory mechanisms, DeepVerse maintains long-term spatial consistency in complex environments. Geo4D [55] predicts some additional physical information including point maps, depth maps, and ray maps, to achieve a more complete understanding of the scene. Zhen et al. [135] proposed TesserAct, which leverages RGB-D-normal supervision and consistency regularization to enforce spatiotemporal coherence in generated 4D embodied worlds. Future prediction and unified perception are also emphasized. Wu et al. [114] introduced Geometry Forcing, which aligns implicit diffusion features with external geometric priors such as VGGT, showing that explicit 3D annotations are not always required to learn structured geometry. More recently, Shang et al. [93] proposed RoboScape, a physics-informed world model that integrates temporal depth prediction and keypoint dynamics learning. By encoding both geometric consis-

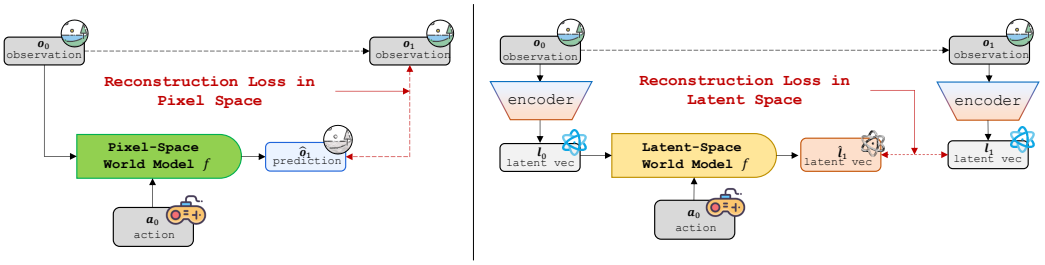


Figure 9: Comparative schematic of pixel-space world models and latent-space world models.

tency and physical properties (such as object shape and material), RoboScape enhances the fidelity of video rendering and improves complex motion modeling.

Overall, 3D reconstruction-enhanced world models advance along two trajectories: cascade models, which incorporate reconstruction as priors or memory for generation, and unified models, which learn geometry and dynamics jointly in an end-to-end manner. While cascaded approaches emphasize leveraging external models, unified approaches highlight the intrinsic coupling of representations. Together, they greatly enhance temporal coherence, geometric fidelity, and physical reasoning, paving the way for scalable world models that can both understand and simulate the physical world.

3.3 Latent Space World Models

When constructing world models from temporal snapshots (such as video frames or game screens), modeling environment dynamics directly at the pixel level incurs prohibitive computational cost during feedback and often causes the model to overfit redundant details, which hinders generalization and reasoning [121]. Consequently, a growing body of work has focused on mapping temporal data into a latent space and building world models therein. Representative works from recent years are summarized in Table 3. We present a brief comparative schematic of pixel-based reconstruction world models and latent-space world models, as shown in Figure 9.

Hafner *et al.* [39] introduced PlaNet, a purely model-based agent that learns latent dynamics directly from pixel observations and performs online planning within the latent space. PlaNet employs a Recurrent State-Space Model (RSSM) that combines deterministic and stochastic latent transitions, enabling robust long-term predictions while maintaining computational efficiency. A key innovation is the *latent overshooting* objective, which regularizes multi-step predictions in latent space without requiring costly pixel reconstructions. Empirically, PlaNet achieves competitive performance on continuous control tasks from pixels with up to 200× fewer environment interactions than model-free baselines.

Building upon PlaNet, Hafner *et al.* [38] proposed Dreamer, which extends latent dynamics modeling to policy optimization. Dreamer decouples world-model learning and policy learning: the world model is first trained via latent overshooting and pixel reconstruction, after which an actor-critic agent is trained entirely within the latent imagination. By back-propagating analytic value gradients through imagined latent trajectories, Dreamer attains state-of-the-art data efficiency on 20 visual control tasks, outperforming both model-based and model-free methods while requiring less training time. Recognizing the limitations of Gaussian latent variables in capturing multi-modal environment transitions, Hafner *et al.* [40] presented DreamerV2, which replaces Gaussian latents with categorical variables optimized via straight-through gradients. DreamerV2 additionally introduces KL balancing to stabilize training by decoupling the learning rates of the prior and posterior distributions. On the Atari 2600 benchmark, DreamerV2 becomes the first latent-imagination agent to achieve human-level performance across 55 tasks using a single GPU, surpassing prior model-free baselines such as Rainbow and IQN. Most recently, Hafner *et al.* [42] introduced DreamerV3, a fully general-purpose algorithm that learns robust world models under fixed hyperparameters across more than 150 tasks spanning continuous and discrete control, visual and proprioceptive inputs, and sparse and dense rewards. DreamerV3 incorporates several robustness techniques—symlog transformations, two-hot regression, percentile-based return normalization, and KL balancing with free bits—that collectively enable stable learning across vastly different signal scales and reward structures. Notably, DreamerV3 is the first agent to collect diamonds in Minecraft from scratch without

human demonstrations or curricula, addressing a long-standing open challenge in open-world exploration.

While the Dreamer lineage demonstrated remarkable success in simulated domains, its reliance on dense reward signals and high-fidelity action labels rendered direct transfer to real-world settings challenging. In 2022, LeCun [59] articulated the Joint-Embedding Predictive Architecture (JEPA) as a conceptual shift: instead of reconstructing pixels or optimizing task-specific rewards, the agent learns a latent space in which future observations are predictable from past ones. This idea was first instantiated for images by Assran *et al.* [2] with I-JEPA, which masks large spatial blocks in a single image and trains a ViT to predict the representation of the masked regions from the visible context. By dispensing with handcrafted augmentations and pixel-level losses, I-JEPA yields semantically rich features that transfer strongly to ImageNet classification and low-level vision tasks while using an order-of-magnitude fewer GPU-hours than comparable generative approaches.

The temporal analogue was concurrently developed by Bardes *et al.* [8] in V-JEPA. Operating on short video clips sampled from 2 million publicly available videos, V-JEPA masks spatio-temporal cubes and trains a predictive transformer to infer the missing latent representations. Importantly, the objective is purely predictive—no action labels, text, or pixel reconstructions are used—yet the resulting frozen backbone attains competitive accuracy on Kinetics-400 and Something-Something-v2, surpassing prior self-supervised video methods under linear probing. These results established that large-scale, action-free video datasets can suffice for learning general-purpose world representations. Motivated by the scalability of masked prediction, Assran *et al.* [3] subsequently introduced V-JEPA 2, which enlarges the encoder to 1 billion parameters and trains on 22 million hours of curated internet video plus 1 million images. A progressive-resolution schedule allows efficient absorption of temporal context without prohibitive compute. Combined with attentive probes, V-JEPA 2 sets new state-of-the-art on motion-centric benchmarks such as Something-Something-v2 and Epic-Kitchens action anticipation, while remaining competitive on appearance-centric tasks like ImageNet. Crucially, V-JEPA 2's latent space can be repurposed for embodied planning with minimal additional supervision. By freezing the pretrained encoder and training a lightweight, action-conditioned predictor on just 62 hours of unlabeled teleoperation data from the Droid dataset, the authors obtain V-JEPA 2-AC, a latent world model that performs zero-shot model-predictive control on Franka arms for pick-and-place tasks in unseen laboratory environments. This constitutes one of the first demonstrations of a vision-only world model trained primarily on passive internet video yet capable of closed-loop robotic manipulation without task-specific rewards or demonstrations.

