Review

Not peer-reviewed version

Predicting the World via Video Representation: A Comprehensive Survey on Video World Models

Jiaxin Yan , Chaoning Zhang * , Xudong Wang , Pengcheng Zheng , Ya Wen , Qigan Sun , Jiaxin Huang , Shuxu Chen , Yang Yang , Hyundong Shin *

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0435.v1

Keywords: video world models; reinforcement learning; self-supervised learning; imitation learning; diffusion models

 doi:10.20944/preprints202605.0435.v1

Review

Predicting the World via Video Representation: A Comprehensive Survey on Video World Models

Jiaxin Yan [1], Chaoning Zhang [1], Xudong Wang [1], Pengcheng Zheng [1], Ya Wen [1], Qigan Sun [1], Jiaxin Huang [2], Shuxu Chen [3], Yang Yang [1] and Hyundong Shin [3,*]

[1] University of Electronic Science and Technology of China, Cheng Du, China
[2] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[3] Kyung Hee University, Yongin-si, Republic of Korea
* Correspondence: hshin@khu.ac.kr

Abstract

Video world models have emerged as a critical framework, offering a powerful approach to modeling dynamic environments through lens of video data, and serving as a key tool for understanding and predicting complex systems. While prior papers have focused on specific domains such as 3D modeling, autonomous driving, and robotics, they have largely overlooked the growing importance of video modality in the development of future world models. These papers often concentrate on particular data representations, failing to account for how video-based representations can bridge the gap between perception, prediction, and decision-making in intelligent systems. This paper aims to fill this gap by providing a standardized and systematic classification of video world models. We introduce a comprehensive taxonomy that distinguishes between implicit state deduction which, focuses on learning compact latent representations and explicit visual modeling, which emphasizes frame level video processing. Additionally, we analyze indepth review of experimental setups, specific applications, and open problems. By focusing on video world models, this paper offers a unified reference that highlights their critical role in the future of world modeling research.
CCS Concepts: **Computing methodologies** →**Artificial intelligence**

Keywords: video world models; reinforcement learning; self-supervised learning; imitation learning; diffusion models

1. Introduction

World Models constitute a fundamental paradigm for environment modeling, characterized by the learning of compact and structured latent representations to capture complex environmental dynamics. This capability enables agents to perform efficient prediction, planning, and control within a latent space. This line of research, epitomized by the seminal work of Ha and Schmidhuber [1], has laid the groundwork for subsequent latent dynamics-based reinforcement learning. As research advances toward general intelligent systems, environment modeling is undergoing a paradigm shift from explicit geometric and state-based representations to dynamics-centric formulations grounded in visual sequences. Against this backdrop, Video World Models have emerged as a pivotal research direction bridging perception, prediction, and decision-making.

Unlike traditional three-dimensional (3D) or four-dimensional (4D) world models that rely on explicit geometric reconstruction [2–10], video world models take two-dimensional (2D) visual inputs as the primary modeling substrate, characterizing environmental dynamics through continuous video sequences. This design aligns with cognitive hypotheses of human perception [11]: humans do not directly observe complete 3D structures but instead infer spatial organization and physical dynamics from sequences of 2D retinal images. Modeling environments through 2D video sequences not only adopts more natural representational assumptions but also substantially reduces the computational and system-level complexity associated with high-dimensional volumetric data [12–18]. Consequently,

 doi:10.20944/preprints202605.0435.v1

video world models provide a concise yet powerful framework for understanding and predicting complex dynamic environments, making them a core direction in general environment modeling research.

The evolution of world models has been marked by notable architectural and methodological developments. The Dreamer family of models, beginning with *DreamerV1* [19], introduces latent state-space models for RL and enables imagination-based planning directly in latent space without relying on pixel-level control. *DreamerV2* [20] further extends this framework to more diverse and challenging task settings by incorporating discrete latent representations and robust normalization strategies. Meanwhile, foundational approaches including *PlaNet* [21] and *SLAC* [22] improve sample efficiency and long-horizon prediction by learning structured latent dynamics.

With the rapid advancement of generative modeling, diffusion models and autoregressive models are increasingly integrated into the video modeling domain. This shift redirects attention from perception-oriented video synthesis toward systematic modeling of temporal dynamics and latent state transitions. Against this backdrop, video world models such as *DriveDreamer* [23] and *Drive-WM* [24] represent a new stage in video-based environment modeling. Rather than emphasizing pixel-level synthesis quality alone, these approaches construct generative simulations that support prediction, planning, and decision-making, showing strong potential in complex dynamic scenarios such as autonomous driving.

Despite these advances, the field still lacks consensus on the definition, modeling boundaries, and core capabilities of video world models [12,25–31]. Some studies emphasize large-scale generative architectures, including diffusion and autoregressive frameworks, for long-horizon modeling and multimodal generation [32–38], while others focus on self-supervised or weakly supervised learning to improve data efficiency and generalization [39]. Existing papers typically categorize methods by methodological focus or by application domains such as RL, robotics, and autonomous driving [40–46]. This task-centric perspective fragments the field, obscuring shared principles and fundamental distinctions across paradigms, and ultimately limiting systematic analysis within a unified framework as well as deeper insights into cross-task and cross-domain generalization.

To address these challenges, this paper reviews video world models from a high-level perspective and proposes a unified analytical framework. Specifically, we introduce a taxonomy that groups existing approaches into two classes: ***Implicit State Deduction*** and ***Explicit Visual Modeling***. The former learns compact latent state representations that support efficient reasoning and decision-making, whereas the latter generates high-fidelity visual observations to explicitly simulate future environmental states. This taxonomy is independent of specific model architectures, learning algorithms, or application scenarios; instead, it is grounded in the fundamental question of how environmental states are represented and evolved over time. This perspective clarifies the role of video world models in broader intelligent system research and provides a principled reference for developing general and scalable environment modeling frameworks.

