Article

Not peer-reviewed version

Learning to Model the World: A Survey of World Models in Artificial Intelligence

Jiahua Dong †, Qi Lyu †, Baichen Liu *, Xudong Wang, Wenqi Liang, Duzhen Zhang, Jiahang Tu, Hongliu Li, Hanbin Zhao, Henghui Ding, Yulun Zhang, Zhi Han *, Nicu Sebe, Fahad Shahbaz Khan, Salman Khan, Mubarak Shan, Philip Torr, Ming-Hsuan Yang, Dacheng Tao

Posted Date: 10 March 2026

doi: 10.20944/preprints202603.0739.v1

Keywords: world models; simulation; planning; decision-making; general-purpose intelligent systems

 doi:10.20944/preprints202603.0739.v1

Article

Learning to Model the World: A Survey of World Models in Artificial Intelligence

Jiahua Dong [1,†], Qi Lyu [2,†], Baichen Liu [2,*], Xudong Wang [2], Wenqi Liang [3], Duzhen Zhang [1], Jiahang Tu [4], Hongliu Li [5], Hanbin Zhao [4], Henghui Ding [6], Yulun Zhang [7], Zhi Han [2,*], Nicu Sebe [3], Fahad Shahbaz Khan [1], Salman Khan [1], Mubarak Shah [8], Philip Torr [9], Ming-Hsuan Yang [10] and Dacheng Tao [11]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirate
[2] State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
[3] University of Trento, Trento, Italy
[4] Zhejiang University, Hangzhou, China
[5] Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong, China
[6] Fudan University, Shanghai, China
[7] Shanghai Jiao Tong University, Shanghai, China
[8] Center for Research in Computer Vision, University of Central Florida, Orlando, USA
[9] University of Oxford, Oxford, United Kingdom
[10] University of California at Merced, Merced, USA, and also with the Google, Mountain View, USA
[11] Nanyang Technological University, Singapore
* Correspondence: liubaichen@sia.cn (B.L.); hanzhi@sia.cn (Z.H.)
† These authors contributed equally to this work.

Abstract

World models (WMs) provide a unified approach for modeling how environments evolve over time by learning predictive representations of states and observations. Recent advances in large-scale generative modeling and multimodal foundation models have substantially broadened their applicability across a wide range of interactive and multimodal domains; however, existing research remains fragmented across modeling paradigms, application domains, and evaluation protocols. This survey provides a systematic and in-depth review of WMs in artificial intelligence. Based on the world modeling paradigms of existing methods, we first categorize WMs into four branches with formal mathematical formulations: observation-level generative, latent space, reinforcement learning-based, and object-centric WMs. We further review a broad range of WM applications spanning robotics, autonomous driving, scientific discovery, game simulation, GUI-based agents, as well as interpretability and trustworthiness, and analyze benchmarks, new evaluation metrics, simulation platforms, and comparative results across WMs. Finally, we discuss key challenges, including long-horizon consistency, and generalization, and outline promising directions for future research. This survey provides an actively updated GitHub Repository to track developments in WMs and aims to offer a unified reference for understanding, comparing, and advancing WMs.

Keywords: world models; simulation; planning; decision-making; general-purpose intelligent systems

1. Introduction

World models (WMs) [1–5] have emerged as a fundamental mechanism for building intelligent systems capable of simulation, planning, and decision-making in complex environments [6]. Instead of reacting solely to instantaneous observations, a world model seeks to learn internal representations of the environment that capture its underlying structure [7,8], temporal dynamics [9–11], and uncertainty [12,13]. Such an internal model can predict future states, simulate alternative outcomes, and evaluate candidate actions without requiring direct interaction with the real world [14], which is particularly important in scenarios where data collection is expensive, limited, or safety-critical

 doi:10.20944/preprints202603.0739.v1

[15,16]. Meanwhile, the limitations of purely model-free methods [17], including low sample efficiency and poor generalization, have further motivated renewed interest in WMs [18,19]. Thus, WMs are increasingly viewed as general-purpose intelligent systems that bridge perception, cognition, and control [20].

The development of WMs has progressed through several stages [21], driven by advances in representation learning [8], dynamics modeling [22], and large-scale training [23]. Early classical model-based approaches, such as Gaussian processes [24,25] and control-oriented dynamics [26], rely on explicit or low-dimensional dynamics with strong assumptions about environment structure, which limits their scalability to complex real-world scenarios. Recent advances in deep learning [27,28], especially in large-scale generative modeling such as video prediction [29–31] and multimodal foundation models [32,33], have significantly accelerated research on WMs. Modern WMs [34–36] increasingly adopt latent representations to model high-dimensional observations, enabling a clearer separation between perception [37] and dynamics prediction [38]. Meanwhile, transformer frameworks [39] and slot encoding [40] are employed to learn interpretable object-centric embeddings, enabling WMs to model latent dynamics through interacting entities [40]. This shift lays the foundation for neural WMs and broadens their applicability to visual and continuous-control domains [41,42].

Based on their modeling mechanisms and intended usage, as shown in Figure 1, most existing WMs [43–47] can be mainly categorized into four categories. **1)** Observation-level generative WMs directly predict future visual [48], language [42], 3D or 4D observations [49] from past environmental observations and actions in high-dimensional spaces, typically using autoregressive [50,51], NeRF [52] or diffusion-based generative architectures [47,53]. **2)** Latent embedding-based WMs [7,54] first encode high-dimensional observations into compact latent representations. They then learn transition dynamics in the latent space [32,55,56], allowing efficient prediction, and decision-making without operating on raw observations. **3)** Reinforcement learning (RL)-based WMs [9,46,57,58] learn a predictive model of environment dynamics from agent-environment interactions. This learned model then performs planning or policy optimization through imagined rollouts [59], reducing the need for direct real-world interaction. **4)** Object-centric WMs [45,60] represent the environment as a collection of discrete objects with structured attributes. They learn object-level dynamics and relational dependencies [61], enabling compositional reasoning and generalization across scenes and tasks.

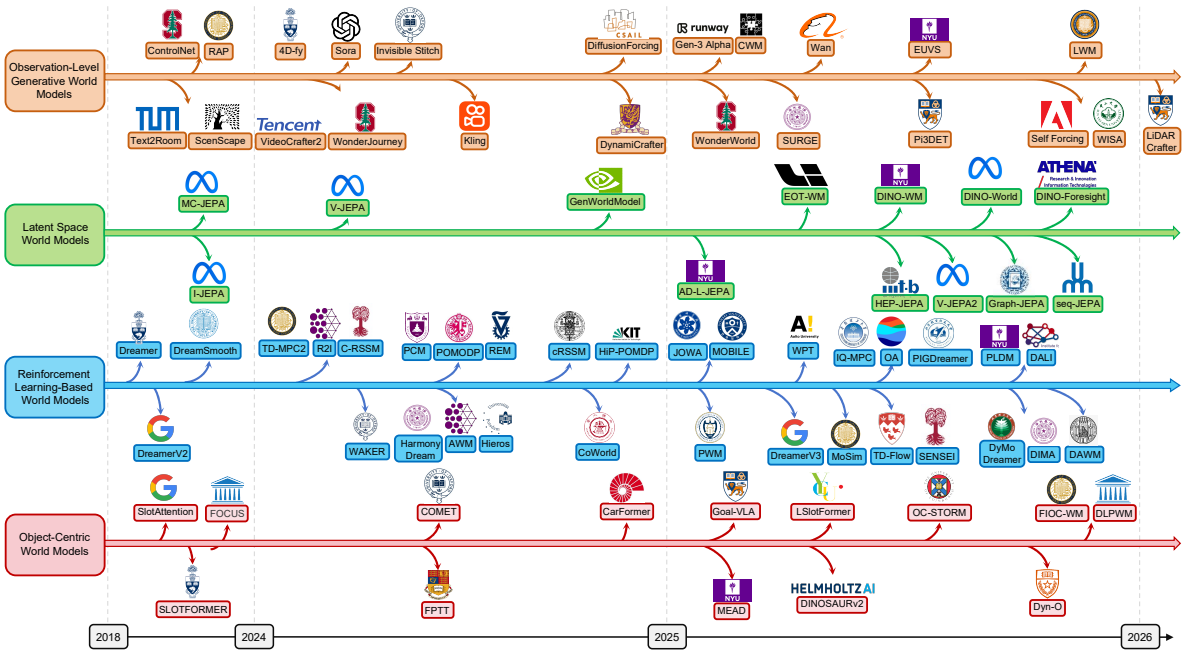

Figure 1. The recent timeline of foundational world models (WMs), covering core methodologies across different categories.

 doi:10.20944/preprints202603.0739.v1

Over the past few years, WMs [62–65] have played an increasingly important role across a broad range of application domains, including robotics [66], autonomous driving [67], virtual game simulation [68], GUI-based agents [69], and scientific discovery [70]. For example, in robotics and autonomous driving, they are used to simulate future scenarios [71], predict the behavior of surrounding agents [72] or physical systems [66], and support risk-aware planning [73] and decision-making [6], thereby improving robustness in uncertain environments. Beyond control-oriented applications, WMs also excel at modeling physical [56], biological [70], and medical treatment [74], accelerating scientific discovery through simulation and hypothesis testing. Evidently, they not only advance the development of intelligent systems, but also contribute to broader societal benefits, including safer transportation, more capable robots, and faster scientific innovation [6].

Contributions: Existing surveys on WMs can be broadly grouped into two branches, as summarized in Table 1. The first branch of surveys largely restricts its scope to specific application domains, including autonomous driving [75–80], 3D/4D modeling [81,82], video generation [84,85], safety [83], and embodied intelligence [86–89], particularly in vision-language manipulation and navigation [90,91]. The second branch of surveys [92–94] primarily focuses on the basic definitions of WMs, as well as on major application domains of WMs. Such studies are often framed as literature overviews or introductory conceptual discussions. However, there is a lack of in-depth and systematic analyses of world modeling paradigms, simulation platforms, comparative experiments across WMs, and interpretability, as well as new evaluation metrics for measuring the generalization, causal reasoning, and long-horizon consistency of WMs. This survey differentiates itself from prior works and makes the following major contributions to the literature:

- Instead of focusing on specific applications or introductory conceptual discussions, we offer an in-depth and systematic overview of world modeling paradigms, methodologies, key functions, and their relationships.
- We summarize the key evolutionary developments of existing major WMs and their core mathematical formulations across branches from a broader perspective.
- Beyond concentrating solely on commonly studied application domains, we comprehensively review all application areas of world models explored to date. These domains include robotics, autonomous driving, scientific discovery, virtual game simulation, GUI-based agents, as well as interpretability and trustworthiness.
- We provide a more thorough and inclusive overview of benchmark datasets, evaluation metrics, simulation platforms, and comparative experiments across WMs.

 doi:10.20944/preprints202603.0739.v1

Table 1. Key differences between existing surveys and our survey.

Survey	Scope	Characteristic
[75,76] [77,78] [79,80]	Autonomous Driving	Focus solely on application-level discussions in autonomous driving.
[81,82]	3D/4D Modeling	Focus solely on application-level discussions in 3D and 4D modeling.
[83]	Safety	Focus solely on safety discussions.
[84,85]	Video Generation	Focus solely on application-level discussions in video generation.
[86,87] [88,89] [90,91]	Embodied Intelligence	Focus solely on application-level discussions in embodied intelligence, particularly in manipulation and navigation.
[92,93] [94]	General WMs	Introductory definitions of WMs, with limited systematic analyses of modeling paradigms, a narrow application scope, and a lack of new metrics and experimental evaluation.
Ours	General WMs	In-depth and systematic analyses of world modeling paradigms, broader applications, simulation platforms, comparative results across WMs, and interpretability.

Structure of This Survey: The overall structure of this survey is introduced in Figure 2. Specifically, Section 2 introduces four major paradigms of WMs: including observation-level generative WMs (Section 2.1), latent space WMs (Section 2.2), RL-based WMs (Section 2.3), and object-centric WMs (Section 2.4). Section 3 reviews application domains of WMs, including robotics (Section 3.1), autonomous driving (Section 3.2), scientific discovery (Section 3.3), virtual game simulation (Section 3.4), GUI-based agents (Section 3.5), as well as interpretability and trustworthiness (Section 3.6). Section 4 summarizes benchmark datasets & evaluation metrics (Section 4.1), simulation platforms (Section 4.2), and comparative performance (Section 4.3). Finally, Section 5 discusses key challenges that limit the deployment and generalization of WMs and outlines future directions emphasizing unified scientific modeling, causal reasoning, scalable training, and real-world validation.

 doi:10.20944/preprints202603.0739.v1

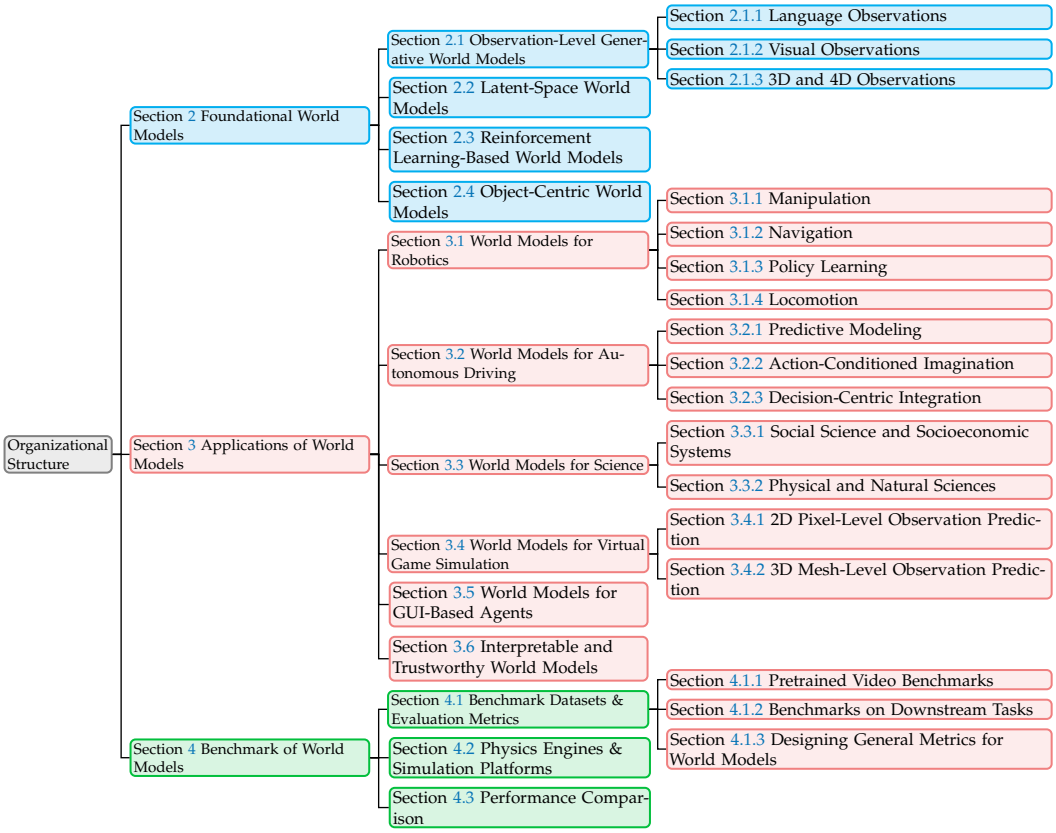

Figure 2. The overall structure layout of the survey.

2. Foundational World Models

WMs [3,9] aim to learn latent environment dynamics that capture the current state of the world and predict its evolution over time, enabling simulation, planning, and decision-making. Generally, at time t, a WM can be viewed as a function that leverages historical states, along with optional agent actions, to predict future environment states:

$$\mathbf{z}_{t+1} \sim \pi_\theta(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{a}_t, \mathbf{o}_{t+1}), \tag{1}$$

where $\pi_\theta(\cdot)$ denotes the WM. $\mathbf{z}_t$ and $\mathbf{z}_{t+1}$ are the latent environment states at time t and $t+1$, respectively. $\mathbf{a}_t$ denotes the optional agent action at time t, and $\mathbf{o}_{t+1}$ represents the observation at time $t+1$. Building upon the formulation in Equation (1), we illustrate the core functions of WMs and the relationships among them, as shown in Figure 3. Specifically, in WMs, simulation generates a set of imagined future trajectories $\{\zeta_i\}_{i=1}^{\phi}$ using learned dynamics, where $\zeta_i^t = \{\eta_i^t\}_{t=1}^{T}$ denotes the i-th predicted trajectory, consisting of T environment states $\{\eta_i^t\}_{t=1}^{T}$. ϕ and T are the numbers of trajectories and time steps, respectively. Here, η_i^t corresponds to $\mathbf{z}_t$ in Equation (1) for simulation. Additionally, planning evaluates or optimizes over these simulated future trajectories to identify desirable trajectory and obtain its action sequences $\{\mathbf{a}_t\}_{t=1}^{T}$, while decision-making selects and executes the final action $\mathbf{a}_T$ based on the planning outcomes.

 doi:10.20944/preprints202603.0739.v1

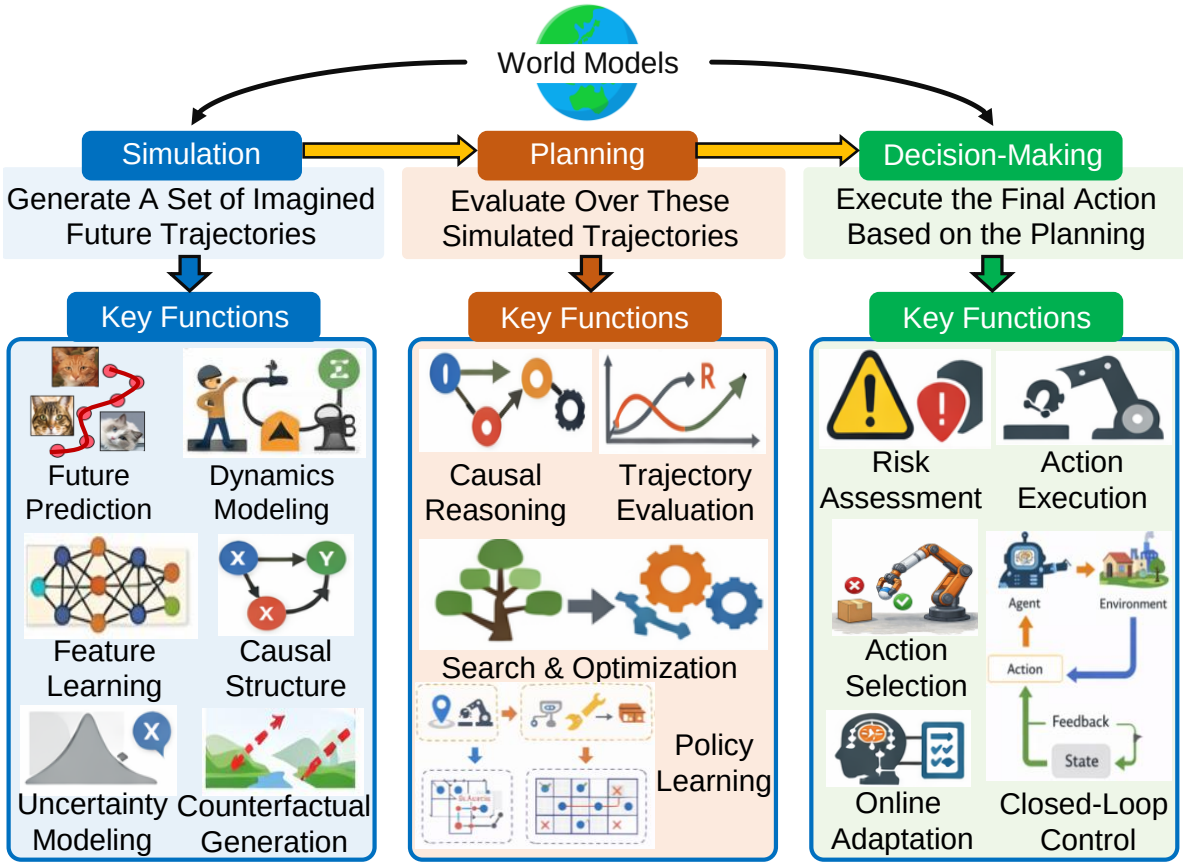

Figure 3. Illustration of the key functions of WMs and their relationships. WMs consist of three consecutive stages: simulation, planning, and decision-making. Each stage incorporates distinct key functions.