Concurrent with the above developments, a complementary line of work has explored task-oriented latent dynamics and offline-to-online finetuning to bridge simulation and real-world deployment. Hansen *et al.* [44] propose TD-MPC, a model-based RL algorithm that performs local trajectory optimization in a learned latent space while jointly training a terminal value function via temporal-difference learning. TD-MPC departs from pixel-reconstruction objectives, instead learning latent representations directly from reward signals, yielding superior sample efficiency on high-dimensional continuous control tasks. Building upon TD-MPC, Feng *et al.* [25] introduce TD-MPC-offline, which integrates uncertainty-aware planning and offline-pretrained world models to enable few-shot finetuning on real robots. Specifically, they pretrain a latent world model on offline robotic datasets and then finetune it with as few as 20 online trials using a test-time regularizer that balances estimated returns against epistemic model uncertainty. This framework demonstrates robust transfer from offline data to online adaptation. Extending this line of research, Hansen *et al.* [43] present TD-MPC2, a scalable and robust generalist world model that learns across 104 diverse continuous control tasks—including high-dimensional locomotion, dexterous manipulation, and physiologically accurate musculoskeletal control—using a single set of hyperparameters. TD-MPC2 leverages implicit, decoder-free latent dynamics and a normalized task-embedding space to generalize across embodiments and action spaces, achieving state-of-the-art performance while maintaining training stability at scales up to 317 million parameters.

Taken together, the progression from Dreamer to JEPA delineates a clear temporal trajectory: early latent-dynamics research focused on mastering simulators with dense rewards, whereas the JEPA family has shifted attention to learning general-purpose world representations from the open-ended visual stream of the internet. The emergence of TD-MPC and its successors further complements this trajectory by providing a principled framework for transferring latent world models from offline pretraining to online real-world control, thereby narrowing the gap between simulated and physical intelligence. The latest developments suggest that, once equipped with scalable predictive objectives

Table 3: Summary of representative latent space world models.

Model	Application Area	World Model Type / Contribution
PlaNet [39]	Visual control	RSSM; latent overshooting; data-efficient planning.
Dreamer [38]	Model-based RL	Latent imagination actor-critic; analytic value gradients.
DreamerV2 [40]	Model-based RL	Categorical latents; KL balancing; human-level Atari.
DreamerV3 [42]	Model-based RL	Robust training (symlog, two-hot, KL free-bits); 150+ tasks incl. Minecraft.
I-JEPA [2]	Image learning	Masked prediction; semantic latent features.
V-JEPA [8]	Video learning	Spatio-temporal masking; predictive transformer; strong transfer.
V-JEPA 2 [3]	Video learning	1B-parameter encoder; progressive resolution; SOTA motion benchmarks.
V-JEPA 2-AC [3]	Robotic control	Frozen encoder + lightweight predictor; zero-shot MPC for robotics.
TD-MPC [44]	Continuous control	Task-oriented latent dynamics; TD-learning for MPC.
TD-MPC-offline [25]	Robotic control	Offline-to-online finetuning with uncertainty-regularized planning.
TD-MPC2 [43]	Multi-task continuous control	Scalable implicit world model; single hyperparameter set across 104 tasks.

and modest interaction data, such representations can bootstrap effective planning in the physical world, narrowing the gap between simulated and real-world intelligence.

4 Training Paradigm of Embodied World Models

Depending on the different applications of the world model, the training paradigms also differ. When the world model serves as a data generation engine, its goal is to generate high-quality observational data for downstream policy learning. The training paradigms of the world model mainly include instruction-conditioned training, action-conditioned training, style transfer-based training, and physics-informed training. When the world model serves as an edge-side action planner, its goal is to guide the agent in generating policies. At this point, the training paradigm of the world model is mainly action-conditioned training and video-action joint training. In addition, the existing world models still have deficiencies in generalization and action controllability. Thus, an increasing number of researchers have begun to use the reinforcement learning-based training method to train world models. This is because the RL training paradigm can make up for the characteristics that the pixel fitting learning target generated solely by video is difficult to learn, such as motion controllability and geometric accuracy. A schematic diagram of the world model training paradigm is shown in Fig. 10.

4.1 Instruction-conditioned Training

In this training paradigm, the input of the world model consists of historical observation sequences and text-based control instructions. This training paradigm evolved from the traditional controllable video generation model owing to the fact that textual instruction can provide high-level control signals to the generation process.

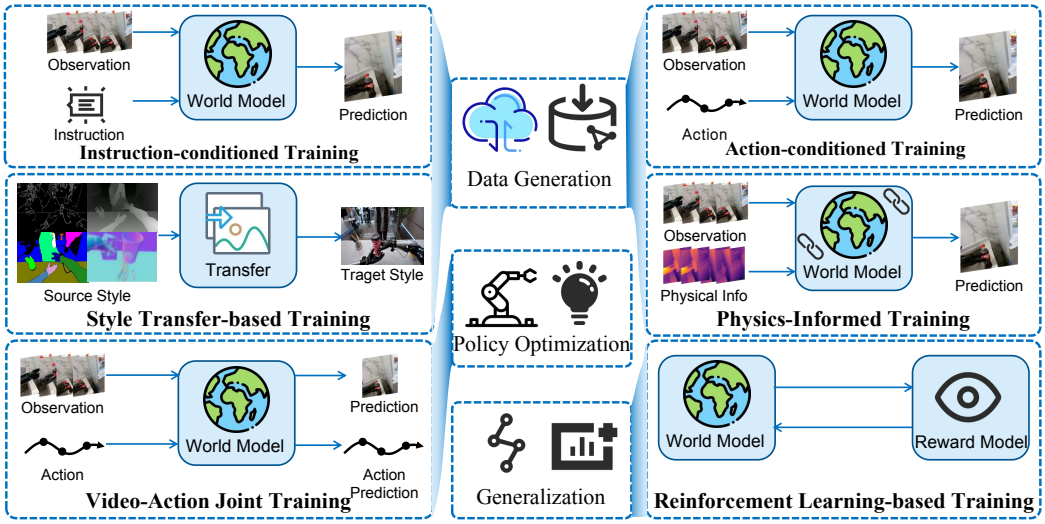


Figure 10: Training Paradigms of Embodied World Models: Instruction-conditioned training, action-conditioned training, physical-informed training, video-action joint training, and reinforcement learning training.

Table 4: Summary of models with different training paradigms. **ICT**: Instruction-conditioned Training, **ACT**: Action-conditioned Training, **PIT**: Physical-informed Training, **VAT**: Video-Action Joint Training, **RLT**: Reinforcement Learning Training.

Method	Paradigm	Contribution
Sora [70]	ICT	Text-driven video generation model
RoboDreamer [139]	ICT	Decompose text instructions into fine-grained phrases
Pandora [119]	ICT	Real-time text control generation
Cosmos [1]	ICT	Instruction controllable world simulator
Vid2World [48]	ACT	Adding the embeddings of actions and observations
UWM [140]	ACT	Concatenating the embeddings of actions and observation
EnerVerse-AC [54]	ACT	Multi-channel aciton injection using cross-attention
FLARE [137]	ACT	Generate action tokens using diffusion
RoboScape [93]	PIT	Key point tracking and depth as physical information
TesserAct [135]	PIT	Decouple geometry, materials, and motion for 4D modeling
HMA [107]	VAT	Using Transformer to predict observation and action
UVA [61]	VAT	Symmetrical encoder and decoupled diffusion head
WorldVLA [12]	VAT	World model head renders observations with action head
RLVR-World [116]	RLT	Training the world model using a RL paradigm

When Sora [70] was first proposed, it was regarded as a kind of world model because it could generate long and high-quality videos based on text instructions. However, it cannot interact in real time with control signals. RoboDreamer [139] breaks complex instructions into fine-grained short phrases and stitches corresponding video segments, enabling zero-shot composition of unseen tasks. Furthermore, Pandora [119] further augments the autoregressive backbone with diffusion-based frame synthesis, allowing real-time textual edits while guaranteeing cross-domain visual consistency via a history buffer. Cosmos [1] unifies a hierarchical diffusion renderer with a learned Newtonian physics module, enabling text-driven generation of long-horizon, physically-plausible scenes that can be steered on the fly by continuous control signals.

4.2 Action-conditioned Training

Although high-level control of text information can control the model to generate different videos, it is difficult to provide fine-grained control signals. The action sequence, due to containing fine-grained control information, can be used to enhance the fine-grained perception and generation capabilities of the world model. In the paradigm of action-conditioned training, the world model

receives the input of the action sequence and the observation sequence and generates the future observations based on the action control signal. The action sequence is regarded as a sequence of control signals to the object, which can affect the object and is controllable, which usually refers to the target pose of the embodiments in embodied AI.