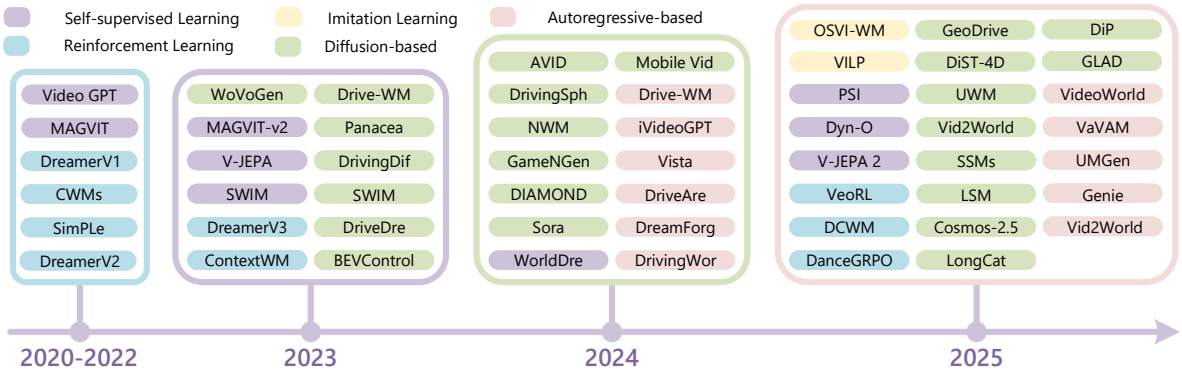

Figure 1. A comprehensive overview of the evolution of video world models across different learning paradigms. The timeline (2020-2025) categorizes notable world model approaches based on their learning paradigms and architectural designs.

 doi:10.20944/preprints202605.0435.v1

The primary contributions of this paper are as follows:

- To the best of our knowledge, this paper presents the first comprehensive review that systematically studies video world models from a unified perspective. We consolidate their core concepts, modeling assumptions, architectural foundations, learning paradigms, and evaluation practices, and provide an organized overview of this rapidly developing research area.
- We introduce a unified taxonomy for video world models, categorizing existing approaches into implicit state deduction and explicit visual modeling. This taxonomy further clarifies how different methods represent, evolve, and utilize environmental states, offering a structured lens for understanding the design space of video-based environment modeling.
- We systematically review representative benchmarks and evaluation protocols for video world models, summarizing commonly used datasets, metrics, and experimental settings, and identifying key limitations of current evaluations in measuring long-horizon consistency, physical plausibility, and decision-relevant fidelity.
- We discuss key open challenges and promising future directions, including scalable long-term simulation, physical grounding, efficient deployment, and cross-domain generalization, to guide future research toward more general and robust environment modeling systems.

The remainder of this paper is organized as follows. Section 2 introduces the background of video world models and presents our unified taxonomy, which decomposes the field into implicit state deduction and explicit visual modeling. Section 3 reviews implicit state deduction approaches, categorizing methods by learning paradigm, including reinforcement learning, self-supervised learning, and imitation learning. Section 4 examines explicit visual modeling frameworks, focusing on diffusion-based and autoregressive architectures for spatiotemporal generation. Section 5 summarizes benchmark datasets, evaluation metrics, and experimental protocols. Section 6 discusses open challenges and future research directions. Section 8 concludes the paper.

2. Background and Taxonomy of Video World Models

2.1. From Video Generation to Video World Model

Early video modeling research has largely centered on the synthesis of visually plausible video sequences. Under this formulation, videos are represented as temporally ordered image sequences, and the modeling task is framed as predicting future frames from past observations. These approaches extend image generation architectures to the temporal domain by incorporating recurrent neural networks, convolutional LSTMs, or predictive coding mechanisms. Despite promising short-term performance, these methods rely heavily on pixel-level supervision and implicit temporal correlations, resulting in rapid error accumulation and degraded temporal consistency over long horizons. This exposes fundamental limitations in modeling complex real-world dynamics. To address uncertainty in future video evolution, stochastic latent variable models have been adopted to capture temporal dynamics within compact latent spaces [47–49]. By decoupling high-dimensional visual observations from underlying dynamics, these approaches improve predictive diversity and robustness. Representative methods include World Models and its successors, which integrate variational autoencoders with recurrent dynamics to learn latent representations for imagination and prediction [50]. *PlaNet* and *SLAC* further extend this paradigm by supporting planning and control directly from raw pixels [51]. Despite their conceptual significance, these models remain largely constrained by reconstruction-driven objectives and exhibit limited scalability to semantically rich, long-horizon environments.

The rapid advancement of large-scale generative modeling has ushered in a new phase of video modeling research. Diffusion-based and autoregressive frameworks have significantly improved video synthesis, enabling the production of high-fidelity, temporally coherent sequences across diverse domains [52]. Advances in diffusion-based approaches achieve remarkable visual realism through progressive denoising [53], whereas autoregressive token-based methods reformulate video modeling as a sequence prediction problem over discrete visual tokens . These paradigms enhance temporal

 doi:10.20944/preprints202605.0435.v1

modeling and abstraction, facilitating the development of powerful conditional video generation systems. Despite their strong perceptual performance, most of these models remain primarily optimized for video generation, with limited emphasis on learning explicit environment representations or supporting long-term reasoning and decision making.

As video models are applied to real-world scenarios such as autonomous driving, robotics, and embodied intelligence, the limitations of purely generative objectives become increasingly evident [54]. In these settings, agents reason over latent environment states, predict action outcomes, and operate under partial observability and extended temporal dependencies [55]. These requirements have driven a conceptual shift from video generation to video world models, which learn compact and structured representations of dynamic environments rather than focusing solely on visual synthesis. Representative approaches, such as the *Dreamer* family, demonstrate that learned latent dynamics support planning and control, positioning video world models as a foundational component for integrating perception, prediction, and decision making in complex systems.