From a mathematical modeling standpoint, as summarized in Table 2, we roughly classify most existing WMs into four categories, a taxonomy that differs significantly from the simple conceptual introductions used in existing surveys [75,79,83,85–87]. **1) Observation-level generative WMs** decode latent environment states into future observation-level predictions, with a primary focus on simulation capabilities. This branch has achieved high generation quality and semantic consistency through large-scale supervised training. However, observation-level generative methods still face challenges regarding long-horizon stability, controllability, and high reconstruction costs. **2)** Unlike observation-level generative WMs that reconstruct original observations, **latent-space WMs** primarily emphasize simulation and planning capabilities by predicting future latent representations and capturing semantics in a compressed space. Although they avoid dependence on large-scale labeled datasets, such methods often sacrifice interpretability and fine-grained details. **3) RL-based WMs** introduce a reward predictor to learn latent dynamics, primarily supporting planning and decision-making. They can improve sample efficiency yet suffer from real-world distribution shifts and compounding errors, making them more suitable for specific planning tasks rather than modeling a general world. **4) Object-centric WMs** represent environment scenes as sets of interacting entities (object slots) and model object-level dynamics, thereby facilitating better planning and decision-making. They improve interpretability, compositional generalization, and efficient reasoning, but reliable tracking and policy learning are challenging.

 doi:10.20944/preprints202603.0739.v1

Table 2. Comparison of the main focus and modeling paradigms of foundational WMs. **Tag**: S= Simulation, P= Planning, D= Decision-making.

Foundational WMs	Main Focus	Modeling Paradigms
Observation-Level Generative World Models	S	Apply a decoder to map latent representations to observations.
Latent-Space World Models	S P	Model the dynamics of environments in the high-dimensional latent spaces.
Reinforcement Learning-Based World Models	P D	Maximize action rewards using a dynamic model.
Object-Centric World Models	P D	Employ slots to represent the world as a set of object-level embeddings.

2.1. Observation-Level Generative World Models

In general, observation-level generative WMs [29,42,95–97] leverage strong generative priors to model environment dynamics conditioned on actions and additional inputs such as language prompts [42], visual images [18], or camera trajectories [98]. They primarily enable faithful simulation by generating future observations directly at the observation level. Building upon the general definition of WMs in Equation (1), observation-level generative WMs typically introduce a decoder $\Phi_\theta(\cdot)$ that projects latent environment states $\mathbf{z}_{t+1}$, conditioned on agent action $\mathbf{a}_t$ and additional inputs $\mathbf{c}_t$ (*e.g.*, camera motion parameters, depth, language, or visual prompts), to future observation $\widehat{\mathbf{o}}_{t+1}$. Therefore, their objective can be formally formulated as follows:

$$\widehat{\mathbf{o}}_{t+1} = \Phi_\theta(\mathbf{z}_{t+1}), \;\; \mathbf{z}_{t+1} \sim \pi_\theta(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{a}_t, \mathbf{c}_t), \tag{2}$$

where the predicted observation $\widehat{\mathbf{o}}_{t+1}$ at time $t+1$ is a future observation. When performing simulation, observation-level generative WMs synthesize multiple imagined trajectories $\{\zeta_i\}_{i=1}^{\phi}$, each ζ_i including a sequence of observation-level environment states $\{\widehat{\mathbf{o}}_i\}_{i=1}^{T}$ over T time steps.

Contemporary generative models, such as large language models (LLMs) [99], diffusion models [53], flow matching models [51], and Gaussian splatting [100], can be viewed as an early and rudimentary form of WMs. Nevertheless, as formulated in Equation (3), they cannot incorporate agent actions that interact with the underlying world state to understand the world, and instead rely solely on additional inputs $\mathbf{c}_t$ to achieve realistic observation-level generation:

$$\widehat{\mathbf{o}}_{t+1} = \Phi_\theta(\mathbf{z}_{t+1}), \;\; \mathbf{z}_{t+1} \sim \pi_\theta(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{c}_t). \tag{3}$$

Therefore, these commonly-used generative models cannot capture the causal dynamics of the environment, causing poor long-horizon simulation stability and limited controllability. At present, there remains a substantial gap in transitioning from Equation (3) to Equation (2) to realize ideal observation-level generative WMs. Based on the type of observations, we categorize this branch of WMs into language (Section 2.1.1), visual (Section 2.1.2), and 3D and 4D observations (Section 2.1.3).

2.1.1. Language Observations

Language observations enable abstract transition modeling but can suffer from brittle long-horizon rollouts when errors compound [101]. Leveraging LLMs (*e.g.*, ChatGPT [99] and Llama [102]), RAP [42] improves decision-making by framing reasoning as planning within a WM and by searching over intermediate outcomes rather than relying on a single forward pass. Reliability can be further improved by injecting explicit structure, including precondition and effect knowledge, as well as causal abstractions for interventions [34,95]. Furthermore, a related line of work studies execution-like rollouts over formal structures; for example, SURGE characterizes boundary cases of surrogate execution in

 doi:10.20944/preprints202603.0739.v1

code-related settings [103]. Context scaling remains central, and long-context architectures extend modeling to million-length video and language sequences [97]. Overall, integrating long-language modeling, structure and planning is crucial for robust and scalable language-level WMs.

2.1.2. Visual Observations

Video generators learn visual observations over plausible futures and serve as strong initializations for world modeling [29,33,44]. In Figure 4, Emu3 [33] exhibits high generative fidelity and temporal coherence by training a unified multimodal transformer with a next-token prediction objective. Recent foundational video models [18,104] demonstrate strong visual fidelity and coherence, making them natural priors for controllable simulators. Wan [104] adopts a LLaVA-style architecture [105] to enhance visual understanding via a low-rank mechanism. To enable better planning and control, recent work adapts video generation models into interactive simulators by injecting actions and enforcing state persistence, with Vid2World [106] providing a general recipe, and real-robot settings revealing robustness challenges under distribution shift [1]. Physics-aware guidance reduces implausible dynamics [2], while explicit memory improves long-horizon consistency under varying viewpoints [44]. CoLA-World [107] aims to advance controllable WMs by jointly learning latent actions and world dynamics. Recent work emphasizes visual realism, while the incorporation of physical and motion-driven mechanisms will help build controllable WMs in the future development.

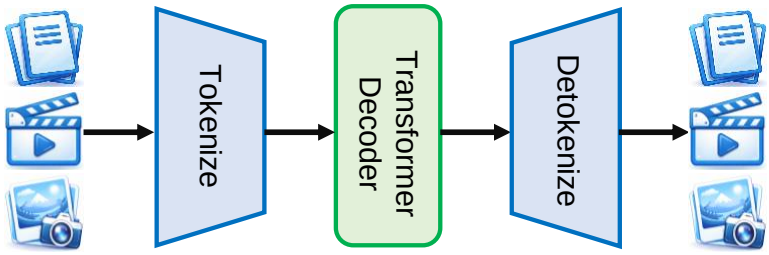

Figure 4. Illustration of Emu3 [33], which unifies the modeling of diverse modalities to enable generative world modeling.

2.1.3. 3D and 4D Observations

While visual observations offer strong appearance fidelity, 3D and 4D observations [98,108,109] expose stable, geometry-grounded states across viewpoint changes. To achieve controllable 3D scene generation, Text2Room [110] employs iterative fusion of multi-view images to directly generate high-quality meshes from textual prompts. Extending this capability to the temporal dimension, 4D-fy [96] alternately integrates supervisory signals from 3D perception models as well as text-to-image and text-to-video models. Furthermore, WonderWorld [19] achieves efficient generation of interactive 3D worlds while maintaining visual quality and geometric consistency. Invisible Stitch [111] addresses geometric inconsistencies in 3D scene generation caused by monocular depth estimation by introducing depth inpainting. PointWorld [112] unifies state and action into a 3D point flow space to forecast full-scene dynamics from a few RGB-D observations. In summary, 3D and 4D WMs can construct high-fidelity and interactive worlds, but still need to improve efficiency and physical consistency.

2.2. Latent-Space World Models

In contrast to observation-level generative WMs [106], latent-space WMs use high-dimensional latent representations to model environmental dynamics, enabling them to better capture underlying semantic structures and motion, thereby primarily supporting simulation and planning. Based on the self-supervised foundational frameworks of JEPA [54] and V-JEPA [55], V-JEPA 2 [7] extends these approaches to construct latent-space WMs, as shown in Figure 5. Specifically, V-JEPA 2 [7] introduces a latent space predictor $\Omega_\theta(\cdot)$ to predict the future latent representation $\widehat{\mathbf{z}}_{t+1}$ at time $t+1$, and then

 doi:10.20944/preprints202603.0739.v1

minimizes the difference between $\hat{\mathbf{z}}_{t+1}$ and the corresponding ground-truth representation $\mathbf{z}_{t+1}$. The major optimization objective of V-JEPA 2 [7] is defined as follows:

$$\hat{\mathbf{z}}_{t+1} \sim \Omega_\theta(\hat{\mathbf{z}}_{t+1}|\mathbf{z}_{1:t}, \mathbf{s}_{1:t}, \mathbf{a}_{1:t}), \tag{4}$$

$$\min \mathcal{L} = \mathbb{E}[\|\hat{\mathbf{z}}_{t+1} - \mathbf{z}_{t+1}\|^2], \tag{5}$$

where $\mathbf{z}_{1:t}$, $\mathbf{s}_{1:t}$, and $\mathbf{a}_{1:t}$ represent the latent representations, the end-effector state sequence (*e.g.*, robotic poses), and the agent action sequence from time 1 to t, respectively.

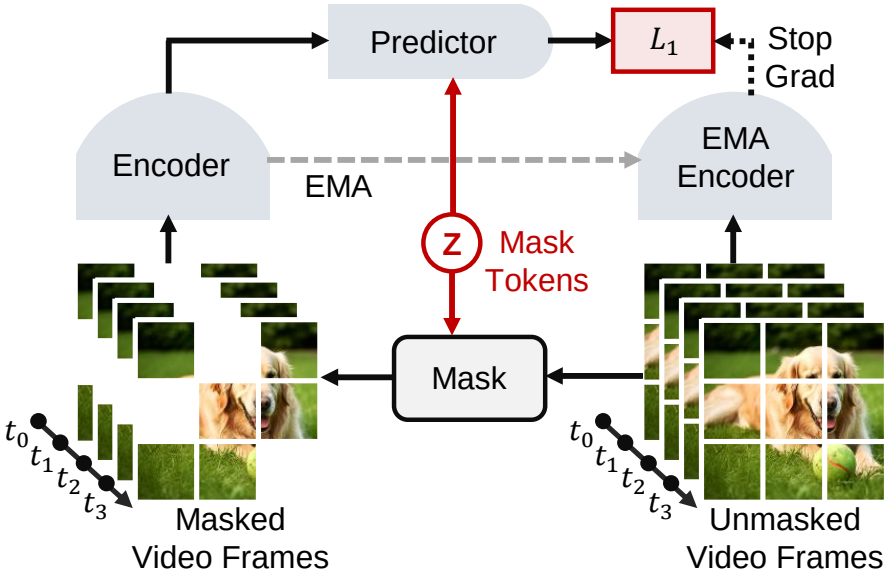

Figure 5. Illustration of V-JEPA 2 [7]. It uses a joint-embedding predictive objective trained on internet-scale data, aligning the predictor's output with targets from an exponential moving average (EWA) encoder.

Following V-JEPA 2 [7], which scales training to over a million hours of video and introduces action conditioning, seq-JEPA [8] disentangles equivariant and invariant features to improve representation distinctness, while MC-JEPA [113] further enhances motion modeling by jointly training optical flow estimation. Another notable trend leverages pretrained visual foundation models. Approaches such as DINO-WM [27], DINO-World [114], and DINO-Foresight [32] use pretrained DINO models to treat spatial patch features as environmental states, enabling effective zero-shot planning and the prediction of evolving semantic features via masked transformers. Additionally, Delliaux *et al.* [115] incorporate geometric priors to further improve transition prediction and representation disentanglement.

2.3. Reinforcement Learning (RL)-Based World Models

RL-based WMs [46,116] use learned dynamics to facilitate planning and policy decision-making. Building upon Equation (1), a reward model $q_\theta(\cdot)$ is leveraged to predict the reward $r_t \sim q_\theta(r_t|\mathbf{z}_t, \mathbf{a}_t)$ at time t. Their objective over T time steps is:

$$\max \mathbb{E}\Big[\sum_{t=1}^{T} q_\theta(r_t|\mathbf{z}_t, \mathbf{a}_t)\Big]. \tag{6}$$

Based on the formulation in Equation (6), the dominant latent-dynamics backbone for RL-based WMs is the recurrent state-space model (RSSM) [66]. As shown in Figure 6, RSSM combines deterministic recurrent states with latent variables to capture uncertainty. Motivated by RSSM [66], Dreamer [9] introduces latent imagination as an alternative to expensive environment interaction by learning policies from imagined rollouts. Moreover, DreamerV2 [10] demonstrates that discrete latents stabilize learning and scale to pixel-based Atari control, while DreamerV3 [3] emphasizes robustness across diverse domains using a single configuration. Subsequent variants [22,117,118] further improve

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2026 doi:10.20944/preprints202603.0739.v1

stability and inductive bias through reward smoothing, loss balancing, dynamic modulation, and the incorporation of privileged information under the safety-oriented settings.

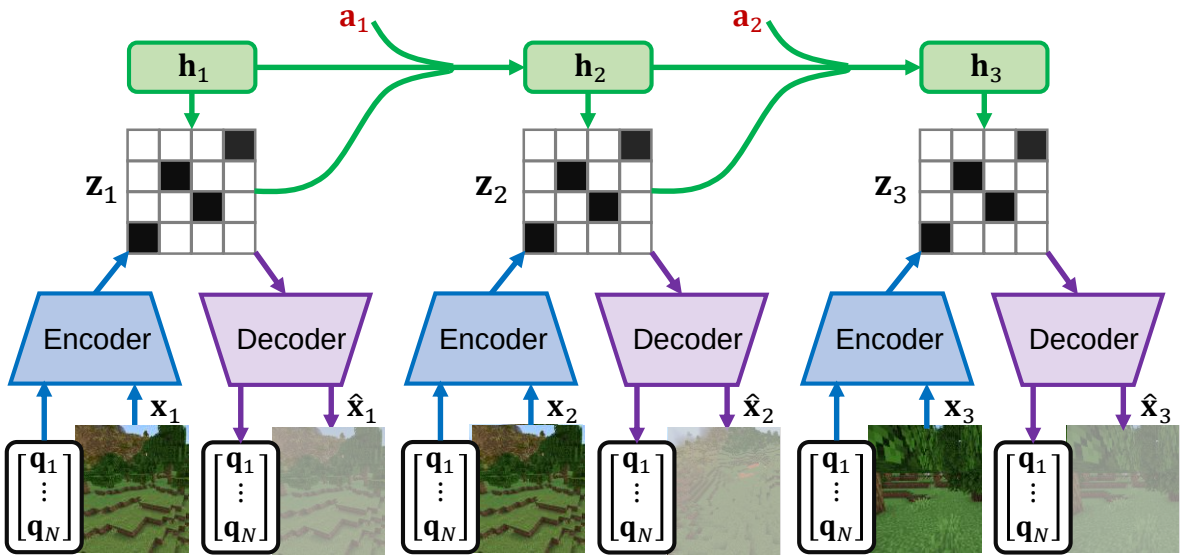

Figure 6. Illustration of DreamerV3 [3]. It mainly consists of a recurrent state-space model (RSSM), which learns environment dynamics by performing state inference in the latent space conditioned on the actions.

While model-based rollouts enable better planning and decision-making capabilities, distribution shift and robustness still remain challenging in closed-loop settings [35,59]. Recent RL-based WMs [119, 120] address these limitations by tightening the model-policy loop through collaborative or policy-conditioned modeling, as well as stabilized objectives for value estimation and planning [59]. However, these models still rely on separate reward models. Some approaches instead train robust world models under reward-free conditions [121,122]. Additionally, token-based world models, such as REM [116], convert environmental observations into discrete token sequences resembling language and utilize sequence models to predict environmental dynamics in latent space. Motivated by these advances, recent generalization trends increasingly emphasize contextual transfer, continual adaptation, and multi-task pretraining [123] guided by foundational models.