In this training paradigm, a key challenge is how to inject action signals so that they are both causal and controllable. Vid2World [48] represents the robot's 6-DoF pose trajectory as a dense action tensor and injects it into the diffusion process by direct element-wise addition with the visual latent features before every U-Net decoder block. A causal mask is applied in the temporal attention layers to ensure that predictions at time t depend only on actions up to the current timestamp, preserving temporal causality. Unified World Models [140] decouple the diffusion timesteps of actions and images, allowing flexible marginalization or conditioning. Instead of direct addition, they concatenate action tokens with latent image patches and let a transformer jointly denoise both modalities via cross-attention. EnerVerse-AC [54] introduces a multi-level action-conditioning mechanism in which spatial action maps rendered from end-effector poses are concatenated with image latents, while delta-action embeddings are fused via cross-attention layers, enabling finer spatial and temporal control. FLARE [137] avoids dense pixel-level injection altogether. It aligns a small set of learnable future tokens in the diffusion transformer with latent embeddings of future observations conditioned on the action sequence, achieving implicit world modeling without extra architectural complexity. Furthermore, there are also some action injection methods, such as using an Action-Control Transformer Block that handles continuous and discrete action signals [125]. Continuous vectors are channel-wise concatenated after spatio-temporal compression. The discrete codes are treated as key/value pairs in cross-attention, allowing a single pre-trained diffusion backbone to merge the action information.

4.3 Physics-informed Training

Although diffusion and autoregressive models can produce visually compelling videos, they frequently violate basic physical laws—objects float, liquids pass through solids, or human joints bend impossibly. Physics-informed training explicitly injects physics-related priors into the learning objective to close the sim-to-real gap. These physical priors mainly include inherent properties of objects, such as hardness, elasticity, color, material, etc, spatial structure information, such as depth, normal, and point cloud, dynamic interaction information, such as key point trajectory and motion speed or acceleration, and environmental physical parameters, such as light, force or moment direction, collision constraints, and relative motion. Although there are many physical constraints, the key issue in this research lies in how to select the appropriate physical laws and effectively integrate them into the world model.

RoboScape [93] exemplifies the paradigm by fusing RGB and depth branches within a shared autoregressive backbone: predicted depth is injected layer-wise as a physical prior, while self-supervised key-point trajectories capture contact-driven dynamics; together they enforce spatial coherence and cut sim-to-real policy gaps. TesserAct [135] employs a 4D entity modeling approach, extracting the underlying codes for geometry, materials, and motion, and then routing the masks for specific tasks, thereby enabling the implementation of video prediction, 4D reconstruction, and visual planning through a single model.

4.4 Video-action Joint Training

The video-action joint training paradigm can simultaneously predict future observation results and actions. Such methods achieve this by using the multi-task learning approach to enable the model to better learn the correspondence between video and actions, thereby enhancing the prediction performance. The core design is a shared latent space into which visual and action features are projected via distinct encoders, then decoded by separate diffusion or autoregressive heads.

HMA [107] introduces Heterogeneous Pre-training (HPT) to map actions from different robot morphologies into a common embedding, and employs masked autoregression so that the same Transformer parameters simultaneously predict next frames and next joint angles. UVA [61] proposes a symmetrical encoder followed by decoupled diffusion heads: one head synthesizes video, the other regresses continuous actions, while a masking schedule flexibly switches between reconstruction, forecasting, and pure action generation objectives. WorldVLA [12] instantiates the paradigm with

three encoders (image, text, action) feeding a joint latent space; a world-model head renders future observations conditioned on the action head's predicted controls.

4.5 RL-based Training

Over the past period of time, an increasing number of researchers have begun to utilize the RL framework to train LLMs and VLMs, in order to enhance the understanding and reasoning capabilities of large models in handling corner cases. Then a natural question arises: *Can we use the methods of RL to train the world model?*

RLVR-World [116] was the first to introduce the RL-based training paradigm into the world model, and used a metric for video generation quality as the reward. It formulates the world model itself as an autoregressive strategy, whose action is the next sequence of latent state tokens, and directly uses visual metrics such as LPIPS as rewards to optimize the world model. The success demonstrates that reinforcement learning can indeed sculpt world-model dynamics toward downstream utility without catastrophic forgetting of generative fidelity.

Based on these early successful results, the next research frontier lies in making the reward mechanism not merely slightly improve the quality of video generation. We expect it to have the potential to become a major indicator for controlling behavioral controllability and may better guide the world model to follow the laws of physics. The training method based on RL shapes the underlying dynamic mechanisms, enabling small changes in continuous joint angles to be transformed into reliable and physically logical future outcomes.

5 Applications of Embodied World Models

The recent improvements in architectural design and training paradigms have catalyzed significant progress in embodied world models, revealing a promising future for their widespread application. These models, which learn to simulate the dynamics of the physical world from data, offer powerful solutions to several long-standing challenges in robotics. This section introduces the motivation, implementation, and benefits of applying embodied world models in four domains: 1) Robotic Data Generation, 2) Reinforcement Learning Environment Simulation, 3) Robotic Policy Evaluation, and 4) Action Planning in Agents, as shown in Figure 11. Representative works in embodied world model application are listed in Table 5.

5.1 Offline Robotic Data Generation Engine

Training robust and generalizable robotic policies requires vast and diverse data [103, 82, 57]. One of the most significant bottlenecks challenging the scalability of robot learning is data scarcity [93]. Traditional methods for collecting robotic data often involve human teleoperation of a physical robot, which is usually time-consuming, labor-intensive, and fraught with safety risks. Furthermore, data collected this way often lacks diversity, reflecting the biases and limited behaviors of human operators. These critical limitations have motivated the use of embodied world models as powerful offline data generation engines.

An embodied world model functions by taking an initial environmental state and a sequence of actions or text instructions to effectively predict the subsequent states of the robot and its environment. This predictive capability allows for the creation of robotic trajectories in an efficient way to train a downstream robotic policy [52]. Leveraging extensive world knowledge acquired during pretraining on large-scale, diverse corpora, many embodied world models can generate a rich variety of trajectories that encompass different robotic behaviors, environmental conditions, and task outcomes [52, 67]. These introductions of synthesized diversity data are proven helpful to unlock behavior and environment generalization in robot learning. Moreover, world models can also be leveraged to generate failed trajectories with the least cost, helping to improve the policy's generalization [54].

The use of embodied world models as data generation engines offers a scalable solution to the challenges of data scarcity and homogenization. It dramatically reduces the labor, time, and financial costs associated with human teleoperation, clearing a path toward more powerful and generalizable robotic systems. Empirical results demonstrate that data synthesized by world models, whether

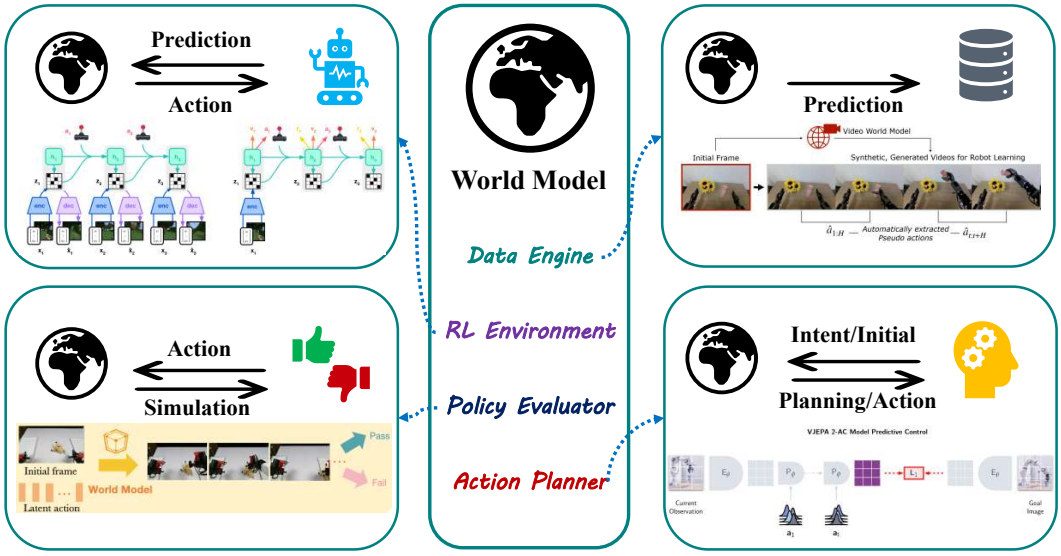


Figure 11: Illustrations of four main applications of embodied world models. The model can be deployed as: **1) A Data Generation Engine** to overcome data scarcity by synthesizing diverse trajectories. **2) A Reinforcement Learning Environment** to enable efficient and safe policy training in imagination. **3) A Policy Evaluator** to provide scalable and reproducible assessment of robotic behaviors. **4) A Planning Module** to facilitate long-horizon reasoning by simulating potential action outcomes.

