From Seeing to Knowing the World: A Survey of Vision World Models

Xiao Yu, Yichen Zhang, Mingzhang Wang, Shifang Zhao, Weizhe Liu, Yuyang Yin, Zhongwei Ren, Ning An, Xinglong Wu, Hao Liu, Houwen Peng, Yao Zhao, *IEEE Fellow,* Jianchao Yang, Jiashi Feng, Shuicheng Yan, *IEEE Fellow,* Yunchao Wei , Xiaojie Jin

Abstract—Acquiring world knowledge directly from visual observation is fundamental to Artificial General Intelligence (AGI). To support this capability, the Vision World Model (VWM) has emerged as a key paradigm, which learns how the world evolves over time from visual streams. However, recent progress has been driven by diverse research communities, resulting in inconsistent problem formulations, disconnected taxonomies, and divergent evaluation protocols. We argue that addressing this gap requires a conceptual shift: vision should not be treated merely as an input modality, but as the primary driver shaping how world models are represented, learned, and evaluated. Guided by this vision-centric perspective, we introduce a unified framework that organizes VWM research into three core components: vision encoding, knowledge learning, and controllable simulation, and use it to analyze existing model designs and evaluation methodologies. Finally, we outline future research directions that emphasize stronger physical and causal grounding, more meaningful evaluation beyond visual appearance, and scaling toward more general and reliable world modeling capabilities. The project page is at AIWorldLab.github.io/survey.

Index Terms—Survey, World Model, Vision World Model

1 Introduction

Consider a person instinctively reaching out to catch a falling glass before it shatters, or a driver braking in anticipation of a child chasing a ball onto the street. These everyday scenarios reveal a fundamental aspect of human intelligence: reliable decision-making depends not only on recognizing what is currently happening, but also on forming a robust understanding of how the world changes over time and using it to anticipate what may happen next. By imagining how situations are likely to develop, humans can act in advance and adjust their behavior before events actually occur. Such anticipatory decision-making is widely regarded as a core ingredient of intelligence and a central objective in the pursuit of Artificial General Intelligence (AGI) [1, 2, 3].

To equip artificial systems with such capability, the research community has gradually shifted toward the paradigm of the *World Model* [4, 5, 6, 7]. Instead of treating the world as a black-box system, this paradigm posits that an agent should learn fundamental knowledge about how states of the world evolve over time, e.g., how physical processes unfold and how actions influence future outcomes. By constructing an internal model of these world dynamics, an agent can anticipate possible future states before acting, enabling more informed and safer decisions without excessive reliance on costly real-world trial and error [8, 9, 10, 11].

- *X. Yu, Y. Zhang, M. Wang, S. Zhao, W. Liu, Y. Yin, Z. Ren, Y. Zhao, Y. Wei, and X. Jin are with Beijing Jiaotong University, Beijing, China. X. Yu and X. Jin contributed equally to the paper. (xiaoyu@bjtu.edu.cn)*
- *X. Wu is with Bytedance, Beijing, China. H. Peng and J. Feng are with Tencent, Shenzhen, China. N. An is with the Core Research Institutes for Coal, Beijing, China. H. Liu is with Midea Research Lab, Shanghai, China. S. Yan is with the School of Computing, National University of Singapore, Singapore.*
- *Corresponding author: **Xiaojie Jin** (xjjin0731@gmail.com) and **Yunchao Wei** (wychao1987@gmail.com)*
- *Project lead: **Xiaojie Jin***

The effectiveness of a world model, however, largely depends on how the world itself is revealed and perceived, particularly how physical phenomena such as motion, interaction and causal consequences of actions/events unfold over time. Recent works [12, 13, 14, 15, 16] have explored world modeling through text-based reasoning, often leveraging Large Language Models (LLMs) to infer outcomes from linguistic descriptions. While such approaches are effective at capturing high-level concepts, they cannot access the detailed processes that describe real-world change, e.g., language can state that "a glass shatters," but it cannot specify how fragments scatter or how material interactions unfold after impact. This limitation arises because the physical world does not present itself as symbols or predefined rules, but as continuous, high-dimensional sensory signals. In contrast, visual streams directly encode these signals, jointly capturing appearance, dynamics and underlying causal relationships. As a result, visual observation provides the most direct and comprehensive evidence of how the physical world behaves.

Building on this insight, learning world knowledge[1] directly from visual signals becomes a critical direction. Visual learning mirrors how humans develop an intuitive grasp of the world: through visual observation, humans gradually form a coherent understanding of the principles that govern how the world changes. Supported by the availability of internet-scale video data [17, 18], the field has increasingly converged on a *unified* modeling paradigm, which we refer to as the **Vision World Model (VWM)**, a paradigm that learns world knowledge directly from visual signals and enables interaction-conditioned future simulation.

1. In this work, we use the term "knowledge" to refer to what governs how world states evolve over time and respond to interventions, broadly encompassing physical principles, causal mechanisms, structural constraints, and other forms of dependency that determine the behavior of the world.

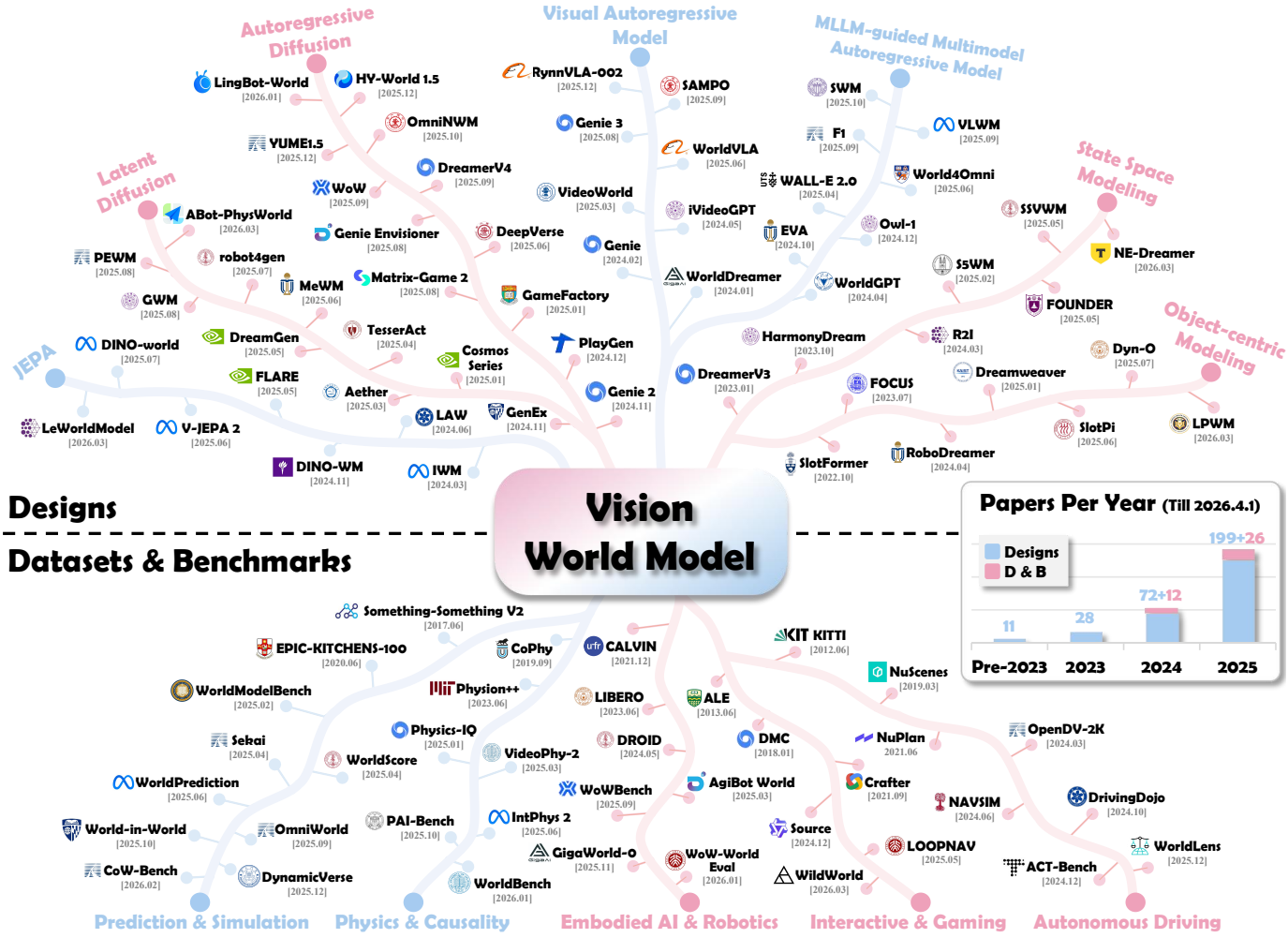

Fig. 1: Landscape of Vision World Model (VWM) research. Design architectures are grouped into four major families in the upper panel, and datasets and benchmarks are organized by application domains in the lower panel.

In recent years, research on VWMs has advanced rapidly, spanning areas such as generative modeling, representation learning, and embodied intelligence. Although these efforts originate from different research communities, their goals largely overlap in the shared pursuit of understanding and simulating the world. As illustrated in Fig. 1, existing approaches can be organized according to how they represent visual signals and predict future states. Autoregressive and diffusion-based methods primarily focus on synthesizing visually coherent future content [19, 20, 21, 22, 23, 24, 25, 26, 27]; embedding prediction approaches emphasize learning predictive structure in representation space [28, 29, 30, 31, 32]; and state transition methods abstract visual input into compact latent states for long-horizon modeling [11, 33, 34, 35, 36, 37]. Despite capturing complementary aspects of world modeling, these paradigms are often developed in isolation. This separation leads to inconsistent terminology, disconnected evaluation practices, and limited cross-paradigm comparison, making it difficult to assess progress in a unified manner.

Several surveys have attempted to summarize existing works. They fall into two broad categories. Application-focused surveys analyze world models within specific domains, such as robotics or autonomous driving, offering domain-specific insights but treating VWMs as supporting components instead of the core framework [38, 39, 40, 41, 42, 43, 44, 45]. Other surveys provide high-level overviews of world modeling, but often treat vision merely as an input modality, without examining how the unique properties of vision fundamentally shape representation, learning objectives, and evaluation [2, 46, 47, 48, 49]. As a result, the field still lacks a clear, vision-centric roadmap that explains how diverse modeling choices relate to one another and how progress should be assessed.

To address this gap, we advocate a vision-centric perspective that treats vision not merely as an input modality, but as a key factor of world modeling. Under this perspective, we present a unified framework for understanding and organizing VWMs.

Our primary contributions are summarized as follows:

- We propose a unified framework for VWMs that structures world modeling around three core components: *vision encoding*, *knowledge learning*, and *controllable simulation*. This framework provides a principled view for analyzing how VWMs represent, learn, and simulate the world.
- Based on this framework, we organize existing VWM approaches into four major architectural families: sequential generation, diffusion-based generation, embedding prediction, and state transition. Our analysis clarifies their relationships and highlights key design trade-offs.
- We systematically review evaluation protocols of VWMs, summarizing how different metrics and benchmarks assess visual

quality, physical plausibility, and task performance. We further examine how these evaluation criteria relate to different views of a reliable world model.
- We identify key open challenges and future research directions, focusing on strengthening physical and causal grounding, advancing evaluation protocols beyond appearance-level assessment, and scaling models toward more general and reliable world modeling capabilities.

Scope. This survey focuses on world models that learn how the world changes over time directly from visual signals. We review methods that learn world knowledge from visual data and use it to predict future outcomes of the environment. Methods that operate purely on symbolic states without grounding in vision are outside the scope of this survey [50, 51, 52, 53, 54]. In addition, while applications such as robotics, autonomous driving, and interactive environments are referenced to motivate design choices and evaluation criteria, our analysis remains focused on VWMs themselves rather than downstream control or policy optimization tailored for each application scenario.

Roadmap. Section 2 introduces the definition of VWM and presents the unified framework. Section 3 organizes existing VWM architectures according to this framework and analyzes their design trade-offs. Section 4 reviews evaluation metrics, datasets, and benchmarks. Finally, Section 5 discusses open challenges and future directions before concluding in Section 6.

2 A Unified Framework for Vision World Models

To provide a unified view of Vision World Models (VWMs), we introduce a framework that summarizes common design principles across existing approaches. As illustrated in Fig. 2, a VWM can be decomposed into three core components: *Vision Encoding* (Section 2.2) describes how raw visual data are transformed into representations suitable for modeling world change. *Knowledge Learning* (Section 2.3) focuses on how models learn world knowledge from visual data. *Controllable Simulation* (Section 2.4) describes how a VWM generates possible future world states conditioned on actions or instructions.

2.1 Definition

At its core, we define a Vision World Model (VWM) as follows:

A Vision World Model (VWM) is an AI model that learns world knowledge from visual data and generates future world states conditioned on interaction.

Specifically, a VWM learns how the world changes over time by capturing the underlying principles and mechanisms from visual data, and uses this knowledge to perform future simulation conditioned on interaction signals. A formal description of a VWM is given by a probabilistic model f_θ that parameterizes the distribution of future world states given observed visual context and interaction conditions:

$$p(\mathcal{S}_{t+1:T} \mid v_{0:t}, c_t) = f_\theta\big(\mathcal{E}(v_{0:t}), c_t\big), \tag{1}$$

where $v_{0:t}$ denotes visual data from time 0 to t, and c_t denotes interaction conditions such as agent actions, language instructions, or other control commands. $\mathcal{E}(\cdot)$ denotes a visual encoder that transforms raw visual data into visual representations. $\mathcal{S}_{t+1:T}$ denotes future world states, which may take different forms depending on the modeling paradigm, including future frames, latent states, or other meaningful attributes (e.g., depth, flow, occupancy, 3D primitives, or trajectories).

Compared to traditional world models [3, 41, 55], which often employ low-dimensional or predefined state spaces and treat vision only as an input modality, VWMs instead rely on high-dimensional visual data as the foundation for modeling world change, fundamentally reshaping how world knowledge is represented and learned.

2.2 Vision Encoding: Learning Representations from Visual Data

Raw visual data provide rich information about the world, but they also entangle multiple factors of variation. While they contain essential cues about objects, motion, and interactions, these cues are intertwined with incidental variations such as camera jitter, background clutter, and sensor noise. Such entanglement makes it difficult to isolate the underlying principles that govern world change and can hinder the learning of reliable world knowledge. The role of vision encoding in VWMs is therefore to transform raw visual data into representations that disentangle relevant factors for modeling world change while suppressing irrelevant variation.

This transformation determines what aspects of visual information are retained, the level of abstraction at which they are encoded, and how they support modeling world knowledge. In this section, we analyze vision encoding from two perspectives: the types of visual inputs commonly used for world modeling (Section 2.2.1), and the forms of representations into which these inputs are encoded (Section 2.2.2).

2.2.1 *Visual Inputs for World Modeling*

VWMs can be trained on a wide range of visual inputs, from standard RGB images/videos to specialized modalities such as depth maps, optical flow, and bird's-eye-view (BEV) data. Standard RGB images/videos [19, 24, 26, 28, 56] are widely used due to their availability and broad coverage of real-world environments. Unlike more structured data, RGB inputs do not explicitly encode geometry, motion, or scene layout. Therefore, models must infer such information directly from pixels, making RGB-based modeling highly general yet challenging.

To ease this difficulty, many approaches incorporate additional visual modalities that provide more explicit environment attributes. Depth maps, point clouds, and multi-view geometry [37, 57, 58, 59, 60, 61] expose 3D geometric information, supporting representations that preserve spatio-temporal consistency and coherence. Similarly, optical flow [62, 63] encodes motion information explicitly, facilitating the separation of dynamic entities from static background. In domains such as autonomous driving, BEV data [64, 65, 66, 67, 68, 69] organize spatial information within a unified coordinate frame, enabling more convenient spatial reasoning and long-horizon prediction. Multi-view static camera setups and egocentric recordings [70, 71, 72, 73, 74, 75] further provide complementary perspectives that reinforce 3D consistency and align visual input with agent actions.

2.2.2 *Forms of Visual Representation*

Visual representations differ in the aspect of visual information they retain, thereby determining what world knowledge can be learned in VWMs. We group existing approaches according to how they organize visual information.

Continuous Latent Representations. Many world models encode observations into continuous latent states using convolutional

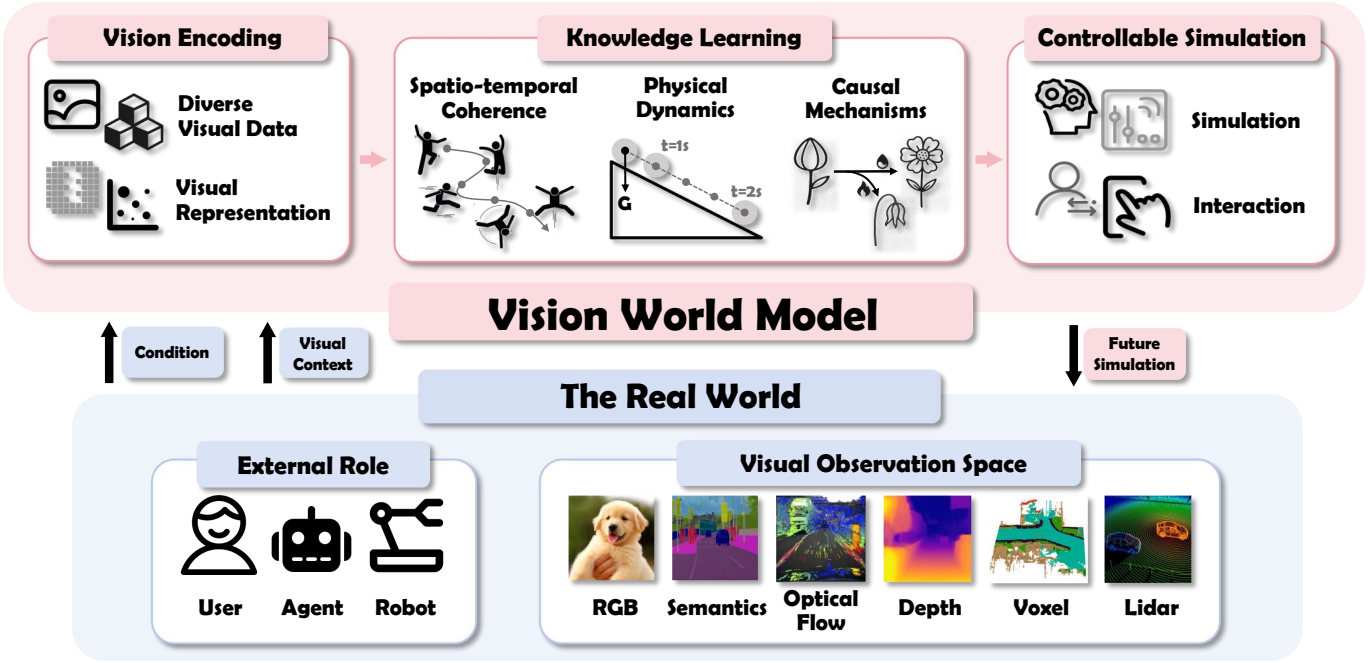

Fig. 2: A unified framework for Vision World Models (VWMs). A VWM encodes visual observations, learns structured world knowledge about how the world changes, and performs future simulation conditioned on interaction.

networks or vision transformers [10, 11, 76, 77]. These representations form compact, continuous state spaces that evolve smoothly over time, and are well-suited for modeling motion, interaction, and long-horizon dynamics.

Discrete Tokenized Representations. Another line of work discretizes visual inputs into a finite vocabulary of tokens [24, 26, 78, 79]. By mapping visual content into a fixed-size dictionary, these approaches enable more computationally efficient modeling compared to continuous representations. Moreover, discrete tokenization bridges between visual modeling and sequence-based generative frameworks. VWMs can therefore adopt autoregressive training objectives and benefit from the appealing scalability properties of Transformer architectures developed for LLMs.

Object- and Entity-centric Representations. Some approaches explicitly build representations around objects or entities with persistent identity. By making entity boundaries explicit, such representations facilitate learning spatio-temporal coherence and causal interactions, and support compositional generalization [34, 36].

Hybrid and Hierarchical Representations. Recent models combine multiple representation forms, such as continuous representations over discrete tokens or object-centric states embedded in latent spaces [80, 81, 82]. These hybrid representations demonstrate better trade-offs between scalability, physical consistency and expressive capacity.

2.3 Knowledge Learning: Structured World Knowledge in VWMs

Built upon the representations introduced in Section 2.2, a VWM must learn the world knowledge that governs how the world changes, including how entities remain consistent over time, how their states transform, and how actions or events produce consequences, etc. We group this knowledge into three complementary aspects: *spatio-temporal coherence*, *physical dynamics*, and *causal mechanisms*.