2.4. Object-Centric World Models

To address limitations in object tracking and interpretable dynamics, object-centric WMs [16,20, 45,132] use slot attention [40] to represent structured scenes as sets of composable entities (object slots) and to model object-level dynamics with improved interpretability and compositional generalization. Here, slots are a set of learnable embeddings that represent objects in a scene or independent visual concepts. This branch of WMs tends to facilitate planning and decision-making in world modeling. Inspired by slot attention [40], object-centric WMs focus on minimizing the difference between the predicted object slots at future time steps and the corresponding ground-truth [60]:

$$\widehat{\mathbf{h}}_{t+1} \sim \lambda_\theta(\widehat{\mathbf{h}}_{t+1} | \mathcal{A}(\mathbf{o}_t), \mathbf{a}_t), \tag{7}$$

$$\min \mathcal{L} = \mathbb{E}[\mathcal{D}(\widehat{\mathbf{h}}_{t+1}, \mathcal{A}(\mathbf{o}_{t+1}))], \tag{8}$$

where $\mathcal{A}(\cdot)$ represents the slot attention function that maps an observation to a set of slots. $\mathcal{A}(\mathbf{o}_t)$ and $\mathcal{A}(\mathbf{o}_{t+1})$ denote the slot sets at times t and $t+1$, respectively. $\lambda_\theta(\cdot)$ is a slot predictor that predicts the future slot set $\widehat{\mathbf{h}}_{t+1}$ at time $t+1$, and $\mathcal{D}(\cdot)$ denotes the commonly-used distance measure.

As illustrated in Figure 7, SlotFormer [39] encodes sequential observations into object-level representations and predicts future object states by learning visual dynamics. To improve slot quality and controllability, Dyn-O [45] introduces a slot-based world model for Procgen-style games by injecting segmentation priors and disentangling static and dynamic factors. Related work constructs

 doi:10.20944/preprints202603.0739.v1

object-centric dynamics to support efficient exploration, multi-step planning, and improved policy learning [16,61]. Zhang *et al.* show that pixel-level objectives may miss decision-critical small entities and propose an object-centric pipeline that identifies key objects via segmentation and models their dynamics for imagined rollouts [20]. Beyond forward prediction, object factorization also enhances latent action discovery and compositional generalization under compute constraints [133]. In robotics [134] and autonomous driving [135], object-centric representations enable semantic conditioning by aligning object states with language and agent actions [136]. Object-centric representations excel in visual reconstruction, but struggle with downstream tasks such as policy learning [132].

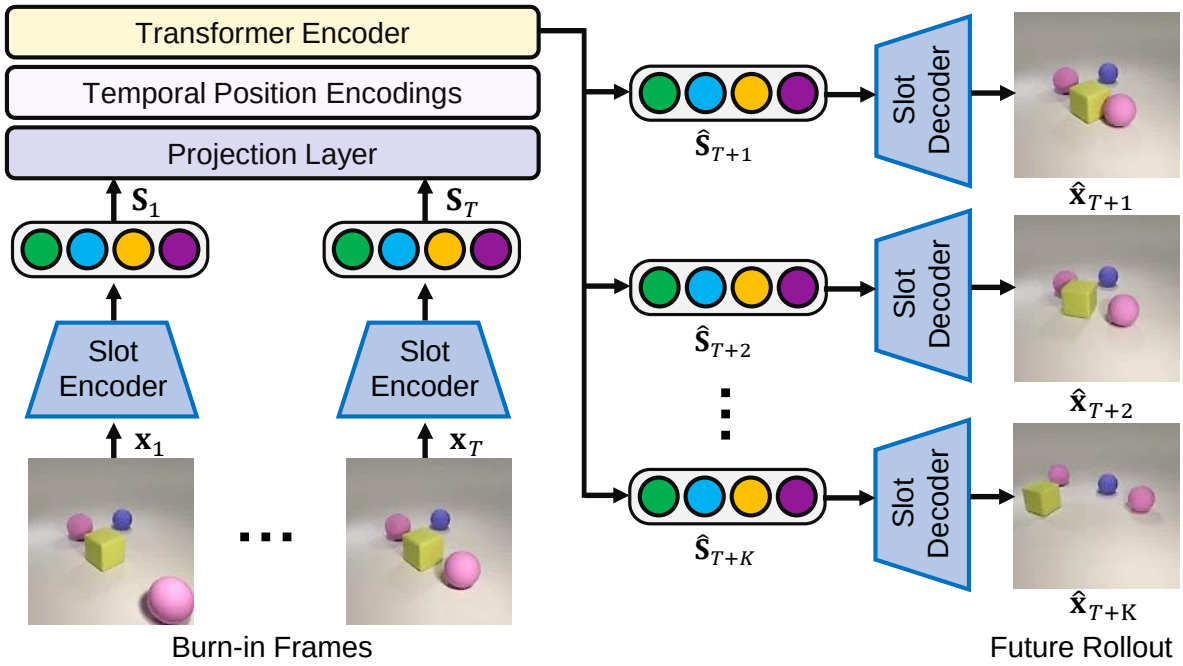

Figure 7. Algorithmic pipeline of SlotFormer [39]. It utilizes a pretrained object-centric model to map multiple video frames into object-level representations and to predict future object states.

2.5. Discussion of Expected World Models

Although existing foundational WMs have made impressive progress, they remain far from envisioned future WMs in how knowledge is represented, updated, and validated. Current WMs primarily optimize predictive accuracy from observational data, without enforcing causal, mechanistic, or falsifiable structure. As summarized in Table 3, future WMs encode knowledge symbolically as equations and formal relations (*e.g.*, Hamiltonian mechanics, and Maxwell's equations) independent of specific datasets or observers. Knowledge is updated cumulatively by extending shared theory rather than through opaque, entangled parameter changes. Such knowledge is socially verifiable: equations and proofs can be checked, and predictions tested through reproducible experiments. These models have explicit domains of validity, and failures arise from principled breakdowns of governing laws, such as classical mechanics at relativistic speeds or the Navier–Stokes equations in turbulent regimes, rather than unpredictable extrapolation errors. In contrast, current WMs often fail opaquely, functioning mainly as sophisticated interpolators rather than explicit, theory-grounded models. Bridging this gap may require integrating symbolic and neural components, enforcing hard physical constraints, and designing protocols that reward causal discovery and falsifiability over reconstruction alone.

 doi:10.20944/preprints202603.0739.v1

Table 3. Paradigm comparison between future and existing WMs.

Future WMs	Existing WMs
Symbolic, readable	Subsymbolic, opaque
Updated by proofs/experiments	Updated by data
Socially verifiable by others	Black box
Known failure modes	Unpredictable failures
Explain "*why*"	Predict "*what*"

3. Applications of World Models in AI

Foundational WMs acquire powerful perception and reasoning capabilities by learning latent environmental dynamics. Leveraging their ability to infer future states, WMs enable agents to perform simulation, planning, and decision-making. In this section, we revisit the applications of foundational world models in downstream domains, including robotics, scientific discovery, and GUI-based agents.

3.1. World Models for Robotics

Artificial intelligence (AI) for robotics [57,137] requires robots to operate under partial observability and rich physical interaction. Recently, this field has been significantly advanced by embodied foundational WMs [23,127,129], which scale world modeling through large multimodal data and unified model interfaces, aiming to integrate perception, prediction, and action into a transferable backbone. Unlike existing surveys on embodied intelligence, particularly those focusing on vision-language navigation and manipulation [86–91], as illustrated in Figure 8, we review nearly all major application tasks of world models in robotics, including manipulation (Section 3.1.1), navigation (Section 3.1.2), policy learning (Section 3.1.3), and locomotion (Section 3.1.4). Additionally, we reanalyze robotic manipulation from a control-perception loop perspective and navigation from a navigation-reasoning loop perspective, rather than focusing solely on vision-language perception.

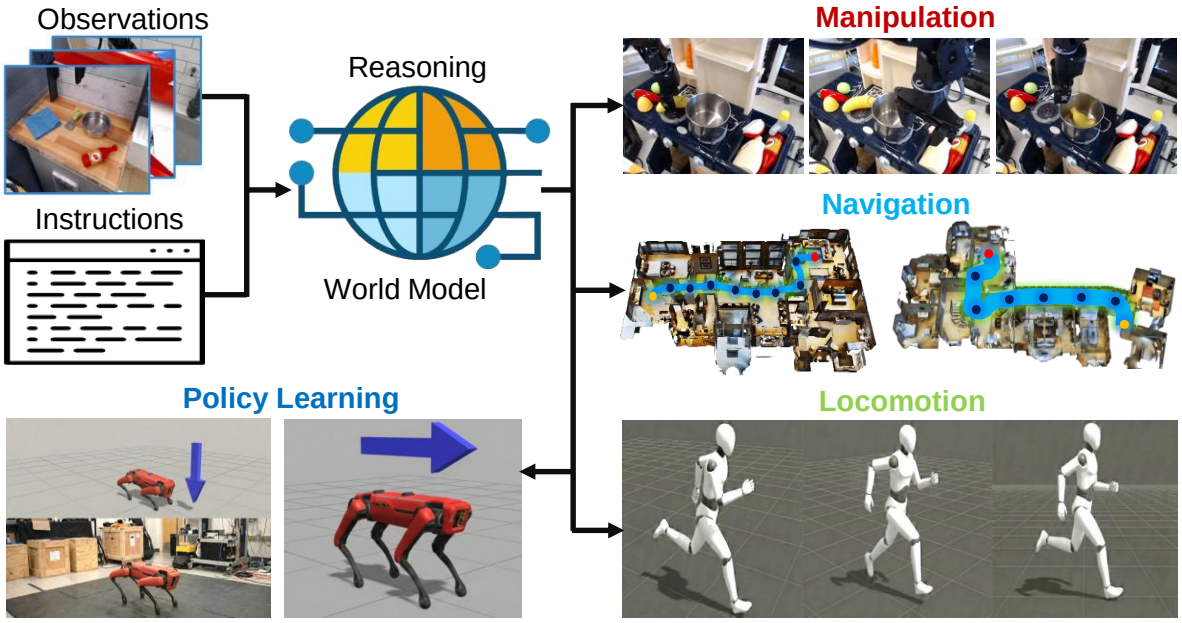

Figure 8. Core applications of world models in AI for robotics.

3.1.1. Manipulation

Robotic manipulation [138] imposes stringent requirements on WMs due to contact-rich dynamics and discontinuous state transitions (*e.g.*, grasp or release). Some previous surveys [87–89,91] typically treat manipulation as vision-language-action (VLA) pipelines, while manipulation-focused reviews commonly organize WMs by architecture. Generally, they focus on specific perception-centric aspects and do not explicitly analyze how WMs interface with manipulation control. In contrast, we adopt a

 doi:10.20944/preprints202603.0739.v1

control-perception loop taxonomy. As presented in Table 4, we categorize manipulation WMs by (a) the control interface they expose (*i.e.*, latent action state imagination and control-oriented planning), (b) the unified world-action interface they provide (*i.e.*, world action modeling), and (c) the specific prediction targets they model (*i.e.*, visual future prediction). Compared with prior surveys [89,91], this survey emphasizes the WM's operational role in the manipulation loop and the associated trade-offs in bias accumulation, latency, and feasibility, rather than primarily cataloging architectures. Specifically, latent imagination is sample-efficient yet can accumulate bias over long rollouts. Control-oriented planning enables constraint-aware model predictive control (MPC) or search via inference-time simulation, but is compute-intensive and depends on accurate short-horizon predictions. World action modeling can unify perception, prediction, and action for transfer, but still faces challenges in computational cost and robustness. Visual future prediction supports planning and supervision but requires aggressive compression to remain consistent.

Table 4. Taxonomy of world models (WMs) for robotic manipulation, categorized by how the WM is used in the control loop and what it predicts.

Category	Representative Works	Architecture	Prediction	Role of WMs	Key Characteristic of WMs
Latent Action State Imagination	Dreamer [9] FLARE [124] AdaWorld [63]	RNN Diffusion Autoregressive	Latent dynamics Future latent Latent actions	Training rollouts Policy regularizer Fast adaptation	Learn and upgrade manipulation skills via latent rollouts during training.
Control-Oriented Planning	Diffuser [6] Decision Diffuser [125] PIVOT-R [126]	Diffusion Diffusion Transformer	Trajectories Trajectories Waypoints	Trajectory sampling Trajectory sampling Waypoint planning	Use WM as an inference-time simulator or generator for trajectory search/MPC.
World Action Modeling	Genie Envisioner [127] WoW [23] iMoWM [128]	Diffusion Diffusion Autoregressive	Video-actions Video-actions Multimodal video	Action generation Action generation Interactive simulation	Incorporate WMs evolution and action generation under a unified interface.
Visual Future Prediction	iVideoGPT [129] FlowDreamer [130] TesserAct [131]	Transformer Diffusion Diffusion	Token video RGB-D video 4D scenes	Visual planning Visual planning 4D planning	Predict future observations for planning or supervision.

Latent Action State Imagination: This branch of approaches uses WMs during training as compact simulators that generate rollouts for policy improvement. As shown in Figure 9, AdaWorld [63] instantiates this idea for manipulation by enabling knowledge transfer across embodiments via adaptive rollouts, while FLARE [124] extends the paradigm to regularize diffusion-based control. More generally, these methods learn latent states and action-conditioned transitions, optimizing policies with RL-style updates driven by imagined returns [66]. OSVI-WM [139] and DEMO^3 [140] aim to synthesize supervision for imitation and demonstration-augmented RL. Besides, pretrained generative models support low-data refinement and grasp selection [141].

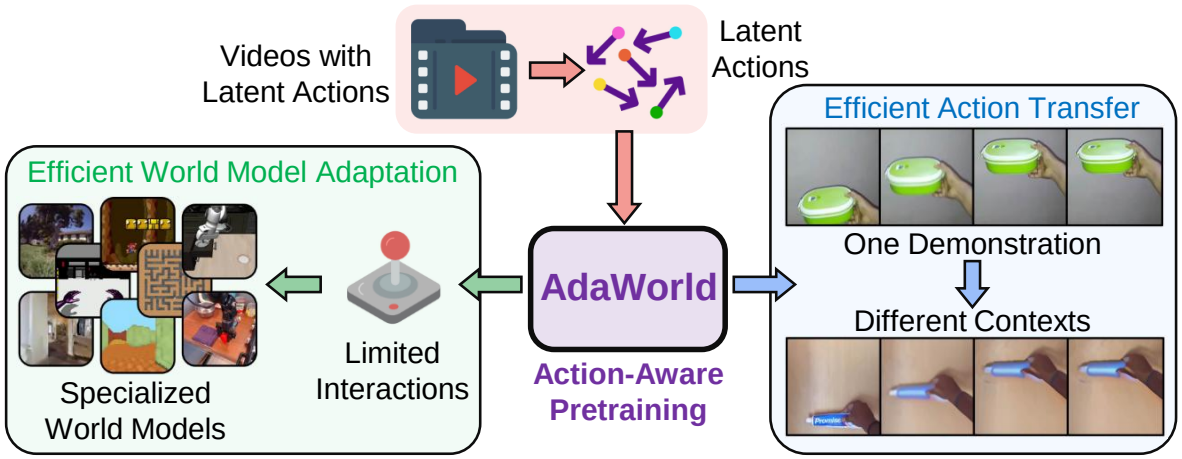

Figure 9. Illustration of AdaWorld [63], which pretrains an autoregressive video WM with self-supervised latent actions to predict futures.

Control-Oriented Planning: In contrast to latent action state imagination, which leverages WMs during training, control-oriented planning methods employ WMs at inference time as roll-

 doi:10.20944/preprints202603.0739.v1

out simulators for receding-horizon trajectory search [142,143]. PETS [73] improves reliability via uncertainty-calibrated ensemble dynamics. TD-MPC [26] and TD-MPC2 [11] accelerate replanning with compact latent dynamics, while TAWM [144] makes rollouts time-aware under varying observation and control rates. In parallel, generative planners synthesize trajectories via denoising in Diffuser [6] and Decision Diffuser [125], or through constraint-guided sampling in potential-based diffusion planning [145]. For long-horizon manipulation, WMs enable structured search with intermediate abstractions: PIVOT-R [126] rolls out waypoints in a waypoint-aware model. Rollout utility is further improved by action-tree visual guidance [146] and multi-view scaffolding [147].

World Action Modeling: Unlike control-oriented planning methods, this branch moves beyond using WMs solely for action search or future prediction by offering a unified interface that couples perception, world evolution, and action generation via tokenized multimodal streams and autoregressive or diffusion-based objectives. Diffusion-based systems such as Genie Envisioner [127] and WoW [23] generate instruction-conditioned interaction futures with closed-loop rollouts and bridge imagination to execution via action decoding or inverse modules. Token-based variants (*e.g.*, PAR [148] and iMoWM [128]) represent observations and actions in a shared discrete space to enable efficient rollout and learning. PhysicalAgent [149] introduces iterative reasoning and replanning, while World4Omni [134] reuses pretrained image generators to propose goal or future states.

Visual Future Prediction: Visual future prediction approaches learn action-conditioned predictors for future observations and utilize visual rollouts for planning-by-prediction and goal-reaching [150, 151]. ViPRA [152] uses world models (WMs) to learn a motion-centric latent action space and decodes it into robot controls with limited supervision. To improve physical plausibility, FlowDreamer [130], TesserAct [131], and ORV [153] incorporate 3D/4D structure (*e.g.*, flow and 4D representations), while WristWorld [154] and RoboMaster [155] address viewpoint shifts and contact-rich interactions. Moreover, RoboDreamer [150] and Vidar [156] exploit rollouts for compositional generalization and action reasoning. Predicted futures can supervise action learning (Video2Action [151], LaDi-WM [157]) or reduce rollout cost via key-frame generation (KeyWorld [158]).