2.2. Taxonomy of Video World Models

The rapid development of video world models has produced a diverse set of architectures and learning paradigms, varying in their representation of environment states and modeling of temporal dynamics. A coherent taxonomy is essential for clarifying conceptual relationships among existing methods and enabling systematic comparisons across application domains [56]. From a modeling perspective, most approaches can be distinguished by their latent state representations and the mechanisms used to learn and propagate state transitions over time. These design choices directly affect the capacity of a model to perform long-horizon prediction, reasoning, and decision making in complex environments.

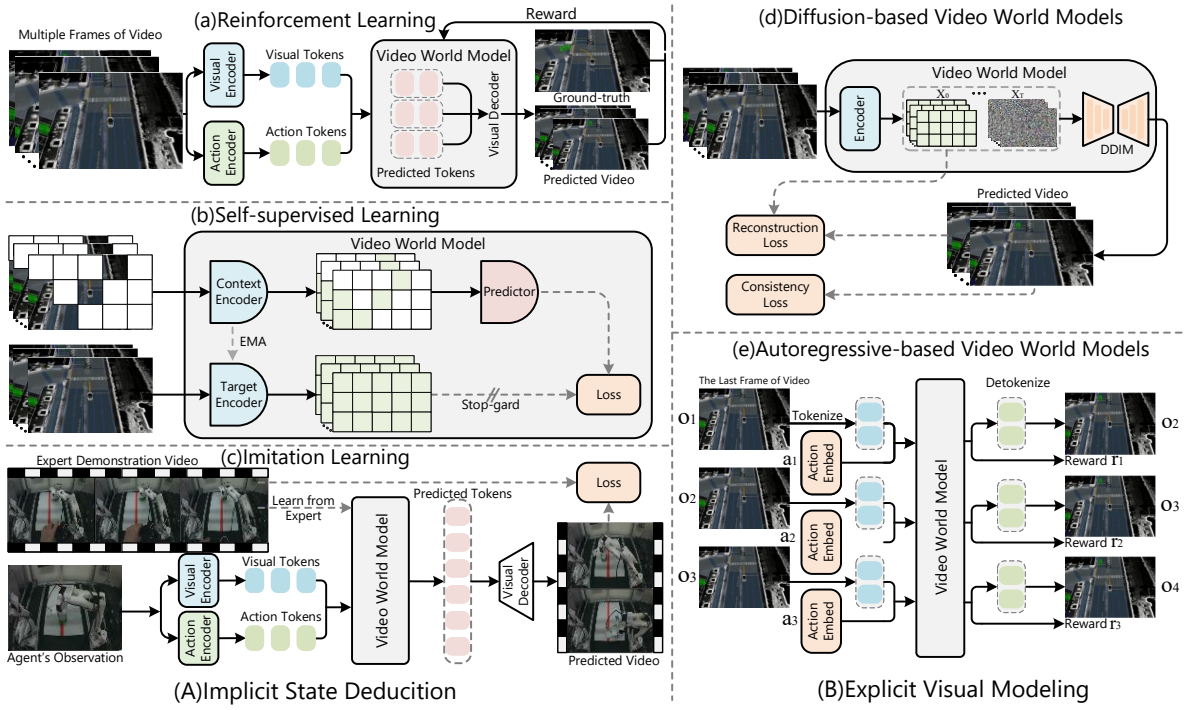

Figure 2. A comprehensive overview of video world models across different learning paradigms. The figure illustrates the corresponding model architectures and learning processes for each method: (a) Reinforcement Learning, (b) Self-Supervised Learning, (c) Imitation Learning, (d) Diffusion-based Video World Models, (e) Autoregressive-based Video World Models. This figure also systematically presents the diverse techniques used for video world modeling, categorizing them into (A) Implicit State Deduction and (bottom left) Explicit Visual Modeling (bottom right).

 doi:10.20944/preprints202605.0435.v1

Based on these considerations, existing video world models can be broadly categorized by implicit state representation and dynamics learning strategy. Some methods infer continuous latent variables that compactly encode environment states, typically optimized with prediction-based objectives. Other approaches discretize visual observations into tokenized representations and cast video modeling as a sequence prediction task, enabling scalable long-range temporal modeling through autoregressive objectives. Within both categories, differences emerge in the incorporation of actions, rewards, or external constraints, reflecting varying degrees of interaction and control awareness. This taxonomy offers a structured overview of representative video world models and provides an organizing framework for the detailed analysis in the following sections.

3. Implicit State Deduction

3.1. Continuous Latent Variable Representations

3.1.1. Reinforcement Learning-based

These methods address POMDPs from pixels by learning a world model within the MBRL paradigm. The prevalent Recurrent State-Space Model (RSSM) learns three components: an encoder $q_\phi(z_t|o_t)$, a dynamics model with deterministic state $h_t = f_\psi(h_{t-1}, z_{t-1}, a_{t-1})$ and stochastic output $p_\theta(z_t|h_t)$, and a decoder $p_\xi(o_t|s_t)$ with reward predictor $p_\zeta(r_t|s_t)$ where $s_t = (z_t, h_t)$. The generative process is:

$$p(o_{1:T}, r_{1:T}, s_{1:T} \mid a_{1:T}) = \prod_{t=1}^{T} p_\xi(o_t \mid s_t)\; p_\zeta(r_t \mid s_t)\; p(s_t \mid h_t), \tag{1}$$

with $p(s_t \mid h_t) \equiv p_\theta(z_t \mid h_t)\, \delta(h_t - f_\psi(h_{t-1}, z_{t-1}, a_{t-1}))$. This factorization enables *latent-space imagination*: rolling out future trajectories (s_τ, r_τ) in the compact latent space without pixel decoding. Policy optimization (e.g., actor-critic) on these rollouts decouples high-dimensional rendering from control, yielding high sample efficiency. The objective combines a variational bound on observations with reward maximization for end-to-end training [57].