2.3.1 *Spatio-temporal Coherence*

Spatio-temporal coherence refers to the consistent existence and identity of entities across space and time, forming the foundational basis upon which other aspects of world modeling are built. Although coherence ultimately relates to physical constraints, we treat it separately because it reflects the structural consistency directly observable in visual data. By maintaining stable entities and their continuity, spatio-temporal coherence provides the "scaffold" for modeling more explicit dynamics and causal relations, which can be analyzed along both spatial and temporal dimensions.

At the spatial level, coherence requires models to maintain consistent object identity across varying viewpoints. This includes multi-view consistency, where the same object is recognized under different camera perspectives [83, 84, 85, 86, 87], as well as geometric stability, where an object's shape and structural integrity remain coherent rather than arbitrarily collapsing or deforming [37, 59, 88].

At the temporal level, coherence requires that entities persist over time with stable identity and consistent state transitions. This includes object permanence, where entities remain present despite temporary occlusion, and smooth state progression, where changes follow plausible trajectories rather than abrupt discontinuities [33, 89]. Together, these properties ensure entity identity and consistency across long temporal horizons.

2.3.2 *Physical Dynamics*

Given coherent entities, physical dynamics describe how their states change under physical constraints. This form of world knowledge ensures that the predicted motion and interaction remain physically plausible rather than merely visually convincing.

Physical dynamics require models to adhere to fundamental physical constraints, such as gravity, contact, and material resistance. By capturing these constraints, a VWM can avoid common physically inconsistent artifacts observed in other video generative

models, such as objects unrealistically intersecting solid surfaces due to a lack of physical understanding.

Physical dynamics can be understood at multiple levels of complexity. At a basic level, models capture macroscopic motion governed by classical mechanics, such as rigid-body movement and object interaction [90, 91]. At more advanced levels, models draw on principles from continuum mechanics to address scenarios involving deformable materials and fluid behavior, where material properties determine how substances undergo deformation and flow [92]. Across these settings, conservation principles such as energy and momentum provide unifying constraints that maintain physical consistency, even when underlying micro-scale forces are not explicitly modeled.

2.3.3 Causal Mechanisms

Causal mechanisms describe how actions and events produce outcomes. Rather than modeling state transitions in isolation, they organize world knowledge around fundamental action–outcome relations that generalize across different scenarios.

Models that rely primarily on statistical correlation, including many video generation approaches [93, 94, 95], tend to predict what typically follows from observed patterns. Such models may fail when contexts shift or rare events occur. For example, a model trained only on normal driving data may struggle to anticipate the consequences of a collision, since the underlying relation between impact and deformation is not captured from nominal driving observations alone.

In contrast, a VWM requires learning the fundamental causal relations between actions and outcomes, enabling reliable prediction across a broader range of conditions [96, 97]. For instance, understanding that high-speed impact leads to structural deformation allows the model to anticipate crash consequences even in unseen environments. Therefore, grounding modeling in causal relations can provide a stable foundation for reliable prediction across diverse and unseen conditions.

A key capability enabled by learning causal mechanisms is counterfactual reasoning: considering how outcomes would differ under alternative actions or interventions. This capability allows a VWM to evaluate hypothetical scenarios and make more reliable decisions in previously unseen settings [98, 99].

Causal mechanisms also extend beyond purely physical processes. In human-centered environments, world behavior is also shaped by social norms, conventions, and shared intentions [100]. For example, traffic behavior may be influenced by temporary human instructions or contextual rules. Modeling such human-governed relations is essential for reliable interaction in complex real-world settings.

2.4 Controllable Simulation: Simulating the Future through Interaction

A Vision World Model uses the world knowledge it has learned to simulate how the world may evolve under different interaction conditions. Given the current visual context, simulation generates possible future world states, while interaction (e.g., actions or instructions) specifies how the world is expected to respond. Together, simulation and interaction define how a VWM produces future rollouts conditioned on external inputs. Accordingly, we analyze this capability from two perspectives: simulation (Section 2.4.1), which describes the forms of future states being generated, and interaction (Section 2.4.2), which describes how external conditions guide these predictions.

2.4.1 Simulation

Simulation refers to the generation of future world states over time based on the current visual context and learned world knowledge. Instead of replaying observed data, simulation constructs plausible future trajectories that reflect how the world may change. The form of simulated future states (denoted as $\mathcal{S}_{t+1:T}$ in Eq. 1) varies across modeling paradigms and can be broadly categorized into three types:

• **Latent States.** Simulation can be carried out in a compressed latent space, where future trajectories are generated over compact continuous representations rather than pixels. This form of simulation is computationally efficient and well-suited for planning and reasoning tasks that prioritize decision-making over visual details [11, 28, 35, 101, 102].

• **Visual States.** Simulation can also be performed directly in the visual space by decoding predicted representations into images, videos [103, 104, 105], or explicit geometric forms [65, 88, 106]. While more computationally demanding, visual simulation is important for human interpretability, synthetic data generation, and closed-loop evaluation.

• **Structured Outputs.** Beyond visual outputs, simulation can produce structured future representations, such as object attributes, spatial configurations, or action trajectories. These representations provide more direct inputs for control or planning, as they summarize task-relevant aspects of the environment without requiring additional processing [36, 107, 108, 109, 110, 111].

2.4.2 Interaction

Interaction specifies how simulated futures respond to external conditions. A VWM therefore captures how different actions or prompts lead to distinct outcomes given the same visual context.

• **Action Signals.** A common form of interaction is conditioning simulation on control signals, such as robot motor commands [112, 113] or keyboard and mouse inputs [23, 24, 114]. This enables a VWM to anticipate the consequences of specific control choices in both continuous [10, 115] and discrete [116] action spaces. More recent work explores latent action representations learned directly from video, reducing reliance on explicitly labeled action data [22, 63, 80, 117, 118].

• **Multimodal Interaction.** To handle complex goals, such as long-horizon tasks and multi-step planning, models can use multimodal prompts to condition simulation. Early approaches used language to enrich state descriptions [71, 119], while more recent Vision-Language-Action frameworks [120, 121, 122] provide unified interfaces for specifying task-level objectives. By integrating language guidance, simulation becomes more interpretable and accessible, enabling more natural and human-friendly interaction experiences [123, 124, 125].

3 Designs of Vision World Modeling

This section organizes VWM designs into four architectural families, covering seven representative sub-designs. We group methods by architectural form, which largely shapes how world knowledge is represented, learned, and simulated, as defined in Section 2. Fig. 3 provides a concise overview of these designs. For each sub-design, the upper panel illustrates its typical input–output pipeline, while the lower panel highlights how it instantiates the three core components of our framework, namely Vision Encoding, Knowledge Learning, and Controllable Simulation, thereby clarifying the distinctive design choices of each architecture and

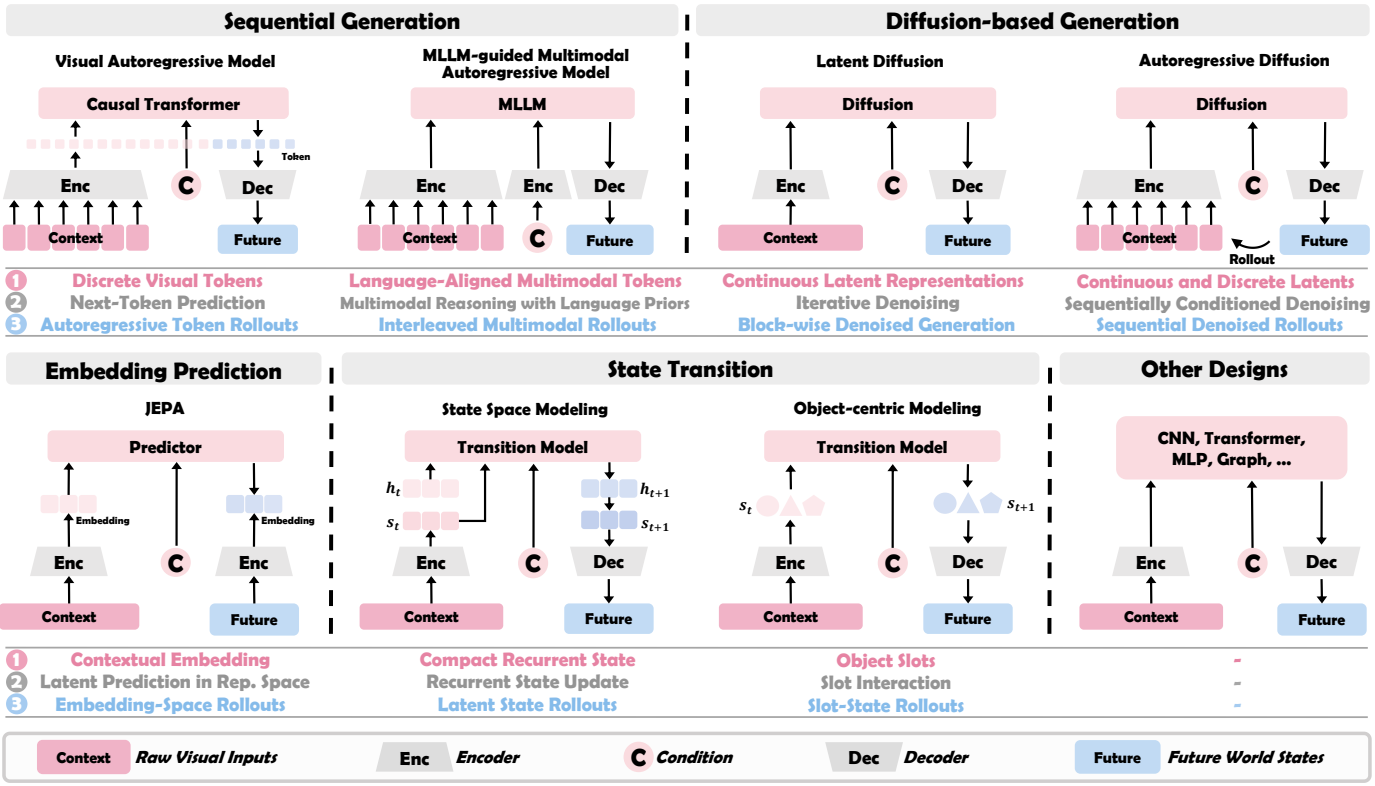

Fig. 3: A taxonomy of VWM designs is organized into four architectural families with seven sub-designs. For each sub-design, the upper panel illustrates its typical input–output pipeline (from visual context and conditions to future outputs). The lower panel highlights its key design choices under the three components of our framework (Section 2): ① *Vision Encoding* (what form observations are encoded into), ② *Knowledge Learning* (through what mechanism structured world knowledge is learned), and ③ *Controllable Simulation* (what form the future rollouts take). *Rep.*: Representation.

facilitating comparison. In the detailed discussion that follows, we examine each sub-design by explaining how it realizes the key components of our framework, consistent with the framework structure established in Section 2.

3.1 Sequential Generation

Sequential generation approaches cast visual world modeling as token sequence generation. They typically encode visual context (and optional conditions) into a token stream, and produce future outcomes step by step conditioned on history tokens. Within this family, we focus on two representative sub-designs: (1) *visual autoregressive model* (Section 3.1.1), which predicts future visual tokens via next-token prediction; and (2) *MLLM-guided multimodal autoregressive model* (Section 3.1.2), which represents visual inputs as LLM-compatible tokens and generates interleaved multimodal rollouts for reasoning and planning.

3.1.1 Visual Autoregressive Model

Visual autoregressive methods first convert videos into discrete sequences and then formulate world modeling as sequential token generation. They can benefit from well-developed training recipes for language modeling and scale to large datasets and long contexts [117, 126]. Table 1 summarizes representative works under this design.

• **Discrete Visual Tokens.** Visual signals are usually converted into discrete tokens via VQ-VAE [78, 147, 148] or VQ-GAN [79]. A critical variation of these methods is *how* videos are tokenized. Early methods such as GAIA-1 [126] and WorldDreamer [128] mainly use spatial tokenization, representing a video as a frame-wise token sequence. More recent methods, such as Genie [80] and iVideoGPT [131], adopt spatio-temporal tokenization, encoding multiple frames into shared tokens to better preserve temporal information in the token space.

Discrete tokenization also extends beyond 2D pixels. For example, OccWorld [127], RenderWorld [134], and OccTENS [68] discretize 3D voxel grids into vocabulary indices. Compared to purely appearance-driven tokens, voxel grids make spatial structure (e.g., geometry and free space) more explicit, which facilitates modeling interaction and long-horizon world change.

• **Next-Token Prediction.** Methods in this category acquire world knowledge through next-token prediction over visual token sequences. This objective encourages the model to capture temporal dependencies that maintain spatio-temporal coherence, including: (1) *entity consistency*, where an object remains identifiable across motion or occlusion; and (2) *temporal continuity*, where state changes evolve smoothly rather than abruptly. Maintaining coherence over long horizons is challenging because prediction errors accumulate during autoregressive rollouts. Increasing the effective context length is a common strategy. For example, LWM [129] extends the context window to mitigate identity drift after long occlusions.

Beyond coherence, many scenarios require *action-conditioned* modeling: the same visual context can evolve into different future sequences depending on what the agent does. Several works intro-

TABLE 1: Summary of visual autoregressive modeling VWMs. For each work, we present the visual encoder for extracting visual representation, the world knowledge it learns, with the format of interaction input and simulation output. *: Genie 3 [117] is believed to use a visual autoregressive architecture for world modeling.

Methods	Visual Encoder	World Knowledge	Interaction Input	Simulation Output
Gaia-1 [126]	VQ-VAE	Traffic dynamics, causal decision-making	Action (speed/curvature), text	Image
OccWorld [127]	VQ-VAE	Spatio-temporal scene evolution	vehicle position/trajectory	Occ-grid, displacement
WorldDreamer [128]	VQ-GAN	General world physics	Action (velocity/steering), text	Image
LWM [129]	VQ-GAN	Long-range spatio-temporal dynamics	Text	Text, video
Genie [80]	VQ-VAE	Interactive environments, 2D physics	Latent action	Image
WHAM [130]	VQ-GAN	3D game physics, combat mechanics	Action (controller signals)	Image, action
iVideoGPT [131]	VQ-GAN	Robotic manipulation, causal relationships	Action, text	Image
VidIT [132]	VQ-VAE	Visual semantic patterns, task imitation	Latent action	Image
LatentDriver [133]	PlanT	Driving decision uncertainty, Agent interactions	Action (waypoints)	State
Renderworld [134]	AM-VAE	4D scene evolution, Driving occ dynamics	State, trajectory	Occ-grid, trajectory
GR-2 [135]	VQ-GAN	Manipulation skills, Object interaction	Text, State (end-effector, gripper)	Image, action
WHALE [136]	VQ-VAE	Manipulation skills, spatiotemporal dependencies	Latent action	Image
Drivingworld [137]	VQ-GAN	Spatio-temporal driving dynamics	Vehicle pose (orientation, location)	Image, Vehicle pose
UVA [138]	VAE	Forward and inverse dynamics	Action, text	Image, action
DWS [139]	VAE, VQ-GAN	Motion-centric dynamics	Action	Image
SurgWM [140]	VQ-VAE	Surgical tool dynamics, instrument persistence	Latent action	Image
VideoWorld [26]	MAGVIT-v2	Game strategy, robotic manipulation	Latent action	Image
MineWorld [141]	VQ-VAE	Minecraft physics, block interactions	Action (keyboard/mouse signals)	Image
RLVR-World [142]	VQ-GAN	State transition dynamics	Action, Text	State, image, text
UniVLA [123]	DINOv2, SigLip	Generalizable manipulation skills	Latent action, text	Action, visual tokens
RoboScape [143]	MAGVIT-v2	3D consistency, contact-rich physics	Keypoint Trajectory	Image, depth
WorldVLA [125]	VQ-GAN	Manipulation mechanics, underlying physics	Action	Image, action
Xray2Xray [144]	VQ-GAN	Transition dynamics of X-ray projections	Action	Xray tokens
I^2-World [145]	RQ-VAE-like	4D dynamics, driving physics, scene persistence	Action (Trajectories, speed, angle)	Occ-grid
Genie 3* [117]	*Unknown*	Intuitive physics, 3D consistency	*Unknown*	Image
OccTENS [68]	2D CNN	Spatio-temporal occ dynamics	Ego motion, pose	Occ-grid, trajectory
SAMPO [27]	VQ-GAN	Coarse-to-fine spatial structure	Action, motion prompts	Image
RynnVLA-001 [120]	VQ-GAN	Ego-centric manipulation dynamics	Text, state (joint/keypoint positions)	Image, action
iMoWM [146]	VQ-GAN	3D geometric and spatio-temporal dynamics	Robot end-effector pose	Images, depth, masks

duce latent action learning to model this dependency, where action variables are inferred from videos and then used to condition prediction [26, 80, 132]. This makes the model relate control inputs (e.g., steering or pushing) to their predicted consequences, rather than predicting the next states solely from past observations.

• **Autoregressive Token Rollouts.** Simulation proceeds as an autoregressive rollout in token space, where predicted tokens are decoded into images/videos or structured outputs (e.g., voxel grids). Interaction is implemented by inserting condition tokens (e.g., action tokens) into the context, allowing different control inputs to guide subsequent generation [123, 125]. This design supports diverse control modalities, ranging from continuous control signals in autonomous driving [68, 133, 137] to discrete keyboard/mouse inputs in interactive environments [80, 141]. VidIT [132] further shows that action-relevant behaviors can sometimes emerge even without explicit action labels, when suitable conditioning mechanisms are introduced during video-only training.

• **Strengths and Limitations.** A key advantage of autoregressive token modeling is its scalability and flexibility in producing rollouts of arbitrary length. However, long-horizon generation is vulnerable to error accumulation, and discrete tokenization may limit fine-grained geometric detail required for precise physical interactions. Moreover, physical dynamics and causal relations are primarily induced from large-scale data rather than enforced by explicit constraints, which can reduce robustness under distribution shifts or rare interaction scenarios.

3.1.2 MLLM-guided Multimodal Autoregressive Model

When tasks involve goal descriptions or instructions expressed in language, purely visual autoregressive models become less human-friendly to deploy, as their representations are confined to visual token spaces. MLLM-guided approaches extend sequential generation by projecting visual observations into language-compatible tokens and producing multimodal token streams that support reasoning and decision-related outputs. Table 2 summarizes representative methods.

• **Language-Aligned Multimodal Tokens.** These methods map visual observations into token sequences that can be processed jointly with text by large language models, typically via learned connectors or adapters [149, 150]. The resulting representation is a multimodal token stream that combines visual tokens with text (and sometimes action tokens) within a unified autoregressive context. For example, ADriver-I [151] and Doe-1 [152] construct mixed sequences of visual tokens, textual descriptions, and discrete action tokens to support joint multimodal prediction. In 3D settings, OccLLaMA [64] discretizes semantic voxel grids into a scene-level token vocabulary, providing structured spatial tokens that can be processed within the same autoregressive framework.

• **Multimodal Reasoning with Language Priors.** Instead of relying solely on visual token statistics, these approaches leverage the knowledge of pretrained language models as semantic priors. This is especially useful when visual evidence is incomplete or ambiguous, since language knowledge about object properties and likely effects can guide predictions toward plausible outcomes [153]. For spatio-temporal coherence, linguistic object concepts can help maintain identity across occlusion and long-horizon evolution [154, 155]. For causal relations, the multimodal backbone can generate explanations linking actions to their predicted consequences; GR00T [156] and ADriver-I [151] illustrate how predicted futures can be accompanied by textual reasoning about why certain outcomes occur. Some works further introduce additional constraints through neuro-symbolic components [107] or semantic planners [111] to improve long-horizon consistency.

• **Interleaved Multimodal Rollouts.** The generated future often

TABLE 2: Summary of MLLM-based VWMs works. **".V"** denotes the visual part.