used alone or in conjunction with real data, can significantly improve the performance of policies across various robotic tasks [52, 54, 67]. Despite these advantages, the simulation-to-reality gap [87] still remains a persistent challenge. While current works of world models are striving to minimize this gap, the discrepancies between the generated data and physical reality can cause performance degradation when a policy is deployed on a physical robot.

5.2 Environment Substitute for Reinforcement Learning

Reinforcement Learning has demonstrated remarkable success in enabling agents to learn complex behaviors automatically. However, its practical application in robotics is often hindered by profound sample inefficiency [127]. Training robotic policies through direct trial-and-error interactions in the real world can require millions of steps to converge, a process that is slow, expensive, and can cause significant wear and tear on physical hardware. Furthermore, the physical embodied environment and agents are not necessarily independent and identically distributed (i.i.d.), bringing instability into the learning process. Model-based RL offers a powerful alternative paradigm that addresses these issues by using a learned model of the environment.

This process can be described as ”learning in imagination” [38]. Instead of interacting with the physical world, the agent interacts with the world model, exploring different action sequences and learning from the predicted outcomes safely and efficiently. This process can be massively parallelized and executed much faster than real-time, allowing the agent to accumulate a vast amount of experience in a short period. Once a competent policy is learned, it can be transferred to the real environment for deployment.

For example, the GenRL framework [75] proposes a system that connects the latent representations of pretrained Vision-Language Models with the latent space of a generative world model. After grounding vision-language prompts into embodied domains, agents can be trained for corresponding behaviors in imagination. Similarly, iVideoGPT [115] introduces an autoregressive transformer framework that integrates multimodal signals (visual observations, actions, and rewards) as a world model. This model can then be used to conduct visual model-based reinforcement learning, achieving competitive performance in various robotic manipulation tasks. To overcome the need for costly and time-consuming real-world interaction, RoboScape-R [96] employs a unified embodied world model as a proxy. This model predicts future observations and reward signals, enabling efficient

online reinforcement learning to train robotic policies within a purely simulated environment. For a world model to be an effective substitute for the real world, it must learn a representation of the environment's dynamics that is both accurate and generalizable. An inaccurate or biased world model is a critical limitation, as the agent may exploit flaws in the model's physics or logic, leading to policies that are effective in imagination but fail catastrophically upon real-world deployment.

5.3 Robotic Policy Evaluator

The development of diverse and increasingly complex robotic policies presents a significant evaluation challenge. Directly assessing these policies in the real world is a laborious, time-consuming, and often expensive process. In addition, real-world evaluation may suffer from a lack of reproducibility and stability, as minor, uncontrollable variations in physical conditions—such as lighting, object position, or surface friction—can lead to inconsistent outcomes and noisy assessments. Embodied world models provide a compelling solution to these evaluation challenges by serving as a high-fidelity proxy for the real world [62, 54].

Instead of being executed in a physical embodied environment, a policy's actions are fed into a world model for predicting the consequences. The world model then simulates the interaction and generates a video of the predicted outcome. This approach allows the evaluation process in an efficient environment, reducing the need for physical robotic operations. It is also inherently reproducible and safe since the evaluations are conducted virtually. For instance, RoboScape [93] can act as a competitive robotic policy evaluator without the need for actual robotic experiments.

This method enables not just qualitative assessment but also the extraction of precise, quantitative metrics. Furthermore, world models facilitate powerful counterfactual evaluation, providing deeper insights into a policy's generalization capabilities and failure modes.

5.4 Action Planner as Embodied Agents

A core challenge for contemporary AI is enabling systems to learn about the world and develop behavioral capabilities primarily through observation [59]. For embodied agents, visual representation is vital for understanding, predicting, and planning within their physical environment [46]. Research into human cognition indicates that we use internal mental models to forecast the outcomes of our actions and formulate plans [26, 86, 74, 13, 37]. Embodied world models aim to provide this same capability to AI agents, addressing common struggles like long-term reasoning and planning. Two primary architectural designs have emerged for integrating world models into action planners:

Bi-model Architecture: In this design, the world model and the policy planning model are separate components [46, 83, 92]. The world model simulates the outcomes of many different action sequences based on prior knowledge. Then the world model can be used to enhance the target policy model. This process improves performance, especially in novel scenarios and complex tasks. For example, MoE-WM [92] proposes a hybrid approach by combining latent-space and visual-space world models to predict future observations and collectively decode actions.

One-model Architecture: This approach utilizes the world model's inherent capability for environment transition modeling to plan actions, a principle exemplified by the V-JEPA 2 architecture [3]. It is built upon a pretrained action-free joint-embedding-predictive model which is trained in a self-supervised manner. The model's action planning process involves sampling candidate action trajectories and then optimizing them based on minimizing an energy function to derive a predicted trajectory. Based on large-scale self-supervised training and post-training, the latent action-conditioned world model gains robot skills like planning in tasks with visual goal specification and can generalize zero-shot to new environments.

6 Benchmarks of Embodied World Models

Benchmarks for embodied world models aim to evaluate whether generative models can serve not merely as visual synthesizers but as reliable simulators of physical and interactive environments. Unlike traditional video generation metrics that emphasize visual fidelity or language alignment, benchmarks in this domain must capture dimensions essential for embodied intelligence, including physical plausibility, task execution, policy evaluation, and data scalability. To structure this

Table 5: Representative Works in Embodied World Model Applications.

Application Area	Method	World Model Type
Offline Robotic Data Generation	DreamGen [52]	Finetuned WAN 2.1 [104]
	RoboTransfer [67]	Multi-view, geometry, and appearance conditioned diffusion model
	EnerVerse-AC [54]	UNet-based Video Diffusion Model
RL Environment Substitute	GenRL [75]	GRU-based Recurrent Architecture
	iVideoGPT [115]	Autoregressive Transformer over quantized multimodal tokens
	RoboScape-R [96]	Autoregressive Transformer predicting both observation and reward
	DreamerV3 [42]	Recurrent State-Space Model with discrete latent representations
Robotic Policy Evaluator	WorldEval [62]	Finetuned WAN 2.1 [104]
	EnerVerse-AC [54]	UNet-based Video Diffusion Model
Action Planner as Embodied Agents	GPC [83]	UNet-based Video Diffusion Model
	VPP [46]	Finetuned Stable Video Diffusion [9]
	MoE-WM [92]	Mixture of latent and pixel space world models
	V-JEPA 2-AC [3]	Joint-Embedding Predictive Architecture [2]

landscape, we review four complementary perspectives. The first examines generated data quality, reconceptualizing fidelity as a multi-dimensional construct that integrates perceptual, semantic, and physical consistency. The second shifts focus to end-to-end manipulation evaluation, where world models are assessed as planners capable of producing valid action trajectories. The third addresses evaluation reliability, asking whether world-model-based assessments faithfully predict real-world policy performance. Finally, the fourth perspective considers data scaling, measuring how synthetic rollouts support efficient, generalizable, and transferable policy learning. Together, these benchmarks establish a systematic framework for assessing embodied world models as both generators of physically grounded data and enablers of embodied decision-making. Representative works and associated metrics are summarized in Table 6.