Early work by Hafner et al. [21] introduced the *PlaNet* model, which learns a latent dynamics model with both deterministic and stochastic components from pixels and performs planning via the Cross-Entropy Method (CEM) in *SLAC*, which integrates representation learning and RL into a single end-to-end framework. It learns a stochastic latent sequential model and trains an actor-critic directly in the learned latent space, improving stability and sample efficiency. Concurrently, Assran et al. proposed learning *Deep Structured Representations for Model-Based Reinforcement Learning (SOLAR)*, which explicitly structures the latent space to support simple (e.g., linear) dynamics models and classical control methods such as iterative LQR, thereby enabling deployment on real-world robotic tasks.

The *Dreamer* series significantly advanced this line. Hafner et al. [19] introduced the *DreamerV1* model, which learns an RSSM from pixels and uses latent trajectory imagination to train policy and value functions purely from imagined futures. This approach achieved high performance on 20 visual continuous control tasks while avoiding expensive pixel-level planning. Later, Wu et al. [58] present *ContextWM*, which pre-trains continuous context-aware world models from in-the-wild videos and transfers them to RL tasks, demonstrating improved generalization. Together, these works illustrate a trajectory from initial pixel-based video prediction models to sophisticated, scalable discrete token world models that enable sample-efficient reinforcement learning, robust cross-domain generalization, and novel applications like curriculum generation and offline RL.

3.1.2. Self-Supervised Learning

Self-supervised learning has emerged as a core paradigm for continuous latent state modeling in video world models, enabling the discovery of compact and predictive representations directly from raw video streams. By exploiting temporal continuity and predictive objectives, these methods learn latent world states that discard pixel-level redundancy while retaining information critical for modeling future dynamics [59].

 doi:10.20944/preprints202605.0435.v1

An early and influential framework in this line of research is *World Models*, which formulates world modeling as a modular decomposition of perception, dynamics, and control under self-supervision. In this formulation, each visual observation o_t is first encoded into a continuous latent variable z_t using a variational autoencoder:

$$z_t \sim q_\phi(z_t \mid o_t), \tag{2}$$

where the encoder is trained through reconstruction of video frames. Temporal evolution is then captured by a recurrent dynamics model parameterized as a mixture density network, which predicts the distribution over future latent states and rewards conditioned on the current latent state and action:

$$p_\theta(z_{t+1}, r_t \mid z_t, a_t). \tag{3}$$

This design demonstrates that long-horizon prediction and planning can be performed entirely within a learned continuous latent space. Although initially studied in interactive control settings, World Models establishes a foundational principle for video world models in which implicit world states are induced through self-supervised temporal prediction rather than explicit semantic annotation, a principle that later continuous latent dynamics models further formalize and extend.

Building upon this idea, subsequent work such as *SWIM* [60] emphasizes stronger temporal coherence and motion-aware representation learning. Instead of relying solely on frame-wise reconstruction, *SWIM* introduces predictive consistency objectives that align latent representations across time. By encouraging predicted latent states to match future observations, the learned representations encode object-level structure and motion patterns that persist over multiple frames. Such designs improve robustness to appearance variation and enhance the model's ability to capture long-term dependencies in video dynamics, which is critical for world modeling beyond short-term prediction. More recent approaches depart from pixel-level generation entirely and focus on representation prediction in latent space. *V-JEPA* [61] extends joint-embedding predictive architectures to video by learning through masked spatiotemporal prediction. Given a video clip, the model observes a subset of context blocks and predicts the representations of disjoint target blocks from the same clip:

$$\mathcal{L} = \sum_i \left\| f_\theta(x_i^{\text{target}}) - g_\phi(x^{\text{context}}) \right\|_2^2, \tag{4}$$

where the target and context encoders are trained to align their outputs without reconstructing pixels. A key empirical finding of *V-JEPA* is that semantic abstraction arises from the masking strategy itself. Predicting sufficiently large target regions forces the model to capture high-level semantics, while ensuring that context regions are spatially distributed preserves global scene structure. When combined with transformer-based architectures, *V-JEPA* scales efficiently to large models and yields strong performance across a range of downstream visual and video understanding tasks. From the perspective of video world modeling, this demonstrates that implicit state deduction can be achieved through predictive representation learning rather than explicit frame synthesis.

Scaling this paradigm to internet-scale video data, *V-JEPA 2* further reveals the potential of self-supervised latent prediction. Large-scale pretraining on massive collections of uncurated videos produces world representations with strong motion understanding and anticipation capabilities, even in the absence of action supervision. When aligned with language models, the representations learned by *V-JEPA 2* support video reasoning tasks, indicating that self-supervised visual world models can serve as a foundation for multimodal understanding and decision making. Complementary to representation-centric approaches, object-centric self-supervised world models such as *Dyn-O* [62] introduce structural inductive biases into continuous latent spaces. By decomposing scenes into object-level latent variables that persist over time, *Dyn-O* learns world dynamics through pixel-level reconstruction and temporal prediction losses alone. Such factorized latent representations improve interpretability and facilitate structured video prediction and planning, highlighting the role of inductive bias in shaping the geometry of learned world states [63].

 doi:10.20944/preprints202605.0435.v1

Overall, self supervised learning for continuous latent world modeling has evolved from reconstruction driven dynamics learning, exemplified by *World Models*, to large scale predictive representation learning, represented by *V-JEPA* and *V-JEPA 2*. Despite differences in architecture and objectives, these methods share a unifying principle: latent world states are not directly observed but are inferred through temporal prediction constraints imposed by video data itself. This principle underpins implicit state deduction in modern video world models, serving as a foundational link between perception and long-horizon reasoning.

3.1.3. Imitation Learning

Within the video world modeling paradigm, imitation learning serves as a mechanism for guiding implicit state deduction when direct interaction or explicit rewards are limited. Rather than reproducing demonstrated videos at the pixel level, imitation learning leverages pretrained video world models to reason about latent environment dynamics and generate task-consistent future evolution. In this setting, demonstrations provide high-level behavioral constraints, while the world model supplies a predictive environment for planning and simulation.