Methods	Visual Encoder	MLLM Backbone
ADriver-I [151]	CLIP	Vicuna [157]
3D-VLA [149]	CLIP, DINO	BLIP-2 FlanT5 [158]
WorldGPT [159]	LanguageBind	Vicuna [157]
OccLLaMA [64]	Occ Tokenizer	LLaMA 3.1 [160]
PIVOT-R [161]	CLIP	LLaVA [162]
Eva [155]	CLIP, VAE	Chat-UniVi [163]
Doe-1 [152]	Lumina-mGPT**.V**	Lumina-mGPT [164]
Owl-1 [154]	VQVAE	Chameleon [165]
HERMES [150]	CLIP, BEVFormer	InternVL 2 [166]
GR00T N1 [156]	SigLIP-2	Eagle-2 [167]
WALL-E 2.0 [107]	Textual JSON Encoder	GPT-4 / GPT-3.5 [168]
World4Omni [169]	GPT-4o**.V**	GPT-4o [170]
DreamVLA [122]	MAE-ViT	GPT-2 Medium [171]
VLWM [111]	Perception Encoder	PerceptionLM [172]
F1 [121]	SigLIP, Residual VQ	PaliGemma [173]
OccVLA [69]	SigLIP	PaliGemma 2 [174]
SWM [153]	SigLIP	PaliGemma [173]
UniWM [175]	Anole**.V**	Anole [176]

takes the form of an interleaved multimodal sequence containing predicted visual outputs, textual explanations and actions. Interaction is expressed through language instructions or high-level goals that condition the rollout and determine which future outcome is produced. In practice, models may generate visual futures for grounding (e.g., frames [155] or voxel grids [69]) while simultaneously producing plans or reasoning steps [121, 151], improving interpretability and supporting task-oriented simulation.

• **Strengths and Limitations.** MLLM-guided autoregressive modeling benefits from the broad knowledge learned by pretrained language models, flexible language-based interaction, and improved interpretability through textual outputs. However, projecting visual content into language-compatible tokens may discard fine-grained information, and language-level knowledge can bias predictions when visual evidence deviates from common patterns. Moreover, long-horizon rollouts remain sensitive to accumulated prediction errors, similar to visual autoregressive models.

3.2 Diffusion-based Generation

Diffusion-based approaches model future world states through iterative denoising in a continuous latent space. Instead of predicting tokens sequentially, these methods generate future visual content by progressively refining noisy representations. Within this family, we distinguish two common designs: (1) *Latent Diffusion* (Section 3.2.1), which performs denoising in a compressed latent space and typically generates future frames jointly; and (2) *Autoregressive Diffusion* (Section 3.2.2), which incorporates temporal dependency into the denoising process to extend generation over longer horizons.

3.2.1 Latent Diffusion

Latent diffusion applies diffusion modeling in a compressed continuous latent space, enabling high-quality visual generation while reducing pixel-level computation. The red part of Fig. 4 summarizes representative works under this design.

• **Continuous Latent Representations.** These methods first encode visual observations into continuous latent features, typically via a variational autoencoder (VAE) [177]. Unlike discrete tokenization, continuous latents avoid hard quantization and preserve fine-grained visual information. Early methods such as DrivingDiffusion [86] and Panacea [178] compress multi-view inputs into 2D spatio-temporal latents. More recent works increasingly adopt 3D-aware representations. For example, GAIA-2 [83] constructs latent tokens that are not only semantically meaningful but also spatially grounded, preserving spatial alignment and geometric detail in the token space. In contrast, GWM [106] and WoVoGen [179] represent scenes using explicit 3D representations, such as Gaussian splats or voxel grids. Compared to purely 2D latents, these 3D-aware representations make scene layout and object geometry more explicit, which helps the model maintain occlusion and contact relationships and better reflect how actions or control inputs change the 3D scene during generation.

• **Iterative Denoising.** Knowledge learning in this paradigm is driven by the denoising objective, where the model learns to recover future latent states from progressively corrupted inputs. Because denoising is performed over an entire spatio-temporal block, diffusion models are naturally encouraged to model dependencies across both space and time within the generated window, which often improves short-term coherence. Physical and action-conditioned dynamics are commonly injected through conditioning signals that specify scene structure or control intent. For example, DriveDreamer [207] and DOME [67] condition generation on structured spatial cues such as 3D box layouts or voxel grids, which guide the model toward spatially plausible scene configurations. Beyond injecting physical structure, GeoDrive [191] shows that manipulating geometric control conditions can produce counterfactual futures (e.g., the ego car swerving under an alternative control), linking control inputs to different future world outcomes.

• **Block-wise Denoised Generation.** Latent diffusion performs simulation over a temporal window as a whole, rather than frame by frame. A spatio-temporal latent block is jointly denoised and decoded into a block of future frames (or a 4D representation) in a single sampling process. Compared to step-wise autoregressive rollout, this block-wise generation strategy models space–time dependencies within the window more coherently, which often improves intra-clip consistency.

Interaction is supported through conditional diffusion. Control signals such as actions, text instructions, or scene layouts are provided as conditioning inputs to steer the denoising process toward different plausible futures. This conditioning interface has been widely used in autonomous driving to support closed-loop evaluation or scenario generation, where agents can test decision-making under diverse imagined situations [72, 83, 181, 208, 210]. In robotics, diffusion-based designs have also been extended to action-conditioned manipulation, for example by jointly generating video futures and action sequences, enabling task-level rollouts under control [196, 209].

• **Strengths and Limitations.** A main strength of latent diffusion is high visual quality, benefiting from continuous latents and iterative refinement. Compared to discrete token generation, continuous tokens can preserve finer visual detail, while denoising over an entire temporal window can reduce short-term drift within the generated clip [19]. At the same time, diffusion inference is computationally expensive due to its iterative sampling process, which can limit real-time interaction. Another limitation is temporal scalability: generation is typically tied to a fixed window length, and extending to very long or open-ended rollouts remains challenging without additional mechanisms (e.g., autoregressive extension or memory).

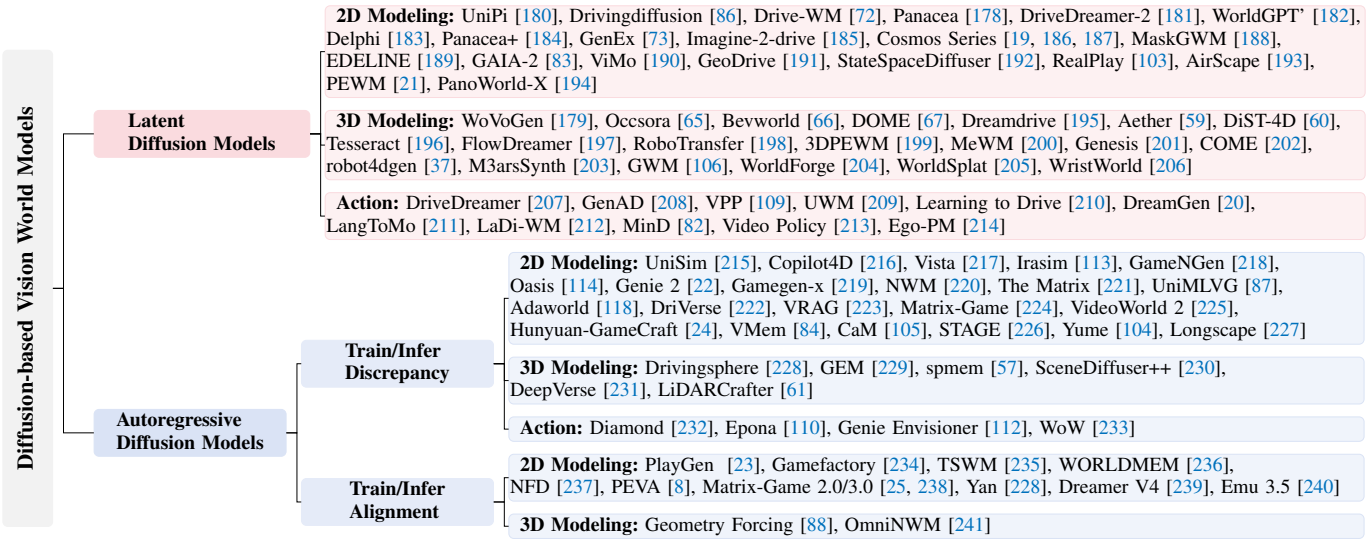

Fig. 4: Summary of representative methods of diffusion-based VWMs. Methods are divided into Latent Diffusion and Autoregressive Diffusion. They are categorized by output type into *2D Modeling*, *3D Modeling*, and action-extended generation (*Action*). Autoregressive Diffusion approaches are further classified based on the alignment of training and inference distributions into *Train/Inference Discrepancy* and *Train/Inference Alignment* paradigms.

3.2.2 *Autoregressive Diffusion*

Autoregressive diffusion combines sequential generation (Section 3.1) with latent diffusion (Section 3.2.1). While latent diffusion jointly denoises a fixed temporal window in a single process, autoregressive diffusion propagates generation beyond that window by conditioning each denoising step on previously generated outputs. This design produces future states sequentially, enabling extended rollouts while retaining diffusion-based visual refinement. The blue part of Fig. 4 summarizes representative works under this design.

However, introducing sequential dependency into diffusion creates a training–inference mismatch. During training, denoising is conditioned on ground-truth history, whereas during inference it relies on previously generated states that may deviate from the true distribution. Such deviations can accumulate over time and lead to drift in long rollouts. To address this mismatch, recent methods such as PlayGen [23], GameFactory [234], WORLDMEM [236], and Geometry Forcing [88] introduce noise augmentation, autoregressive rollout simulation, or memory mechanisms during training. These strategies encourage the model to handle self-generated historical inputs more robustly and reduce error accumulation in long sequences.

• **Continuous and Discrete Latents.** Autoregressive diffusion inherits representation choices from both autoregressive token models and latent diffusion. Some designs [110, 113, 217, 221, 228, 231] adopt continuous VAE latents to preserve fine visual detail. Others [22, 216, 234, 236] employ discrete token sequences to leverage transformer-based scalability. The choice reflects a trade-off between visual precision and architectural efficiency.

• **Sequentially Conditioned Denoising.** Methods in this design learn world knowledge through sequentially conditioned denoising, where the denoising process at each step is conditioned on previously generated history to predict future states sequentially rather than jointly within a fixed window. This step-by-step conditioning makes it possible to extend generation over long horizons, but it also means that small deviations in earlier predictions can propagate and accumulate, leading to long-horizon inconsistency.

To improve robustness over long-horizon rollouts, many works introduce auxiliary mechanisms that strengthen the use of historical context or impose stronger geometric guidance. For example, memory modules in WORLDMEM [236] and VMem [84] retain additional past information beyond the immediate context window, helping the model recover relevant scene state and reduce long-horizon inconsistency. Other approaches such as Geometry Forcing [88] incorporate geometry-based guidance to encourage spatially consistent generation, for instance maintaining coherent 3D layout and reducing physically implausible interpenetration or geometry collapse.

• **Sequential Denoised Rollouts.** Simulation proceeds as a sequential rollout of denoised latent states. At each step, new future states are generated conditioned on prior outputs and control inputs. This design supports diverse interaction modalities, including high-level language goals [215] and low-level control signals such as keyboard/mouse inputs [25, 219]. Improvements in inference efficiency have enabled near real-time interactive generation. For example, GameNGen [218] and Yan [250] demonstrate playable neural environments operating at interactive frame rates (20+ FPS).

• **Strengths and Limitations.** Autoregressive diffusion combines high visual fidelity with the ability to generate extended future sequences. By applying denoising at each step while conditioning on previously generated states, it can produce visually consistent futures beyond a fixed temporal window. Nevertheless, this hybrid design inherits challenges from both components. Sequential conditioning increases sensitivity to error accumulation over long horizons, and diffusion sampling remains computationally intensive. While memory mechanisms and geometry-based conditioning mitigate these issues, achieving reliable long-horizon consistency and stable physical behavior remains an open problem.

3.3 Embedding Prediction

Embedding prediction methods depart from pixel-level generation and instead model world change directly in representation space.

TABLE 3: Summary of embedding prediction-based VWMs.

Paper	Visual Encoder	World Knowledge	Interaction Input
IWM [242]	ViT	Photometric transformations	Transformation parameters
V-JEPA [29]	ViT	Spatio-temporal coherence, motion understanding	Video mask
LAW [32]	Swin, ResNet	Scene persistence, ego dynamics	Action, trajectory
DINO-WM [243]	DINOv2	Spatio-temporal dynamics, object interactions	Action
DINO-Foresight [244]	DINOv2	Scene dynamics, geometry	Latent action
AD-L-JEPA [245]	3D CNN	Spatial geometry in BEV, scene persistence	Spatial mask
EchoWorld [246]	ViT	Heart anatomy, motion dynamics	Ultrasound probe movement and pose
OSVI-WM [247]	ResNet	Environment dynamics, imitation dynamics	Waypoint
FLARE [30]	SigLIP-2	Policies alignment, long-term consequences	Robot state, text
seq-JEPA [248]	ResNet	Invariant-equivariant representations	Action (rotation, position)
V-JEPA 2 [28]	ViT	Motion, action anticipation, causal physics	Robot action, pose
SIVWM [249]	DINO	Spatio-temporal coherence, efficient planning	Action
DINO-world [31]	DINOv2	Temporal dynamics, intuitive physics	Action, trajectory

Rather than reconstructing images or videos, they predict future embeddings that encode task-relevant spatio-temporal and semantic information. This idea is exemplified by the Joint Embedding Predictive Architecture (JEPA) series [28, 29, 251], where the objective is to predict future feature representations instead of full visual signals. Table 3 summarizes representative works in this category.

• **Contextual Embeddings.** In this design, visual observations are encoded into contextual embeddings produced by vision foundation models such as DINOv2, CLIP, or SigLIP [77, 252, 253]. The term *contextual* refers to embeddings computed from the observed past or current frames, which serve as the conditioning context for predicting future representations. This distinguishes them from the predicted target embeddings, which represent future states.

A common strategy is to freeze the visual encoder and train only the predictive module. For example, DINO-WM [243] and DINO-World [31] reuse pretrained encoders to obtain semantically rich features, reducing the burden of learning visual representations from scratch. This separation simplifies training and allows the predictive module to focus on modeling how world state changes over time and how control inputs influence future representations. Moreover, the same embedding-based framework can be applied to different input modalities beyond RGB images, as long as they can be mapped into a shared embedding space. AD-L-JEPA [245] applies the approach to LiDAR point clouds, while EchoWorld [246] adapts it to ultrasound data, showing that contextual embeddings can serve as a common representation across different sensory modalities.

• **Latent Prediction in Representation Space.** Instead of reconstructing pixels, these methods train a predictor to approximate future embeddings given the current context. Training typically follows a mask-and-predict strategy, where parts of the input are masked and the model predicts their embeddings from visible context [28, 29]. This encourages the model to maintain spatio-temporal coherence in representation space, such as preserving object identity and motion continuity when parts of the input are occluded.

Beyond coherence, action-conditioned variants such as DINO-WM [243] and FLARE [30] extend embedding prediction to model causal relations. By incorporating control inputs into the prediction process, they capture how different actions lead to different future embeddings. The same framework can also be applied to domain-specific settings. For example, LAW [32] models driving behaviors and traffic conventions directly in representation space, while OSVI-WM [247] learns task-relevant state transitions from limited demonstrations.

• **Embedding-Space Rollouts.** Simulation in this design takes the form of predicted target embeddings rather than rendered images. Future world states are represented as sequences of feature vectors, which can be used directly for planning or policy evaluation.

Because visual decoding is not required, planning and action evaluation can be performed entirely in representation space. For example, DINO-WM [243] evaluates multiple candidate action sequences within embedding space to select promising behaviors. SIVWM [249] further improves efficiency by selectively updating only spatial regions relevant to the task. In addition, embedding representations are compact and modality-agnostic, allowing integration of heterogeneous inputs such as robot proprioception and language instructions [30, 245].

• **Strengths and Limitations.** A primary advantage of embedding prediction is computational efficiency, as rollouts involve lightweight operations in feature space rather than expensive video synthesis. This makes the approach particularly suitable for long-horizon planning.

However, the absence of explicit visual decoding reduces interpretability, since predicted futures cannot be directly inspected as images. Furthermore, reliance on frozen foundation models can constrain representational capacity, as the predictive module is bounded by the capability of the pretrained encoder.

3.4 State Transition

Unlike generation-focused designs that simulate futures as pixels or token sequences (Sections 3.1 and 3.2), and unlike embedding prediction that forecasts future embeddings without an explicit state update (Section 3.3), state-transition approaches represent the world as a compact latent state and model how this state evolves over time. The core idea is to maintain an evolving internal state that summarizes relevant history and supports long-horizon prediction under interaction.

We categorize these designs by how the latent state is structured: *State Space Modeling* (Section 3.4.1) maintains a single global recurrent state, whereas *Object-Centric Modeling* (Section 3.4.2) factorizes the state into a set of entity slots to model interactions among individual objects.

3.4.1 State Space Modeling

State space modeling is a classical design in world modeling, exemplified by the Dreamer line of work [10, 11, 116]. The core idea is to encode visual observations into a recurrent latent state and update this state over time, so that future prediction can be

TABLE 4: Summary of VWMs based on state space modeling. **".V"** denotes the visual part.

Methods	Visual Encoder	Transition Model
PlaNet [76]	CNN	RSSM
DreamerV1 [10]	CNN	RSSM
DreamerV2 [116]	CNN	RSSM
CADDY [56]	CNN	ConvLSTM
DayDreamer [254]	CNN	RSSM
MWM [255]	CNN, ViT	RSSM
IRIS [256]	CNN	Transformer
DreamerV3 [11]	CNN	RSSM
MV-MWM [70]	MAE	RSSM
TWM [257]	CNN	Transformer
CoWorld [258]	CNN	RSSM
SafeDreamer [259]	CNN	RSSM
SWIM [260]	NVAE	RSSM
MoDem-V2 [261]	CNN, MLP	MLP
TD-MPC2 [115]	CNN, MLP	MLP
HarmonyDream [102]	CNN	RSSM
STORM [262]	VAE	Transformer
MUVO [263]	CNN	GRU
Think2drive [264]	CNN	RSSM
REM [265]	VQ-VAE	RetNet
R2I [266]	CNN	S3M (SSM variant)
Puppeteer [81]	CNN+MLP	MLP
DynaLang [119]	CNN	RSSM
DriveWorld [137]	CNN	Memory SSM
GenRL [267]	InternVideo2**.V**	GRU
LS-Imagine [101]	CNN	Dual-branch RSSM
RoboHorizon [71]	MAE	RSSM
AdaWM [268]	CNN	RSSM
S5WM [269]	CNN	S5 (SSM variant)
DMWM [270]	CNN	RSSM + LINN
Simulus [271]	VQ-VAE	RetNet
Dywa [124]	PointNet++	MLP
WoTE [272]	ResNet	Transformer
ReDRAW [273]	CNN	MLP
FOUNDER [34]	InternVideo2**.V**	RSSM
SSVWM [33]	3D VAE	Mamba SSM
Raw2Drive [274]	BEVFormer	RSSM
NavMorph [275]	CLIP	RSSM
EMERALD [35]	VAE	Transformer SSM
LPS [63]	CNN	RSSM
GASv2 [276]	VAE	RSSM

performed efficiently in latent space. Table 4 summarizes representative systems, including their encoders and transition models.

• **Compact Recurrent State.** These methods compress visual observations (e.g., images or videos) into a compact recurrent state that retains information most relevant for prediction and decision-making. Early designs typically rely on CNN/ResNet encoders [277] for feature extraction [10, 56, 76, 116]. More recent work replaces early CNN encoders with stronger backbone models, often based on different variants of Vision Transformers (ViTs), to better capture global context and spatial relationships. For example, MWM [255] first trains a ViT-based autoencoder as a visual encoder before learning the recurrent transition model. In self-driving settings, methods such as Think2Drive, WoTE, and DriveWorld [137, 264, 272] aggregate multi-camera inputs into a unified BEV representation. Domain-specific pipelines, such as GASv2 [276], further adapt the encoder to process stereoscopic surgical images for precise manipulation.

• **Recurrent State Transition.** Knowledge learning is realized through the state transition model $P(s_{t+1}|s_t, a_t)$, which updates the latent state s_t conditioned on action a_t. A key capability of this design is long-horizon modeling: by repeatedly applying the transition update, the model can propagate predictive information over many steps. As the sequence grows, the transition model is required to effectively retain and utilize historical information, which motivates architectures that improve long-range state propagation. For example, R2I [266] replaces GRU-style updates with a structured state-space mechanism for learning long-range dependencies, and SSVWM [33] uses mamba-style transitions to maintain coherence over long sequences even under occlusion. Hierarchical transitions in LS-Imagine [101] further support longer-range rollouts by operating at multiple temporal scales.

• **Latent State Rollouts.** Simulation is performed as rollouts in the latent state space. This is conceptually close to embedding prediction methods in that futures are generated in a compact representation space rather than pixel space; however, state-transition models explicitly maintain a recurrent state and update it step by step. This enables fast long-horizon rollouts for evaluating candidate action sequences, as demonstrated in the Dreamer family [10, 11]. Interaction can range from low-level motor commands [254] to higher-level action abstractions [56]. Hierarchical control is often supported by stacking controllers at different abstraction levels; for example, Puppeteer [81] uses a high-level world model to guide lower-level skills, while systems such as FOUNDER [34] and RoboHorizon [71] incorporate language models to break down instructions into sub-goals.