3.1.2. Navigation

Robotic navigation [167] imposes distinct challenges on WMs, primarily requiring long-range spatial consistency and robust operation in dynamic environments. Some previous surveys on embodied AI [89,91] categorize navigation within broad vision-language-action (VLA) frameworks, often overlooking the explicit role of WMs, from internal simulation to the reasoning layer. These discussions typically focus on reactive policy architectures and do not systematically analyze how WMs serve as a structured reasoning interface to bridge raw perception with complex navigation control. In contrast, as presented in Table 5, this survey organizes navigation WMs by their functional roles within the navigation-reasoning loop, rather than under the broad VLA paradigm. Accordingly, we categorize existing methods into five groups from a navigation-reasoning loop perspective. Specifically, generative imagination enables high-fidelity foresight but suffers from long-horizon drift, while persistent memory mitigates partial observability yet demands efficient compression and updating. Neuro-symbolic frameworks provide interpretable and socially compliant navigation at increased architectural cost, whereas test-time adaptation and belief-guided modeling emphasize robustness and knowledge transfer in novel environments.

 doi:10.20944/preprints202603.0739.v1

Table 5. Taxonomy of WMs for robotic navigation, organized by their roles in the pipeline and prediction targets.

Category	Representative Works	Architecture	Prediction
Generative Imagination Navigation	NWM [159]	Diffusion (DiT)	Future video
	MindJourney [160]	Video Diffusion	3D scene rollouts
	Scene-Graph [161]	GNN/Transformer	Graph evolution
Persistent Memory Representation	NavMorph [162]	RSSM (Recurrent)	Latent states
	UniWM [163]	Autoregressive	Multimodal tokens
	RECON [164]	VAE/InfoBottleneck	Latent goals
Neuro-Symbolic Modeling	WMNav [165]	VLM + Map	Curiosity map
	NaVi-WM [166]	LLM Agent	Symbolic subgoals
	Neuro-Sym [167]	Hybrid	Logic constraints
Test-Time Adaptation	MindJourney [160]	Diffusion	Spatial search
	Scale-Infer [168]	Transformer	Value estimates
	Kinodynamic [169]	Physics-based	Feasible motion
Belief-Guided Modeling	X-Mobility [170]	Autoregressive	Action-State
	FalconWing [171]	Gaussian Splat	Photoreal renders
	Abs-Sim2Real [172]	Dynamics Model	Physical params

Generative Imagination Navigation: These generative imagination methods explicitly simulate future observations to support navigation planning. As illustrated in Figure 10, the Navigation World Model (NWM) [159] employs a large-scale conditional diffusion transformer to predict visual rollouts, enabling decision-making via imagined future observations. Unified WM [163] integrates visual foresight with memory to support long-horizon planning. In addition, MindJourney [160] couples vision-language models (VLMs) with generative architectures to imagine future views for spatial reasoning. Besides, scene graph world models [161] further integrate structured graphs, allowing robotic agents to simulate object-level relations and spatial configurations.

 doi:10.20944/preprints202603.0739.v1

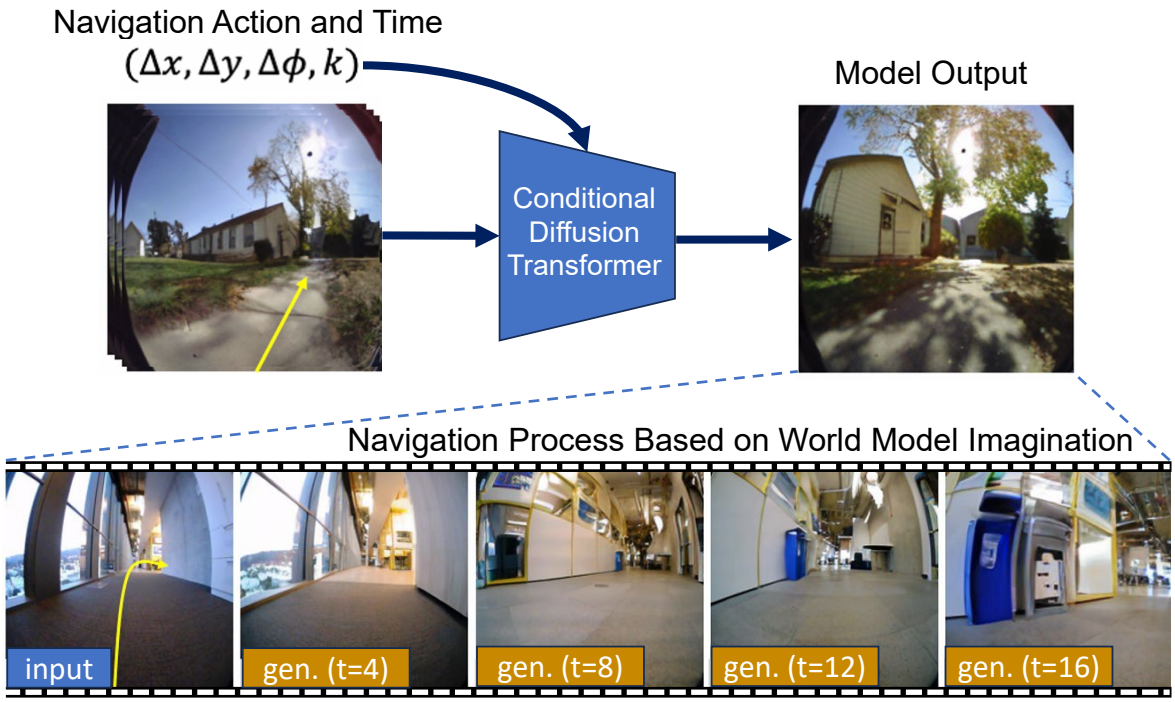

Figure 10. Illustration of navigation world model (NWM) [159].

Persistent Memory Representation: Addressing partial observability, this branch of research leverages WMs to maintain evolving internal states. After Persistent Embodied WM [173] proposes to explore a 3D latent representation that accumulates spatial semantic knowledge over time, NavMorph [162] introduces a self-evolving model with contextual memory, enabling continuous updates during exploration. Notably, Unified WM [163] also utilizes memory for foresight. Moreover, RECON [164] employs latent goal models to guide rapid exploration in open worlds, effectively acting as a goal-directed WM for robotic navigation.

Neuro-Symbolic Modeling: Neuro-symbolic modeling approaches extend dynamics modeling with linguistic or symbolic constraints. WMNav [165] integrates VLMs to ground language into environment memory for navigation. To address social constraints, NaVi-WM [166] augments WMs with deductive chain-of-thought reasoning. Similarly, Neuro-Symbolic WM [167] combines neural perception with symbolic reasoning to ensure socially compliant and interpretable navigation. NavMorph [162] further embeds language grounding directly into its continuous WM updates.

Test-Time Adaptation: This category employs WMs to handle imperfect priors and novel environments when executing navigation tasks. Representative methods include guiding inference-time search via vision value models [168] and dynamically implanting WM components for rapid adaptation [174]. To address the challenge of model inconsistency, Kinodynamic Planning [169] proposes robust methods that tolerate mismatches across conflicting models. MindJourney [160] further demonstrates test-time scaling by coupling VLMs with controllable video diffusion WMs.

Belief-Guided Modeling: Rather than explicit imagination, belief-guided modeling regard WMs as decision-sufficient belief abstractions. Deep active inference [175] uses multi-timescale belief dynamics to guide diffusion policies. Theoretical work in Abs-Sim2Rea [172] demonstrates that models preserving information states enable robust knowledge transfer. In addition, X-Mobility [170] utilizes a probabilistic world model and multi-stage training to mitigate challenges such as data scarcity and poor robustness.

3.1.3. Policy Learning

WMs function as imagination engines, action-free models, and evaluation proxies for robotic policy learning by allowing policies to be optimized through imagined interactions rather than costly real-world trials. Prior work falls into three categories. Imagination-based optimization enables

 doi:10.20944/preprints202603.0739.v1

zero-shot policy refinement via synthesized rollouts, yet risks exploiting model inaccuracies during extended optimization. Action-free modeling optimizes sample efficiency by leveraging unlabeled or human videos for knowledge transfer, but face challenges in aligning latent representations with precise robot actions. Evaluation proxy provides scalable, safe benchmarking environments without real-world costs, though their utility depends on strictly preserving relative policy rankings under distribution shift.

Imagination-Based Optimization: WMs are increasingly adopted not merely as predictors, but as optimization substrates that enable robotic policy learning directly in imagination. WMPO [176] formulates world-model-based policy optimization for vision-language-action models, enabling policy gradients to be computed over imagined trajectories rather than real interactions. Diffusion-based WMs improve rollout fidelity and stability, enabling policies to be refined entirely within generated environments. This is demonstrated by World4RL [177] and DAWM [178], both of which achieve substantial gains in offline and online RL settings.

Action-Free Modeling: Another key trend is to learn control policies from unlabeled or weakly labeled observation sequences through WMs, without relying on expensive explicit action supervision. LAWM [179] and PreLAR [180] demonstrate that aligning inferred latent actions with limited ground-truth control signals enables scalable policy learning from passive observations. WMs also extract transferable dynamics from heterogeneous supervision. For instance, DexWM [181] and cross-embodiment WMs [182] use large-scale human videos for zero-shot transfer to robot policies, while TraceGen [183] and 3DFlowAction [184] abstract dynamics into embodiment-agnostic representations to support policy adaptation across morphologies.

Evaluation Proxy: Beyond the policy optimization methods discussed above, controllable and simulator-like methods explicitly position the WM as a proxy simulator for evaluation. Ctrl-World [185] and Robotic WM [186] enable policy-in-the-loop training, comparison, and evaluation without incurring costly real-world rollouts. Recent works such as WorldGym [187] and WorldEval [188] further demonstrate that WMs can preserve relative policy rankings, establishing them as reliable proxies for real-world policy evaluation.

3.1.4. Locomotion

WMs have recently demonstrated strong potential for learning robust locomotion behaviors by explicitly capturing contact-rich dynamics, terrain geometry, and wide-range physical consistency. Rather than relying solely on reactive policies, they serve as structured predictors that reconstruct or predict locomotion-relevant state transitions, enabling policy optimization under complex constraints. Despite high performance on specific hardware and effective modeling of contact uncertainty, they struggle with the high-frequency discontinuities of real-world impacts and remain tightly coupled to distinct morphologies, leaving the scaling of existing WMs to diverse interactions an open challenge.

Specifically, WMP [196] learns a predictive visual WM to ground perception for legged control, significantly improving robustness on uneven and partially observed terrains. WMR [197] introduces a reconstructive WM that regularizes policy learning by enforcing physically consistent latent dynamics, leading to stable humanoid locomotion under severe distribution shifts. Beyond visual prediction, works such as ProTerrain [198] and DWL [199] incorporate physical structure and terrain uncertainty into WMs to better capture locomotion dynamics. In addition, WMs have been extended to support contact reasoning and whole-body motion synthesis. For example, Ego [200] integrates egocentric perception with latent dynamics and control, enabling real-time, contact-aware motion generation for humanoid robots.

3.2. World Models for Autonomous Driving

WMs facilitate autonomous driving systems [79,80] by modeling traffic dynamics and agent interactions, allowing vehicles to anticipate future scenarios and plan actions accordingly in uncertain environments [75–77]. While some surveys [75–80] investigate recent advances in autonomous driving, we reanalyze this domain from a new perspective on the involvement of WMs in the decision-making

 doi:10.20944/preprints202603.0739.v1

loop, providing a clearer understanding of the role of WMs in autonomous driving. Accordingly, as summarized in Table 6, we categorize WMs for autonomous driving into predictive modeling, action-conditioned imagination, and decision-centric integration. Specifically, predictive modeling approaches excel at capturing scene dynamics for pretraining and data generation. Action-conditioned imagination frameworks introduce action-aware foresight, enabling explicit comparison of alternative driving behaviors under diverse intents. Decision-centric integration further strengthens autonomy by using imagined futures to guide planning and policy learning. However, current approaches remain limited by data scarcity, computational efficiency, and challenges in ensuring reliable and physically consistent wide-range simulations under diverse real-world driving environments.

Table 6. Taxonomy of WMs for autonomous driving, categorized by the degree of their involvement in the decision-making process.

Category	Representative Works	Architecture	Input Modality
Predictive Modeling	Copilot4D [189]	Diffusion	Point cloud
	UniWorld [190]	Transformer	Image
	UNO [72]	Transformer	Point cloud
Action-Conditioned Imagination	Drive-WM [191]	Diffusion	Image, Action, Layout
	Vista [67]	Diffusion	Image, Action, Trajectory
	GAIA-1 [192]	Transformer	Image, Action
Decision-Centric Integration	Think2Drive [193]	RNN	BEV, Sign
	AdaWM [194]	RNN	BEV, Action
	DriveVLA-W0 [195]	Diffusion	Image, Action, Text

3.2.1. Predictive Modeling

Predictive modeling for autonomous driving treats WMs as environment forecasters that predict future world states from past observations, without conditioning on ego actions or control commands [62,201]. These methods formulate world modeling as a video or scene generation problem, exploring how traffic scenes evolve over time and providing reusable representations for simulation and generation [202,203]. Copilot4D [189] uses discrete diffusion to model coherent 4D scene evolution without action supervision, while UNO [72] proposes that observation-only occupancy prediction can capture complex traffic dynamics. Other predictive modeling approaches [72,190,203] have been comprehensively reviewed in recent surveys [78,92], and are therefore not discussed in detail here. From a broader perspective, predictive modeling WMs in autonomous driving are conceptually aligned with observation-driven dynamics models used in other domains, including scientific forecasting tasks such as climate prediction [204,205], where future system evolution is inferred from historical states. Such models emphasize learning latent spatiotemporal regularities of the environment, and the effectiveness of 3D/4D predictive representations in driving suggests broader applicability to long-horizon forecasting.

3.2.2. Action-Conditioned Imagination

The action-conditioned imagination branch [206,207] extends the above predictive modeling by conditioning future world evolution on ego actions, trajectories, or high-level commands, enabling counterfactual reasoning. Specifically, WMs function as imagination modules that produce future outcomes for candidate behaviors [208–210]. GAIA-1 [192] and Drive-WM [191] are representative works of this line of research, both introducing explicit action conditioning into observation-level generative WMs to facilitate long-horizon, behavior-aware future generation. Other action-conditioned imagination approaches for autonomous driving have been comprehensively discussed in existing

 doi:10.20944/preprints202603.0739.v1

surveys [76,78] and are therefore not elaborated here. Action-conditioned imagination WMs in autonomous driving share close commonalities with WMs in robotic manipulation [63,211] and navigation [162]. In these domains, future environment states are explicitly simulated under alternative action hypotheses to enable counterfactual reasoning and behavior-aware evaluation. This formulation positions action-conditioned imagination as an intermediate layer between environment modeling and decision-making.

3.2.3. Decision-Centric Integration

Decision-centric integration approaches [193,212] embed WMs in the decision loop, using imagined rollouts to evaluate actions, guide planning, or update policies. In this setting, WMs are no longer treated as standalone generators, but function as core components of action selection and control. Think2Drive [193] exemplifies this paradigm by leveraging imagination-based reinforcement learning, enabling policies to be improved through internal rollouts with reduced reliance on external simulators. DriveVLA [195] further demonstrates that large-scale WMs can directly support policy learning via data scaling, positioning the WM itself as a central substrate for closed-loop decision-making. Decision-centric WMs in autonomous driving exemplify a broader paradigm shared with world modeling approaches in games [213], reinforcement learning [46], and other sequential decision-making domains [186]. In these domains, imagined rollouts are embedded into planning or policy learning processes, enabling action evaluation under uncertainty without exhaustive real-world trials. From this perspective, WMs transition from passive predictors to active decision substrates, a role that is evident in the closed-loop autonomy requirements of driving.

3.3. World Models for Science

WMs for science transition the focus from interaction-centric simulation to data-driven scientific modeling of complex, partially observed systems. As shown in Figure 11(a), WMs for science can be mainly divided into two categories: social and socioeconomic systems (Section 3.3.1) and physical and natural sciences (Section 3.3.2). In social and socioeconomic domains, WMs enable large-scale simulation of collective behavior and policy evaluation, but remain highly sensitive to data bias and are difficult to rigorously validate due to complex emergent dynamics. In physical and natural sciences, WMs serve as physically constrained simulators for long-horizon prediction and planning under uncertainty, yet continue to face some essential challenges related to physical fidelity, generalization beyond training regimes, and interpretability.

 doi:10.20944/preprints202603.0739.v1

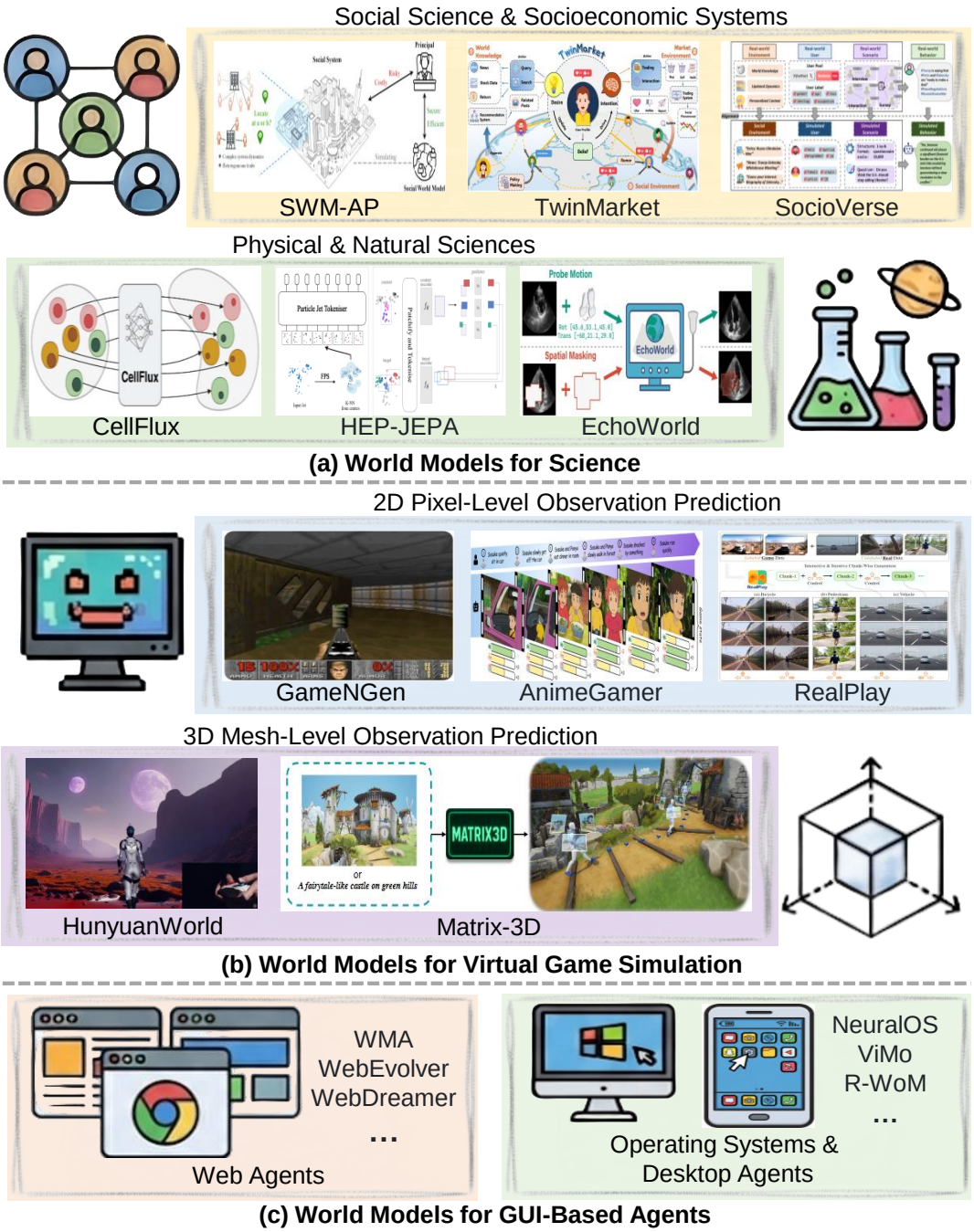

Figure 11. Taxonomy of WMs for science (a), virtual game simulation (b) and GUI-based agents (c).