VILP [64] illustrates this paradigm by treating demonstrations as evidence of task intent rather than visual trajectories to be copied. A video world model is used to internally simulate how the environment may evolve under different action hypotheses, allowing the agent to select trajectories that align with the demonstrated objective while remaining physically plausible. Here, imitation learning operates at the level of latent state transitions, shaping the use of the world model without constraining it to exact visual reproduction. *OSVI-WM* [65] further reinforces this view by emphasizing the role of video world models as internal simulators for long-horizon reasoning. This model views demonstration videos as clues about environment dynamics and uses imitation to guide future predictions in the world model. The resulting behavior emerges from reasoning over predicted environment evolution, rather than from frame-wise matching to demonstrations. These examples show how imitation learning guides video world models to generate task-consistent and physically plausible video. By operating on implicit world states rather than raw pixels, imitation learning enables video world models to enable planning, generalization, and decision making in embodied and interactive scenarios.

3.2. Discrete Token Representations

3.2.1. Reinforcement Learning-based

These methods apply the MBRL paradigm within a discrete latent space to solve POMDPs. They first learn a tokenizer (e.g., VQ-VAE) that maps images o_t to discrete codes z_t, and then train a dynamics model $p_\theta(z_{t+1}|z_{\leq t}, a_{\leq t})$ on the token sequence. The resulting world model defines the joint probability $p(z_{1:T} \mid a_{1:T}) = \prod_{t=1}^{T} p_\theta(z_t \mid z_{<t}, a_{<t})$, upon which a policy $\pi_\psi(a_t \mid z_{\leq t})$ is optimized via RL. This discrete formulation provides key benefits: it creates a natural information bottleneck to reduce noise, enhances compositional generalization through token recombination, and mitigates compounding prediction error due to a bounded state space. These properties enable substantially more robust and efficient latent-space planning compared to continuous representations.

Early exploration of discrete world models for RL includes *SimPLe*, which used a pixel-level video prediction model as a world model for sample-efficient RL in the Atari 100k benchmark, setting a baseline for future work. [66] then proposed using VQ-VAE to compress pixels into discrete codes and trained a convolutional LSTM dynamics model on these codes, demonstrating that smaller, discrete models could remain sample-efficient. A major breakthrough came with *DreamerV2*, which replaced the continuous Gaussian latents in RSSM with multiple discrete categorical variables. This discrete latent representation significantly improved world model accuracy on Atari games and achieved human-level performance from pixels. The subsequent *DreamerV3* [67] scaled this approach with robust normalization techniques, enabling a single configuration to work across diverse domains and underscoring the generality of discrete latent world models.

 doi:10.20944/preprints202605.0435.v1

Recent works have further expanded the scope. *DART* [68] tokenizes observations using VQ and employs a Transformer-decoder as an autoregressive token-level world model, with a Transformer-encoder policy, showing strong sample efficiency on Atari. *Genie* [69] learned a foundational, generative interactive environment from large-scale unsupervised web videos using discrete video tokens and latent action models, enabling the training of RL agents in synthesized environments. For curriculum RL, *CQM* [70] used VQ-VAE discrete representations as a semantic goal space to automatically generate training curricula, improving exploration in goal-conditioned RL. Moreover, *DCWM* [71] systematically studied discrete codebook world models for continuous control tasks, showing advantages in efficiency and robustness over continuous counterparts. Finally, *VeoRL* [72] demonstrated how tokenized world models learned from offline video data can enhance offline RL policy optimization and value correction.

Together, these works illustrate a trajectory from initial pixel-based video prediction models to sophisticated, scalable discrete token world models that enable sample-efficient reinforcement learning, robust cross-domain generalization, and novel applications like curriculum generation and offline RL.

3.2.2. Self-Supervised Learning

Self-supervised learning has emerged as a central paradigm for learning discrete latent state spaces in video world models, as it enables scalable training on large collections of unlabeled videos while preserving temporal structure and dynamics. Within this paradigm, video sequences are first compressed into sequences of discrete visual tokens, after which temporal dependencies are learned via sequence modeling objectives. This formulation transforms video modeling into a latent-state prediction problem, where discrete tokens serve as abstract representations of implicit world states.

VideoGPT [73] represents one of the earliest and most influential self-supervised approaches that formulate video modeling as an autoregressive prediction problem over discrete latent tokens. The core idea is to decouple spatial perception from temporal modeling by introducing a learned vector-quantized encoder. Given an input video sequence $\{x_t\}_{t=1}^{T}$, each frame is encoded into a grid of discrete latent variables $z_t \in \mathcal{Z}^{H\times W}$ using a VQ-VAE, where $\mathcal{Z}$ denotes a finite codebook. Temporal dynamics are then modeled by an autoregressive Transformer that factorizes the joint distribution over latent tokens as

$$p(z_{1:T}) = \prod_{t=1}^{T} p(z_t \mid z_{<t}). \tag{5}$$

By operating entirely in the discrete latent space, VideoGPT significantly reduces computational cost while retaining high-level spatiotemporal structure. From a world modeling perspective, the discrete tokens implicitly represent latent environment states, and autoregressive prediction corresponds to learning state transition dynamics. Although VideoGPT is trained without explicit action inputs, it shows that self-supervised next-token prediction on discretized video representations is sufficient to learn meaningful temporal regularities, thereby laying the groundwork for subsequent discrete world models.