• **Strengths and Limitations.** A core strength of state space modeling is efficient long-horizon rollout in latent space, making it suitable for settings where many candidate futures must be evaluated under tight compute budgets. The ability to propagate the latent state over many steps also allows the model to retain information about past observations and actions across extended horizons. Limitations include reduced human interpretability, since predicted futures are represented as latent states rather than directly visualized as images. In addition, compact state representations may struggle to preserve fine-grained spatial or geometric detail compared to models that operate on richer visual or 3D-aware representations.

TABLE 5: Summary of VWMs based on object-centric modeling.

Methods	Visual Encoder	Object Decomposition
G-SWM [278]	CNN, ViT	Background, foreground objects
HOWM [279]	Slot Attention	background, foreground slots
SlotFormer [280]	SAVi	Different slots
FOCUS [281]	SAM, CNN	Individual scene objects
COSMOS [282]	SAM, ResNet, CLIP	Neuro-symbolic pairs
SSWM [108]	Slot Attention	Object-centric slots
RoboDreamer [283]	VQ-VAE	Semantic components
slotSSM [284]	Slot Attention	Object-centric slots
MEAD [285]	CNN	Item identity, attribute
Dreamweaver [286]	CNN	Static and dynamic factors
OC-STORM [287]	Cutie+VAE	Object feature, background
DisWM [288]	β-VAE	Disentangled factors
LSlotFormer [289]	SAVi	Objects, background
SlotPi [290]	SAVi	Physical properties
Dyn-o [36]	Cosmos	Static and dynamic propertiess

3.4.2 *Object-Centric Modeling*

Object-centric designs represent the world as a set of entity slots, providing a factored latent state for modeling world change. By updating object-level states rather than modeling the entire scene as a single monolithic vector, these methods aim to improve interpretability and compositional generalization. Table 5 summarizes representative systems, including their encoders and slot construction mechanisms.

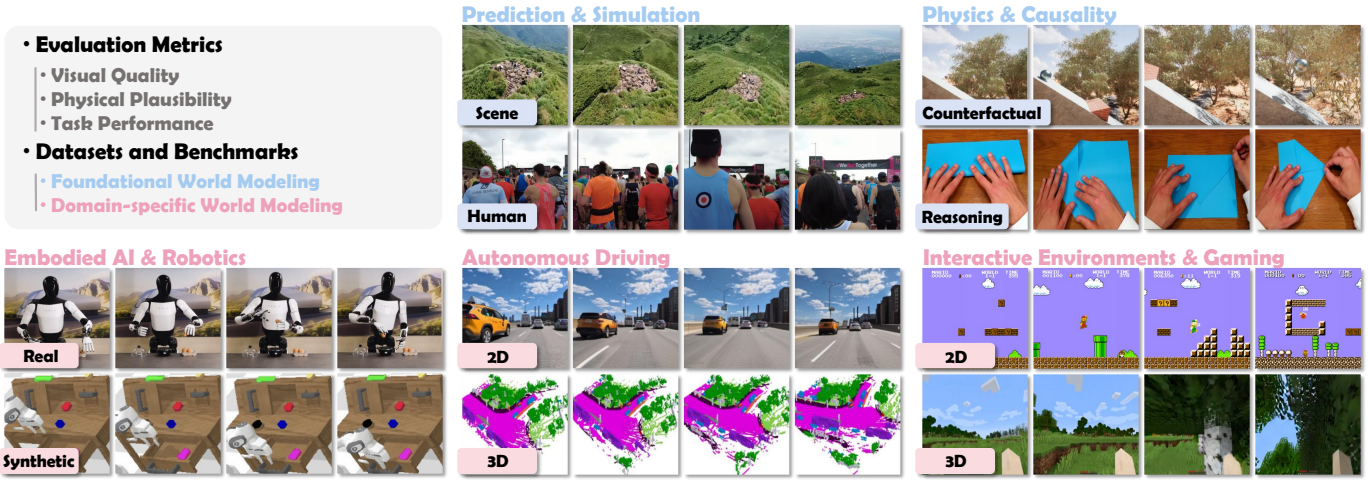

Fig. 5: Overview of the evaluation system of VWMs.

• **Object Slots.** The basic unit is an object slot, a vector representation intended to capture the state of a single entity (e.g., identity/attributes and spatial extent). Slots are typically obtained via unsupervised binding mechanisms such as Slot Attention [291] or SAVi [292], which group pixels into entity-centric components. Recent works incorporate stronger backbones before binding, for example tokenizing patches with ViTs (SlotFormer [280], SlotSSMs [284]) or separating foreground from background (G-SWM [278]). To incorporate semantic information, COSMOS [282] augments continuous slot vectors with symbolic attributes (e.g., shape or color) and aligns them with representations from CLIP [252]. This alignment enables slot representations to correspond to interpretable concepts.

• **Slot Interaction.** Knowledge learning is realized by modeling interactions among slots over time, capturing how entities influence one another. Instead of a single global state update, these models learn interaction modules that propagate information between slots (e.g., attention or graph-based interactions). SlotFormer [280] and LSlotFormer [289] use attention-based slot dynamics to predict long-horizon trajectories. Physical constraints can be incorporated to encourage more realistic interactions; for example, SlotPi [290] introduces Hamiltonian structure to regularize energy exchange, and G-SWM [278] models occlusion/collision-related interactions via structured networks. These interaction-centric updates also support compositional generalization, enabling learned interaction patterns to transfer across novel combinations of objects [279, 282].

• **Slot-State Rollouts.** Simulation produces rollouts in slot space, updating entity states over time according to learned interactions. This disentangled form of rollout supports controllable manipulation of entity attributes. Dreamweaver [286] demonstrates object-level editing by swapping or altering properties, and Dyn-O [36] leverages *both* static and dynamic disentanglement to maintain identity while simulating complex motions. Interaction can be multimodal; for example, LSlotFormer [289] fuses slots with T5 embeddings [293] for text-conditioned manipulation, and MEAD [285] evaluates outcomes of targeted object perturbations.

• **Strengths and Limitations.** A key strength is compositional generalization: object-centric factorization can handle scenes with new combinations of known entities more robustly than holistic latent states. The explicit slot structure also improves interpretability, as predictions can be traced to specific entities. A major limitation is the binding problem in complex real-world videos: slot assignment can be ambiguous under clutter, heavy occlusion, or fine-grained texture, and many methods rely on a fixed number of slots. Therefore, scaling robust slot discovery and interaction modeling to unconstrained natural videos remains challenging.

3.5 Other Emerging Directions

Beyond the major architectural families discussed above, several alternative modeling directions have been explored. Some works adopt lightweight architectures (e.g., CNN-based or feed-forward designs) for specific application settings. Others experiment with alternative transition mechanisms, including attention-based updates, graph neural networks for explicit entity interaction modeling, and flow-matching formulations as generative alternatives [294, 295, 296, 297, 298]. These approaches remain relatively specialized and have not yet formed large, unified research branches. Nevertheless, they reflect ongoing exploration of architectural choices for modeling world knowledge.

4 Evaluation

Building upon the unified framework in Section 2 and the architectural designs in Section 3, this section reviews how VWMs are evaluated in practice. As illustrated in Fig. 5, we organize the discussion into two parts: *Evaluation Metrics* (Section 4.1), which define what aspects of world modeling are measured, and *Datasets and Benchmarks* (Section 4.2), which introduce the data resources and standard test suites used in these evaluations.

4.1 Evaluation Metrics

We organize evaluation metrics into three groups: *Visual Quality*, which measures the visual fidelity of generated images or videos; *Physical Plausibility*, which evaluates adherence to physical dynamics, geometric structure, and long-horizon spatio-temporal consistency; and *Task Performance*, which assesses whether the model enables reliable task completion and reflects the causal dependencies required to achieve those tasks. Table 6 summarizes representative metrics under each category.

• **Visual Quality.**

TABLE 6: Summary of representative metrics for VWM, categorized by core dimensions and their corresponding sub-dimensions.

Sub-dimension	Metric
Visual Quality	
Objective Fidelity	PSNR [299], SSIM [300], FVD [301], FID [302]
Perceptual Alignment	LPIPS [303], DOVER Score [304], DreamSim [305]
Physical Plausibility	
Kinematic Accuracy	ADE [208], FDE [208], RPE [306], MPJPE [307], MPJVE [307], PA-MPJPE [307], Camera Pose Loss [234], Flow Error [234]
Geometric Validity	2D Reprojection Error [308], Chamfer Distance [243], AbsRel [143]
Spatio-temporal Consistency	Revisit Error [309], Scene Revisit Consistency [310]
Task Performance	
Process-level Evaluation	Raw Return [10], Driving Score [311], Human Normalized Score [116], PDMS Score [312]
Goal Completion	Success Rate [313], Contact Rate [281], MTLC Success Rate [314], Grasping Score [276], Collision Rate [315]
Perception/Control Accuracy	Top-K Accuracy [28], Recall [34], Precision [34], F1-Score [141], Translation Error [254], Rotation Error [246]

Visual quality measures the visual fidelity of generated outputs. For models that synthesize images or videos [11, 117], this provides a direct assessment of image sharpness, temporal smoothness, and overall visual realism. We group these metrics into objective fidelity and perceptual alignment.

Objective fidelity metrics quantify similarity between generated and reference data. Low-level measures such as PSNR [299] and SSIM [300] evaluate pixel-wise differences and local structural similarity. Distributional metrics like FID [302] (for images) and FVD [301] (for videos) compare feature distributions of real and generated samples to measure global appearance similarity.

Because pixel-level similarity does not always reflect human judgment, perceptual alignment metrics such as LPIPS [303] and DreamSim [305] compute distances in pretrained feature spaces to better align with human preference. No-reference metrics such as DOVER Score [304] further evaluate visual realism by estimating artifact level, naturalness, and perceptual quality without ground-truth comparison.

• **Physical Plausibility.**

Physical plausibility evaluates whether predicted motion and spatial structure remain consistent with fundamental physical constraints across space and time. We group these metrics into three aspects: kinematic accuracy, geometric validity, and spatio-temporal consistency. Kinematic accuracy measures the correctness of predicted motion states such as trajectories, poses, and velocities. Metrics including ADE and FDE [208] evaluate trajectory deviation; RPE [306] and Camera Pose Loss [234] assess ego-motion estimation; MPJPE, PA-MPJPE, and MPJVE [307] quantify articulated pose accuracy; and Optical Flow Error [197] measures short-term motion precision.

Geometric validity assesses whether the predicted 3D structure remains physically feasible. Chamfer Distance [316] measures structural similarity in point clouds; AbsRel [143] evaluates depth prediction accuracy; and 2D Reprojection Error [308] tests multi-view geometric alignment.

Spatio-temporal consistency evaluates whether scene structure remains stable over extended rollouts. Scene Revisit Consistency (SRC) [310] and Revisit Error (RVE) [309] measure whether returning to the same spatial location yields a structurally consistent scene. Together, these metrics assess whether a model maintains physically plausible motion and stable structure over time.

• **Task Performance.**

Task performance measures whether a world model supports successful downstream task completion. When the model captures the task-critical physical constraints and causal dependencies, agents using its rollouts should achieve higher success rates and more reliable behavior.

We group task metrics into process-level evaluation, goal completion, and perception/control accuracy. Process-level metrics, as commonly employed in reinforcement learning, evaluate cumulative reward or driving score, such as Raw Return [10], Human Normalized Score [116], Driving Score [311], and PDMS Score [312]. Goal completion metrics directly assess whether predefined objectives are achieved. Success Rate [313], MTLC Success Rate [314], Contact Rate [281], Grasping Score [276], and Collision Rate [315] quantify task reliability and safety.

Perception and control accuracy further measure downstream prediction and control precision, including Precision, Recall, F1-Score [34, 141], Top-K Accuracy [28], Translation Error [254], and Rotation Error [246]. Collectively, these metrics evaluate whether the learned world model supports accurate, reliable, and causally grounded task execution.

4.2 Datasets and Benchmarks

This section reviews datasets and benchmarks used to evaluate world modeling capability. For clarity, when applicable, we first describe datasets mainly used for training, followed by benchmarks designed for evaluation, and finally resources that serve both purposes.

We organize the discussion into two groups. *Foundational World Modeling* (Section 4.2.1) summarizes datasets and benchmarks targeting general world modeling ability, while *Domain-specific World Modeling* (Section 4.2.2) focuses on evaluation settings tailored to specific application domains. Tables 7 and 8 summarize representative datasets and benchmarks in these two groups. For each benchmark, we list its original metrics and add tags mapping them to the three metric categories in Section 4.1 (Visual Quality / Physical Plausibility / Task Performance) to provide a unified view across heterogeneous evaluation protocols.

4.2.1 Foundational World Modeling

Table 7 summarizes representative datasets and benchmarks for foundational evaluation. We group them into two categories: *General World Prediction and Simulation*, which emphasizes long-horizon forecasting and controllability, and *Physics and Causality Benchmarks*, which assess whether models adhere to physical constraints and produce predictions that follow plausible causal relationships.

• **General World Prediction and Simulation.**

Large-scale video datasets such as Something-Something V2 (SSV2) [317] and Ego4D [74] are widely used to pretrain visual encoders and temporal representations, providing broad supervision for learning structured world behavior from video. Several benchmarks evaluate general world modeling capability under long-horizon prediction and controllability. WorldModelBench [318] reports Instruction, Physics, and Commonsense

TABLE 7: Foundational world modeling datasets and benchmarks. Each entry is annotated by its primary role: (D) indicates datasets mainly used for training, and (B) indicates benchmarks primarily designed for evaluation. When applicable, benchmark metrics are tagged according to the three evaluation categories defined in Section 4.1: (**V**) Visual Quality, (**P**) Physical Plausibility, and (**T**) Task Performance. Representative downstream works are listed for reference.

Name	Year	Role	Metric	Downstream Work
General World Prediction and Simulation				
SSV2 [317]	2017	D	-	[31, 118, 159]
Ego4D [74]	2021	D	-	[155, 159, 220]
WorldModelBench [318]	2025	B	Instruction Score (**T**), Physics Score (**P**), Commonsense Score (**V, P**)	[233, 319]
WorldScore [308]	2025	B	Controllability Score (**T**), Quality Score (**V**), Dynamics Score (**P**)	[59, 320]
WorldPrediction [321]	2025	B	World Modeling Score (**P**), Procedural Planning Score (**T**)	[111, 322]
Sekai [323]	2025	D&B	VBench's metrics (**V, P**), TransErr (**T**), RotErr (**T**)	[104, 324, 325]
OmniWorld [326]	2025	D&B	Camera Parameter Metrics (**P, T**), FVD (**V**)	[327]
Others: UCF101 [328], YouTube-8M [329], Kinetics [330], COIN [331], HowTo100M [332], WebVid-2M [333], InternVid [17], Ego-Exo4D [334], EgoVid-5M [335], OpenHumanVid [336], MM-OR [337]				
Physics and Causality Benchmark				
CoPhy [338]	2019	B	MSE (**P**)	[339]
Physion++ [340]	2023	B	Object Contact Prediction Accuracy (**P**)	[280, 290]
Physics-IQ [341]	2025	B	Spatial IoU/Spatio-temporal IoU (**P**), MSE (**P**)	[21, 342]
Vbench-2.0 [343]	2025	B	VQA-based Physics and Commonsense Score (**V, P**)	[112, 154, 231]
VideoPhy-2 [92]	2025	B	Semantic Adherence (**V**), Physical Commonsense (**P**), Physical Rule Violation (**P**)	[344, 345]
IntPhys 2 [9]	2025	B	Surprise Score (**P**)	[31]
PAI-Bench [346]	2025	B	Quality Score (**V, P**), Domain Score (**P**), Control Fidelity (**T**)	[187, 319]
Others: VoE Dataset [347], InfLevel [348], VIDEOPHY [349], Phybench [350], PhyGenBench [351], PBench [352], PhyCoBench [353], PisaBench [354], WISA-32K [355], T2VPhysBench [356], PhysVidBench [357], VideoVerse [345]				

Scores to assess world understanding under prompts; WorldScore [308] evaluates controllability and generation quality via Controllability, Quality, and Dynamics scores; and WorldPrediction [321] measures usefulness for downstream decision-making with a World Modeling Score and a Procedural Planning Score. Recent works provide both large-scale datasets and standard evaluation protocols to test structural consistency. Sekai [323] evaluates 4D spatio-temporal consistency using VBench [358], and OmniWorld [326] further incorporates camera-parameter-based metrics to test viewpoint-consistent generation.

• **Physics and Causality Benchmarks.**

Works in this category are primarily evaluation benchmarks, assessing whether models adhere to physical constraints and follow causal relationships. Early benchmarks emphasize controlled prediction and counterfactual evaluation. CoPhy [338] introduces paired factual and counterfactual scenes, and Physion++ [340] requires inferring latent physical properties (e.g., mass, friction) before forecasting outcomes, evaluated via metrics such as Object Contact Prediction Accuracy. IntPhys 2 [9] adopts a violation-of-expectation protocol and uses a Surprise Score to measure whether models distinguish plausible from physically impossible events.

The emergence of large-scale video generation models has prompted a shift in benchmarking, from static predictive evaluation toward automated testing of physical consistency in synthesized rollouts. Physics-IQ [341] quantifies spatio-temporal alignment using IoU-based measures. VideoPhy-2 [92] evaluates compliance with physical rules and commonsense constraints in generated videos. VBench-2.0 [343] and PAI-Bench [346] further employ VLM-based judges to scale evaluation of physical plausibility, commonsense consistency, and controllability in generated worlds.

4.2.2 Domain-specific World Modeling

Table 8 summarizes representative datasets and benchmarks in three major domains: *Embodied AI and Robotics*, *Autonomous Driving*, and *Interactive Environments and Gaming*.

• **Embodied AI and Robotics.**

Large-scale robotics datasets such as DROID [362] provide diverse real-world interaction trajectories for training policies and world models. Simulation platforms such as RLBench [359] and Meta-World+ [365] offer controlled manipulation environments, where task success rate serves as a primary metric across predefined task suites. CALVIN [314] and LIBERO [361] further extend evaluation to longer-horizon and multi-step tasks, assessing skill composition and transfer through metrics including MTLC success, Forward Transfer (FWT), and Area Under the Curve (AUC). More recent platforms combine large-scale real-world data with closed-loop evaluation tailored to world modeling. AgiBot World [363] reports real-world task completion under closed-loop control, while WoWBench [233] evaluates planning ability, physical constraint adherence, and instruction following in real manipulation scenarios.

• **Autonomous Driving.**

Driving datasets such as KITTI [376] and Waymo [378] provide large-scale real-world videos for perception and trajectory prediction. nuScenes [377] introduces richer 3D annotations and supports evaluation with metrics including L2 error and collision-related measures. Video datasets such as OpenDV-2K [379] further expand coverage of complex traffic scenes for training generative world models. Recent benchmarks emphasize closed-loop and action-conditioned evaluation. NAVSIM [312] evaluates planning quality using PDMS. Act-bench [383] assesses action-conditioned future prediction, combining Instruction–Execution Consistency with trajectory metrics such as ADE and FDE to measure long-term consequence prediction. DrivingDojo [381] serves as both a dataset and a benchmark for action-conditioned generation, reporting FID/FVD and instruction-following errors to evaluate visual quality and controllability.

• **Interactive Environments and Gaming.**

TABLE 8: Domain-specific world modeling datasets and benchmarks. Each entry is annotated by its primary role: (D) indicates datasets mainly used for training, and (B) indicates benchmarks primarily designed for evaluation. When applicable, benchmark metrics are tagged according to the three evaluation categories defined in Section 4.1: (**V**) Visual Quality, (**P**) Physical Plausibility, and (**T**) Task Performance. Representative downstream works are listed for reference.