3.3.1. Social Science and Socioeconomic Systems

A growing body of work studies WMs for social simulation and planning. Social WMs formalize social systems as learnable models of collective dynamics and agent interactions [214], which has been extended to WM-augmented mechanism design for policy learning and economic decision-making [64]. SocioVerse [215] scales social simulation by combining LLM-based agents to model population-level behavior and emergent dynamics (*e.g.*, politics, economics and news). Moreover, related efforts apply WMs to specific domains, including financial markets [216], virtual-reality-based behavioral analysis [217], and influence-driven diffusion processes [218]. Methodological advances in multivariate social time series and structured sequence modeling provide foundational tools for scalable social WMs [219].

 doi:10.20944/preprints202603.0739.v1

3.3.2. Physical and Natural Sciences

WMs act as data-driven, physically constrained simulators for modeling systems in natural and physical sciences. In biology, CellFlux enables continuous-time simulation of cellular dynamics via flow matching [70], while ODesign frames biomolecular interaction design as world modeling for generative molecular exploration [220]. In healthcare, image- and state-based WMs support diagnosis and planning, including tumor evolution modeling [74], radiograph world models [65,221], cardiac ultrasound guidance [222], and surgical procedure modeling [223]. Another major direction links world modeling with forecasting and planning in physical systems, such as spatiotemporal forecasting via generative WMs and model-based RL [224], physics-informed time-series modeling [225], dynamics reconstruction from partial observations [226], and modular architectures for compositional reasoning. Some key applications further extend to large-scale scientific domains, including collider physics through foundation-style models [56].

3.4. World Models for Virtual Game Simulation

Virtual game simulation is a natural and historically important application of WMs, as games offer fully observable and interactive environments. While early game-oriented WMs act mainly as auxiliary predictors for RL, recent efforts in large-scale observation-level generative WMs have shifted them toward standalone interactive simulation. Based on the type of predicted observations, we group existing methods into 2D pixel-level (Section 3.4.1) and 3D mesh-level (Section 3.4.2), as shown in Figure 11(b). Pixel-level prediction [47] offers flexible and visually realistic simulation and enables scalable interactive gameplay, but often struggles with long-horizon consistency and faithful game dynamics. In contrast, 3D mesh-level WMs [227] use stronger object permanence, structural and spatial priors that improve immersion and physical grounding, yet face challenges in scalability and real-time interaction. From a broader perspective, 3D mesh-level WMs are better viewed as complementary to 2D pixel-based ones rather than direct replacements. Emerging trends suggest hybrid architectures that combine 3D structure with neural rendering or generative visual models, aiming to balance geometric priors with generative flexibility.

3.4.1. 2D Pixel-Level Observation Prediction

Most existing game WMs focus on predicting future observations directly at the pixel level, conditioned on past visual frames and agent actions. Representative examples include DIAMOND, which enables accurate multi-step Atari rollouts via diffusion-based dynamics modeling [47], and GameNGen, which supports real-time interactive gameplay without explicit simulators [228]. Beyond arcade-style settings, pixel-level observation prediction has been extended to more open-ended and expressive virtual environments. Oasis [229] explores sequence-based universe modeling with multi-step visual coherence, while AnimeGamer [68] demonstrates stylized, infinite-life simulation tailored to specific aesthetic domains. These works reflect a shift from task-centric simulation toward more general-purpose and generative world modeling. More recently, this branch has begun to converge with the paradigm of foundation-style interactive models. Matrix-Game and Matrix-Game 2.0 [230,231] aim to unify perception, dynamics, and interaction within a single real-time generative framework, while RealPlay [232] and GameFactory [213] emphasize embodiment, generative game creation, and multi-step consistency.

3.4.2. 3D Mesh-Level Observation Prediction

While pixel-level observation prediction dominates current game WM research, there is growing interest in models that operate on explicit 3D geometric representations, such as meshes or structured scene representations. HunyuanWorld 1.0 exemplifies this trend by generating explorable and interactive 3D worlds from text or images, bridging large-scale 3D content generation with WM learning [227]. Matrix-3D advances this paradigm by enabling omnidirectional, explorable 3D world generation with globally consistent spatial structure [233]. As foundation models increasingly integrate multimodal

 doi:10.20944/preprints202603.0739.v1

perception and action, 3D mesh-level observation prediction will be important for enabling physically grounded and semantically coherent virtual worlds.

3.5. World Models for GUI-Based Agents

WMs significantly enhance graphical user interface (GUI)-based agents by enabling predictive simulation of interface dynamics, allowing agents to plan, evaluate alternatives, and assess actions before execution in stochastic environments. Unlike physical or game domains [68], these environments are highly stochastic and governed by complex symbolic and visual dynamics, making explicit modeling of environment transitions crucial for long-horizon decision-making [41]. We categorize GUI-based agents by their primary interaction environments, distinguishing web agents from operating system and desktop agents. As presented in Figure 11(c), model-based simulation enables foresighted planning, alternative exploration, and error correction for both web and desktop agents, but relies on accurate prediction of complex visual and symbolic transitions. While iterative search and reasoning-integrated WMs improve multi-step decision-making, prediction errors can accumulate under partial observability and non-stationary interfaces.

Web Agents: Web navigation poses unique challenges due to large state spaces and delayed rewards. WebDreamer exploits LLMs as implicit web WMs and demonstrates model-based planning via simulated rollouts [234], while complementary work explicitly learns web environment dynamics to guide decision-making [41]. Recent methods emphasize iterative improvement and search-based planning, including WebSynthesis, which combines world models with MCTS for efficient user interface trajectory synthesis [235], WebEvolver, which coevolves agents and WMs through self-play [236], and WALL-E 2.0, which aligns WMs with symbolic constraints to improve planning reliability [21]. Several approaches explore unified architectures that tightly integrate reasoning, acting, and simulation. SimuRA [237] introduces an LLM-based simulative reasoning framework that enables goal-directed planning via internal rollouts, while Dyna-Think [238] further formalizes this integration by jointly optimizing reasoning, action, and WM simulation to improve multi-step execution performance.

Operating Systems (OS) & Desktop Agents: Several works equip agents with internal simulators of operating systems or desktop applications. NeuralOS models OS-level dynamics for anticipatory planning without execution [239], while ViMo builds generative visual WMs for mobile and desktop applications to support foresighted GUI interaction [240]. FPWC further combines WM simulation with code execution for model-driven device control [241]. Retrieval-augmented approaches enhance fidelity and generalization, including R-WoM [242], which integrates external retrieval into prediction, and WKM [69], which learns world knowledge for planning across tasks. They underscore the value of combining latent dynamics with explicit knowledge.

3.6. Interpretable and Trustworthy World Models

Beyond empirical success, recent studies scrutinize the theoretical foundations of WMs, focusing on their emergent causal structures and safety guarantees in critical tasks. In this section, we categorize these efforts into two directions: interpretability and safety-oriented trustworthiness. Interpretable and trustworthy WMs reveal a core tension between predictive accuracy and causal fidelity. Interpretability studies show that structured internal models can emerge for long-horizon generalization, yet strong prediction often relies on shortcut correlations, making true causal verification difficult. Safety-oriented work finds that scaling improves stability but does not prevent failures under distribution shift, adversarial perturbation, or partial observability. While strong safety and interpretability guarantees enhance reliability, they often reduce the scalability of WMs.

Interpretability: Some theoretical research suggests that internal WMs are necessary for generalizing to long-horizon goals beyond inductive biases [243]. Although transformers can learn representations aligned with causal structure [244,245], high predictive accuracy does not guarantee genuine internal simulation, as models may rely on shortcut correlations that fail under intervention [246]. Notably, this phenomenon has motivated mechanistic and causal probing to verify whether internal states encode true open-world geometry rather than superficial patterns [247,248].

 doi:10.20944/preprints202603.0739.v1

To develop such genuine internal simulations, extensive environmental interaction is required. However, real-world data collection is prohibitively costly, and current Sim-to-Real approaches [249–251] struggle to achieve scalable, high-quality modeling. To address this limitation, recent efforts [252–254] employ self-evolutionary learning by introducing internal mechanisms such as self-verification [255,256], self-reward [257], and self-play [258,259]. These mechanisms enable WMs to maintain stable, continuous optimization in environments without external supervision, relying solely on internal predictive capabilities and self-updating mechanisms [260,261]. This shift reflects a gradual transition from dependence on empirical phenomena toward constructing reproducible, explainable, and mechanistic paradigms. Additionally, the successful application of self-evolutionary learning across diverse downstream tasks demonstrates its capacity to impose observable constraints on the stability of closed-loop systems [262,263]. Both theoretically and empirically, self-evolutionary learning supports the effectiveness and robustness of constraining strategy divergence through internal predictive signals, thereby providing structural and mechanistic foundations for WMs as a basis for stable intelligent evolution [4].

Safety-Oriented Trustworthiness: The practical deployment of WMs is further limited by stability and observability constraints. Although scaling laws indicate that increased capacity improves representation stability [264], it cannot ensure robustness to distribution shifts or adversarial perturbations, nor does it overcome theoretical limits in partially observable settings [265]. From a safety perspective, incorrect WM assumptions can bias information acquisition, leading to unsafe downstream decisions [15]. Thus, empirical performance should be accompanied by formal verification to mitigate risks in safety-critical environments.

3.7. Limitations of WMs in Downstream Applications

While WMs have achieved promising gains on downstream applications (*e.g.*, robotics [88,91] and autonomous driving [201]), significant challenges remain. First, the training objectives of WMs primarily focus on representation metrics such as reconstruction or generation, whereas downstream applications emphasize decision quality and policy learning effectiveness. This inherent misalignment between objectives creates a difficult trade-off between "world knowledge acquired through pretraining" and "downstream performance enhancement". Additionally, existing WMs generally lack explicit modeling of causal structures, increasing susceptibility to error accumulation and prediction drift during closed-loop interactions and long-horizon reasoning. These issues can trigger planning and decision instability while introducing the risk of unexplained failures. Moreover, WMs are constrained by the computational and latency costs of inference and planning, preventing direct deployment on edge devices such as robots or in-vehicle systems.

4. Benchmark of World Models

Establishing comprehensive benchmarks plays a vital role in facilitating progress in WMs. In this section, we systematically analyze existing benchmark datasets, evaluation metrics, simulation platforms, and comparative performance.

4.1. Benchmark Datasets & Evaluation Metrics

WMs are typically pretrained on large-scale, broad, weakly supervised videos and interaction traces, and then adapted to downstream tasks for prediction and control. We first introduce pretrained video benchmarks and commonly used datasets for downstream tasks, along with their corresponding evaluation metrics. As shown in Table 7, common WM benchmarks can be categorized based on their functional roles into simulation, planning, and decision-making. We then analyze the weaknesses of existing benchmarks and propose three new general metrics to measure the generalization, causal reasoning, and long-horizon consistency of WMs, thereby guiding their future development and design.

 doi:10.20944/preprints202603.0739.v1

Table 7. Benchmark datasets categorized by their primary functions in WM research. **Tag**: S= Simulation, P= Planning, D= Decision-making.

Tasks	Benchmark	Main Focus	Classical Metrics
Pretraining	WebVid-10M [266]	S	Recall@K, MedR
	Panda-70M [267]	S	Recall@K, FVD, CLIPSim
	Ego4D [268]	S	Recall@K, mAP, Accuracy
	HowTo100M [269]	S	Recall@K, MedR
	WorldScore [270]	S	WS-S, WS-D, CCon, OCon
Robotics	Open X-Embodiment [271]	D	SR
	Room-Across-Room [272]	P	SR, NE, PL
	EWMBench [273]	S	SceneC, SHD
	Meta-World [274]	D	SR
Autonomous Driving	NuScenes [275]	S	mAP, NDS, AMOTA
	ACT Bench [276]	P	ADE, FDE
Science	JUMP Cell Painting [277]	S	FID
	HCC-TACE-Seg [278]	D	AP, F1-score, JI, recall
Game Simulation	Arcade Learning Environment [279]	D	MedS, NS
	MineRL [280]	D	Episode return
GUI-Based Agents	OSWorld [281]	P	SR
	WindowsAgentArena [282]	P	SR

4.1.1. Pretrained Video Benchmarks

For pretraining on text-video pairs, WebVid-10M [266] provides 10M web video-caption pairs scraped from stock-footage sites, and Panda-70M [267] further scales supervision by automatically generating captions for 70M video clips. Ego4D [268] adds long-horizon egocentric videos with rich annotations, while HowTo100M [269] supplies large narrated instructional videos for weakly-supervised text-video pretraining. The core metrics for video-text pretraining benchmarks include text-video alignment (*e.g.*, Recall@K, Median Rank (MedR) and CLIPSim) and visual quality (*e.g.*, Fréchet Video Distance (FVD)) [267,269]. To tackle the limitations of single-modal approaches and evaluate WMs through multiple pathways, Worldscore [270] uses a multi-step next-scene protocol for simulation, reporting aggregate and component metrics spanning controllability, quality, and dynamics. The core metrics of Worldscore include worldscore-static (WS-S), worldscore-dynamic (WS-D), camera controllability (CCon), object controllability (OCon), content alignment (CAli), 3D consistency (3DCo), photometric consistency (PCon), style consistency (SCon), subjective quality (SQua), motion accuracy (MAcc), motion magnitude (MMag), and motion smoothness (MSmo).

4.1.2. Benchmarks on Downstream Tasks

1) **Robotics:** In manipulation, Open X-Embodiment [271] aggregates cross-robot trajectories to train generalist policies, typically evaluated by success rate (SR). For navigation tasks, Room-Across-Room [272] tests robotic agents on natural language instruction following, with performance evaluated by SR, navigation error (NE), and success weighted by path length (PL). EWMBench [273] complements physical benchmarks by measuring scene consistency (SceneC) and symmetric hausdorff distance (SHD). Meta-World [274] is a standard multi-task testing platform used to evaluate the effectiveness of policy learning, whose core metric is SR.

2) **Autonomous Driving:** NuScenes [275] is a widely used benchmark dataset for perception and prediction in autonomous driving, utilizing metrics such as mean average precision (mAP), the nuScenes detection score (NDS), and average multi-object tracking accuracy (AMOTA). To assess action fidelity, ACT Bench [276] evaluates the consistency between executed and target trajectories via average displacement error (ADE) and final displacement error (FDE).

 doi:10.20944/preprints202603.0739.v1

3) **Science**: JUMP Cell Painting [277] provides hyperscale high-content imaging for cellular morphology and phenotypic drug discovery, with metrics including Fréchet Inception Distance (FID). In the medical domain, HCC-TACE-Seg [278] serves as a benchmark for action prediction, where models are evaluated against ground-truth using metrics such as AP, F1-score, Jaccard index (JI), and recall.

4) **Game Simulation**: The Arcade Learning Environment (ALE) dataset [279] evaluates Atari 2600 agents using median scores (MedS) and normalized scores (NS). The MineRL dataset [280] enhances RL from large-scale demonstrations in Minecraft, where the evaluation metric is episode return.

5) **GUI-Based Agents**: OSWorld [281] evaluates multimodal agents on open-ended desktop, using execution-based SR as the primary metric. WindowsAgentArena [282] assesses GUI-based agents in a reproducible Windows environment, focusing on SR across multi-step execution tasks.

4.1.3. Designing General Metrics for World Models

As is well known, a universal WM should possess cross-domain generalization capabilities, causal reasoning abilities, and long-horizon consistency. While existing datasets measure different aspects of a model's performance, they still fail to provide a unified evaluation of generalization, causal reasoning, and long-horizon consistency. Therefore, in this subsection, we propose three new mathematical metrics to assess these capabilities and support the foundational development and future design directions of unified WMs.