6.1 Generated Data Quality

In evaluating embodied world models, the quality of generated data is a central prerequisite, as only videos that are visually convincing, semantically faithful, and physically coherent can provide reliable substrates for reasoning and control. Recent benchmarks reconceptualize data quality as a multi-dimensional construct rather than a single score. VBench [51] introduced sixteen disentangled dimensions, separating perceptual video quality (temporal smoothness, subject consistency, aesthetics) from condition consistency with text prompts (semantic fidelity, spatial relations, style). T2V-CompBench [94] extended this view to spatio-temporal compositionality, assessing dynamic attribute binding, motion and action binding, interactions, and numeracy through multimodal language models, detection methods, and motion tracking. VBench-2.0 [136] further advanced the field by focusing on intrinsic faithfulness, measuring human fidelity, creativity, controllability, physics, and commonsense via hybrid pipelines of VLM–LLM alignment, video-based question answering, and anomaly detection. Collectively, these works trace a progression from surface fidelity, to compositional coherence, to intrinsic world faithfulness, showing that data quality must integrate perceptual, semantic, and physical-consistency dimensions to support embodied intelligence.

Within this progression, adherence to physical laws has emerged as a particularly crucial dimension, since only physically plausible videos can underpin reliable reasoning and control. VideoPhy [5]

Table 6: Evaluation perspectives, representative works, and metrics for embodied world models.

Perspectives	Representative Works	Metrics
Generated Data Quality	VBench [51]	Temporal quality, frame-wise quality, semantics consistency and style consistency.
	T2V-CompBench [94]	Video LLM-based metrics, spatial relationships, generative numeracy and dense optical tracking.
	VBench-2.0 [136]	Human fidelity, creativity, controllability, physics and commonsense.
	VideoPhy [5]	Semantic adherence and physical commonsense.
	VideoPhy-2 [6]	Semantic adherence, physical commonsense and physical rules.
	PhyGenBench [77]	Key physical phenomena detection, physics order verification and naturalness evaluation.
	WorldModelBench [60]	Instruction following, physics adherence and commonsense.
	EWMbench [129]	Scene consistency, motion correctness, and semantic alignment & diversity.
End-to-end Manipulation Evaluation	DreamerV3 [41]	Planning performance on downstream tasks.
	V-JEPA 2 [3]	QA accuracy, action anticipation recall and planning success rate.
	WorldSimBench [84]	Task success rate and execution accuracy.
Evaluation Reliability towards Policy Model	WPE [85]	Evaluating In-Distribution and Out-of-Distribution policies.
	WorldEval [62]	Pearson correlation coefficient and Mean Maximum Rank Violation.
	RoboScape [93]	Pearson correlation and R^2 between world models and the ground-truth simulator.
Data Scaling in Downstream Policy Model	DreamGen [52]	Instruction following and physics alignment.
	RoboTransfer [67]	Multi-view consistency, geometric consistency, and semantic consistency.
	GenSim [106]	Success rates.
	WorldGPT [32]	Cosine similarity for state prediction, task accuracy, and generation efficiency.
	Traj-LLM [58]	Average displacement error, final displacement error, and miss rate.
	RoboScape [93]	Pearson correlation and R^2.

first foregrounded this issue through human judgments of commonsense plausibility in material interactions, while VideoPhy-2 [6] scaled evaluation to sports and object-interaction scenarios with rule-level annotations of laws such as gravity and momentum conservation, coupled with automatic evaluators distilled from human feedback. PhyGenBench [77] establishes a law-grounded evaluation paradigm by linking prompts to 27 physical laws and introducing a three-tier framework that assesses phenomenon recognition, temporal ordering, and naturalness through vision–language models. WorldModelBench [60] broadened this perspective to application-driven domains such as robotics and driving, integrating instruction following, commonsense, and fine-grained physics adherence categories. Finally, EWMBench [129] specializes in the evaluation of embodied world models in robotic manipulation, combining scene consistency, trajectory-based motion correctness, and semantic alignment with task instructions. Taken together, these benchmarks trace a trajectory from intuitive plausibility to explicit law-grounded testing to task-specific embodied settings, establishing physical evaluation as a systematic pillar of generated data quality.

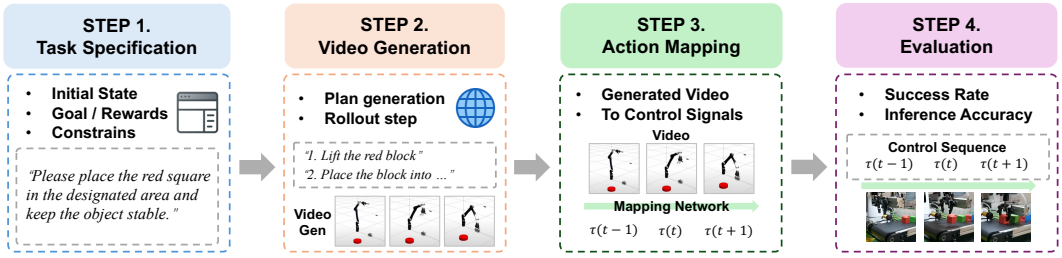


Figure 12: Workflow of benchmarking world models for planning accuracy. The process evaluates whether the world model can directly generate valid trajectories to achieve specified goals.

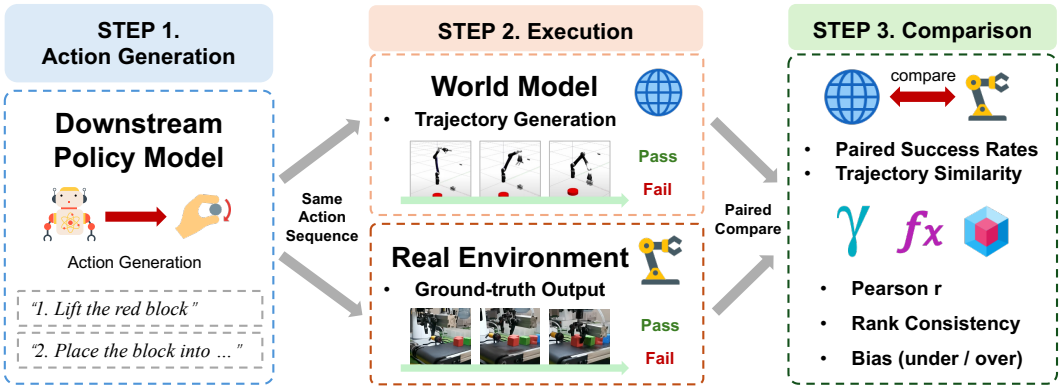


Figure 13: Workflow of benchmarking evaluation reliability towards policy models. Candidate policies are tested both in the world model and in the real world, with their outcomes compared using correlation- and ranking-based metrics.

6.2 End-to-end Manipulation Evaluation

The evaluation of action planning accuracy treats the world model not as a training environment but as the agent itself. The key question is whether the model can predict the outcomes of actions and generate trajectories that effectively achieve the specified goals. This assesses the model's capacity for decision-making and long-horizon planning. Figure 12 illustrates the workflow of this benchmark, where the world model itself serves as a planner to generate trajectories, and the resulting rollouts are evaluated against ground-truth references.

More recently, large-scale self-supervised approaches have advanced the frontier of planning accuracy. V-JEPA 2 [3] leverages internet-scale video pretraining to learn predictive representations of physical dynamics, and with only $\sim$62 hours of robot interaction data, fine-tunes an action-conditioned module that supports zero-shot robotic planning. The model demonstrates pick-and-place manipulation on Franka arms in novel environments without task-specific rewards, underscoring the benchmark value of evaluating zero-shot planning success.