Name	Year	Role	Metric	Downstream Work
Embodied AI and Robotics				
RLBench [359]	2019	D&B	Success Rate (**T**)	[71, 283, 360]
CALVIN [314]	2021	D&B	Success Rate, MTLC Success Rate (**T**)	[26, 149, 212]
LIBERO [361]	2023	D&B	FWT (**T**), AUC (**T**), Success Rate (**T**)	[112, 122, 125]
DROID [362]	2024	D	-	[20, 363, 364]
AgiBot World [363]	2025	D&B	Task Completion Score (**T**)	[20, 118, 122]
Meta-World+ [365]	2025	B	Success Rate (**T**)	[106]
WoWBench [233]	2025	B	Video Quality, Planning Reasoning, Physical Law, and Instruction Understanding Score (**V, P, T**)	[364]
Others: RoboNet [366], Isaac Gym [367], BC-Z [368], RT-1 Dataset [369], VP^2 [370], RH20T [371], BridgeData V2 [372], OXE [18], MimicGen [373], BEHAVIOR-1K [374], RoboCasa [375]				
Autonomous Driving				
KITTI [376]	2012	D	-	[31, 208, 216]
nuScenes [377]	2019	D&B	nuScenes Detection Score (**P, T**), L2 Error (**P**), Collision Rate (**T**)	[32, 150, 207]
Waymo [378]	2019	D	-	[87, 208, 217]
OpenDV-2K [379]	2024	D	-	[188, 217, 229]
NAVSIM [312]	2024	B	Predictive Driving Model Score (**T**)	[32, 110, 380]
DrivingDojo [381]	2024	D&B	Action Instruction Following Errors (**T**), FID/FVD (**V**)	[229, 382]
Act-bench [383]	2024	B	Instruction-Execution Consistency (**T**), ADE (**P**), FDE (**P**)	[384]
Others: CARLA [311], ApolloScape [385], BDD100K [386], SemanticKITTI [387], INTERACTION [388], Argoverse [389], A2D2 [390], Lyft Level 5 [391], WOMD [392], ONCE [393], KITTI-360 [394], Argoverse 2 [395], OpenOccupancy [396], Occ3D [397], ZOD [398], Cardreamer [399], DriveArena [400]				
Interactive Environments and Gaming				
ALE [401]	2013	B	Game Score (**T**)	[11, 116, 256]
DMC [402]	2018	B	Raw Return (**T**)	[11, 115]
Crafter [403]	2021	B	Raw Return (**T**), Score of Achievement (**T**), Success Rate (**T**)	[11, 256]
Source [221]	2024	D	-	[59, 118, 236]
LOOPNAV [404]	2025	D&B	FVD, LPIPS, SSIM (**V**)	[220, 232]
Matrix-Game-MC [224]	2025	D&B	GameWorld Score (**V, P, T**)	[104, 324]
Others: Procgen [405], NLE [406], MineRL [407], Counter-Strike Deathmatch [408], Mars [409], OGameData [219], GF-Minecraft [234], JARVIS-VLA [410]				

Early platforms such as ALE [401] and the DMC Suite [402] provide closed-loop environments for evaluating control and prediction using gameplay scores (e.g., raw return). Crafter [403] introduces procedural generation and evaluates systematic generalization through raw return, achievement score, and success rate. Large-scale game-derived datasets such as Source [221] provide high-fidelity visual data for training generative world models. More recent benchmarks evaluate long-horizon consistency and interactive controllability. LOOPNAV [404] tests spatial memory and revisit consistency using metrics such as FVD, LPIPS, and SSIM. Matrix-Game-MC [224] integrates visual quality, physical plausibility, and controllability into a composite GameWorld Score for interactive world modeling.

5 Future Directions

In this section, we discuss key directions for advancing VWMs beyond current paradigms. Our discussion is organized around three complementary themes (Fig. 6): **Re-grounding** the learned knowledge in stronger physical and causal foundations, **Re-evaluating** progress with protocols that reflect versatile world modeling capability across diverse tasks and settings, and **Re-scaling** training and inference to unlock broader generalization and stronger reasoning under interaction.

5.1 Re-grounding: Strengthening the Knowledge Foundation

While current VWMs can generate visually plausible futures, their outputs can become unreliable when interactions are complex, rare events occur, or intervention effects vary across environments. A key next step is therefore **Re-grounding**: strengthening the world knowledge learned by VWMs (Section 2.3) so that models more faithfully capture how states change under physical constraints and interventions.

Two complementary directions are highlighted. First, broadening the scope of world knowledge (Section 5.1.1), moving beyond simplified physical settings toward richer natural and human environments. Second, enhancing architectural support for grounding (Section 5.1.2), so that model designs better preserve geometric structure and incorporate physical and causal priors.

5.1.1 Broadening the Scope of World Knowledge

Current VWM research has mainly emphasized a subset of physical settings and interaction patterns. To operate robustly in open-world environments, the knowledge scope should expand along two directions: richer physical interactions in the natural world, and human-centered rules that mediate action–effect relations in social scenarios.

- **Richer Physical Interactions in the Natural World.**

Current benchmarks and datasets emphasize relatively clean dynamics, such as rigid objects and simple motion, and underrepresent regimes where physical behavior depends on subtle interaction effects. In realistic settings, contact-rich manipulation, deformable materials, and surface-dependent motion often determine outcomes. For instance, success in robotic manipulation may hinge on precise contact and friction, while navigation reliability can vary with surface conditions. Improving coverage of such

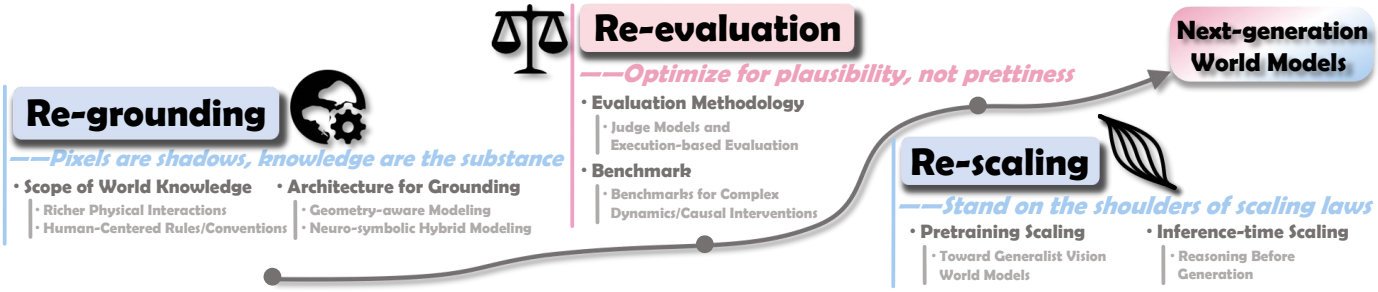

Fig. 6: A structured view of open challenges and future directions for Vision World Models (VWMs).

regimes requires models to capture these fundamental constraints in the real world, moving beyond simplified benchmark settings toward more realistic and reliable world modeling.

• **Human-Centered Rules and Conventions.**

Beyond physical dynamics, behavior in human environments is shaped by social norms, conventions, and shared intentions formed by cultural and institutional practices. Current models often struggle to capture the principles that govern how social context modifies the effects of actions (e.g., exceptions or priority rules). For instance, a model may learn the correlation "red light → stop" from training data, yet fail in scenarios where human factors alter this rule, such as temporary traffic control or human-directed overrides. For future VWMs, modeling such human-centered social norms and contextual rules is essential to support reliable and safe interaction with people.

5.1.2 Architectures for Stronger Grounding

Expanding the scope of world knowledge also requires corresponding advances in model design. We advocate two complementary directions: (i) preserving geometric or 3D structure to strengthen spatio-temporal coherence; and (ii) incorporating physical and causal constraints to improve modeling of complex dynamics and causal relations.

• **Geometry-aware Modeling.**

A common limitation of current VWMs is weak geometric grounding: many models lack an explicit representation of 3D structure and scene layout, which undermines their ability to maintain stable object identity, occlusion relations, and spatial consistency over time. Existing 3D-oriented approaches [411, 412, 413, 414, 415] are largely designed for static scenes and struggle in highly dynamic, interaction-rich settings. Therefore, future VWMs require stronger geometry/3D-aware modeling methods. We advocate two complementary directions: (i) explicitly modeling the world as a time-varying 3D structure, where spatial layout evolves directly over geometric primitives (e.g., 4D Gaussian representations [204, 412]); or (ii) injecting geometry-aware constraints into existing architectures (e.g., multi-view consistency or camera-aware conditioning) without full 3D reconstruction. While the former models dynamic 3D structures more directly, the latter integrates geometric inductive biases into existing models with lower overhead.

• **Neuro-symbolic Hybrid Modeling.**

Purely neural architectures may generalize poorly under out-of-distribution (OOD) conditions, such as shifts in physical parameters or unseen intervention patterns. In contrast, symbolic systems (e.g., physics engines or rule-based planners) provide precise rule-based reasoning but lack flexibility for complex visual inputs. Neuro-symbolic hybrids combine these strengths: neural components model perception and variability, while symbolic modules introduce explicit physical or causal constraints. For instance, a differentiable physical solver [90, 416] can model underlying dynamics, while a neural generator captures visual detail; similarly, a rule-based causal planner can guide action–effect reasoning, helping distinguish genuine intervention effects from spurious correlations.

This hybrid design offers several advantages. First, by incorporating structured physical or causal priors, it can improve generalization under distribution shifts, especially when interaction patterns change. Second, it provides more interpretable intermediate variables (e.g., forces, contacts, or causal factors) rather than relying solely on opaque latent vectors. Finally, such architectures may enable the extraction of explicit physical or causal structure from visual data, offering a path toward more structured and explainable world modeling.

5.2 Re-evaluation: Toward Versatile and Reliable Evaluation

Evaluation of VWMs remains a critical bottleneck: current protocols mainly employ metrics from adjacent fields (e.g., video generation), often emphasizing appearance quality while overlooking a VWM's ability to capture fundamental physical and causal principles [299, 300]. We discuss two complementary directions: improving evaluation methodology and designing benchmarks that better reflect core world modeling capability.

• **Judge Models and Execution-based Evaluation.**

A key limitation of current evaluation is the lack of holistic mechanisms that directly assess world modeling capability. One promising direction is to develop dedicated judge models for VWMs. Such models can be trained to evaluate whether predicted futures satisfy physical constraints and respond correctly to interaction conditions. These judge models can be further refined through preference learning or reinforcement-based alignment to better align with task requirements.

Complementing model-based judgment, evaluation through execution offers a more decisive signal of world modeling quality. By placing the world model in the execution loop, agents can plan and act using simulated rollouts, and the resulting task performance becomes a direct indicator of whether the model captures sufficient world knowledge. When performance degrades or planning breaks down, such failures provide concrete evidence of where the model's physical or causal understanding is incomplete, offering a more diagnostic signal than static scoring alone.

• **Benchmarks for Complex Dynamics and Causal Interventions.**

Benchmark design should incorporate more realistic scenarios, such as contact-rich manipulation, deformable materials and

friction-dependent motion. These settings better reflect real-world complexity and test whether models remain reliable under intricate physical constraints.

Beyond complex physical constraints, benchmarks should also probe causal behavior under controlled interventions. Starting from the same initial context, evaluations can vary an action or environmental condition and examine whether predicted futures change in the correct direction. Counterfactual settings are particularly useful in this regard [347]. For example, given a planted seed, the model should generate distinct growth patterns under drought versus adequate watering. Such tests directly assess whether the model captures stable causal relationships that govern how the world evolves, rather than merely copying observed patterns.

5.3 Re-scaling: Scaling Laws for Generalization and Reasoning

Scaling behavior will likely be the deciding factor in whether VWMs evolve toward more general and reliable capabilities. Empirical evidence suggests that increasing model size mainly improves visual fidelity [117, 233], while improvements in physical and causal knowledge remain limited across diverse settings. Re-scaling therefore concerns how scaling strategies can better promote stronger world modeling capabilities rather than visual quality alone. We discuss two complementary directions: pretraining scaling and inference-time scaling.

• **Pretraining Scaling: Toward Generalist VWMs.**

A promising direction is to scale VWMs under a unified modeling interface, so that a single model can support diverse world tasks and interaction settings, and potentially exhibit emergent capabilities such as cross-domain generalization, longer-horizon reasoning, and improved robustness under novel interactions. Realizing this potential requires scaling not only model capacity, but also the breadth and structure of training data and objectives [240]. Training corpora must cover diverse interaction patterns and long-horizon processes, while objectives should encourage the learning of fundamental physical and causal relations, rather than overfitting to superficial correlations. Because visual data are highly redundant across space and time, naive scaling can be computationally inefficient. Therefore, designing more efficient spatio-temporal tokenizers and scalable conditioning methods is critical for improving the effectiveness of scaling [417, 418].

• **Inference-time Scaling: Reasoning Before Generation.**

A second direction is to allocate additional test-time compute for better planning and causal reasoning. Instead of generating futures in one shot, a model may perform intermediate deliberation, such as proposing candidate outcomes, checking physical/causal constraints, or iteratively refining a rollout under intervention. This parallels recent trends in multimodal models [419, 420] where extra inference compute can improve reliability. For VWMs, inference-time scaling may be especially valuable for rare physical events, complex contact dynamics, and counterfactual reasoning, where a single forward pass is prone to instability.

6 Conclusions

In this survey, we provide a systematic analysis of Vision World Models (VWMs), a paradigm aimed at enabling artificial systems to understand, predict, and interact with the physical world by learning from visual data. We establish a unified framework that decomposes VWM into three core components: Vision Encoding, Knowledge Learning, and Controllable Simulation. Through this framework, we present a structured taxonomy and in-depth analysis of four major architectural families: sequential generation, diffusion-based generation, embedding prediction, and state transition models. We further review evaluation methodologies, organizing metrics, datasets, and benchmarks that shape the development of the field. Despite recent progress, substantial challenges remain in strengthening physical and causal grounding, developing more reliable and versatile evaluation protocols, and designing effective scaling strategies for next-generation world models. Addressing these challenges is essential for moving VWMs beyond specialized generative systems toward more general and reliable components of embodied intelligence, ultimately enabling artificial agents to act robustly in complex real-world environments.

References

[1] D. Hendrycks, D. Song, C. Szegedy, H. Lee, Y. Gal, E. Brynjolfsson, S. Li, A. Zou, L. Levine, B. Han *et al.*, "A definition of agi," *arXiv preprint arXiv:2510.18212*, 2025.

[2] J. Ding, Y. Zhang, Y. Shang, Y. Zhang, Z. Zong, J. Feng, Y. Yuan, H. Su, N. Li, N. Sukiennik *et al.*, "Understanding world or predicting future? a comprehensive survey of world models," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–38, 2025.

[3] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, vol. 2, no. 3, 2018.

[4] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.

[5] J. Del Ser, J. L. Lobo, H. Müller, and A. Holzinger, "World models in artificial intelligence: Sensing, learning, and reasoning like a child," *arXiv preprint arXiv:2503.15168*, 2025.

[6] E. Xing, M. Deng, J. Hou, and Z. Hu, "Critiques of world models," *arXiv preprint arXiv:2507.05169*, 2025.

[7] Y. Hu, L. Wang, X. Liu, L.-H. Chen, Y. Guo, Y. Shi, C. Liu, A. Rao, Z. Wang, and H. Xiong, "Simulating the real world: A unified survey of multimodal generative models," *arXiv preprint arXiv:2503.04641*, 2025.

[8] Y. Bai, D. Tran, A. Bar, Y. LeCun, T. Darrell, and J. Malik, "Whole-body conditioned egocentric video prediction," *arXiv preprint arXiv:2506.21552*, 2025.

[9] F. Bordes, Q. Garrido, J. T. Kao, A. Williams, M. Rabbat, and E. Dupoux, "Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments," *arXiv preprint arXiv:2506.09849*, 2025.

[10] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[11] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[12] J. Xiang, T. Tao, Y. Gu, T. Shu, Z. Wang, Z. Yang, and Z. Hu, "Language models meet world models: Embodied experiences enhance language models," *Advances in neural information processing systems*, vol. 36, pp. 75 392–75 412, 2023.

[13] K. Xie, I. Yang, J. Gunerli, and M. Riedl, "Making large language models into world models with precondition and effect knowledge," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 7532–7545.

[14] S. Zhou, T. Zhou, Y. Yang, G. Long, D. Ye, J. Jiang, and C. Zhang, "Wall-e: World alignment by rule learning improves world model-based llm agents," *arXiv preprint arXiv:2410.07484*, 2024.

[15] H. Tang, D. Key, and K. Ellis, "Worldcoder, a model-based llm agent: Building world models by writing code and interacting with the environment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 70 148–70 212, 2024.

[16] Y. Li, H. Wang, J. Qiu, Z. Yin, D. Zhang, C. Qian, Z. Li, P. Ma, G. Chen, H. Ji *et al.*, "From word to world: Can large language models be implicit text-based world models?" *arXiv preprint arXiv:2512.18832*, 2025.

[17] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[18] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collab-