- **Generalization:** Given the world model $\pi_\theta(\cdot)$ and the task metric $\Gamma(\cdot)$, we express the metric $\mathcal{G}$ for measuring cross-domain generalization capability as follows:

$$\mathcal{G} = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{\text{tr}}}[\Gamma(\pi_\theta, \mathbf{x})] - \mathbb{E}_{\mathbf{x}'\sim\mathcal{P}_{\text{te}}}[\Gamma(\pi_\theta, \mathbf{x}')], \tag{9}$$

  where $\Gamma(\pi_\theta, \mathbf{x})$ and $\Gamma(\pi_\theta, \mathbf{x}')$ represent the task metrics (*e.g.*, success rate or accuracy) when the WM $\pi_\theta(\cdot)$ is evaluated on the training sample $\mathbf{x}\sim\mathcal{P}_{\text{tr}}$ and the testing sample $\mathbf{x}'\sim\mathcal{P}_{\text{te}}$. $\mathcal{P}_{\text{tr}}$ and $\mathcal{P}_{\text{te}}$ denote the distributions of the training and testing datasets, and they exhibit a significant distribution discrepancy ($\mathcal{P}_{\text{te}} \neq \mathcal{P}_{\text{tr}}$).
- **Causal Reasoning:** Given a set of interventions $\mathcal{I}$ and a distance measure $\mathcal{D}(\cdot)$, inspired by Pearl's do-calculus [283], we adopt a counterfactual intervention operator $do(i)$ ($i\sim\mathcal{I}$) to measure causal reasoning capability $\mathcal{C}$:

$$\mathcal{C} = \mathbb{E}_{i\sim\mathcal{I}}[\mathcal{D}(\mathbf{y}, \mathbf{y}')], \tag{10}$$

  where $\mathbf{y}\sim\pi_\theta(\mathbf{y}|\mathbf{x})$ and $\mathbf{y}'\sim\pi_\theta(\mathbf{y}|do(\mathbf{x}=i))$ denote the expected outputs when the inputs are $\mathbf{x}$ and $do(\mathbf{x}=i)$.
- **Long-Horizon Consistency:** For multi-step execution tasks, we compare the incremental deviation between the actual trajectory and the imagined trajectory under identical conditions and policies to evaluate the long-horizon consistency $\mathcal{H}$ of existing foundational WMs:

$$\mathcal{H} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{D}\Big(\zeta_t,\ \widehat{\zeta}_t\Big), \tag{11}$$

  where ζ_t and $\widehat{\zeta}_t$ denote the ground-truth and predicted trajectory (generated based on Eq. (1)) at time t.

4.2. Physics Engines & Simulation Platforms

WMs are typically trained and evaluated in closed-loop interactions, where actions influence future observations, and physics engines and simulation platforms therefore serve as controllable data generators and evaluation harnesses. Given the high cost and safety risks associated with real-world data collection, high-fidelity simulation has become essential infrastructure for training and validating

 doi:10.20944/preprints202603.0739.v1

WMs. As demonstrated in Table 8, we review existing physics engines and simulation platforms used to evaluate WMs.

Table 8. Summary of physics engines and simulation platforms used for evaluating WMs. OSS: open source software. ODE: open dynamics engine. RB: rigid body. MB: multi-Body. SI: sequential impulse. PGS: projected Gauss seidel. TGS: temporal Gauss seidel. AD: automatic differentiation. Def: deformables objects. GPU accel and Def take the following values: Y (yes), P (partial), L (limited), D (depends), and N (no).

Simulator	Main Focus	Typical Functions	Key Traits	Dynamics and Solver	Language	OSS	Def	GPU
Bullet[1]	S	Contact Rich Robotics, Prototyping	Widely Used, Soft Body Module	RB Constraints, SI	C++	Y	Y	P
Simbody[2]	S	Dynamics Analysis, Biomechanics	High Fidelity Multi-Body Modeling	Multi-Body Dynamics	C++	Y	N	N
PhysX[3]	S	Scalable Robotics, Simulation Backends	Industrial Grade, FEM Modules	RB Constraints, PGS TGS	C++	Y	Y	Y
Gazebo[4]	S P	Navigation, Multi-Robot Evaluation	ROS Integration, System Simulation	Backend Dependent	C++, Python	Y	D	D
Webots[5]	S P	Education, System Tests	Integrated Robots and GUI	ODE RB	hybrid	Y	L	L
Omniverse[6]	S P	Photorealistic Data, Synthetic Sensors	RTX Sensors, Digital Twin Stack	USD Physics to PhysX	Python, C++	N	Y	Y
PyBullet[7]	S P D	Dataset Collection, Baselines	Scripting Utilities, RL Tooling	Bullet Backend	Python	Y	P	L
MuJoCo[8]	S P D	Control Benchmarks, Manipulation	Control Oriented, MJX Accel Path	Multi-Body Dynamics	C, Python	Y	L	P
Isaac Gym[9]	S P D	Massive Parallel Rollouts	Tensor Oriented API, Legacy Status	PhysX on GPU	Python	N	L	Y
Genesis[10]	S P D	Embodied AI Training Data	Research Platform, High Throughput	Multi-Backend	Python	Y	P	Y

Simulation: Effective simulation relies on physics engines capable of replicating real-world properties. Bullet[1] is an open-source physics engine supporting multi-physics simulation, widely applied across various fields. Simbody[2] emphasizes high-precision multibody dynamics and is particularly relevant for biomechanics-oriented embodied modeling. Additionally, PhysX[3] provides an industry-adopted physics foundation. Despite their maturity, recent documentation indicates that large-scale simulations on GPUs and cross-platform determinism still remain challenging.

Simulation & Planning: Constructing WMs necessitates not only high-fidelity physics simulation but also planning capabilities. Early CPU-oriented robotics simulators such as Gazebo[4] and Webots[5] provide mature robotic workflows, but their throughput can be limiting when large-scale interactive experiences are required. Meanwhile, Omniverse[6] offers advanced rendering capabilities that improve visual realism. However, these platforms struggle to ensure reliable multi-agent evaluations in complex environments and also incur significant computational overhead from perception and dense contacts during the planning process.

Simulation & Planning & Decision-Making: To support closed-loop decision-making, modern simulation platforms have further enhanced stability and begun supporting GPUs to accelerate simulation. PyBullet[7] is widely used in early deep RL due to its accessibility and ease of integration; however, complex contact-rich tasks and large-scale parallel training can expose limitations in performance and stability. For model-based RL, stable multibody dynamics and reliable contact handling are essential. MuJoCo[8] is explicitly designed to meet these requirements. Additionally, Isaac Gym[9] demonstrates GPU-resident closed-loop simulation tightly coupled with neural network training. Furthermore, Genesis[10] is designed specifically for embodied AI and has achieved superior speed. However, since closed-loop decision-making demands high-frequency interaction, modern platforms still need to grapple with the trade-off between real-time performance and physical fidelity.

Discussion: The physics engines and simulation platforms provide a controlled environment for generating and evaluating data for WMs. However, large-scale simulations remain challenging in terms of GPU acceleration and cross-platform determinism, leading to insufficient robustness in result reproducibility and rigorous comparisons. Consequently, the validity of evaluations for complex

1 Bullet project site: Github.io. Bullet guide site: Wordpress.
2 Multibody dynamics for biomedical research: PDF.
3 NVIDIA PhysX SDK document: Docs.
4 design and use paradigms for Gazebo: PDF.
5 Professional mobile robot simulation: PDF.
6 Developer overview: Docs. Renderer path tracing: Docs.
7 PyBullet project site: Pybullet.org. PyBullet quickstart guide: PDF.
8 MuJoCo project site: Github.io. MuJoCo guide document: Docs.
9 Isaac Gym project site: Homepage. Technical report: PDF.
10 Genesis project site: Github.io. Genesis quickstart guide: Docs.

 doi:10.20944/preprints202603.0739.v1

tasks is severely compromised by computational overhead. Computational demands arising from multi-agent coordination, high-frequency perception-rendering, and contact-rich interactions often result in slow solution speeds. Consequently, when addressing closed-loop decision-making tasks, a trade-off persists between contact modeling accuracy, numerical stability, and large-scale parallel efficiency.

4.3. Performance Comparison

Table 9 compares the performance of different WMs on the WorldScore benchmark. Additional performance rankings are tracked and updated live through WorldScore-Leaderboard. As shown in Table 9, TeleWorld [284] achieves the strongest and most balanced performance, leading in key metrics including WS-S, WS-D, 3D consistency (3DCo), photometric consistency (PCon), semantic consistency (SCon), and subjective quality (SQua), indicating superior visual fidelity and temporal coherence. 3D-based methods such as Text2Room [110] and WonderJourney [98] substantially excel in controllability-related metrics, particularly camera controllability (CCon), but show near-zero scores on motion-related metrics, highlighting their limitations in dynamic scene modeling. In contrast, video-based models (*e.g.*, Gen-3[11] and T2V-Turbo [285]) demonstrate better motion accuracy and smoothness, yet underperform in fine-grained controllability. Methods such as SceneScape [109] and 4D-fy [96] exhibit consistently low performance across most metrics, reflecting challenges in both consistency and dynamics. Overall, the comparison results reveal a clear trade-off between controllability, visual consistency, and motion realism across existing approaches.

Table 9. Performance comparison of observation-level generative WMs on the WorldScore dataset [270]. The best-performing values are highlighted in **red**, while the second-best results are shown in **blue**.

Model	Type	Accessibility	Ability	WS-S	WS-D	CCon	OCon	CAli	3DCo	PCon	SCon	SQua	MAcc	MMag	MSmo
TeleWorld [284]	Video	Open Source	Image-to-Video	**78.23**	**66.73**	76.58	74.44	**73.20**	87.35	**88.82**	**85.59**	61.66	**53.94**	31.55	34.18
Gen-3[11]	Video	API	Image-to-Video	60.71	57.58	29.47	**62.92**	50.49	68.31	87.09	62.82	**63.85**	**54.53**	27.48	**68.87**
VideoCrafter2 [18]	Video	Open Source	Text-to-Video	52.57	47.49	28.92	39.07	**72.46**	65.14	61.85	43.79	56.74	47.12	30.40	29.39
Wan2.1 [104]	Video	Open Source	Image-to-Video	57.56	52.85	23.53	40.32	45.44	78.74	78.36	77.18	59.38	54.27	**33.26**	38.05
T2V-Turbo [285]	Video	Open Source	Text-to-Video	45.65	40.20	27.80	30.68	69.14	38.72	34.84	49.65	**68.74**	34.87	**40.09**	7.48
DynamiCrafter [286]	Video	Open Source	Image-to-Video	52.09	47.19	25.15	47.36	25.00	72.90	60.95	**78.85**	54.40	41.11	**39.25**	26.92
WonderWorld [19]	3D	Open Source	Image-to-Video	**72.69**	50.88	**92.98**	51.76	71.25	86.87	85.56	70.57	49.81	0.00	0.00	0.00
WonderJourney [98]	3D	Open Source	Image-to-Video	63.75	44.63	84.60	37.10	35.54	80.60	79.03	62.82	**66.56**	0.00	0.00	0.00
Text2Room [110]	3D	Open Source	Image-to-Video	62.10	43.47	**94.01**	38.93	50.79	**88.71**	**88.36**	37.23	36.69	0.00	0.00	0.00
InvisibleStitch [111]	3D	Open Source	Image-to-Video	61.12	42.78	93.20	36.51	29.53	**88.51**	**89.19**	32.37	58.50	0.00	0.00	0.00
SceneScape [109]	3D	Open Source	Text-to-Video	50.73	35.51	84.99	47.44	28.64	76.54	62.88	21.85	32.75	0.00	0.00	0.00
4D-fy [96]	4D	Open Source	Text-to-Video	27.98	32.10	69.92	55.09	0.85	35.47	1.59	32.04	0.89	22.22	22.88	**80.06**

5. Challenges & Future Directions

5.1. Scientific Modeling

Existing approaches, such as Dreamer [234], Dyn-O [45], and NWM [166], often focus on fitting the distribution of observational data to improve visual fidelity, controllability, or motion realism. As a result, they can generate visually plausible rollouts while violating physical laws. As shown in Table 3, future research should focus on developing scientific modeling that not only provides predictive capabilities but also offers strong explanatory power. In scientific modeling, knowledge is expressed through explicit symbolic forms, such as equations and formal relations independent of specific datasets, experiments, or observers. Progress is achieved by incrementally extending a shared theoretical framework rather than modifying opaque and tightly coupled neural parameters. Such knowledge is inherently inspectable, enabling verification through proof checking, experimental replication, and edge-case analysis. Moreover, scientific modeling explicitly specifies the conditions under which it holds, making both applicability and failure modes transparent and principled. A promising direction is to integrate symbolic and neural components, impose hard physical constraints (*e.g.*, conservation laws), and develop protocols that reward causal discovery and falsifiability.

[11] Gen-3 project site: Github.io.

 doi:10.20944/preprints202603.0739.v1

5.2. Long-Horizon Consistency & Causal Reasoning

Maintaining consistency over long temporal horizons remains a major obstacle for WMs. Errors in prediction often accumulate, leading to visual drift, broken physical interactions, or unrealistic future states. Moreover, many WMs, such as Wan [104], WonderJourney [98], and 4D-fy [96], rely on correlational patterns rather than causal understanding, limiting their reliability in closed-loop or multi-step rollouts. Future advances in hierarchical temporal modeling and causal representation learning may improve long-horizon stability. Incorporating explicit mechanisms for state abstraction, memory, and causal intervention could enable WMs to reason beyond short-term correlations. Benchmarks that emphasize multi-step prediction and counterfactual reasoning will further drive progress in this direction.

5.3. Grounding in Physical and Semantic Constraints

Despite impressive visual synthesis, many WMs [2,18,286] lack grounding in physical laws and semantic structure, leading to implausible motions, inconsistent object interactions, and semantically incoherent scene evolution. This limits their applicability to real-world decision-making and robotics. Future research directions include integrating physics-informed priors, differentiable simulators, or symbolic knowledge, as well as combining neural representations with structured semantic graphs or object-centric models to improve physical plausibility and interpretability. Furthermore, bridging neural and symbolic approaches remains a promising path toward grounded world modeling.

5.4. Generalization & Scalability in Real-World

Most current WMs [1,3,7] are trained and evaluated within narrow domains, which raises concerns about their generalization across environments, tasks, and embodiments. Moreover, the high computational and data requirements of training large-scale WMs hinder scalability and real-world deployment. To tackle these challenges, future research may explore foundation WMs trained on diverse, multimodal interaction data, enabling broad generalization. In addition, data-efficient learning, parameter-efficient fine-tuning, and continual learning strategies are likely to be critical for scalability. Emphasis on real-world deployment will further push WMs toward practical AI systems.

6. Conclusions

This survey presents a systematic review of WMs in artificial intelligence. We introduce a unified definition and categorize existing methods into four paradigms: observation-level generative, latent-space, RL-based, and object-centric WMs, and review their applications across robotics, autonomous driving, scientific discovery, virtual game simulation, GUI-based agents, as well as interpretability and trustworthiness. We also summarize benchmark datasets, evaluation metrics, simulation platforms, and comparative results. Moreover, we identify key challenges limiting the deployment and generalization of WMs, including scientific modeling, long-horizon consistency, controllability, robustness, evaluation limitations, and transfer across tasks and embodiments. We have also discussed potential directions that emphasize integrated modeling, scalable training, principled benchmarking, and real-world validation. This survey serves as a foundation for advancing more generalizable, and trustworthy WMs. Looking ahead, WMs should move beyond scaling predictive accuracy toward establishing physical consistency and explanatory structure, with internal representations examinable against physical laws and invariances. Rather than manually encoding physical laws, WMs are trained on large-scale data to autonomously learn representations and predictive capabilities of the physical world through self-evolutionary learning. Advancing toward this goal requires tighter integration of learning, dynamics, and causality, shifting world modeling from large-scale statistical interpolation toward a scientific modeling framework.

References

1. Zhu, F.; et al. Irasim: Learning interactive real-robot action simulators. *arXiv* **2024**.

 doi:10.20944/preprints202603.0739.v1

2. Wang, J.; Ma, A.; Cao, K.; et al. WISA: World simulator assistant for physics-aware text-to-video generation. In Proceedings of the NeurIPS, 2025.
3. Hafner, D.; Pasukonis, J.; et al. Mastering diverse control tasks through world models. *Nature* **2025**, pp. 1–7.
4. Ha, D.R.; Schmidhuber, J. World Models. *arXiv* **2018**.
5. Zhao, C.; Zhang, R.; et al. World Models for Cognitive Agents: Transforming Edge Intelligence in Future Networks. *arXiv* **2025**.
6. Janner, M.; et al. Planning with Diffusion for Flexible Behavior Synthesis. In Proceedings of the ICML, 2022.
7. Assran, M.; Bardes, A.; et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv* **2025**.
8. Ghaemi, H.; et al. seq-JEPA: Autoregressive Predictive Learning of Invariant-Equivariant World Models. In Proceedings of the NeurIPS, 2025.
9. Hafner, D.; Lillicrap, T.; et al. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the ICLR, 2020.
10. Hafner, D.; et al. Mastering Atari with Discrete World Models. In Proceedings of the ICLR, 2021.
11. Hansen, N.; Su, H.; Wang, X. TD-MPC2: Scalable, Robust World Models for Continuous Control. In Proceedings of the ICLR, 2024.
12. Yang, S.; Du, Y.; Dai, B.; et al. Probabilistic Adaptation of Black-Box Text-to-Video Models. In Proceedings of the ICLR, 2024.
13. Kotar, K.; Lee, W.; Venkatesh, R.; et al. World Modeling with Probabilistic Structure Integration. *arXiv* **2025**.
14. Kang, B.; Yue, Y.; Lu, R.; et al. How Far Is Video Generation from World Model: A Physical Law Perspective. In Proceedings of the ICML, 2025.
15. Yang, H. Utilizing World Models for Adaptively Covariate Acquisition Under Limited Budget for Causal Decision Making. In Proceedings of the ICLR Workshop, 2025.
16. Feng, F.; Lippe, P.; Magliacane, S. Learning Interactive World Model for Object-Centric Reinforcement Learning. *arXiv* **2025**.
17. Lee, H.; Lee, Y.; et al. Hyperspherical Normalization for Scalable Deep Reinforcement Learning. In Proceedings of the ICML, 2025.
18. Chen, H.; et al. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In Proceedings of the CVPR, 2024.
19. Yu, H.X.; Duan, H.; et al. WonderWorld: Interactive 3D Scene Generation from a Single Image. *CVPR* **2024**.
20. Zhang, W.; Jelley, A.; McInroe, T.; et al. Objects Matter: Object-Centric World Models Improve Reinforcement Learning in Visually Complex Environments. In Proceedings of the RLC Workshop, 2025.
21. Zhou, S.; Zhou, T.; et al. WALL-E 2.0: World Alignment by NeuroSymbolic Learning improves World Model-based LLM Agents. *arXiv* **2025**.
22. Zhang, B.; Wang, R.; Xiao, W.; et al. DyMoDreamer: World Modeling with Dynamic Modulation. In Proceedings of the NeurIPS, 2025.
23. Chi, X.; Jia, P.o. Wow: Towards a world omniscient world model through embodied interaction. *arXiv* **2025**.
24. Samsami, M.R.; et al. Mastering Memory Tasks with World Models. In Proceedings of the ICLR, 2024.
25. Wang, Y.; Wan, S.; Gan, L.; et al. AD3: Implicit Action is the Key for World Models to Distinguish the Diverse Visual Distractors. In Proceedings of the ICML, 2024.
26. Hansen, N.A.; et al. Temporal Difference Learning for Model Predictive Control. In Proceedings of the ICML, 2022.
27. Zhou, G.; Pan, H.; LeCun, Y.; et al. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In Proceedings of the ICML, 2025.
28. team, F.C.; Copet, J.; et al. CWM: An Open-Weights LLM for Research on Code Generation with World Models. *arXiv* **2025**.
29. Brooks, T.; Peebles, B.; Holmes, C.; et al. Video generation models as world simulators. *Technical Report* **2024**.
30. Acuaviva, P.; et al. From Generation to Generalization: Emergent Few-Shot Learning in Video Diffusion Models. *arXiv* **2025**.
31. Wang, Z.; Wei, X.; Li, B.; et al. VideoVerse: How Far is Your T2V Generator from a World Model? *arXiv* **2025**.
32. Karypidis, E.; Kakogeorgiou, I.; et al. DINO-Foresight: Looking into the Future with DINO. In Proceedings of the NeurIPS, 2025.
33. Wang, X.; Zhang, X.; Luo, Z.; et al. Emu3: Next-Token Prediction is All You Need. *arXiv* **2024**.
34. Gkountouras, J.; et al. Language Agents Meet Causality – Bridging LLMs and Causal World Models. In Proceedings of the ICLR, 2025.