Building upon the discrete tokenization paradigm, *MAGVIT* [74] introduces a more expressive and scalable video tokenizer tailored for high-resolution and long-duration videos. Unlike earlier VQ-based approaches that treat frames independently, *MAGVIT* employs a spatiotemporal tokenizer that jointly encodes spatial and temporal information into compact discrete tokens. Formally, a video clip $x_{1:T}$ is mapped to a sequence of tokens $\{z_k\}_{k=1}^{K}$, where each token represents a local spatiotemporal volume rather than a single frame. The tokenizer is trained using a reconstruction objective combined with perceptual and adversarial losses, ensuring that the discrete representation preserves motion continuity and temporal coherence. Once tokenized, the video is modeled using a self-supervised sequence prediction objective over the discrete token stream. This design shifts the modeling focus from pixel-level redundancy to higher-level temporal abstractions, enabling more stable and semantically meaningful latent dynamics. In the context of video world models, *MAGVIT* can be interpreted as learning a structured discrete state space in which each token encodes a short-term evolution of the visual world, thereby facilitating long-range temporal modeling under self-supervision. *MAGVIT-*

 doi:10.20944/preprints202605.0435.v1

v2 [75] further advances discrete self-supervised video modeling by unifying tokenization and temporal prediction within a more efficient and semantically aligned framework. A key improvement lies in the introduction of factorized and hierarchically organized token representations, which reduces codebook entropy while enhancing representational capacity. Let $\{z_t^{(l)}\}$ denote discrete tokens at hierarchy level l, where lower levels capture fine-grained appearance details and higher levels encode longer-term temporal structure. The overall learning objective jointly optimizes multi-level reconstruction and next-token prediction:

$$\mathcal{L} = \sum_l \Big(\mathcal{L}_{\text{rec}}^{(l)} + \lambda \, \mathcal{L}_{\text{pred}}^{(l)} \Big). \tag{6}$$

This hierarchical formulation encourages the model to learn temporally persistent latent states that evolve smoothly over time, which is particularly desirable for world modeling. By emphasizing temporal abstraction and efficient discrete representations, *MAGVIT-v2* enables large-scale self-supervised training on internet-scale videos and provides a stronger latent foundation for downstream dynamics modeling. As a result, it brings discrete self-supervised video world models closer to practical long-horizon simulation and planning settings.

4. Explicit Visual Modeling/Model Architecture

4.1. Diffusion-based Video World Models

Diffusion models (DMs) are employed to generate and predict video data by progressively introducing noise and then learning to reverse the corruption process. These models maintain temporal consistency and high-fidelity details during generation. The forward diffusion process gradually adds noise to clean video frames, while the reverse denoising process learns to recover the original video data by reversing this noise.

Representative formulations of diffusion models include *Denoising Diffusion Probabilistic Models (DDPMs)* [76], *Noise Conditioned Score Networks (NCSNs)* [77], and *Score-based Stochastic Differential Equations (Score SDEs)* [78]. The following section outlines the forward diffusion and reverse generation processes of each method, highlighting their specific application to video world models.

DDPMs define a discrete-time diffusion process that progressively corrupts data with Gaussian noise. Let $x_0 \sim q(x_0)$ denote a clean sample. The forward process is a Markov chain $x_1, \ldots, x_T$ governed by a Gaussian transition kernel:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\Big), \quad t = 1, \ldots, T, \tag{7}$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule, and $\mathbf{I}$ is the identity matrix. An important feature of this formulation is that it admits a closed-form marginal distribution at any step t:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{8}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$. The reverse (denoising) process is defined by a learnable Gaussian kernel:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{9}$$

The reverse process is typically initialized from a simple prior $p(x_T)$ (e.g., $\mathcal{N}(0, I)$) and carried out through iterative ancestral sampling from $t = T$ down to $t = 1$.

NCSNs adopt a score-based perspective in discrete noise scales. Instead of learning an explicit reverse transition kernel, NCSNs learn a noise-conditioned score network that estimates the score function of perturbed data distributions across a sequence of noise levels. The forward perturbation is typically defined by Gaussian noise injection:

$$q_{\sigma_t}(x_t \mid x_0) = \mathcal{N}\Big(x_t; x_0, \sigma_t^2 \mathbf{I}\Big), \tag{10}$$

 doi:10.20944/preprints202605.0435.v1

where increasing σ_t results in progressively noisier samples. For generation, NCSNs typically utilize annealed Langevin dynamics across decreasing noise levels, where the learned score network $s_\theta(x,t)$ guides reverse-time updates:

$$x \leftarrow x + \eta_t s_\theta(x,t) + \sqrt{2\eta_t}\, z, \qquad z \sim \mathcal{N}(0,\mathbf{I}). \tag{11}$$

where η_t is the step size associated with the noise level. This iterative procedure starts from Gaussian noise and gradually moves samples toward regions of higher data density, eventually converging to the clean data domain.

Score *SDEs* offer a continuous-time formulation of diffusion generative modeling. The forward diffusion process is described by the following stochastic differential equation (SDE):

$$dx = f(x,t)\,dt + g(t)\,dW_t, \tag{12}$$

where $f(x,t)$ and $g(t)$ denote the drift and diffusion coefficients, respectively, and W_t is a standard Wiener process. Given the marginal density $p_t(x)$ at time t, the corresponding reverse-time generative process follows the reverse-time SDE:

$$dx = \Big[f(x,t) - g(t)^2 \nabla_x \log p_t(x)\Big]\,dt + g(t)\,d\overline{W}_t, \tag{13}$$

where $\overline{W}_t$ is a reverse-time Wiener process and $\nabla_x \log p_t(x)$ is the score function that drives the trajectory toward high-density regions. In practice, a time-dependent score model is learned to approximate $\nabla_x \log p_t(x)$, and sampling is performed by numerically solving the reverse-time dynamics from an initial noise state to the data space.

Diffusion-based Video World Models (*VWMs*) represent a fundamental paradigm shift in generative AI, moving from static high-fidelity synthesis toward dynamic, interactive physical simulation. Unlike traditional models that rely on deterministic transitions or simple stochastic latent variables, the diffusion architecture utilizes an iterative denoising process to capture the complex, multi-modal spatiotemporal distributions of the real world. This provides a high-tolerance probabilistic framework for simulating complex environments. The research trajectory in this domain begins with validating the capability of large-scale parameterization to model physical laws. *GAIA-1* [79] pioneers the application of diffusion models in autonomous driving and establishes the fundamental paradigm for multi-agent interactive simulation through action-conditioned generation. Subsequently, Sora [80] demonstrates the scalability of video simulators via a large-scale Diffusion Transformer (DiT) architecture, exhibiting emergent physical properties such as object permanence and basic collision logic. Building on this, *DIAMOND* [81] integrates diffusion models into the reinforcement learning loop, proving that pixel-level fine-grained representations offer superior performance over latent-space models in complex visual decision-making tasks.