oration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.
[19] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding *et al.*, "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025.
[20] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin *et al.*, "Dreamgen: Unlocking generalization in robot learning through video world models," *arXiv preprint arXiv:2505.12705*, 2025.
[21] Q. Sun, L. Yang, W. Tang, W. Huang, K. Xu, Y. Chen, M. Liu, J. Yang, H. Zhu, Y. Wang *et al.*, "Learning primitive embodied world models: Towards scalable robotic learning," *arXiv preprint arXiv:2508.20840*, 2025.
[22] J. Parker-Holder, P. J. Ball, J. Bruce, V. Dasagi, K. Holsheimer, C. Kaplanis, A. Moufarek, G. Scully, J. Shar, J. Shi *et al.*, "Genie 2: A large-scale foundation world model," https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/, 2024, accessed: April 1, 2026.
[23] M. Yang, J. Li, Z. Fang, S. Chen, Y. Yu, Q. Fu, W. Yang, and D. Ye, "Playable game generation," *arXiv preprint arXiv:2412.00887*, 2024.
[24] J. Li, J. Tang, Z. Xu, L. Wu, Y. Zhou, S. Shao, T. Yu, Z. Cao, and Q. Lu, "Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition," *arXiv preprint arXiv:2506.17201*, 2025.
[25] X. He, C. Peng, Z. Liu, B. Wang, Y. Zhang, Q. Cui, F. Kang, B. Jiang, M. An, Y. Ren *et al.*, "Matrix-game 2.0: An open-source, real-time, and streaming interactive world model," *arXiv preprint arXiv:2508.13009*, 2025.
[26] Z. Ren, Y. Wei, X. Guo, Y. Zhao, B. Kang, J. Feng, and X. Jin, "Videoworld: Exploring knowledge learning from unlabeled videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 029–29 039.
[27] S. Wang, J. Tian, L. Wang, Z. Liao, J. Li, H. Dong, K. Xia, S. Zhou, W. Tang, and H. Gang, "Sampo: Scale-wise autoregression with motion prompt for generative world models," *arXiv preprint arXiv:2509.15536*, 2025.
[28] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus *et al.*, "V-jepa 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.
[29] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint arXiv:2404.08471*, 2024.
[30] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne *et al.*, "Flare: Robot learning with implicit world modeling," *arXiv preprint arXiv:2505.15659*, 2025.
[31] F. Baldassarre, M. Szafraniec, B. Terver, V. Khalidov, F. Massa, Y. LeCun, P. Labatut, M. Seitzer, and P. Bojanowski, "Back to the features: Dino as a foundation for video world models," *arXiv preprint arXiv:2507.19468*, 2025.
[32] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan, "Enhancing end-to-end autonomous driving with latent world model," *arXiv preprint arXiv:2406.08481*, 2024.
[33] R. Po, Y. Nitzan, R. Zhang, B. Chen, T. Dao, E. Shechtman, G. Wetzstein, and X. Huang, "Long-context state-space video world models," *arXiv preprint arXiv:2505.20171*, 2025.
[34] Y. Wang, R. Yu, S. Wan, L. Gan, and D.-C. Zhan, "Founder: Grounding foundation models in world models for open-ended embodied decision making," in *Forty-second International Conference on Machine Learning*, 2025.
[35] M. Burchi and R. Timofte, "Accurate and efficient world modeling with masked latent transformers," *arXiv preprint arXiv:2507.04075*, 2025.
[36] Z. Wang, K. Wang, L. Zhao, P. Stone, and J. Bian, "Dyn-o: Building structured world models with object-centric representations," *arXiv preprint arXiv:2507.03298*, 2025.
[37] Z. Liu, S. Li, E. Cousineau, S. Feng, B. Burchfiel, and S. Song, "Geometry-aware 4d video generation for robot manipulation," *arXiv preprint arXiv:2507.01099*, 2025.
[38] Y. Guan, H. Liao, Z. Li, J. Hu, R. Yuan, G. Zhang, and C. Xu, "World models for autonomous driving: An initial survey," *IEEE Transactions on Intelligent Vehicles*, 2024.
[39] S. Tu, X. Zhou, D. Liang, X. Jiang, Y. Zhang, X. Li, and X. Bai, "The role of world models in shaping autonomous driving: A comprehensive survey," *arXiv preprint arXiv:2502.10498*, 2025.
[40] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," *arXiv preprint arXiv:2501.11260*, 2025.
[41] X. Long, Q. Zhao, K. Zhang, Z. Zhang, D. Wang, Y. Liu, Z. Shu, Y. Lu, S. Wang, X. Wei *et al.*, "A survey: Learning embodied intelligence from physical simulators and world models," *arXiv preprint arXiv:2507.00917*, 2025.
[42] W. Liang, R. Zhou, Y. Ma, B. Zhang, S. Li, Y. Liao, and P. Kuang, "Large model empowered embodied ai: A survey on decision-making and embodied learning," *arXiv preprint arXiv:2508.10399*, 2025.
[43] X. Li, X. He, L. Zhang, and Y. Liu, "A comprehensive survey on world models for embodied ai," *arXiv preprint arXiv:2510.16732*, 2025.
[44] L. Baraldi, Z. Zeng, C. Zhang, A. Nayak, H. Zhu, F. Liu, Q. Zhang, P. Wang, S. Liu, Z. Hu *et al.*, "The safety challenge of world models for embodied ai agents: A review," *arXiv preprint arXiv:2510.05865*, 2025.
[45] P.-F. Zhang, Y. Cheng, X. Sun, S. Wang, L. Zhu, and H. T. Shen, "A step toward world models: A survey on robotic manipulation," *arXiv preprint arXiv:2511.02097*, 2025.
[46] X. Chen, L. Chang, X. Yu, Y. Huang, and X. Tu, "A survey on world models grounded in acoustic physical information," *arXiv preprint arXiv:2506.13833*, 2025.
[47] Z. Zhu, X. Wang, W. Zhao, C. Min, N. Deng, M. Dou, Y. Wang, B. Shi, K. Wang, C. Zhang *et al.*, "Is sora a world simulator? a comprehensive survey on general world models and beyond," *arXiv preprint arXiv:2405.03520*, 2024.
[48] K. Kotar, W. Lee, R. Venkatesh, H. Chen, D. Bear, J. Watrous, S. Kim, K. L. Aw, L. N. Chen, S. Stojanov *et al.*, "World modeling with probabilistic structure integration," *arXiv preprint arXiv:2509.09737*, 2025.
[49] J. Bai, Y. Lei, H. Wu, Y. Zhu, S. Li, Y. Xin, X. Li, M. Tao, A. Grover, and M.-H. Yang, "From masks to worlds: A hitchhiker's guide to world models," *arXiv preprint arXiv:2510.20668*, 2025.
[50] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.
[51] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "*pi*_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.
[52] J. Feng, Y. Zhang, C. Zhang, Y. Lu, S. Liu, and M. Wang, "Web world models," *arXiv preprint arXiv:2512.23676*, 2025.
[53] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin *et al.*, "Latent action pretraining from videos," *arXiv preprint arXiv:2410.11758*, 2024.
[54] S. Yin, J. Wu, S. Huang, X. Su, X. He, J. Hao, and M. Long, "Trajectory world models for heterogeneous environments," *arXiv preprint arXiv:2502.01366*, 2025.
[55] NVIDIA Corporation, "World models — nvidia glossary," https://www.nvidia.com/en-us/glossary/world-models/, 2025, accessed: January 8, 2026.
[56] W. Menapace, S. Lathuiliere, S. Tulyakov, A. Siarohin, and E. Ricci, "Playable video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 061–10 070.
[57] T. Wu, S. Yang, R. Po, Y. Xu, Z. Liu, D. Lin, and G. Wetzstein, "Video world models with long-term spatial memory," *arXiv preprint arXiv:2506.05284*, 2025.
[58] J. Zhang, Y. Chen, Y. Xu, Z. Huang, Y. Zhou, Y.-J. Yuan, X. Cai, G. Huang, X. Quan, H. Xu *et al.*, "4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration," *arXiv preprint arXiv:2506.22242*, 2025.
[59] A. Team, H. Zhu, Y. Wang, J. Zhou, W. Chang, Y. Zhou, Z. Li, J. Chen, C. Shen, J. Pang *et al.*, "Aether: Geometric-aware unified world modeling," *arXiv preprint arXiv:2503.18945*, 2025.
[60] J. Guo, Y. Ding, X. Chen, S. Chen, B. Li, Y. Zou, X. Lyu, F. Tan, X. Qi, Z. Li *et al.*, "Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation," *arXiv preprint arXiv:2503.15208*, 2025.
[61] A. Liang, Y. Liu, Y. Yang, D. Lu, L. Li, L. Kong, H. Zhao, and W. T. Ooi, "Lidarcrafter: Dynamic 4d world modeling from lidar sequences," *arXiv preprint arXiv:2508.03692*, 2025.
[62] S. Kim, K. L. Aw, K. Kotar, C. Eyzaguirre, W. Lee, Y. Liu, J. Watrous, S. Stojanov, J. C. Niebles, J. Wu *et al.*, "Taming generative video models for zero-shot optical flow extraction," *arXiv preprint arXiv:2507.09082*, 2025.
[63] Y. Wang, M. Verghese, and J. Schneider, "Latent policy steering with embodiment-agnostic pretrained world models," *arXiv preprint arXiv:2507.13340*, 2025.
[64] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "Occllama: An occupancy-language-action generative world model for autonomous

driving," *arXiv preprint arXiv:2409.03272*, 2024.
[65] L. Wang, W. Zheng, Y. Ren, H. Jiang, Z. Cui, H. Yu, and J. Lu, "Occsora: 4d occupancy generation models as world simulators for autonomous driving," *arXiv preprint arXiv:2405.20337*, 2024.
[66] Y. Zhang, S. Gong, K. Xiong, X. Ye, X. Li, X. Tan, F. Wang, J. Huang, H. Wu, and H. Wang, "Bevworld: A multimodal world simulator for autonomous driving via scene-level bev latents," *arXiv preprint arXiv:2407.05679*, 2024.
[67] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, "Dome: Taming diffusion model into high-fidelity controllable occupancy world model," *arXiv preprint arXiv:2410.10429*, 2024.
[68] B. Jin, S. Gu, X. Hu, Y. Zheng, X. Guo, Q. Zhang, X. Long, and W. Yin, "Occtens: 3d occupancy world model via temporal next-scale prediction," *arXiv preprint arXiv:2509.03887*, 2025.
[69] R. Liu, L. Kong, D. Li, and H. Zhao, "Occvla: Vision-language-action model with implicit 3d occupancy supervision," *arXiv preprint arXiv:2509.05578*, 2025.
[70] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.
[71] Z. Chen, J. Huo, Y. Chen, and Y. Gao, "Robohorizon: An llm-assisted multi-view world model for long-horizon robotic manipulation," *arXiv preprint arXiv:2501.06605*, 2025.
[72] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.
[73] T. Lu, T. Shu, A. Yuille, D. Khashabi, and J. Chen, "Generative world explorer," *arXiv preprint arXiv:2411.11844*, 2024.
[74] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 995–19 012.
[75] K. Xing, X. Jin, L. Li, Y. Yin, H. Liang, G. Luo, C. Fang, J. Wang, K. N. Plataniotis, Y. Zhao *et al.*, "Stereoworld: Geometry-aware monocular-to-stereo video generation," *arXiv preprint arXiv:2512.09363*, 2025.
[76] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.
[77] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
[78] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.
[79] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.
[80] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps *et al.*, "Genie: Generative interactive environments," in *Forty-first International Conference on Machine Learning*, 2024.
[81] N. Hansen, J. SV, V. Sobal, Y. LeCun, X. Wang, and H. Su, "Hierarchical world models as visual whole-body humanoid controllers," *arXiv preprint arXiv:2405.18418*, 2024.
[82] X. Chi, K. Ge, J. Liu, S. Zhou, P. Jia, Z. He, Y. Liu, T. Li, L. Han, S. Han *et al.*, "Mind: Unified visual imagination and control via hierarchical world models," *arXiv preprint arXiv:2506.18897*, 2025.
[83] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado, "Gaia-2: A controllable multi-view generative world model for autonomous driving," *arXiv preprint arXiv:2503.20523*, 2025.
[84] R. Li, P. Torr, A. Vedaldi, and T. Jakab, "Vmem: Consistent interactive video scene generation with surfel-indexed view memory," *arXiv preprint arXiv:2506.18903*, 2025.
[85] T. Hu, H. Peng, X. Liu, and Y. Ma, "Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh," *arXiv preprint arXiv:2506.05554*, 2025.
[86] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model," in *European Conference on Computer Vision*. Springer, 2024, pp. 469–485.
[87] R. Chen, Z. Wu, Y. Liu, Y. Guo, J. Ni, H. Xia, and S. Xia, "Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving," *arXiv preprint arXiv:2412.04842*, 2024.
[88] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian, "Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling," *arXiv preprint arXiv:2507.07982*, 2025.
[89] Y. Chen, Y. Ge, W. Tang, Y. Li, Y. Ge, M. Ding, Y. Shan, and X. Liu, "Moto: Latent motion token as the bridging language for learning robot manipulation from videos," *arXiv preprint arXiv:2412.04445*, 2024.
[90] Y. Yuan, X. Wang, T. Wickremasinghe, Z. Nadir, B. Ma, and S. H. Chan, "Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics," *arXiv preprint arXiv:2509.21309*, 2025.
[91] M.-Q. Le, Y. Zhu, V. Kalogeiton, and D. Samaras, "What about gravity in video generation? post-training newton's laws with verifiable rewards," *arXiv preprint arXiv:2512.00425*, 2025.
[92] H. Bansal, C. Peng, Y. Bitton, R. Goldenberg, A. Grover, and K.-W. Chang, "Videophy-2: A challenging action-centric physical commonsense evaluation in video generation," *arXiv preprint arXiv:2503.06800*, 2025.
[93] OpenAI, "Sora 2: Advancing video generation models," https://openai.com/index/sora-2/, 2025, accessed: October 7, 2025.
[94] DeepMind, "Veo – deepmind's video evolution model," https://deepmind.google/models/veo/, 2025, accessed: October 7, 2025.
[95] K. Team, J. Chen, Y. Ding, Z. Fang, K. Gai, K. He, X. He, J. Hua, M. Lao, X. Li *et al.*, "Kling-motioncontrol technical report," *arXiv preprint arXiv:2603.03160*, 2026.
[96] J. Pearl, *Causality*. Cambridge university press, 2009.
[97] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.
[98] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
[99] J. Richens and T. Everitt, "Robust agents learn causal world models," *arXiv preprint arXiv:2402.10877*, 2024.
[100] X. Zhou, J. Liu, A. Yerukola, H. Kim, and M. Sap, "Social world models," *arXiv preprint arXiv:2509.00559*, 2025.
[101] J. Li, Q. Wang, Y. Wang, X. Jin, Y. Li, W. Zeng, and X. Yang, "Open-world reinforcement learning over long short-term imagination," *arXiv preprint arXiv:2410.03618*, 2024.
[102] H. Ma, J. Wu, N. Feng, C. Xiao, D. Li, J. Hao, J. Wang, and M. Long, "Harmonydream: Task harmonization inside world models," *arXiv preprint arXiv:2310.00344*, 2023.
[103] W. Sun, F. Wei, J. Zhao, X. Chen, Z. Chen, H. Zhang, J. Zhang, and Y. Lu, "From virtual games to real-world play," *arXiv preprint arXiv:2506.18901*, 2025.
[104] X. Mao, S. Lin, Z. Li, C. Li, W. Peng, T. He, J. Pang, M. Chi, Y. Qiao, and K. Zhang, "Yume: An interactive world generation model," *arXiv preprint arXiv:2507.17744*, 2025.
[105] J. Yu, J. Bai, Y. Qin, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu, "Context as memory: Scene-consistent interactive long video generation with memory retrieval," *arXiv preprint arXiv:2506.03141*, 2025.
[106] G. Lu, B. Jia, P. Li, Y. Chen, Z. Wang, Y. Tang, and S. Huang, "Gwm: Towards scalable gaussian world models for robotic manipulation," *arXiv preprint arXiv:2508.17600*, 2025.
[107] S. Zhou, T. Zhou, Y. Yang, G. Long, D. Ye, J. Jiang, and C. Zhang, "Wall-e 2.0: World alignment by neurosymbolic learning improves world model-based llm agents," *arXiv preprint arXiv:2504.15785*, 2025.
[108] J. Collu, R. Majellaro, A. Plaat, and T. M. Moerland, "Slot structured world models," *arXiv preprint arXiv:2402.03326*, 2024.
[109] Y. Hu, P. Guo, Yanjiang xand Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024.
[110] K. Zhang, Z. Tang, X. Hu, X. Pan, X. Guo, Y. Liu, J. Huang, L. Yuan, Q. Zhang, X.-X. Long *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," *arXiv preprint arXiv:2506.24113*, 2025.
[111] D. Chen, T. Moutakanni, W. Chung, Y. Bang, Z. Ji, A. Bolourchi, and P. Fung, "Planning with reasoning using vision language world model," *arXiv preprint arXiv:2509.02722*, 2025.
[112] Y. Liao, P. Zhou, S. Huang, D. Yang, S. Chen, Y. Jiang, Y. Hu, J. Cai, S. Liu, J. Luo *et al.*, "Genie envisioner: A unified world foundation platform for robotic manipulation," *arXiv preprint arXiv:2508.05635*, 2025.
[113] F. Zhu, H. Wu, S. Guo, Y. Liu, C. Cheang, and T. Kong, "Irasim: Learning interactive real-robot action simulators," *arXiv preprint arXiv:2406.14540*, 2024.

[114] E. Decart, Q. McIntyre, S. Campbell, X. Chen, and R. Wachen, "Oasis: A universe in a transformer," *URL: https://oasis-model.github.io*, vol. 2, no. 3, p. 6, 2024.

[115] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," *arXiv preprint arXiv:2310.16828*, 2023.

[116] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

[117] DeepMind, "Genie 3: A new frontier for world models," https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models, 2025, accessed: October 5, 2025.

[118] S. Gao, S. Zhou, Y. Du, J. Zhang, and C. Gan, "Adaworld: Learning adaptable world models with latent actions," *arXiv preprint arXiv:2503.18938*, 2025.

[119] J. Lin, Y. Du, O. Watkins, D. Hafner, P. Abbeel, D. Klein, and A. Dragan, "Learning to model the world with language," *arXiv preprint arXiv:2308.01399*, 2023.

[120] Y. Jiang, S. Huang, S. Xue, Y. Zhao, J. Cen, S. Leng, K. Li, J. Guo, K. Wang, M. Chen *et al.*, "Rynnvla-001: Using human demonstrations to improve robot manipulation," *arXiv preprint arXiv:2509.15212*, 2025.

[121] Q. Lv, W. Kong, H. Li, J. Zeng, Z. Qiu, D. Qu, H. Song, Q. Chen, X. Deng, and J. Pang, "F1: A vision-language-action model bridging understanding and generation to actions," *arXiv preprint arXiv:2509.06951*, 2025.

[122] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, H. Wang, Z. Zhang *et al.*, "Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025.

[123] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Univla: Learning to act anywhere with task-centric latent actions," *arXiv preprint arXiv:2505.06111*, 2025.

[124] J. Lyu, Z. Li, X. Shi, C. Xu, Y. Wang, and H. Wang, "Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation," *arXiv preprint arXiv:2503.16806*, 2025.

[125] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang *et al.*, "Worldvla: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[126] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[127] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *European conference on computer vision*. Springer, 2024, pp. 55–72.

[128] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "Worlddreamer: Towards general world models for video generation via predicting masked tokens," *arXiv preprint arXiv:2401.09985*, 2024.

[129] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with blockwise ringattention," *arXiv preprint arXiv:2402.08268*, 2024.

[130] A. Kanervisto, D. Bignell, L. Y. Wen, M. Grayson, R. Georgescu, S. Valcarcel Macua, S. Z. Tan, T. Rashid, T. Pearce, Y. Cao *et al.*, "World and human action models towards gameplay ideation," *Nature*, vol. 638, no. 8051, pp. 656–663, 2025.

[131] J. Wu, S. Yin, N. Feng, X. He, D. Li, J. Hao, and M. Long, "ivideogpt: Interactive videogpts are scalable world models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 68 082–68 119, 2024.

[132] W. Zhang, J. Guo, T. He, L. Zhao, L. Xu, and J. Bian, "Video in-context learning: Autoregressive transformers are zero-shot video imitators," *arXiv preprint arXiv:2407.07356*, 2024.

[133] L. Xiao, J.-J. Liu, S. Yang, X. Li, X. Ye, W. Yang, and J. Wang, "Learning multiple probabilistic decisions from latent world model in autonomous driving," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1279–1285.

[134] Z. Yan, W. Dong, Y. Shao, Y. Lu, H. Liu, J. Liu, H. Wang, Z. Wang, Y. Wang, F. Remondino *et al.*, "Renderworld: World model with self-supervised 3d label," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 6063–6070.

[135] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[136] Z. Zhang, R. Chen, J. Ye, Y. Sun, P. Wang, J. Pang, K. Li, T. Liu, H. Lin, Y. Yu *et al.*, "Whale: Towards generalizable and scalable world models for embodied decision-making," *arXiv preprint arXiv:2411.05619*, 2024.

[137] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing *et al.*, "Driveworld: 4d pre-trained scene understanding via world models for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 522–15 533.

[138] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025.

[139] H. He, Y. Zhang, L. Lin, Z. Xu, and L. Pan, "Pre-trained video generative models as world simulators," *arXiv preprint arXiv:2502.07825*, 2025.

[140] S. Koju, S. Bastola, P. Shrestha, S. Amgain, Y. R. Shrestha, R. P. Poudel, and B. Bhattarai, "Surgical vision world model," *arXiv preprint arXiv:2503.02904*, 2025.

[141] J. Guo, Y. Ye, T. He, H. Wu, Y. Jiang, T. Pearce, and J. Bian, "Mineworld: a real-time and open-source interactive world model on minecraft," *arXiv preprint arXiv:2504.08388*, 2025.

[142] J. Wu, S. Yin, N. Feng, and M. Long, "Rlvr-world: Training world models with reinforcement learning," *arXiv preprint arXiv:2505.13934*, 2025.

[143] Y. Shang, X. Zhang, Y. Tang, L. Jin, C. Gao, W. Wu, and Y. Li, "Roboscape: Physics-informed embodied world model," *arXiv preprint arXiv:2506.23135*, 2025.

[144] Z. Yang, X. Song, X. Xu, Y. Shi, G. Wang, M. K. Kalra, and P. Yan, "Xray2xray: World model from chest x-rays with volumetric context," *arXiv preprint arXiv:2506.19055*, 2025.

[145] Z. Liao, P. Wei, R. Zhang, S. Chen, H. Wang, and Z. Ren, "2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting," *arXiv preprint arXiv:2507.09144*, 2025.

[146] C. Zhang, Z. Wu, G. Lu, Y. Tang, and Z. Wang, "imowm: Taming interactive multi-modal world model for robotic manipulation," *arXiv preprint arXiv:2510.09036*, 2025.

[147] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, V. Birodkar, A. Gupta, X. Gu *et al.*, "Language model beats diffusion–tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[148] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 523–11 532.

[149] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[150] X. Zhou, D. Liang, S. Tu, X. Chen, Y. Ding, D. Zhang, F. Tan, H. Zhao, and X. Bai, "Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation," *arXiv preprint arXiv:2501.14729*, 2025.

[151] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[152] W. Zheng, Z. Xia, Y. Huang, S. Zuo, J. Zhou, and J. Lu, "Doe-1: Closed-loop autonomous driving with large world model," *arXiv preprint arXiv:2412.09627*, 2024.

[153] J. Berg, C. Zhu, Y. Bao, I. Durugkar, and A. Gupta, "Semantic world models," *arXiv preprint arXiv:2510.19818*, 2025.

[154] Y. Huang, W. Zheng, Y. Gao, X. Tao, P. Wan, D. Zhang, J. Zhou, and J. Lu, "Owl-1: Omni world model for consistent long video generation," *arXiv preprint arXiv:2412.09600*, 2024.

[155] X. Chi, C.-K. Fan, H. Zhang, X. Qi, R. Zhang, A. Chen, C.-m. Chan, W. Xue, Q. Liu, S. Zhang *et al.*, "Eva: An embodied world model for future video anticipation," *arXiv preprint arXiv:2410.15461*, 2024.

[156] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[157] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[158] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[159] Z. Ge, H. Huang, M. Zhou, J. Li, G. Wang, S. Tang, and Y. Zhuang, "Worldgpt: Empowering llm as multimodal world model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7346–7355.

[160] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint*

arXiv:2302.13971, 2023.

[161] K. Zhang, P. Ren, B. Lin, J. Lin, S. Ma, H. Xu, and X. Liang, "Pivot-r: Primitive-driven waypoint-aware world model for robotic manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 54 105–54 136, 2024.