 doi:10.20944/preprints202603.0739.v1

35. Zhang, Y.; et al. Revisiting Multi-Agent World Modeling from a Diffusion-Inspired Perspective. In Proceedings of the NeurIPS, 2025.
36. Mattes, P.; Schlosser, R.; Herbrich, R. Hieros: Hierarchical Imagination on Structured State Space Sequence World Models. In Proceedings of the ICML, 2024.
37. Brito, C.S.; et al. World Models as Reference Trajectories for Rapid Motor Adaptation. In Proceedings of the NeurIPS, 2025.
38. Levy, G.; Colas, C.; et al. WorldLLM: Improving LLMs' world modeling using curiosity-driven theory-making. *arXiv* **2025**.
39. Wu, Z.; Dvornik, N.; et al. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models. In Proceedings of the ICLR, 2023.
40. Locatello, F.; Weissenborn, D.; et al. Object-centric learning with slot attention. In Proceedings of the NeurIPS, 2020.
41. Chae, H.; Kim, N.; iunn Ong, K.T.; et al. Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation. In Proceedings of the ICLR, 2025.
42. Hao, S.; Gu, Y.; Ma, H.; et al. Reasoning with Language Model is Planning with World Model. In Proceedings of the EMNLP, 2023.
43. Foffano, D.; Russo, A.; Proutiere, A. Adversarial Diffusion for Robust Reinforcement Learning. In Proceedings of the NeurIPS, 2025.
44. Wu, T.; Yang, S.; Po, R.; et al. Video World Models with Long-term Spatial Memory. In Proceedings of the NeurIPS, 2025.
45. Wang, Z.; et al. Dyn-O: Building Structured World Models with Object-Centric Representations. In Proceedings of the NeurIPS, 2025.
46. Lee, V.; Abbeel, P.; et al. DreamSmooth: Improving Model-based Reinforcement Learning via Reward Smoothing. In Proceedings of the ICLR, 2024.
47. Alonso, E.; Jelley, A.; et al. Diffusion for world modeling: Visual details matter in atari. *NeurIPS* **2024**.
48. Li, L.; Fan, Z.; Cong, W.; et al. Martian World Model: Controllable Video Synthesis with Physically Accurate 3D Reconstructions. In Proceedings of the neurIPS, 2025.
49. Park, B.; Go, H.; et al. SteerX: Creating Any Camera-Free 3D and 4D Scenes with Geometric Steering. *ICCV* **2025**.
50. Che, H.; et al. GameGen-X: Interactive Open-world Game Video Generation. In Proceedings of the ICLR, 2025.
51. Jin, Y.; et al. Pyramidal Flow Matching for Efficient Video Generative Modeling. In Proceedings of the ICLR, 2025.
52. Barhdadi, M.R.; et al. PhysicsNeRF: Physics-Guided 3D Reconstruction from Sparse Views. In Proceedings of the ICML Workshop, 2025.
53. Chen, B.; Monsó, D.M.; et al. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In Proceedings of the NeurIPS, 2024.
54. Assran, M.; et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In Proceedings of the CVPR, 2023.
55. Bardes, A.; Garrido, Q.; Ponce, J.; et al. Revisiting Feature Prediction for Learning Visual Representations from Video. *arXiv* **2024**.
56. Bardhan, J.; Agrawal, R.; et al. HEP-JEPA: A foundation model for collider physics. In Proceedings of the ICLR Workshop, 2025.
57. Wang, B.; Meng, X.; et al. EmbodieDreamer: Advancing Real2Sim2Real Transfer for Policy Training via Embodied World Modeling. *arXiv* **2025**.
58. Hao, C.; et al. Neural Motion Simulator Pushing the Limit of World Models in Reinforcement Learning. In Proceedings of the CVPR, 2025.
59. Park, K.; Lee, Y. Model-based Offline Reinforcement Learning with Lower Expectile Q-Learning. In Proceedings of the ICLR, 2025.
60. Lin, Z.; Wu, Y.F.; Peri, S.; et al. Improving Generative Imagination in Object-Centric World Models. In Proceedings of the ICML, 2020.
61. GX-Chen, A.; Marino, K.; Fergus, R. Efficient Exploration and Discriminative World Model Learning with an Object-Centric Abstraction. In Proceedings of the ICLR, 2025.
62. Lu, J.; Huang, Z.; Yang, Z.; et al. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In Proceedings of the ECCV, 2024.

 doi:10.20944/preprints202603.0739.v1

63. Gao, S.; Zhou, S.; et al. AdaWorld: Learning Adaptable World Models with Latent Actions. In Proceedings of the ICML, 2025.
64. Zhang, X.; et al. Social World Model-Augmented Mechanism Design Policy Learning. In Proceedings of the NeurIPS, 2025.
65. Yue, Y.; Wang, Y.; et al. CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning. In Proceedings of the CVPR, 2025.
66. Hafner, D.; Lillicrap, T.; Fischer, I.; et al. Learning latent dynamics for planning from pixels. In Proceedings of the ICML, 2019.
67. Gao, S.; Yang, J.; Chen, L.; et al. Vista: A generalizable driving world model with high fidelity and versatile controllability. In Proceedings of the NeurIPS, 2024.
68. Cheng, J.; Ge, Y.; et al. Animegamer: Infinite anime life simulation with next game state prediction. In Proceedings of the ICCV, 2025.
69. Qiao, S.; Fang, R.; et al. Agent planning with world knowledge model. *NeurIPS* **2024**.
70. Zhang, Y.; Su, Y.; et al. CellFlux: Simulating Cellular Morphology Changes via Flow Matching. In Proceedings of the ICML, 2025.
71. Zhao, Z.; et al. From Forecasting to Planning: Policy World Model for Collaborative State-Action Prediction. In Proceedings of the NeurIPS, 2025.
72. Agro, B.; Sykora, Q.; et al. Uno: Unsupervised occupancy fields for perception and forecasting. In Proceedings of the CVPR, 2024.
73. Chua, K.; et al. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In Proceedings of the NeurIPS, 2018.
74. Yang, Y.; et al. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv* **2025**.
75. Feng, T.; Wang, W.; et al. A survey of world models for autonomous driving. *arXiv* **2025**.
76. Fu, A.; Zhou, Y.; Zhou, T.; et al. Exploring the interplay between video generation and world models in autonomous driving: A survey. *arXiv* **2024**.
77. Guan, Y.o. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* **2024**.
78. Tu, S.; Zhou, X.; et al. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv* **2025**.
79. Zhao, J.; Zhao, W.; et al. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications* **2024**.
80. Zablocki, É.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of deep vision-based autonomous driving systems: Review and challenges. *IJCV* **2022**.
81. Kong, L.; Yang, W.; et al. 3D and 4D World Modeling: A Survey. *arXiv* **2025**.
82. Xie, N.; et al. From 2D to 3D Cognition: A Brief Survey of General World Models. *arXiv* **2025**.
83. Baraldi, L.; et al. The Safety Challenge of World Models for Embodied AI Agents: A Review. *arXiv* **2025**.
84. Zhu, Z.; et al. Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond. *arXiv* **2025**.
85. Lin, M.; et al. Exploring the Evolution of Physics Cognition in Video Generation: A Survey. *arXiv* **2025**.
86. Liu, Y.; et al. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI. *arXiv* **2025**.
87. Li, X.; He, X.; et al. A Comprehensive Survey on World Models for Embodied AI. *arXiv* **2025**.
88. Long, X.; et al. A Survey: Learning Embodied Intelligence from Physical Simulators and World Models. *arXiv* **2025**.
89. Fung, P.; Bachrach, Y.; et al. Embodied AI Agents: Modeling the World. *arXiv* **2025**.
90. Zhang, P.F.; et al. A Step Toward World Models: A Survey on Robotic Manipulation. *arXiv* **2025**.
91. Sun, J.; et al. Integrating World Models into Vision Language Action and Navigation: A Comprehensive Survey. *TechRxiv* **2025**.
92. Ding, J.; et al. Understanding World or Predicting Future? A Comprehensive Survey of World Models. *ACM Comput. Surv.* **2025**, *58*.
93. Ser, J.D.; et al. World Models in Artificial Intelligence: Sensing, Learning, and Reasoning Like a Child. *arXiv* **2025**.
94. Xu, K.; Zhao, H.; et al. From Specialist to Generalist: A Comprehensive Survey on World Models. *TechRxiv* **2026**.

 doi:10.20944/preprints202603.0739.v1

95. Xie, K.; et al. Making Large Language Models into World Models with Precondition and Effect Knowledge. *arXiv* **2024**.
96. Bahmani, S.; et al. 4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling. *CVPR* **2023**.
97. Liu, H.; Yan, W.; Zaharia, M.; et al. World model on million-length video and language with blockwise ringattention. *arXiv* **2024**.
98. Yu, H.X.; Duan, H.; Hur, J.; et al. WonderJourney: Going from Anywhere to Everywhere. *CVPR* **2023**.
99. OpenAI.; Achiam, J.; et al. GPT-4 Technical Report. *arXiv* **2024**.
100. Lu, G.; et al. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In Proceedings of the ECCV, 2024, pp. 349–366.
101. Wang, R.; Todd, G.; Xiao, Z.; et al. Can Language Models Serve as Text-Based World Simulators? In Proceedings of the ACL, 2024.
102. Grattafiori, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**.
103. Lyu, B.; Huang, S.; Liang, Z. SURGE: On the Potential of Large Language Models as General-Purpose Surrogate Code Executors. In Proceedings of the EMNLP, 2025.
104. Wang, A.; Ai, B.; Wen, B.; et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv* **2025**.
105. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the NeurIPS, 2023.
106. Huang, S.; Wu, J.; Zhou, Q.; et al. Vid2World: Crafting Video Diffusion Models to Interactive World Models. *arXiv* **2025**.
107. Wang, Y.; Zhang, F.; .; et al. Co-Evolving Latent Action World Models. *arXiv* **2025**.
108. Liang, A.; Liu, Y.; Yang, Y.; et al. LiDARCrafter: Dynamic 4D World Modeling from LiDAR Sequences. In Proceedings of the AAAI, 2025.
109. Fridman, R.; et al. SceneScape: Text-Driven Consistent Scene Generation. In Proceedings of the NeurIPS, 2023.
110. Höllein, L.; Cao, A.; et al. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. *ICCV* **2023**.
111. Engstler, P.; Vedaldi, A.; et al. Invisible Stitch: Generating Smooth 3D Scenes with Depth Inpainting. In Proceedings of the 3DV, 2024.
112. Huang, W.; Chao, Y.W.; et al. PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation. *arXiv* **2026**.
113. Bardes, A.; et al. MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features. *arXiv* **2023**.
114. Baldassarre, F.; et al. Back to the Features: DINO as a Foundation for Video World Models. *arXiv* **2024**.
115. Delliaux, T.; Vu, N.K.; Francois-Lavet, V.; et al. Learning Abstract World Models with a Group-Structured Latent Space. In Proceedings of the EWRL, 2025.
116. Cohen, L.; Wang, K.; Kang, B.; et al. Improving Token-Based World Models with Parallel Observation Prediction. In Proceedings of the ICML, 2024.
117. Ma, H.; Wu, J.; Feng, N.; et al. HarmonyDream: Task Harmonization Inside World Models. In Proceedings of the ICML, 2024.
118. Huang, D.; WANG, J.; Li, Y.; et al. PIGDreamer: Privileged Information Guided World Models for Safe Partially Observable Reinforcement Learning. In Proceedings of the ICML, 2025.
119. Wang, Q.; Yang, J.; Wang, Y.; et al. Making Offline RL Online: Collaborative World Models for Offline Visual Reinforcement Learning. In Proceedings of the NeurIPS, 2024.
120. Chen, R.; Chen, X.H.; et al. Policy-conditioned Environment Models are More Generalizable. In Proceedings of the ICML, 2024.
121. Li, S.; Huang, Z.; Su, H. Reward-free World Models for Online Imitation Learning. In Proceedings of the ICML, 2025.
122. Rigter, M.; Jiang, M.; Posner, I. Reward-Free Curricula for Training Robust World Models. In Proceedings of the ICLR, 2024.
123. Georgiev, I.; Giridhar, V.; Hansen, N.; et al. PWM: Policy Learning with Multi-Task World Models. In Proceedings of the ICLR, 2025.
124. Zheng, R.; Wang, J.; et al. FLARE: Robot Learning with Implicit World Modeling. *arXiv* **2025**.
125. Ajay, A.; Du, Y.; Gupta, A.; et al. Is Conditional Generative Modeling all you need for Decision Making? In Proceedings of the ICLR, 2023.
126. Zhang, K.; et al. PIVOT-R: Primitive-Driven Waypoint-Aware World Model for Robotic Manipulation. In Proceedings of the NeuIPS, 2024.

 doi:10.20944/preprints202603.0739.v1

127. Liao, Y.; Zhou, P.; et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv* **2025**.
128. Zhang, C.; Wu, Z.; Lu, G.; et al. iMoWM: Taming Interactive Multi-Modal World Model for Robotic Manipulation. *arXiv* **2025**.
129. Wu, J.; Yin, S.; et al. ivideogpt: Interactive videogpts are scalable world models. *NeurIPS* **2024**.
130. Guo, J.; Ma, X.; Wang, Y.; et al. FlowDreamer: A RGB-D World Model with Flow-based Motion Representations for Robot Manipulation. *arXiv* **2025**.
131. Zhen, H.; Sun, Q.; et al. TesserAct: Learning 4D Embodied World Models. *arXiv* **2025**.
132. Ferraro, S.; Nakano, A.; et al. When Object-Centric World Models Meet Policy Learning: From Pixels to Policies, and Where It Breaks. *arXiv* **2025**.
133. Kapl, F.; et al. Object-Centric Representations Generalize Better Compositionally with Less Compute. In Proceedings of the ICLR Workshop, 2025.
134. Chen, H.; Wang, B.; et al. World4Omni: A Zero-Shot Framework from Image Generation World Model to Robotic Manipulation. *arXiv* **2025**.
135. Hamdan, S.; Güney, F. CarFormer: Self-driving with Learned Object-Centric Representations. In Proceedings of the ECCV, 2025.
136. Jeong, Y.; Chun, J.; et al. Object-centric world model for language-guided manipulation. *arXiv* **2025**.
137. Liang, W.; Sun, G.; et al. PixelVLA: Advancing Pixel-level Understanding in Vision-Language-Action Model. *arXiv* **2025**.
138. Barcellona, L.; et al. Dream to Manipulate: Compositional World Models Empowering Robot Imitation Learning with Imagination. In Proceedings of the ICLR, 2025.
139. Goswami, R.G.; et al. OSVI-WM: One-Shot Visual Imitation for Unseen Tasks using World-Model-Guided Trajectory Generation. *arXiv* **2025**.
140. López Escoriza, A.; et al. Multi-Stage Manipulation with Demonstration-Augmented Reward, Policy, and World Model Learning. In Proceedings of the ICML, 2025.
141. Qi, H.; Yin, H.; et al. Strengthening Generative Robot Policies through Predictive World Modeling. *arXiv* **2025**.
142. Ajay, A.; Han, S.; et al. Compositional Foundation Models for Hierarchical Planning. In Proceedings of the NeurIPS, 2023.
143. Pezzato, C.; et al. Mobile Manipulation with Active Inference for Long-Horizon Rearrangement Tasks. *arXiv* **2025**.
144. Nhu, A.N.; et al. Time-Aware World Model for Adaptive Prediction and Control. In Proceedings of the ICML, 2025.
145. Luo, Y.; Sun, C.; et al. Potential Based Diffusion Motion Planning. In Proceedings of the ICML, 2024.
146. Li, Y.; Wei, X.; et al. ManipDreamer: Boosting Robotic Manipulation World Model with Action Tree and Visual Guidance. *arXiv* **2025**.
147. Chen, Z.; Huo, J.; Chen, Y.; Gao, Y. RoboHorizon: An LLM-Assisted Multi-View World Model for Long-Horizon Robotic Manipulation. *arXiv* **2025**.
148. Song, Z.; Qin, S.; Chen, T.; Lin, L.; Wang, G. Physical Autoregressive Model for Robotic Manipulation without Action Pretraining. *arXiv* **2025**.
149. Lykov, A.; Sam, J.; et al. PhysicalAgent: Towards General Cognitive Robotics with Foundation World Models. *arXiv* **2025**.
150. Zhou, S.; Du, Y.; Chen, J.; et al. RoboDreamer: Learning Compositional World Models for Robot Imagination. In Proceedings of the ICML, 2024, pp. 61885–61896.
151. Luo, Y.; Du, Y. Grounding Video Models to Actions through Goal Conditioned Exploration. In Proceedings of the ICLR, 2025.
152. Routray, S.; Pan, H.; et al. ViPRA: Video Prediction for Robot Actions. *arXiv* **2025**.
153. Yang, X.; Li, B.; et al. ORV: 4D Occupancy-centric Robot Video Generation. *arXiv* **2025**.
154. Qian, Z.; Chi, X.; et al. WristWorld: Generating Wrist-Views via 4D World Models for Robotic Manipulation. *arXiv* **2025**.
155. Fu, X.; Wang, X.; et al. Learning Video Generation for Robotic Manipulation with Collaborative Trajectory Control. *arXiv* **2025**.
156. Feng, Y.; Tan, H.; et al. Vidar: Embodied Video Diffusion Model for Generalist Bimanual Manipulation. *arXiv* **2025**.