Complementing algorithmic progress, WorldSimBench [84] provides the first benchmark framework explicitly targeting world models as planners. Its Implicit Manipulative Evaluation (measuring whether generated videos can be consistently translated into correct actions)—tests models in three embodied domains: open-ended environments (Minecraft), autonomous driving, and robotic manipulation. This provides a principled way to measure whether a world model's imagined rollouts align with actionable control sequences.

In this perspective, the world model is judged as a planning agent, and benchmarks emphasize decision quality, rollout fidelity, and the capacity to generate trajectories that directly translate into successful embodied actions.

6.3 Evaluation Reliability towards Policy Model

Evaluation reliability highlights whether embodied world models can provide policy assessments that remain predictive of real-world outcomes. Rather than focusing solely on perceptual fidelity, this dimension examines whether the metrics obtained from world-model-based evaluation are trustworthy indicators of actual robotic performance, thereby reducing the long-standing sim-to-real gap. Figure 13 shows how this benchmark operates by comparing policy evaluations conducted in the world model with those obtained from real-world executions, emphasizing correlation, ranking stability, and bias analysis.

Recent work highlights several strategies for enhancing evaluation reliability. On the one hand, Quevedo et al. [85] introduce World-model-based Policy Evaluation (WPE), where an action-conditioned video generation model serves as a proxy for policy testing. By comparing generated rollouts with real-robot executions under the same action sequences, WPE proposes metrics based on pixel-level similarity, perceptual scores, and semantic segmentation. Empirical findings show that WPE tends to underestimate in-distribution policies and overestimate out-of-distribution ones, yet consistently preserves relative policy rankings. This suggests that even imperfect world models can serve as reliable pre-deployment evaluators when real-world testing is expensive or risky.

Building on this idea, Li et al. [62] propose WorldEval, an automated framework that turns a video generation model into a policy evaluator through Policy2Vec, a latent action representation derived directly from the policy network. By injecting these policy-specific embeddings into a pre-trained video model, WorldEval generates policy-following videos that reveal the quality of the underlying controller. Experiments across multiple manipulation tasks demonstrate that WorldEval achieves a strong correlation with real-world success rates, reporting Pearson coefficients above 0.94 and stable ranking consistency. Furthermore, WorldEval extends to safety detection, as collapsed or unsafe policies produce unrealistic or degenerate video outputs, enabling early detection of hazardous behaviors before deployment.

On the other hand, Shang et al. [93] propose RoboScape, a physics-informed embodied world model that augments video generation with temporal depth prediction and adaptive keypoint dynamics learning. By incorporating geometric and motion constraints into the training process, RoboScape addresses common failure modes of purely RGB-based models, such as implausible object deformations or discontinuous motions. Experiments demonstrate that RoboScape not only improves visual and physical fidelity but also achieves high alignment with ground-truth simulators in policy evaluation, yielding a Pearson correlation above 0.95. This underscores the importance of embedding physical priors to ensure reliable evaluation.

Together, these studies show that reliable policy evaluation cannot be achieved by surface-level visual quality alone. Instead, benchmarks must assess both the stability of policy rankings and the extent to which generated rollouts respect physical constraints. Moving forward, evaluation protocols should integrate rank consistency, Pearson correlation, and bias analysis (under- vs. overestimation) with physics-aware modeling objectives. Such practices will help ensure that embodied world models are not only visually convincing but also trustworthy evaluators of robotic policies.

6.4 Data Scaling in Downstream Policy Model

Another key perspective for benchmarking embodied world models is their ability to scale policy learning. In this setting, the world model is treated as a surrogate environment or data generator, and its utility is reflected in improvements in sample efficiency, generalization, and adaptation as the scale of generated data grows. Figure 14 depicts how benchmarks assess the impact of data scaling, where world models generate datasets of different sizes to train downstream policies, and improvements in efficiency, generalization, and transferability are measured.

Recent advances demonstrate multiple pathways through which data scaling enhances downstream policy performance. One representative direction is the use of controllable video world models for robotic data augmentation. Jang et al. propose DreamGen [52], which leverages generative video models to produce diverse yet physically plausible rollouts, enabling agents to acquire robust policies that generalize across tasks and environments. Extending this idea, RoboTransfer [67] introduces a geometry-consistent video diffusion model to improve sim-to-real transfer, leading to higher success rates in real-world robotic deployments. Beyond trajectory augmentation, world models are also utilized as task generators. GenSim [106] employs large language models to synthesize novel

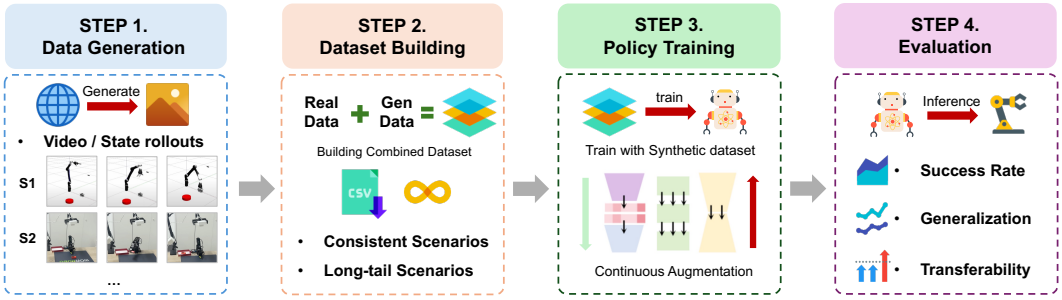


Figure 14: Workflow of benchmarking data scaling in downstream policy models. World models generate datasets of varying scales, which are then used to train policies, and improvements are measured in efficiency, generalization, and transferability.

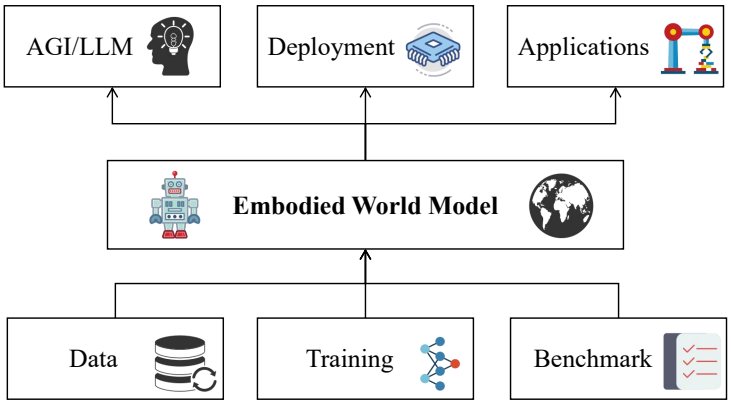


Figure 15: Challenges and Future Directions in Embodied World Models

robotic tasks and simulation scenarios, thereby broadening the training distribution and improving robustness under unseen conditions.

Another line of work investigates multimodal and language-augmented world models. WorldGPT [32] aligns large language models with multimodal scene representations, allowing agents to benefit from world models not only in perception but also in reasoning about task semantics. In trajectory-centric domains, Traj-LLM [58] demonstrates that language-pretrained models can augment world models for trajectory prediction, thereby assisting planning modules to acquire more accurate and generalizable strategies. In addition, RoboScape [93] shows that embedding physical priors into rollouts improves both sample efficiency and transferability when scaling to robotic policy learning.

In summary, data scaling serves as a rigorous perspective for evaluating embodied world models: the extent to which a model can effectively exploit large-scale synthetic or multimodal data to accelerate learning, enhance generalization, and facilitate deployment determines its practical value. However, challenges remain in ensuring that scaled data respects physical laws, avoids distribution collapse, and balances diversity with fidelity.