[162] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[163] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 700–13 710.

[164] D. Liu, S. Zhao, L. Zhuo, W. Lin, Y. Xin, X. Li, Q. Qin, Y. Qiao, H. Li, and P. Gao, "Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining," *arXiv preprint arXiv:2408.02657*, 2024.

[165] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 43 447–43 478, 2023.

[166] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *Science China Information Sciences*, vol. 67, no. 12, p. 220101, 2024.

[167] Z. Li, G. Chen, S. Liu, S. Wang, V. VS, Y. Ji, S. Lan, H. Zhang, Y. Zhao, S. Radhakrishnan *et al.*, "Eagle 2: Building post-training data strategies from scratch for frontier vision-language models," *arXiv preprint arXiv:2501.14818*, 2025.

[168] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[169] H. Chen, B. Wang, J. Guo, T. Zhang, Y. Hou, X. Huang, C. Tie, and L. Shao, "World4omni: A zero-shot framework from image generation world model to robotic manipulation," *arXiv preprint arXiv:2506.23919*, 2025.

[170] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[171] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[172] J. H. Cho, A. Madotto, E. Mavroudi, T. Afouras, T. Nagarajan, M. Maaz, Y. Song, T. Ma, S. Hu, S. Jain *et al.*, "Perceptionlm: Open-access data and models for detailed visual understanding," *arXiv preprint arXiv:2504.13180*, 2025.

[173] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[174] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long *et al.*, "Paligemma 2: A family of versatile vlms for transfer," *arXiv preprint arXiv:2412.03555*, 2024.

[175] Y. Dong, F. Wu, G. Chen, Z.-Q. Cheng, Q. Hu, Y. Zhou, J. Sun, J.-Y. He, Q. Dai, and A. G. Hauptmann, "Unified world models: Memory-augmented planning and foresight for visual navigation," *arXiv preprint arXiv:2510.08713*, 2025.

[176] E. Chern, J. Su, Y. Ma, and P. Liu, "Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation," *arXiv preprint arXiv:2407.06135*, 2024.

[177] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[178] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, "Panacea: Panoramic and controllable video generation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6902–6912.

[179] J. Lu, Z. Huang, Z. Yang, J. Zhang, and L. Zhang, "Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 329–345.

[180] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023.

[181] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 412–10 420.

[182] D. Yang, L. Hu, Y. Tian, Z. Li, C. Kelly, B. Yang, C. Yang, and Y. Zou, "Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs," *arXiv preprint arXiv:2403.07944*, 2024.

[183] E. Ma, L. Zhou, T. Tang, Z. Zhang, D. Han, J. Jiang, K. Zhan, P. Jia, X. Lang, H. Sun *et al.*, "Unleashing generalization of end-to-end autonomous driving with controllable long video generation," *arXiv preprint arXiv:2406.01349*, 2024.

[184] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, "Panacea: Panoramic and controllable video generation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6902–6912.

[185] A. Garg and K. Madhava Krishna, "Imagine-2-drive: High-fidelity world modeling in carla for autonomous vehicles," *arXiv e-prints*, pp. arXiv–2411, 2024.

[186] H. A. Alhaija, J. Alvarez, M. Bala, T. Cai, T. Cao, L. Cha, J. Chen, M. Chen, F. Ferroni, S. Fidler *et al.*, "Cosmos-transfer1: Conditional world generation with adaptive multimodal control," *arXiv preprint arXiv:2503.14492*, 2025.

[187] A. Ali, J. Bai, M. Bala, Y. Balaji, A. Blakeman, T. Cai, J. Cao, T. Cao, E. Cha, Y.-W. Chao *et al.*, "World simulation with video foundation models for physical ai," *arXiv preprint arXiv:2511.00062*, 2025.

[188] J. Ni, Y. Guo, Y. Liu, R. Chen, L. Lu, and Z. Wu, "Maskgwm: A generalizable driving world model with video mask reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 381–22 391.

[189] J.-H. Lee, B.-J. Lin, W.-F. Sun, and C.-Y. Lee, "Edeline: Enhancing memory in diffusion-based world models via linear-time sequence modeling," *arXiv preprint arXiv:2502.00466*, 2025.

[190] D. Luo, B. Tang, K. Li, G. Papoudakis, J. Song, S. Gong, J. Hao, J. Wang, and K. Shao, "Vimo: A generative visual gui world model for app agents," *arXiv preprint arXiv:2504.13936*, 2025.

[191] A. Chen, W. Zheng, Y. Wang, X. Zhang, K. Zhan, P. Jia, K. Keutzer, and S. Zhang, "Geodrive: 3d geometry-informed driving world model with precise action control," *arXiv preprint arXiv:2505.22421*, 2025.

[192] N. Savov, N. Kazemi, D. Zhang, D. P. Paudel, X. Wang, and L. Van Gool, "Statespacediffuser: Bringing long context to diffusion world models," *arXiv preprint arXiv:2505.22246*, 2025.

[193] B. Zhao, R. Tang, M. Jia, Z. Wang, F. Man, X. Zhang, Y. Shang, W. Zhang, C. Gao, W. Wu *et al.*, "Airscape: An aerial generative world model with motion controllability," *arXiv preprint arXiv:2507.08885*, 2025.

[194] Y. Yin, H. Guo, F. Liu, M. Wang, H. Liang, E. Li, Y. Wang, X. Jin, Y. Zhao, and Y. Wei, "Panoworld-x: Generating explorable panoramic worlds via sphere-aware video diffusion," *arXiv preprint arXiv:2509.24997*, 2025.

[195] J. Mao, B. Li, B. Ivanovic, Y. Chen, Y. Wang, Y. You, C. Xiao, D. Xu, M. Pavone, and Y. Wang, "Dreamdrive: Generative 4d scene modeling from street view images," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 367–374.

[196] H. Zhen, Q. Sun, H. Zhang, J. Li, S. Zhou, Y. Du, and C. Gan, "Tesseract: learning 4d embodied world models," *arXiv preprint arXiv:2504.20995*, 2025.

[197] J. Guo, X. Ma, Y. Wang, M. Yang, H. Liu, and Q. Li, "Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation," *arXiv preprint arXiv:2505.10075*, 2025.

[198] L. Liu, X. Wang, G. Zhao, K. Li, W. Qin, J. Qiu, Z. Zhu, G. Huang, and Z. Su, "Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer," *arXiv preprint arXiv:2505.23171*, 2025.

[199] S. Zhou, Y. Du, Y. Yang, L. Han, P. Chen, D.-Y. Yeung, and C. Gan, "Learning 3d persistent embodied world models," *arXiv preprint arXiv:2505.05495*, 2025.

[200] Y. Yang, Z.-Y. Wang, Q. Liu, S. Sun, K. Wang, R. Chellappa, Z. Zhou, A. Yuille, L. Zhu, Y.-D. Zhang *et al.*, "Medical world model: Generative simulation of tumor evolution for treatment planning," *arXiv preprint arXiv:2506.02327*, 2025.

[201] X. Guo, Z. Wu, K. Xiong, Z. Xu, L. Zhou, G. Xu, S. Xu, H. Sun, B. Wang, G. Chen *et al.*, "Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency," *arXiv preprint arXiv:2506.07497*, 2025.

[202] Y. Shi, K. Jiang, Q. Meng, K. Wang, J. Wang, W. Sun, T. Wen, M. Yang, and D. Yang, "Come: Adding scene-centric forecasting control to occupancy world model," *arXiv preprint arXiv:2506.13260*, 2025.

[203] L. Li, Z. Fan, W. Cong, X. Liu, Y. Yin, M. Foutter, P. Pan, C. You,

Y. Wang, Z. Wang *et al.*, "Martian world models: Controllable video synthesis with physically accurate 3d reconstructions," *arXiv preprint arXiv:2507.07978*, 2025.
[204] C. Song, Y. Yang, T. Zhao, R. Li, and C. Zhang, "Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance," *arXiv preprint arXiv:2509.15130*, 2025.
[205] Z. Zhu, Z. Wu, Z. Zhu, L. Zhou, H. Sun, B. Wan, K. Ma, G. Chen, H. Ye, J. Xie *et al.*, "Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving," *arXiv preprint arXiv:2509.23402*, 2025.
[206] Z. Qian, X. Chi, Y. Li, S. Wang, Z. Qin, X. Ju, S. Han, and S. Zhang, "Wristworld: Generating wrist-views via 4d world models for robotic manipulation," *arXiv preprint arXiv:2510.07313*, 2025.
[207] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European conference on computer vision*. Springer, 2024, pp. 55–72.
[208] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo *et al.*, "Generalized predictive model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 662–14 672.
[209] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," *arXiv preprint arXiv:2504.02792*, 2025.
[210] M. Goff, G. Hogan, G. Hotz, A. du Parc Locmaria, K. Raczy, H. Schäfer, A. Shihadeh, W. Zhang, and Y. Yousfi, "Learning to drive from a world model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1964–1973.
[211] K. Ranasinghe, X. Li, C. Mata, J. Park, and M. S. Ryoo, "Pixel motion as universal representation for robot control," *arXiv preprint arXiv:2505.07817*, 2025.
[212] Y. Huang, J. Zhang, S. Zou, X. Liu, R. Hu, and K. Xu, "Ladi-wm: A latent diffusion-based world model for predictive manipulation," *arXiv preprint arXiv:2505.11528*, 2025.
[213] J. Liang, P. Tokmakov, R. Liu, S. Sudhakar, P. Shah, R. Ambrus, and C. Vondrick, "Video generators are robot policies," *arXiv preprint arXiv:2508.00795*, 2025.
[214] B. Zhang and M. Z. Shou, "Ego-centric predictive model conditioned on hand trajectories," *arXiv preprint arXiv:2508.19852*, 2025.
[215] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel, "Learning interactive real-world simulators," *arXiv preprint arXiv:2310.06114*, vol. 1, no. 2, p. 6, 2023.
[216] L. Zhang, Y. Xiong, Z. Yang, S. Casas, R. Hu, and R. Urtasun, "Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion," *arXiv preprint arXiv:2311.01017*, 2023.
[217] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 560–91 596, 2024.
[218] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter, "Diffusion models are real-time game engines," *arXiv preprint arXiv:2408.14837*, 2024.
[219] H. Che, X. He, Q. Liu, C. Jin, and H. Chen, "Gamegen-x: Interactive open-world game video generation," *arXiv preprint arXiv:2411.00769*, 2024.
[220] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 791–15 801.
[221] R. Feng, H. Zhang, Z. Yang, J. Xiao, Z. Shu, Z. Liu, A. Zheng, Y. Huang, Y. Liu, and H. Zhang, "The matrix: Infinite-horizon world generation with real-time moving control," *arXiv preprint arXiv:2412.03568*, 2024.
[222] X. Li, C. Wu, Z. Yang, Z. Xu, D. Liang, Y. Zhang, J. Wan, and J. Wang, "Driverse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment," *arXiv preprint arXiv:2504.18576*, 2025.
[223] T. Chen, X. Hu, Z. Ding, and C. Jin, "Learning world models for interactive video generation," *arXiv preprint arXiv:2505.21996*, 2025.
[224] Y. Zhang, C. Peng, B. Wang, P. Wang, Q. Zhu, F. Kang, B. Jiang, Z. Gao, E. Li, Y. Liu *et al.*, "Matrix-game: Interactive world foundation model," *arXiv preprint arXiv:2506.18701*, 2025.
[225] Z. Ren, Y. Wei, X. Yu, G. Luo, Y. Zhao, B. Kang, J. Feng, and X. Jin, "Videoworld 2: Learning transferable knowledge from real-world videos," *arXiv preprint arXiv:2602.10102*, 2026.
[226] J. Wang, Y. Yao, X. Feng, H. Wu, Y. Wang, Q. Huang, Y. Ma, and X. Zhu, "Stage: A stream-centric generative world model for long-horizon driving-scene simulation," *arXiv preprint arXiv:2506.13138*, 2025.
[227] Y. Shang, L. Jin, Y. Ma, X. Zhang, C. Gao, W. Wu, and Y. Li, "Longscape: Advancing long-horizon embodied world models with context-aware moe," *arXiv preprint arXiv:2509.21790*, 2025.
[228] T. Yan, D. Wu, W. Han, J. Jiang, X. Zhou, K. Zhan, C.-z. Xu, and J. Shen, "Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 531–27 541.
[229] M. Hassan, S. Stapf, A. Rahimi, P. Rezende, Y. Haghighi, D. Brüggemann, I. Katircioglu, L. Zhang, X. Chen, S. Saha *et al.*, "Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 404–22 415.
[230] S. Tan, J. Lambert, H. Jeon, S. Kulshrestha, Y. Bai, J. Luo, D. Anguelov, M. Tan, and C. M. Jiang, "Scenediffuser++: City-scale traffic simulation via a generative world model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1570–1580.
[231] J. Chen, H. Zhu, X. He, Y. Wang, J. Zhou, W. Chang, Y. Zhou, Z. Li, Z. Fu, J. Pang *et al.*, "Deepverse: 4d autoregressive video generation as a world model," *arXiv preprint arXiv:2506.01103*, 2025.
[232] E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. J. Storkey, T. Pearce, and F. Fleuret, "Diffusion for world modeling: Visual details matter in atari," *Advances in Neural Information Processing Systems*, vol. 37, pp. 58 757–58 791, 2024.
[233] X. Chi, P. Jia, C.-K. Fan, X. Ju, W. Mi, K. Zhang, Z. Qin, W. Tian, K. Ge, H. Li *et al.*, "Wow: Towards a world omniscient world model through embodied interaction," *arXiv preprint arXiv:2509.22642*, 2025.
[234] J. Yu, Y. Qin, X. Wang, P. Wan, D. Zhang, and X. Liu, "Gamefactory: Creating new games with generative interactive videos," *arXiv preprint arXiv:2501.08325*, 2025.
[235] S. Kwon, J.-Y. Kim, H. Go, and K. Baek, "Toward stable world models: Measuring and addressing world instability in generative environments," *arXiv preprint arXiv:2503.08122*, 2025.
[236] Z. Xiao, Y. Lan, Y. Zhou, W. Ouyang, S. Yang, Y. Zeng, and X. Pan, "Worldmem: Long-term consistent world simulation with memory," *arXiv preprint arXiv:2504.12369*, 2025.
[237] X. Cheng, T. He, J. Xu, J. Guo, D. He, and J. Bian, "Playing with transformer at 30+ fps via next-frame diffusion," *arXiv preprint arXiv:2506.01380*, 2025.
[238] Skywork AI Matrix-Game Team, "Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory," Technical report, 2026. [Online]. Available: https://github.com/SkyworkAI/Matrix-Game/blob/main/Matrix-Game-3/assets/pdf/report.pdf
[239] D. Hafner, W. Yan, and T. Lillicrap, "Training agents inside of scalable world models," *arXiv preprint arXiv:2509.24527*, 2025.
[240] Y. Cui, H. Chen, H. Deng, X. Huang, X. Li, J. Liu, Y. Liu, Z. Luo, J. Wang, W. Wang *et al.*, "Emu3. 5: Native multimodal models are world learners," *arXiv preprint arXiv:2510.26583*, 2025.
[241] B. Li, Z. Ma, D. Du, B. Peng, Z. Liang, Z. Liu, C. Ma, Y. Jin, H. Zhao, W. Zeng *et al.*, "Omninwm: Omniscient driving navigation world models," *arXiv preprint arXiv:2510.18313*, 2025.
[242] Q. Garrido, M. Assran, N. Ballas, A. Bardes, L. Najman, and Y. LeCun, "Learning and leveraging world models in visual representation learning," *arXiv preprint arXiv:2403.00504*, 2024.
[243] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," *arXiv preprint arXiv:2411.04983*, 2024.
[244] E. Karypidis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis, "Dino-foresight: Looking into the future with dino," *arXiv preprint arXiv:2412.11673*, 2024.
[245] H. Zhu, Z. Dong, K. Topollai, and A. Choromanska, "Ad-l-jepa: Self-supervised spatial world models with joint embedding predictive architecture for autonomous driving with lidar data," *arXiv preprint arXiv:2501.04969*, 2025.
[246] Y. Yue, Y. Wang, H. Jiang, P. Liu, S. Song, and G. Huang, "Echoworld: Learning motion-aware world models for echocardiography probe guidance," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 993–26 003.
[247] R. G. Goswami, P. Krishnamurthy, Y. LeCun, and F. Khorrami, "Osvi-wm: One-shot visual imitation for unseen tasks using world-model-guided trajectory generation," *arXiv preprint arXiv:2505.20425*, 2025.
[248] H. Ghaemi, E. Muller, and S. Bakhtiari, "seq-jepa: Autoregressive predictive learning of invariant-equivariant world models," *arXiv preprint arXiv:2505.03176*, 2025.
[249] J. Chun, Y. Jeong, and T. Kim, "Sparse imagination for efficient visual world model planning," *arXiv preprint arXiv:2506.01392*, 2025.
[250] Y. Team, "Yan: Foundational interactive video generation," *arXiv preprint arXiv:2508.08601*, 2025.

[251] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[252] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[253] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[254] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *Conference on robot learning*. PMLR, 2023, pp. 2226–2240.

[255] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1332–1344.

[256] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are sample-efficient world models," *arXiv preprint arXiv:2209.00588*, 2022.

[257] J. Robine, M. Höftmann, T. Uelwer, and S. Harmeling, "Transformer-based world models are happy with 100k interactions," *arXiv preprint arXiv:2303.07109*, 2023.

[258] Q. Wang, J. Yang, Y. Wang, X. Jin, W. Zeng, and X. Yang, "Making offline rl online: Collaborative world models for offline visual reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 97 203–97 230, 2024.

[259] W. Huang, J. Ji, C. Xia, B. Zhang, and Y. Yang, "Safedreamer: Safe reinforcement learning with world models," *arXiv preprint arXiv:2307.07176*, 2023.

[260] R. Mendonca, S. Bahl, and D. Pathak, "Structured world models from human videos," *arXiv preprint arXiv:2308.10901*, 2023.

[261] P. Lancaster, N. Hansen, A. Rajeswaran, and V. Kumar, "Modem-v2: Visuo-motor world models for real-world robot manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7530–7537.

[262] W. Zhang, G. Wang, J. Sun, Y. Yuan, and G. Huang, "Storm: Efficient stochastic transformer based world models for reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 147–27 166, 2023.

[263] D. Bogdoll, Y. Yang, T. Joseph, M. Yazgan, and J. M. Zollner, "Muvo: A multimodal generative world model for autonomous driving with geometric representations," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025, pp. 2243–2250.

[264] Q. Li, X. Jia, S. Wang, and J. Yan, "Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2)," in *European Conference on Computer Vision*. Springer, 2024, pp. 142–158.

[265] L. Cohen, K. Wang, B. Kang, and S. Mannor, "Improving token-based world models with parallel observation prediction," *arXiv preprint arXiv:2402.05643*, 2024.

[266] M. R. Samsami, A. Zholus, J. Rajendran, and S. Chandar, "Mastering memory tasks with world models," *arXiv preprint arXiv:2403.04253*, 2024.

[267] P. Mazzaglia, T. Verbelen, B. Dhoedt, A. Courville, and S. Rajeswar, "Genrl: Multimodal-foundation world models for generalization in embodied agents," *Advances in neural information processing systems*, vol. 37, pp. 27 529–27 555, 2024.

[268] H. Wang, X. Ye, F. Tao, C. Pan, A. Mallik, B. Yaman, L. Ren, and J. Zhang, "Adawm: Adaptive world model based planning for autonomous driving," *arXiv preprint arXiv:2501.13072*, 2025.

[269] M. Krinner, E. Aljalbout, A. Romero, and D. Scaramuzza, "Accelerating model-based reinforcement learning with state-space world models," *arXiv preprint arXiv:2502.20168*, 2025.

[270] L. Wang, R. Shelim, W. Saad, and N. Ramakrishnan, "Dmwm: Dual-mind world model with long-term imagination," *arXiv preprint arXiv:2502.07591*, 2025.

[271] L. Cohen, K. Wang, B. Kang, U. Gadot, and S. Mannor, "Uncovering untapped potential in sample-efficient world model agents," *arXiv preprint arXiv:2502.11537*, 2025.

[272] Y. Li, Y. Wang, Y. Liu, J. He, L. Fan, and Z. Zhang, "End-to-end driving with online trajectory evaluation via bev world model," *arXiv preprint arXiv:2504.01941*, 2025.

[273] J. Lanier, K. Kim, A. Karamzade, Y. Liu, A. Sinha, K. He, D. Corsi, and R. Fox, "Adapting world models with latent-state dynamics residuals," *arXiv preprint arXiv:2504.02252*, 2025.