 doi:10.20944/preprints202603.0739.v1

157. Huang, Y.; Zhang, J.; et al. LaDi-WM: A Latent Diffusion-based World Model for Predictive Manipulation. *arXiv* **2025**.
158. Li, S.; Hao, Q.; Shang, Y.; Li, Y. KeyWorld: Key Frame Reasoning Enables Effective and Efficient World Models. *arXiv* **2025**.
159. Bar, A.; Zhou, G.; Tran, D.; Darrell, T.; LeCun, Y. Navigation world models. In Proceedings of the CVPR, 2025.
160. Yang, Y.; Liu, J.; Zhang, Z.; et al. MindJourney: Test-Time Scaling with World Models for Spatial Reasoning. *arXiv* **2025**.
161. Hu, Y.; et al. Imaginative World Modeling with Scene Graphs for Embodied Agent Navigation. *arXiv* **2025**.
162. Yao, X.; et al. NavMorph: A Self-Evolving World Model for Vision-and-Language Navigation in Continuous Environments. *arXiv* **2025**.
163. Dong, Y.; et al. Unified World Models: Memory-Augmented Planning and Foresight for Visual Navigation. *arXiv* **2025**.
164. Shah, D.; et al. Rapid Exploration for Open-World Navigation with Latent Goal Models. *arXiv* **2021**.
165. Nie, D.; et al. WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation. *arXiv* **2025**.
166. Wang, W.; et al. Deductive Chain-of-Thought Augmented Socially-aware Robot Navigation World Model. *arXiv* **2025**.
167. Alcedo, K.; et al. Perspective-Shifted Neuro-Symbolic World Models: A Framework for Socially-Aware Robot Navigation. *arXiv* **2025**.
168. Li, H.; et al. Scaling Inference-Time Search with Vision Value Models for Improved Visual Comprehension. In Proceedings of the ICLR Workshop, 2025.
169. Damm, E.R.; et al. Kinodynamic Motion Planning for Mobile Robot Navigation across Inconsistent World Models. In Proceedings of the RSS Workshop on Resilient Off-road Autonomous Robotics, 2025.
170. Liu, W.; et al. X-mobility: End-to-end generalizable navigation via world modeling. In Proceedings of the ICRA, 2025, pp. 7569–7576.
171. Miller, T.; et al. FalconWing: An Ultra-Light Fixed-Wing Platform for Indoor Aerial Applications. In Proceedings of the NeurIPS Workshop, 2025.
172. Deng, Y.; Hanna, J.P. Abstract Sim2Real through Approximate Information States. In Proceedings of the NeurIPS Workshop, 2025.
173. Zhou, S.; et al. Learning 3D Persistent Embodied World Models. *arXiv* **2025**.
174. Yoo, M.; et al. World Model Implanting for Test-time Adaptation of Embodied Agents. *arXiv* **2025**.
175. Yokozawa, R.; et al. Deep Active Inference with Diffusion Policy and Multiple Timescale World Model for Real-World Exploration and Navigation. *arXiv* **2025**.
176. Zhu, F.; Yan, Z.; .; et al. WMPO: World Model-based Policy Optimization for Vision-Language-Action Models. *arXiv* **2025**.
177. Jiang, Z.; Liu, K.; Qin, Y.; et al. World4RL: Diffusion World Models for Policy Refinement with Reinforcement Learning for Robotic Manipulation. *arXiv* **2025**.
178. Li, Z.; Han, X.; .; et al. DAWM: Diffusion Action World Models for Offline Reinforcement Learning via Action-Inferred Transitions. *arXiv* **2025**.
179. Alles, M.; et al. Latent Action World Models for Control with Unlabeled Trajectories. *arXiv* **2025**.
180. Zhang, L.; Kan, M.; et al. Prelar: World model pre-training with learnable action representation. In Proceedings of the ECCV, 2024.
181. Goswami, R.G.; Bar, A.; et al. World Models Can Leverage Human Videos for Dexterous Manipulation. *arXiv* **2025**.
182. He, Z.; Ai, B.; et al. Scaling Cross-Embodiment World Models for Dexterous Manipulation. *arXiv* **2025**.
183. Lee, S.; Jung, Y.; et al. TraceGen: World Modeling in 3D Trace Space Enables Learning from Cross-Embodiment Videos. *arXiv* **2025**.
184. Zhi, H.; Chen, P.; et al. 3DFlowAction: Learning Cross-Embodiment Manipulation from 3D Flow World Model. *arXiv* **2025**.
185. Guo, Y.; Shi, L.X.; et al. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv* **2025**.
186. Li, C.; .; et al. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv* **2025**.

 doi:10.20944/preprints202603.0739.v1

187. Quevedo, J.; Sharma, A.K.; et al. WorldGym: World Model as An Environment for Policy Evaluation. *arXiv* **2025**.
188. Li, Y.; et al. WorldEval: World Model as Real-World Robot Policies Evaluator. *arXiv* **2025**.
189. Zhang, L.; Xiong, Y.; Yang, Z.; et al. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. In Proceedings of the ICLR, 2024.
190. Min, C.; Zhao, D.; Xiao, L.; Nie, Y.; Dai, B. Uniworld: Autonomous driving pre-training via world models. *arXiv* **2023**.
191. Wang, Y.; He, J.; Fan, L.; et al. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In Proceedings of the CVPR, 2024.
192. Hu, A.; Russell, L.; Yeo, H.; et al. Gaia-1: A generative world model for autonomous driving. *arXiv* **2023**.
193. Li, Q.; Jia, X.; Wang, S.; et al. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In Proceedings of the ECCV, 2024.
194. Wang, H.; Ye, X.; et al. Adawm: Adaptive world model based planning for autonomous driving. *ICLR* **2025**.
195. Li, Y.; Shang, S.; et al. DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving. *arXiv* **2025**.
196. Lai, H.; Cao, J.; et al. World model-based perception for visual legged locomotion. In Proceedings of the ICRA, 2025.
197. Sun, W.; Chen, L.; .; et al. Learning humanoid locomotion with world model reconstruction. *arXiv* **2025**.
198. Raja, G.; Agishev, R.; et al. ProTerrain: Probabilistic Physics-Informed Rough Terrain World Modeling. *arXiv* **2025**.
199. Gu, X.; Wang, Y.J.; et al. Advancing Humanoid Locomotion: Mastering Challenging Terrains with Denoising World Model Learning. *RSS* **2024**.
200. Liu, H.; Gao, Y.; et al. Ego-Vision World Model for Humanoid Contact Planning. *arXiv* **2025**.
201. Wu, Z.; Ni, J.; et al. Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving. *arXiv* **2024**.
202. Huang, Z.; Zhang, J.; et al. Neural volumetric world models for autonomous driving. In Proceedings of the ECCV, 2024.
203. Yan, Z.; Dong, W.; Shao, Y.; et al. Renderworld: World model with self-supervised 3d label. In Proceedings of the ICRA, 2025.
204. Zhang, H.; et al. Machine learning methods for weather forecasting: A survey. *Atmosphere* **2025**.
205. Zhang, Y.; et al. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* **2023**.
206. Min, C.; Zhao, D.; Xiao, L.; et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In Proceedings of the CVPR, 2024.
207. Zheng, W.; Chen, W.; et al. Occworld: Learning a 3d occupancy world model for autonomous driving. In Proceedings of the ECCV, 2024.
208. Wang, X.; et al. Drivedreamer: Towards real-world-drive world models for autonomous driving. In Proceedings of the ECCV, 2024.
209. Zhao, G.; Wang, X.; et al. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In Proceedings of the AAAI, 2025.
210. Guo, X.; Ding, C.; et al. Infinitydrive: Breaking time limits in driving world models. *arXiv* **2024**.
211. Lyu, J.; Li, Z.; et al. DyWA: Dynamics-adaptive World Action Model for Generalizable Non-prehensile Manipulation. *arXiv* **2025**.
212. Zheng, W.; Xia, Z.; et al. Doe-1: Closed-loop autonomous driving with large world model. *arXiv* **2024**.
213. Yu, J.; et al. Gamefactory: Creating new games with generative interactive videos. *arXiv* **2025**.
214. Zhou, X.; Liu, J.; et al. Social World Models. *arXiv* **2025**.
215. Team, F.D. SocioVerse: A World Model for Social Simulation Powered by LLM Agents and a Pool of 10 Million Real-World Users. *arXiv* **2025**.
216. Yang, Y.; et al. TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets. In Proceedings of the NeurIPS, 2025.
217. Zhang, C.; Shi, J.; Sui, Y. A Virtual Reality-Integrated System for Behavioral Analysis in Neurological Decline. In Proceedings of the ICLR Workshop, 2025.
218. Sun, L.; Huang, H.; et al. Bidding for Influence: Auction-Driven Diffusion Image Generation. In Proceedings of the ICML Workshop, 2025.
219. Cao, D.Y.; et al. Effectively Designing 2-Dimensional Sequence Models for Multivariate Time Series. In Proceedings of the ICLR Workshop, 2025.

 doi:10.20944/preprints202603.0739.v1

220. Lab, L.; et al. ODesign: A World Model for Biomolecular Interaction Design. *arXiv* **2025**.
221. Yang, Z.; Song, X.; et al. Xray2Xray: World Model from Chest X-rays with Volumetric Context. *arXiv* **2025**.
222. Yue, Y.; Wang, Y.; Jiang, H.; et al. EchoWorld: Learning Motion-Aware World Models for Echocardiography Probe Guidance. In Proceedings of the CVPR, 2025, pp. 25993–26003.
223. Koju, S.; Bastola, S.; et al. Surgical vision world model. In Proceedings of the MICCAI Workshop, 2025.
224. Wu, H.; Gao, Y.; et al. Spatiotemporal Forecasting as Planning: A Model-Based Reinforcement Learning Approach with Generative World Models. *arXiv* **2025**.
225. Park, K.; et al. PINT: Physics-Informed Neural Time Series Models with Applications to Long-term Inference on WeatherBench 2m-Temperature Data. In Proceedings of the ICLR Workshop, 2025.
226. Luo, X.; et al. Reconstructing Dynamics from Steady Spatial Patterns with Partial Observations. In Proceedings of the ICLR Workshop, 2025.
227. Team, H.; et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv* **2025**.
228. Valevski, D.; Leviathan, Y.; et al. Diffusion Models Are Real-Time Game Engines. In Proceedings of the ICLR, 2025.
229. Decart, E.; McIntyre, Q.; et al. Oasis: A universe in a transformer. *Technical Report* **2024**.
230. Zhang, Y.; Peng, C.; et al. Matrix-Game: Interactive World Foundation Model. *arXiv* **2025**.
231. He, X.; Peng, C.; et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv* **2025**.
232. Sun, W.; Wei, F.; et al. From Virtual Games to Real-World Play. *arXiv* **2025**.
233. Yang, Z.; et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv* **2025**.
234. Gu, Y.; Zhang, K.; Ning, Y.; et al. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. *Transactions on Machine Learning Research* **2025**.
235. Gao, Y.; Ye, J.; Wang, J.; Sang, J. Websynthesis: World-model-guided mcts for efficient webui-trajectory synthesis. *arXiv* **2025**.
236. Fang, T.; Zhang, H.; et al. WebEvolver: Enhancing Web Agent Self-Improvement with Coevolving World Model. *arXiv* **2025**.
237. Deng, M.; Hou, J.; et al. SimuRA: Towards General Goal-Oriented Agent via Simulative Reasoning Architecture with LLM-Based World Model. *arXiv* **2025**.
238. Yu, X.; Peng, B.; et al. Dyna-Think: Synergizing Reasoning, Acting, and World Model Simulation in AI Agents. *arXiv* **2025**.
239. Rivard, L.; Sun, S.; et al. Neuralos: Towards simulating operating systems via neural generative models. *arXiv* **2025**.
240. Luo, D.; et al. ViMo: A Generative Visual GUI World Model for App Agents. *arXiv* **2025**.
241. Yin, X.; Luo, X.; et al. Unlocking Smarter Device Control: Foresighted Planning with a World Model-Driven Code Execution Approach. *arXiv* **2025**.
242. Mei, K.; et al. R-WoM: Retrieval-augmented World Model For Computer-use Agents. *arXiv* **2025**.
243. Richens, J.; et al. General agents need world models. In Proceedings of the ICML, 2025.
244. Spies, A.F.; et al. Transformers Use Causal World Models in Maze-Solving Tasks. In Proceedings of the ICLR Workshop, 2025.
245. Rohekar, R.Y.; et al. A Causal World Model Underlying Next Token Prediction: Exploring GPT in a Controlled Environment. *arXiv* **2024**.
246. Zhang, T.; et al. When Do Neural Networks Learn World Models? *arXiv* **2025**.
247. Tehenan, M.; et al. Linear Spatial World Models Emerge in Large Language Models. *arXiv* **2025**.
248. Yuan, Y.; et al. Revisiting the Othello World Model Hypothesis. *arXiv* **2025**.
249. Zhao, W.; et al. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. *IEEE Symposium Series on Computational Intelligence* **2020**.
250. OpenAI.; et al. Solving Rubik's Cube with a Robot Hand. *arXiv* **2019**.
251. Tobin, J.; et al. Domain randomization for transferring deep neural networks from simulation to the real world. *IROS* **2017**.
252. Tao, Z.; et al. A Survey on Self-Evolution of Large Language Models. *arXiv* **2024**.
253. Wu, T.; Yuan, W.; et al. Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. In Proceedings of the EMNLP, 2025.
254. Madaan, A.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In Proceedings of the NeurIPS, 2023.

 doi:10.20944/preprints202603.0739.v1

255. Weng, Y.; et al. Large Language Models are Better Reasoners with Self-Verification. In Proceedings of the EMNLP, 2023.
256. Fu, S.; et al. Self-Verification Provably Prevents Model Collapse in Recursive Synthetic Training. In Proceedings of the NeurIPS, 2025.
257. Yuan, W.; et al. Self-Rewarding Language Models. In Proceedings of the ICML, 2024.
258. Chen, Z.; et al. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. In Proceedings of the ICML, 2024.
259. Wang, Y.; et al. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence* **2023**.
260. Wang, Y.; Kordi, Y.; et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the ACL, 2022.
261. Wu, Y.; et al. Self-Play Preference Optimization for Language Model Alignment. In Proceedings of the ICLR, 2025.
262. Fu, S.; Wang, Y.; et al. A Theoretical Perspective: How to Prevent Model Collapse in Self-consuming Training Loops. In Proceedings of the ICLR, 2025.
263. DeepSeek-AI.; et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **2025**.
264. Pearce, T.; et al. Scaling Laws for Pre-training Agents and World Models. *arXiv* **2024**.
265. Radji, W.; et al. How Hard is it to Confuse a World Model? *arXiv* **2025**.
266. Bain, M.; Nagrani, A.; Varol, G.; et al. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In Proceedings of the ICCV, 2021.
267. Chen, T.S.; et al. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *CVPR* **2024**.
268. Grauman, K.; Westbury, A.; et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In Proceedings of the CVPR, 2022.
269. Miech, A.; et al. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *ICCV* **2019**.
270. Duan, H.; Yu, H.X.; Chen, S.; et al. WorldScore: A Unified Evaluation Benchmark for World Generation. *arXiv* **2025**.
271. Padalkar, A.; Pooley, A.; Jain, A.; et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0. *ICRA* **2024**.
272. Ku, A.; Anderson, P.; et al. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In Proceedings of the EMNLP, 2020.
273. Yue, H.; Huang, S.; et al. EWMBench: Evaluating Scene, Motion, and Semantic Quality in Embodied World Models. *arXiv* **2025**.
274. Yu, T.; et al. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. *arXiv* **2019**.
275. Caesar, H.; Bankiti, V.; Lang, A.H.; et al. nuScenes: A Multimodal Dataset for Autonomous Driving. *CVPR* **2020**.
276. Arai, H.; Ishihara, K.; et al. ACT-Bench: Towards Action Controllable World Models for Autonomous Driving. *arXiv* **2024**.
277. Chandrasekaran, S.N.; Ackerman, J.; et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv* **2023**.
278. Morshid, A.; Elsayes, K.M.; et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence* **2019**.
279. Bellemare, M.G.; et al. The Arcade Learning Environment: An Evaluation Platform for General Agents. *ArXiv* **2012**.
280. Guss, W.H.; Houghton, B.; et al. MineRL: A Large-Scale Dataset of Minecraft Demonstrations. In Proceedings of the IJCAI, 2019.
281. Xie, T.; Zhang, D.; et al. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In Proceedings of the NeurIPS, 2024.
282. Bonatti, R.; Zhao, D.; et al. Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale. In Proceedings of the ICML, 2025.
283. Hüyük, A.; et al. Reasoning Elicitation in Language Models via Counterfactual Feedback. In Proceedings of the ICLR, 2025.
284. Xiang, X.; Chen, Y.; et al. Macro-from-Micro Planning for High-Quality and Parallelized Autoregressive Long Video Generation. *arXiv* **2025**.

 doi:10.20944/preprints202603.0739.v1

285. Li, J.; Feng, W.; et al. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. In Proceedings of the NeurIPS, 2024.
286. Xing, J.; Xia, M.; et al. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. *ECCV* **2024**.