As the requirement for precise control across diverse task scenarios becomes more pressing, the research focus shifts toward transforming pre-trained generative priors into controllable dynamics simulators. *AVID* [82] proposes a parameter-efficient adapter scheme that injects action conditions without compromising the original generative capacity. *Vid2World* [83] and *Vidar* [84] further explore the causalization of the denoising process. By aligning robotic morphological features, these models generate environmental feedback corresponding to specific action commands, thereby constructing a robust "state-action-state" interaction loop. In response to cumulative drift and physical distortion in long-term predictions, the integration of explicit geometric priors becomes crucial for enhancing simulation fidelity. *PA-VDM* [85] and *Epona* [86] effectively mitigate temporal degradation in long-sequence generation through improved attention mechanisms and noise scheduling strategies. To further constrain object trajectories, *Geometry Forcing* [87] and *VerseCrafter* [88] incorporate 3D structures, such as point clouds and Gaussian trajectories, into the diffusion process to ensure 4D spatiotemporal

 doi:10.20944/preprints202605.0435.v1

consistency. This transition from 2D pixel interpolation to geometry-guided simulation provides a reliable foundation for complex manipulation tasks, as seen in *DWM* [89] and *LaDi-WM* [90].

The core value of video world moedls lies in empowering agents with planning and reasoning capabilities. *UWM* [91] and *World4RL* [92] validate the efficacy of using virtual trajectories generated by diffusion models for policy refinement. *DAWM* [93] and *DUST* [94] further optimize the alignment precision between visual features and control signals via dual-stream architectures, while *Astra* [95] introduces general interactive capabilities through autoregressive denoising to enhance the universality of the simulator. In parallel, *CWMDT* [96] introduces digital twin conditioning to support counterfactual reasoning. This capability elevates video world models from simple evolutionary predictors to sophisticated decision engines capable of hypothetical reasoning and counterfactual inference, providing theoretical support for policy evaluation and safety verification in complex dynamic environments.

4.2. Autoregressive-based Video World Models

General-purpose video generation task primarily aims at visual fidelity and alignment with textual or conditional inputs, focusing on synthesizing high-quality and temporally coherent short video clips. In contrast, video world models emphasize causal dynamics, controllability, and long-term consistency, properties that naturally align with autoregressive modeling paradigms. Concretely, autoregressive models factorize the joint distribution of a sequence into a series of temporally ordered conditional predictions, thereby modeling causal temporal evolution through step-by-step generation. Under this modeling paradigm, autoregressive video world models characterize world dynamics via causal sequence decomposition: given historical observations and actions, the model incrementally generates future visual observations. From a global perspective, autoregressive video world models can be abstracted as learning a causal rollout distribution over future observations conditioned on actions:

$$p(o_{t+1:t+H} \mid o_{\leq t}, a_{t:t+H-1}) = \prod_{k=1}^{H} p(o_{t+k} \mid o_{\leq t+k-1}, a_{t:t+k-1}). \tag{14}$$

Here, o_t denotes the visual observation at time t (e.g., video frames or discrete visual tokens), a_t denotes the action applied at time t, and H is the prediction horizon. This autoregressive factorization is not merely a modeling choice; Rather, it can be interpreted as an explicit temporal rollout mechanism that enables controllable simulation under action conditioning. Such a mechanism facilitates long-horizon consistency and supports imagination-based planning and decision making.

Early work represented by *VideoGPT* first demonstrates the feasibility of combining discrete visual tokens with autoregressive prediction. By compressing video frames into discrete tokens via a VQ-VAE and performing next-frame prediction using an autoregressive Transformer in latent space, *VideoGPT* achieves efficient unconditional video generation and preliminary dynamic modeling, laying the foundation for sequence-based video world models. Subsequently, iVideoGPT [73] further integrates generation with planning by introducing actions and rewards as conditional tokens. It constructs a scalable autoregressive Transformer architecture that supports action-conditioned interactive rollouts and is pretrained on millions of robotic trajectories, significantly improving visual planning and model-based reinforcement learning performance. Consequently, autoregressive models not only predict future observations but also support imagination-driven decision making.

Building on this line of work, researchers explicitly strengthen action conditioning to achieve more robust and controllable world evolution, particularly in highly interactive and dynamic scenarios such as autonomous driving. For example, *Vista* [97] incorporates future dynamic priors through a latent replacement mechanism and supports multimodal action control (e.g., trajectories or natural language commands). It enables high-fidelity autoregressive rollouts of up to 15 seconds and substantially outperforms general-purpose video generators on benchmarks such as nuScenes, demonstrating the generalization ability and temporal consistency of the autoregressive framework in complex driving environments. Furthermore, the *Genie* series developed by DeepMind pushes autoregressive modeling

 doi:10.20944/preprints202605.0435.v1

toward longer temporal horizons and stronger interactivity. *Genie* learns interactive environments from internet-scale videos without action annotations by combining spatiotemporal video tokenization, an autoregressive dynamics model, and a latent action model. *Genie 2* scales this approach into a large foundation model capable of real-time 3D environment generation, while *Genie 3* achieves real-time interaction at 720p and 24 fps with consistency over several minutes, bringing autoregressive world models closer to practical world simulators that can be directly invoked by intelligent agents.

Despite these advances, attention-based autoregressive models still face fundamental limitations in computational efficiency and memory capacity when processing long sequences, which hinders sustained interaction beyond minute-level horizons. To address this challenge, Po et al. [98] propose a *Long-Context State-Space Video World Model* that integrates state-space models (SSMs) into the architecture. By leveraging the linear-time complexity of SSMs and combining block-wise scanning with local attention, their approach balances long-term memory retention and frame-level coherence. Experiments on long-horizon interactive datasets demonstrate strong long-term memory retention and efficient inference in practice, providing a scalable solution for ultra-long temporal scenarios.