[274] Z. Yang, X. Jia, Q. Li, X. Yang, M. Yao, and J. Yan, "Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2)," *arXiv preprint arXiv:2505.16394*, 2025.

[275] X. Yao, J. Gao, and C. Xu, "Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments," *arXiv preprint arXiv:2506.23468*, 2025.

[276] H. Lin, B. Li, and K. W. S. Au, "Visuomotor grasping with world models for surgical robots," *arXiv preprint arXiv:2508.11200*, 2025.

[277] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[278] Z. Lin, Y.-F. Wu, S. Peri, B. Fu, J. Jiang, and S. Ahn, "Improving generative imagination in object-centric world models," in *International conference on machine learning*. PMLR, 2020, pp. 6140–6149.

[279] L. Zhao, L. Kong, R. Walters, and L. L. Wong, "Toward compositional generalization in object-oriented world modeling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 841–26 864.

[280] Z. Wu, N. Dvornik, K. Greff, T. Kipf, and A. Garg, "Slotformer: Unsupervised visual dynamics simulation with object-centric models," *arXiv preprint arXiv:2210.05861*, 2022.

[281] S. Ferraro, P. Mazzaglia, T. Verbelen, and B. Dhoedt, "Focus: object-centric world models for robotic manipulation," *Frontiers in Neurorobotics*, vol. 19, p. 1585386, 2025.

[282] A. Sehgal, A. Grayeli, J. J. Sun, and S. Chaudhuri, "Neurosymbolic grounding for compositional world models," *arXiv preprint arXiv:2310.12690*, 2023.

[283] S. Zhou, Y. Du, J. Chen, Y. Li, D.-Y. Yeung, and C. Gan, "Robodreamer: Learning compositional world models for robot imagination," *arXiv preprint arXiv:2404.12377*, 2024.

[284] J. Jiang, F. Deng, G. Singh, M. Lee, and S. Ahn, "Slot state space models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 11 602–11 633, 2024.

[285] A. GX-Chen, K. Marino, and R. Fergus, "Efficient exploration and discriminative world model learning with an object-centric abstraction," *arXiv preprint arXiv:2408.11816*, 2024.

[286] J. Baek, Y.-F. Wu, G. Singh, and S. Ahn, "Dreamweaver: Learning compositional world models from pixels," *arXiv preprint arXiv:2501.14174*, 2025.

[287] W. Zhang, A. Jelley, T. McInroe, and A. Storkey, "Objects matter: object-centric world models improve reinforcement learning in visually complex environments," *arXiv preprint arXiv:2501.16443*, 2025.

[288] Q. Wang, Z. Zhang, B. Xie, X. Jin, Y. Wang, S. Wang, L. Zheng, X. Yang, and W. Zeng, "Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning," *arXiv preprint arXiv:2503.08751*, 2025.

[289] Y. Jeong, J. Chun, S. Cha, and T. Kim, "Object-centric world model for language-guided manipulation," *arXiv preprint arXiv:2503.06170*, 2025.

[290] J. Li, H. Wan, N. Lin, Y.-L. Zhan, R. Chengze, H. Wang, Y. Zhang, H. Liu, Z. Wang, F. Yu *et al.*, "Slotpi: Physics-informed object-centric reasoning models," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 1376–1387.

[291] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *Advances in neural information processing systems*, vol. 33, pp. 11 525–11 538, 2020.

[292] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, "Conditional object-centric learning from video," *arXiv preprint arXiv:2111.12594*, 2021.

[293] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[294] T. Feng, Y. Wu, G. Lin, and J. You, "Graph world model," *arXiv preprint arXiv:2507.10539*, 2025.

[295] Y. Yang, Z. Zhang, X. Zhang, Y. Zeng, H. Li, and W. Zuo, "Physworld: From real videos to world models of deformable objects via physics-aware demonstration synthesis," *arXiv preprint arXiv:2510.21447*, 2025.

[296] M. Q. Ali, A. Sridhar, S. Matiana, A. Wong, and M. Al-Sharman, "Humanoid world models: Open world foundation models for humanoid robotics," *arXiv preprint arXiv:2506.01182*, 2025.

[297] T. Liu, S. Zhao, and N. Rhinehart, "Towards foundational lidar world models with efficient latent flow matching," *arXiv preprint arXiv:2506.23434*, 2025.

[298] A. Mousakhan, S. Mittal, S. Galesso, K. Farid, and T. Brox, "Orbis:

Overcoming challenges of long-horizon prediction in driving world models," *arXiv preprint arXiv:2507.13162*, 2025.
[299] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
[300] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[301] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.
[302] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
[303] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
[304] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 144–20 154.
[305] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint arXiv:2306.09344*, 2023.
[306] J. Sturm, W. Burgard, and D. Cremers, "Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark," in *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, vol. 13, 2012, p. 6.
[307] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim *et al.*, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *European Conference on Computer Vision*. Springer, 2024, pp. 445–465.
[308] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu, "Worldscore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.
[309] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian, "Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling," *arXiv preprint arXiv:2507.07982*, 2025.
[310] Y. Hong, B. Liu, M. Wu, Y. Zhai, K.-W. Chang, L. Li, K. Lin, C.-C. Lin, J. Wang, Z. Yang *et al.*, "Slowfast-vgen: Slow-fast learning for action-driven long video generation," *arXiv preprint arXiv:2410.23277*, 2024.
[311] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.
[312] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone *et al.*, "Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking," *Advances in Neural Information Processing Systems*, vol. 37, pp. 28 706–28 719, 2024.
[313] N. Hansen, X. Wang, and H. Su, "Temporal difference learning for model predictive control," *arXiv preprint arXiv:2203.04955*, 2022.
[314] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.
[315] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang *et al.*, "Drivedreamer4d: World models are effective data machines for 4d driving scene representation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 015–12 026.
[316] T. Khurana, P. Hu, D. Held, and D. Ramanan, "Point cloud forecasting as a proxy for 4d occupancy forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1116–1124.
[317] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.
[318] D. Li, Y. Fang, Y. Chen, S. Yang, S. Cao, J. Wong, M. Luo, X. Wang, H. Yin, J. E. Gonzalez *et al.*, "Worldmodelbench: Judging video generation models as world models," *arXiv preprint arXiv:2502.20694*, 2025.
[319] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.
[320] Y. Dai, F. Jiang, C. Wang, M. Xu, and Y. Qi, "Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction," *arXiv preprint arXiv:2509.21657*, 2025.
[321] D. Chen, W. Chung, Y. Bang, Z. Ji, and P. Fung, "Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning," *arXiv preprint arXiv:2506.04363*, 2025.
[322] D. Chen, M. Shukor, T. Moutakanni, W. Chung, J. Yu, T. Kasarla, A. Bolourchi, Y. LeCun, and P. Fung, "Vl-jepa: Joint embedding predictive architecture for vision-language," *arXiv preprint arXiv:2512.10942*, 2025.
[323] Z. Li, C. Li, X. Mao, S. Lin, M. Li, S. Zhao, Z. Xu, X. Li, Y. Feng, J. Sun *et al.*, "Sekai: A video dataset towards world exploration," *arXiv preprint arXiv:2506.15675*, 2025.
[324] X. Mao, Z. Li, C. Li, X. Xu, K. Ying, T. He, J. Pang, Y. Qiao, and K. Zhang, "Yume-1.5: A text-controlled interactive world generation model," *arXiv preprint arXiv:2512.22096*, 2025.
[325] Y. Zhu, J. Feng, W. Zheng, Y. Gao, X. Tao, P. Wan, J. Zhou, and J. Lu, "Astra: General interactive world model with autoregressive denoising," *arXiv preprint arXiv:2512.08931*, 2025.
[326] Y. Zhou, Y. Wang, J. Zhou, W. Chang, H. Guo, Z. Li, K. Ma, X. Li, Y. Wang, H. Zhu *et al.*, "Omniworld: A multi-domain and multi-modal dataset for 4d world modeling," *arXiv preprint arXiv:2509.12201*, 2025.
[327] J. Chen, T. He, Z. Fu, P. Wan, K. Gai, and W. Ye, "Vino: A unified visual generator with interleaved omnimodal context," *arXiv preprint arXiv:2601.02358*, 2026.
[328] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[329] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.
[330] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[331] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1207–1216.
[332] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640.
[333] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.
[334] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 383–19 400.
[335] X. Wang, K. Zhao, F. Liu, J. Wang, G. Zhao, X. Bao, Z. Zhu, Y. Zhang, and X. Wang, "Egovid-5m: A large-scale video-action dataset for egocentric video generation," *arXiv preprint arXiv:2411.08380*, 2024.
[336] H. Li, M. Xu, Y. Zhan, S. Mu, J. Li, K. Cheng, Y. Chen, T. Chen, M. Ye, J. Wang *et al.*, "Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 7752–7762.
[337] E. Özsoy, C. Pellegrini, T. Czempiel, F. Tristram, K. Yuan, D. Bani-Harouni, U. Eck, B. Busam, M. Keicher, and N. Navab, "Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 378–19 389.
[338] F. Baradel, N. Neverova, J. Mille, G. Mori, and C. Wolf, "Cophy: Counterfactual learning of physical dynamics," *arXiv preprint arXiv:1909.12000*, 2019.
[339] Z. Chen, K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum, and C. Gan, "Comphy: Compositional physical reasoning of objects and events from videos," *arXiv preprint arXiv:2205.01089*, 2022.
[340] H.-Y. Tung, M. Ding, Z. Chen, D. Bear, C. Gan, J. Tenenbaum, D. Yamins, J. Fan, and K. Smith, "Physion++: Evaluating physical scene understanding that requires online inference of different physical prop-

erties," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 048–67 068, 2023.
[341] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos, "Do generative video models understand physical principles?" *arXiv preprint arXiv:2501.09038*, 2025.
[342] F. O'Mahony, R. Cipolla, and A. Tewari, "Vdaworld: World modelling via vlm-directed abstraction and simulation," *arXiv preprint arXiv:2512.11061*, 2025.
[343] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, L. Gu, Y. Zhang, J. He, W.-S. Zheng *et al.*, "Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness," *arXiv preprint arXiv:2503.21755*, 2025.
[344] C. Wang, C. Chen, Y. Huang, Z. Dou, Y. Liu, J. Gu, and L. Liu, "Physctrl: Generative physics for controllable and physics-grounded video generation," *arXiv preprint arXiv:2509.20358*, 2025.
[345] Z. Wang, X. Wei, B. Li, Z. Guo, J. Zhang, H. Wei, K. Wang, and L. Zhang, "Videoverse: How far is your t2v generator from a world model?" *arXiv preprint arXiv:2510.08398*, 2025.
[346] F. Zhou, J. Huang, J. Li, D. Ramanan, and H. Shi, "Paibench: A comprehensive benchmark for physical ai," *arXiv preprint arXiv:2512.01989*, 2025.
[347] A. Dasgupta, J. Duan, M. H. Ang Jr, Y. Lin, S.-h. Wang, R. Baillargeon, and C. Tan, "A benchmark for modeling violation-of-expectation in physical reasoning across event categories," *arXiv preprint arXiv:2111.08826*, 2021.
[348] L. Weihs, A. Yuile, R. Baillargeon, C. Fisher, G. Marcus, R. Mottaghi, and A. Kembhavi, "Benchmarking progress to infant-level physical reasoning in ai," *Transactions on Machine Learning Research*, 2022.
[349] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover, "Videophy: Evaluating physical commonsense for video generation," *arXiv preprint arXiv:2406.03520*, 2024.
[350] F. Meng, W. Shao, L. Luo, Y. Wang, Y. Chen, Q. Lu, Y. Yang, T. Yang, K. Zhang, Y. Qiao *et al.*, "Phybench: A physical commonsense benchmark for evaluating text-to-image models," *arXiv preprint arXiv:2406.11802*, 2024.
[351] F. Meng, J. Liao, X. Tan, W. Shao, Q. Lu, K. Zhang, Y. Cheng, D. Li, Y. Qiao, and P. Luo, "Towards world simulator: Crafting physical commonsense-based benchmark for video generation," *arXiv preprint arXiv:2410.05363*, 2024.
[352] NVIDIA Research, "PBench: A benchmark for evaluating generative models," https://research.nvidia.com/labs/dir/pbench/, 2024, accessed: October 13, 2025.
[353] Y. Chen, X. Zhu, and T. Li, "A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction," *arXiv preprint arXiv:2502.05503*, 2025.
[354] C. Li, O. Michel, X. Pan, S. Liu, M. Roberts, and S. Xie, "Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop," *arXiv preprint arXiv:2503.09595*, 2025.
[355] J. Wang, A. Ma, K. Cao, J. Zheng, Z. Zhang, J. Feng, S. Liu, Y. Ma, B. Cheng, D. Leng *et al.*, "Wisa: World simulator assistant for physics-aware text-to-video generation," *arXiv preprint arXiv:2503.08153*, 2025.
[356] X. Guo, J. Huo, Z. Shi, Z. Song, J. Zhang, and J. Zhao, "T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation," *arXiv preprint arXiv:2505.00337*, 2025.
[357] E. Sanli, B. S. Tezcan, A. Erdem, and E. Erdem, "Can your model separate yolks with a water bottle? benchmarking physical commonsense understanding in video generation models," *arXiv preprint arXiv:2507.15824*, 2025.
[358] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.
[359] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.
[360] T. Yu, G. Lu, Z. Yang, H. Deng, S. S. Chen, J. Lu, W. Ding, G. Hu, Y. Tang, and Z. Wang, "Manigaussian++: General robotic bimanual manipulation with hierarchical gaussian world model," *arXiv preprint arXiv:2506.19842*, 2025.
[361] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.
[362] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.
[363] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.
[364] Y. Guo, L. X. Shi, J. Chen, and C. Finn, "Ctrl-world: A controllable generative world model for robot manipulation," *arXiv preprint arXiv:2510.10125*, 2025.
[365] R. McLean, E. Chatzaroulas, L. McCutcheon, F. Röder, T. Yu, Z. He, K. Zentner, R. Julian, J. Terry, I. Woungang *et al.*, "Meta-world+: An improved, standardized, rl benchmark," *arXiv preprint arXiv:2505.11289*, 2025.
[366] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.
[367] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
[368] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.
[369] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.
[370] S. Tian, C. Finn, and J. Wu, "A control-centric benchmark for video prediction," *arXiv preprint arXiv:2304.13723*, 2023.
[371] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," *arXiv preprint arXiv:2307.00595*, 2023.
[372] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.
[373] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," *arXiv preprint arXiv:2310.17596*, 2023.
[374] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. Martinez *et al.*, "Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation," *arXiv preprint arXiv:2403.09227*, 2024.
[375] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv preprint arXiv:2406.02523*, 2024.
[376] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
[377] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
[378] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.
[379] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo *et al.*, "Genad: Generalized predictive model for autonomous driving," *arXiv preprint arXiv:2403.09630*, 2024.
[380] C. Shi, S. Shi, K. Sheng, B. Zhang, and L. Jiang, "Drivex: Omni scene modeling for learning generalizable world knowledge in autonomous driving," *arXiv preprint arXiv:2505.19239*, 2025.
[381] Y. Wang, K. Cheng, J. He, Q. Wang, H. Dai, Y. Chen, F. Xia, and Z.-X. Zhang, "Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 13 020–13 034, 2024.
[382] Z. Zhou and D. Negrut, "Chronodreamer: Action-conditioned world model as an online simulator for robotic planning," *arXiv preprint arXiv:2512.18619*, 2025.
[383] H. Arai, K. Ishihara, T. Takahashi, and Y. Yamaguchi, "Act-bench:

Towards action controllable world models for autonomous driving," *arXiv preprint arXiv:2412.05337*, 2024.
[384] B. Yu and D. Wang, "A trajectory-guided diffusion model for consistent and realistic video synthesis in autonomous driving," *Computer Modeling in Engineering & Sciences*, vol. 146, no. 1, 2026.
[385] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.
[386] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
[387] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.
[388] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.
[389] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.
[390] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn *et al.*, "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:2004.06320*, 2020.
[391] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*. PMLR, 2021, pp. 409–418.
[392] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9710–9719.
[393] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li *et al.*, "One million scenes for autonomous driving: Once dataset," *arXiv preprint arXiv:2106.11037*, 2021.
[394] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.
[395] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.
[396] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.
[397] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.
[398] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 178–20 188.
[399] D. Gao, S. Cai, H. Zhou, H. Wang, I. Soltani, and J. Zhang, "Cardreamer: Open-source learning platform for world model based autonomous driving," *IEEE Internet of Things Journal*, 2024.
[400] X. Yang, L. Wen, Y. Ma, J. Mei, X. Li, T. Wei, W. Lei, D. Fu, P. Cai, M. Dou *et al.*, "Drivearena: A closed-loop generative simulation platform for autonomous driving," *arXiv preprint arXiv:2408.00415*, 2024.
[401] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of artificial intelligence research*, vol. 47, pp. 253–279, 2013.
[402] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq *et al.*, "Deepmind control suite," *arXiv preprint arXiv:1801.00690*, 2018.
[403] D. Hafner, "Benchmarking the spectrum of agent capabilities," *arXiv preprint arXiv:2109.06780*, 2021.
[404] K. Lian, S. Cai, Y. Du, and Y. Liang, "Toward memory-aided world models: Benchmarking via spatial consistency," *arXiv preprint arXiv:2505.22976*, 2025.
[405] K. Cobbe, C. Hesse, J. Hilton, and J. Schulman, "Leveraging procedural generation to benchmark reinforcement learning," in *International conference on machine learning*. PMLR, 2020, pp. 2048–2056.
[406] H. Küttler, N. Nardelli, A. Miller, R. Raileanu, M. Selvatici, E. Grefenstette, and T. Rocktäschel, "The nethack learning environment," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7671–7684, 2020.
[407] W. H. Guss, M. Y. Castro, S. Devlin, B. Houghton, N. S. Kuno, C. Loomis, S. Milani, S. Mohanty, K. Nakata, R. Salakhutdinov *et al.*, "The minerl 2020 competition on sample efficient reinforcement learning using human priors," *arXiv preprint arXiv:2101.11071*, 2021.
[408] T. Pearce and J. Zhu, "Counter-strike deathmatch with large-scale behavioural cloning," in *2022 IEEE Conference on Games (CoG)*. IEEE, 2022, pp. 104–111.
[409] X. Tang, J. Li, Y. Liang, S.-c. Zhu, M. Zhang, and Z. Zheng, "Mars: Situated inductive reasoning in an open-world environment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 17 830–17 869, 2024.
[410] M. Li, Z. Wang, K. He, X. Ma, and Y. Liang, "Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse," *arXiv preprint arXiv:2503.16365*, 2025.
[411] T. Huang, W. Zheng, T. Wang, Y. Liu, Z. Wang, J. Wu, J. Jiang, H. Li, R. W. Lau, W. Zuo, and C. Guo, "Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation," *arXiv preprint arXiv:2506.04225*, 2025.
[412] Z. Yang, W. Ge, Y. Li, J. Chen, H. Li, M. An, F. Kang, H. Xue, B. Xu, Y. Yin *et al.*, "Matrix-3d: Omnidirectional explorable 3d world generation," *arXiv preprint arXiv:2508.08086*, 2025.
[413] World Labs, "World labs blog," https://www.worldlabs.ai/blog, 2025, accessed: October 7, 2025.
[414] ——, "Rtfm: A real-time frame model," https://www.worldlabs.ai/blog/rtfm, 2025, accessed: October 18, 2025.
[415] S. Li, C. Yang, J. Fang, T. Yi, J. Lu, J. Cen, L. Xie, W. Shen, and Q. Tian, "Worldgrow: Generating infinite 3d world," *arXiv preprint arXiv:2510.21682*, 2025.
[416] Z. Gao, J. Mao, H.-X. Yu, H. Lou, E. Y.-T. Jia, J. Barbic, J. Wu, and Y. Wang, "Seeing the wind from a falling leaf," *arXiv preprint arXiv:2512.00762*, 2025.
[417] H. Zhang, Y. Wang, Y. Tang, Y. Liu, J. Feng, and X. Jin, "Flash-vstream: Efficient real-time understanding for long video streams," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2025, pp. 21 059–21 069.
[418] X. Yu, Y. Fang, X. Jin, Y. Zhao, and Y. Wei, "Prefm: Online audio-visual event parsing via predictive future modeling," *arXiv preprint arXiv:2505.23155*, 2025.
[419] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. B. Hashimoto, "s1: Simple test-time scaling," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 20 286–20 332.
[420] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, "Pixellm: Pixel reasoning with large multimodal model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 374–26 383.