7 Challenges and Future Works

7.1 Effective Data Collection

Generative world models in AI are systems that learn to simulate and predict real-world dynamics, often used in reinforcement learning (RL), robotics, or video generation. However, the data collection for training the world model itself is also quite a challenge. Although there are a very large number of Internet videos, few of them are specified for the embodied tasks and thus hard to use when training world models [45]. The primary issue is the scarcity of high-quality, robot-relevant data tailored to embodied tasks, where models require sequential, action-conditioned datasets that

capture physical interactions, sensorimotor feedback, and environmental transitions—elements that are expensive and labor-intensive to gather in real-world settings, often resulting in datasets orders of magnitude smaller than those used for language or vision models (e.g., tens of thousands of hours versus billions) [68]. Bias and underrepresentation in training data further exacerbate challenges, as skewed datasets—often drawn from limited sources—can perpetuate unfair outcomes or fail to cover diverse scenarios, limiting adaptability across robotic domains [82]. That is, the data collection is still the major direction in the future for better and stronger embodied world models.

7.2 Effective Architecture Designs and Causal Training

Another significant challenge in building generative world models for embodied robotics lies in devising effective architecture designs that incorporate causal training, which refers to structuring neural networks to learn and represent cause-effect relationships rather than mere statistical correlations, enabling models to predict the outcomes of interventions and counterfactual scenarios in dynamic physical environments. Some potential approaches for addressing this challenge are as follows. First, we can incorporate causal structures directly into model architectures. For example, it may enhance the utility to employ causal transformers [76] (e.g., with causal masks) for sequential reasoning in motion planning, focusing on relevant features while embedding cause-effect priors. Second, some hybrid architectures with generation and action is also promising. For example, it can enable differentiable planning and causal simulation of trajectories, by building modular VLAs or JEPA-based planners that condition on prompts for latent forecasting. Third, the causal relation is actually a kind of physical regulation, and the method which can embed the physical laws are also a potential solution. For example, one recent work [93] embed physical priors into generative models for joint RGB video generation and physics prediction, using unified architectures to ensure causal consistency. In conclusion, it is promising to explore more on the causal architecture, hybrid modules, physical laws, etc.

7.3 Effective Benchmark Construction

Another limitation in current embodied world model research is the absence of effective benchmark construction, which refers to the lack of comprehensive, standardized evaluation frameworks tailored specifically to assess the unique capabilities and requirements of generative world models in embodied AI systems, such as those used in robotics for simulating real-world dynamics, predicting action outcomes, and enabling long-horizon planning [122]. This challenge arises because existing benchmarks often focus on narrow aspects like visual perception or basic navigation, failing to holistically evaluate critical elements such as physical realism, motion dynamics, semantic alignment with tasks, causal reasoning, and generalization across diverse environments, leading to inconsistent comparisons between models and hindered progress in identifying true advancements or weaknesses. To solve it, there are some promising future directions. First, it is essential to well bridge the gap between the simulation and the real deployment. This calls for the development of more easy-to-use and widely-used sim2real toolkits and systems, such as constructing benchmarks with high-fidelity digital twins and physics engines [14]. Second, for the embodied task itself, the assessment and evaluation are already very diverse, leading to various evaluation methods for the embodied world model. The researchers should define the metrics of the embodied world model, based on the critical abilities to support the embodied agents, rather than the perspective of AI-based content generation [21]. Third, it is promising to involve more embodied agents combining both LLM/VLM and traditional methods [30], to autonomously interact with the simulation engine and generated world model, to collect benchmark data. The agents should be empowered with adaptive policy to detect the weakness of the existing embodied world model, and consider more corner cases during the data collection process. In conclusion, the future work should consider more on sim2real gap, task-based evaluation, autonomous collection, etc.

7.4 Relation with Large Language Models

The advancement of large models serves as a key catalyst for the current development of embodied intelligence, with early work demonstrating embodied reasoning and task-solving based on the internal world models of LLMs [50]. Although current research on embodied world models primarily focuses on a generative perspective [69, 72, 64], enhancing their capabilities from an understanding-centric viewpoint has become an equally critical direction [20]. The necessity for this focus is

particularly highlighted by findings that many existing multi-modal large language models exhibit significant deficiencies in core capabilities such as fine-grained visual understanding and spatial reasoning [105, 109]. Therefore, how to effectively improve the spatial understanding and reasoning abilities of these models [110], and how to translate these advancements in understanding into substantial improvements in generative performance [66], have emerged as key open questions. Addressing these issues will be instrumental in directly promoting the progress of embodied world models.

7.5 Real-world Deployment and Applications

The real-world deployment and application of embodied world models represent a critical yet relatively overlooked aspect of technological development, with key challenges in safety, generalization, and efficiency standing as major bottlenecks to their maturation and practical use. Regarding safety, current models often function as black-box controllers [69, 72, 64], making it difficult to ensure reliability in extreme or anomalous situations; consequently, investigating effective safety strategies to construct inherently safe embodied world models is a crucial research direction. In terms of generalization, the performance of existing models remains significantly insufficient, heavily depending on the diversity of training data [82]. While collecting more varied data is a viable short-term strategy, the long-term solution lies in developing robust learning mechanisms that enable models to grasp fundamental principles for true generalization. Concerning efficiency, the reliance on large models results in prohibitively large parameter sizes, low inference efficiency, high operational costs, and unacceptable latency, making it essential to significantly improve their generation and inference speed to render them practical and widely deployable.

7.6 Towards Universal and Cross-Scale Physical World Models

While current world models have found significant applications in virtual domains such as gaming [4, 35] and are gaining traction in real-world scenarios like autonomous driving [78, 134] and embodied intelligence [91, 93, 133], the field remains fragmented, where models for different domains are largely independent. This poses a significant challenge for achieving a universal model. For instance, a driving world model primarily focuses on large-scale dynamics and external environmental changes, whereas an embodied world model emphasizes the fine-grained details of object manipulation. Another critical challenge lies in the heterogeneity of action control formats. Different world models require varying types of action instructions, including text commands [1], motion trajectories [7], and discrete joint states [93]. Unifying these diverse action spaces to control a single generative model is a challenging research problem. Looking ahead, we aim to develop a domain-agnostic physical world model that can generate and simulate different scenarios, from large-scale locomotion to fine-grained manipulation. Such a model would represent a major step towards a generalized embodied AI capable of understanding and interacting with our complex, multi-scale world.

8 Conclusion

In this survey, we systematically review the rapid progress in embodied world models, a field critical for the future of artificial general intelligence. To navigate this complex landscape, we introduced a novel technical taxonomy that provides a clear, structured framework. This framework organizes the field across four core dimensions: model architectures, training methodologies, application scenarios, and evaluation approaches. We first detailed vision-based generative world models and their latent-space counterparts, along with distinct training paradigms. We then explored their diverse applications, from serving as scalable cloud simulators to acting as on-device robot brains. We also summarized key evaluation dimensions for effective benchmarking. Ultimately, this survey distills complex research into a clear, navigable guide. We conclude by outlining key challenges and promising future directions, hoping to inspire further innovation in this vital domain.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model

platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[4] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025.

[5] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

[6] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

[7] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[8] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[10] Rodney A Brooks. New approaches to robotics. *Science*, 253(5025):1227–1232, 1991.

[11] 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.

[12] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[13] Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

[14] Peng Chang and Taşkin Padif. Sim2real2sim: Bridging the gap between simulation and real-world in flexible object manipulation. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 56–62. IEEE, 2020.

[15] Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025.

[16] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

[17] Zixuan Chen, Jing Huo, Yangtao Chen, and Yang Gao. Robohorizon: An llm-assisted multi-view world model for long-horizon robotic manipulation. *arXiv preprint arXiv:2501.06605*, 2025.

[18] DeepMind. Genie 2: A large-scale foundation world model. DeepMind Discover blog, December 2024.

[19] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

[20] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2024.

[21] 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.

[22] Jonathan St BT Evans. *Hypothetical thinking: Dual processes in reasoning and judgement*. Psychology Press, 2007.

[23] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

[24] Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

[25] Yunhai Feng, Nicklas Hansen, Ziyan Xiong, Chandramouli Rajagopalan, and Xiaolong Wang. Finetuning offline world models in the real world. *arXiv preprint arXiv:2310.16029*, 2023.

[26] Jay W Forrester. Counterintuitive behavior of social systems. *Theory and decision*, 2(2):109–140, 1971.

[27] Maria Fox and Derek Long. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124, 2003.

[28] Gene F Franklin, J David Powell, Abbas Emami-Naeini, and J David Powell. *Feedback control of dynamic systems*, volume 4. Prentice hall Upper Saddle River, 2002.

[29] Ao Fu, Yi Zhou, Tao Zhou, Yi Yang, Bojun Gao, Qun Li, Guobin Wu, and Ling Shao. Exploring the interplay between video generation and world models in autonomous driving: A survey. *arXiv preprint arXiv:2411.02914*, 2024.

[30] 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.

[31] Shenyuan Gao, Jiazhi 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. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

[32] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355, 2024.

[33] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

[34] Jun Guo, Xiaojian Ma, Yikai Wang, Min Yang, Huaping Liu, and Qing Li. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *arXiv preprint arXiv:2505.10075*, 2025.

[35] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

[36] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[37] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[38] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[39] 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.

[40] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[41] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[42] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

[43] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

[44] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

[45] Yingdong Hu, Fanqi Lin, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *1st Workshop on X-Embodiment Robot Learning*, 2024.

[46] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[47] Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. Simulating the real world: A unified survey of multimodal generative models. *arXiv preprint arXiv:2503.04641*, 2025.

[48] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

[49] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

[50] 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. In *Conference on Robot Learning*, pages 1769–1782. PMLR, 2023.

[51] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[52] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025.

[53] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

[54] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

[55] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

[56] Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

[57] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[58] Zhengxing Lan, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren, and Zhiyong Cui. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Transactions on Intelligent Vehicles*, 2024.

[59] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[60] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

[61] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[62] Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

[63] Ying Li, Xiaobao Wei, Xiaowei Chi, Yuming Li, Zhongyu Zhao, Hao Wang, Ningning Ma, Ming Lu, and Shanghang Zhang. Manipdreamer: Boosting robotic manipulation world model with action tree and visual guidance. *arXiv preprint arXiv:2504.16464*, 2025.

[64] Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large model empowered embodied ai: A survey on decision-making and embodied learning. *arXiv preprint arXiv:2508.10399*, 2025.

[65] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

[66] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[67] Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiaxiong Qiu, Zheng Zhu, Guan Huang, and Zhizhong Su. Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. *arXiv preprint arXiv:2505.23171*, 2025.

[68] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6875–6885, 2025.

[69] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

[70] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[71] Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4d video generation for robot manipulation. *arXiv preprint arXiv:2507.01099*, 2025.

[72] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025.

[73] Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation. *arXiv preprint arXiv:2508.17600*, 2025.

[74] Gerrit W Maus, Jason Fischer, and David Whitney. Motion-dependent representation of space in area mt+. *Neuron*, 78(3):554–562, 2013.

[75] Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. Genrl: Multimodal-foundation world models for generalization in embodied agents. *Advances in neural information processing systems*, 37:27529–27555, 2024.

[76] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pages 15293–15329. PMLR, 2022.

[77] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

[78] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15522–15533, 2024.

[79] Chaojun Ni, Jie Li, Haoyun Li, Hengyu Liu, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Boyuan Wang, Chenxin Li, Guan Huang, et al. Wonderfree: Enhancing novel view quality and cross-view consistency for 3d scene exploration. *arXiv preprint arXiv:2506.20590*, 2025.

[80] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025.

[81] OpenAI. Sora: Creating video from text. `https://openai.com/sora`, 2024. (Accessed on 06/05/2025).

[82] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[83] Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling. *arXiv preprint arXiv:2502.00622*, 2025.

[84] 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.

[85] Julian Quevedo, Percy Liang, and Sherry Yang. Evaluating robot policies in a world model. *arXiv preprint arXiv:2506.00613*, 2025.

[86] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[87] Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 9:153171–153187, 2021.

[88] Jürgen Schmidhuber. *Making the world differentiable: on using self supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments*, volume 126. Inst. für Informatik, 1990.

[89] Jürgen Schmidhuber. An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In *1990 IJCNN international joint conference on neural networks*, pages 253–258. IEEE, 1990.

[90] Jürgen Schmidhuber. Reinforcement learning in markovian and non-markovian environments. *Advances in neural information processing systems*, 3, 1990.

[91] Yu Shang, Lei Jin, Yiding Ma, Xin Zhang, Chen Gao, Wei Wu, and Yong Li. Roboscape-long: An efficient world model for long-horizon embodied video generation. `https://github.com/tsinghua-fib-lab/Roboscape-long`, 2025.

[92] Yu Shang, Yangcheng Yu, Xin Zhang, Wei Wu, and Yong Li. Moewm: Learning composable world models for embodied action planning. `https://github.com/tsinghua-fib-lab/MoE-WM`, 2025.

[93] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

[94] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025.

[95] Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

[96] Yinzhou Tang, Yu Shang, Yinuo Chen, Bingwen Wei, Chao Yu, Wei Wu, and Yong Li. Roboscape-r: Reinforcement learning with embodied world models for policy learning. `https://github.com/tsinghua-fib-lab/RoboScape-R`, 2025.

[97] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[98] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

[99] Yan Team. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025.

[100] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[101] Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498*, 2025.

[102] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer*, pages 23–65. Springer, 2007.

[103] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[104] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[105] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

[106] 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.

[107] Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning real-world action-video dynamics with heterogeneous masked autoregression. *arXiv preprint arXiv:2502.04296*, 2025.

[108] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024.

[109] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24669–24679, 2025.

[110] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[111] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[112] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37:41051–41075, 2024.

[113] WorldLabs. Generating worlds. `https://www.worldlabs.ai/blog`, 2024.

[114] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[115] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

[116] Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025.

[117] 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.

[118] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

[119] 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.

[120] Ningwei Xie, Zizi Tian, Lei Yang, Xiao-Ping Zhang, Meng Guo, and Jie Li. From 2d to 3d cognition: A brief survey of general world models. *arXiv preprint arXiv:2506.20134*, 2025.

[121] Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025.

[122] Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024.

[123] Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Xin Jin, Hang Zhao, et al. Orv: 4d occupancy-centric robot video generation. *arXiv preprint arXiv:2506.03079*, 2025.

[124] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[125] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

[126] Tengbo Yu, Guanxing Lu, Zaijia Yang, Haoyuan Deng, Season Si Chen, Jiwen Lu, Wenbo Ding, Guoqiang Hu, Yansong Tang, and Ziwei Wang. Manigaussian++: General robotic bimanual manipulation with hierarchical gaussian world model. *arXiv preprint arXiv:2506.19842*, 2025.

[127] Yang Yu. Towards sample efficient reinforcement learning. In *IJCAI*, pages 5739–5743, 2018.

[128] Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, et al. Lumos-1: On autoregressive video generation from a unified model perspective. *arXiv preprint arXiv:2507.08801*, 2025.

[129] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

[130] Kaidong Zhang, Pengzhen Ren, Bingqian Lin, Junfan Lin, Shikui Ma, Hang Xu, and Xiaodan Liang. Pivot-r: Primitive-driven waypoint-aware world model for robotic manipulation. *Advances in Neural Information Processing Systems*, 37:54105–54136, 2024.

[131] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

[132] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Jingkai Sun, Jiahang Cao, Jiaxu Wang, et al. Occupancy world model for robots. *arXiv preprint arXiv:2505.05512*, 2025.

[133] Baining Zhao, Rongze Tang, Mingyuan Jia, Ziyou Wang, Fanghang Man, Xin Zhang, Yu Shang, Weichen Zhang, Chen Gao, Wei Wu, et al. Airscape: An aerial generative world model with motion controllability. *arXiv preprint arXiv:2507.08885*, 2025.

[134] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12015–12026, 2025.

[135] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[136] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[137] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[138] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[139] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

[140] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

[141] 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.

[142] 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.

[143] Shaobin Zhuang, Zhipeng Huang, Ying Zhang, Fangyikang Wang, Canmiao Fu, Binxin Yang, Chong Sun, Chen Li, and Yali Wang. Video-gpt via next clip diffusion. *arXiv preprint arXiv:2505.12489*, 2025.

[144] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6772–6781, 2025.