Beyond improving physical dynamics simulation, another line of research aims to elevate autoregressive video world models toward foundation-model capabilities. This direction focuses on learning transferable abstract rules, reasoning, and planning abilities directly from pure visual data, enabling autoregressive video world models to serve as a core component of intelligent agents. In this context, *VideoWorld* [99] explores large-scale training on unlabeled videos by combining FSQ-quantized codebooks with latent dynamics modeling. It advances autoregressive video world models into general-purpose foundation models through a unified next-token sequence modeling interface that integrates perception, prediction, and planning. Experiments on video-based Go and robotic control tasks demonstrate the model's ability to acquire complex rules, reasoning, and planning behaviors from visual observations alone, marking a significant step toward autoregressive video world models as foundational building blocks for AGI.

5. Benchmarks and Evaluation Protocols

Evaluating Video World Models requires a multi-dimensional approach that disentangles spatial visual quality from temporal dynamics. We categorize the evaluation metrics into three domains: Visual Fidelity, Spatiotemporal Dynamics, and Reconstruction Alignment.

5.1. Benchmarks

The evaluation of video world models has evolved alongside the growing ambition of the field. Early benchmarks primarily focused on action recognition or short-term temporal understanding, such as the Something-Something video database [100], which emphasizes fine-grained motion reasoning through human-object interactions. While influential for temporal modeling, such datasets are limited to discriminative evaluation and do not directly assess generative prediction or environment simulation. Subsequent benchmarks began to target broader video understanding and generation capabilities. *VideoGLUE* [101] introduced a unified evaluation suite spanning multiple video-language and video understanding tasks, reflecting the growing interest in general-purpose video representations. In parallel, benchmarks such as *VBench* [102] shifted attention toward generative evaluation, systematically assessing video generation models across dimensions including visual quality, temporal coherence, motion smoothness, and subject consistency. These benchmarks provide standardized tools for comparing large-scale video generators, but remain primarily focused on perceptual realism rather than environment-level correctness.As video world models increasingly target embodied and interactive settings, newer benchmarks have explicitly exposed the limitations of traditional video generation metrics. *EWMBench* [103] is designed to evaluate embodied world models by emphasizing three critical aspects: visual scene consistency, motion correctness, and semantic alignment with task intent. Its findings reveal that models optimized for perceptual quality often fail to satisfy the structured requirements of embodied tasks, highlighting a mismatch between generic video generation objectives and environment-centric evaluation.

 doi:10.20944/preprints202605.0435.v1

More recent efforts further extend evaluation toward long-horizon prediction and physical plausibility. *EVA-Bench* [104] evaluates embodied video predictors trained under multi-stage paradigms, combining autoregressive rollout with high-fidelity generation. Extensive experiments demonstrate that such predictors can generalize to longer sequences and downstream robotics-related tasks, providing empirical evidence that large-scale pretrained video models can support realistic video prediction in real-world scenarios. Complementary to dataset-based benchmarks, *Fréchet Video Motion Distance* [105] introduces a motion-focused metric that explicitly measures the realism of temporal dynamics, addressing a key blind spot of appearance-driven scores such as FID and FVD.The most recent benchmarks reflect a paradigm shift from video quality assessment toward world-level simulation fidelity. *WorldSimBench* [106] and *WorldScore* [107] aim to evaluate whether generated videos are consistent with underlying environment dynamics, rather than merely visually plausible. These benchmarks assess aspects such as long-horizon consistency, state transition validity, and environment coherence, offering a unified evaluation framework for world generation. Together, they signal a transition from perceptual evaluation to environment-centric benchmarking, aligning evaluation protocols more closely with the core objectives of video world modeling.

5.2. Evaluation Analysis

5.2.1. Video Quality Analysis

Based on the benchmarks shown in Table 1, it is clear that Diffusion-based models such as *GAIA-1* generally outperform others in terms of image quality and motion smoothness, as indicated by their low FID and FVD values. For instance, GAIA-1 achieves an exceptionally low FID of 2.15 and FVD of 12.8, which suggests that it produces visually coherent and temporally consistent videos. This is a key advantage of diffusion models, which excel at generating high-fidelity visual content with smooth motion transitions. However, models like *Vid2World*, also based on diffusion, have lower FID values but suffer from higher FVD, indicating that while individual frames may be of high quality, they struggle with maintaining long-term temporal consistency.In contrast, models incorporating Reinforcement Learning-based alignment methods, such as *CogVidX-2B + HALO*, excel in tasks that require interaction and long-term coherence, particularly for human action generation, which scores a near-perfect 99.00. This model shows strong performance in maintaining motion smoothness and human action coherence, highlighting the advantages of RL-based strategies in ensuring that video sequences remain contextually relevant. However, this is achieved at the expense of slightly lower image quality (61.90), which suggests that while RL alignment enhances action and behavior representation, it may not prioritize pixel-level fidelity as much as diffusion models.Autoregressive models, such as Vista, demonstrate strength in maintaining motion smoothness over extended time horizons, but they tend to sacrifice some image quality compared to the diffusion models. While *Vista* performs exceptionally well in generating temporally coherent sequences, it does not match the FID and FVD results seen in diffusion-based approaches. This suggests that autoregressive models, though powerful in modeling long-term dynamics, may not always achieve the same level of visual fidelity as diffusion-based models, particularly when fine-grained image quality is critical.

 doi:10.20944/preprints202605.0435.v1

Table 1. Success rate (in %) comparison of different paradigm video world models on the Meta-World [108] and Push-T simulation benchmarks.

Method	Paradigm	Meta-World	Push-T
DAML [109]	IL	6	–
T-OSVI [110]	IL	28.5	–
AWDA [110]	IL	42.5	–
OSVI-WM [65]	IL	83.8	–
IRIS [111]	IL	100	–
VILP [64]	IL	–	73.5
DreamerV3 [67]	RL	–	30
Sparse Imagination [112]	RL	–	78.3
DINO-WM [113]	SSL	–	98
DDP-WM [114]	SSL	–	90

5.2.2. Downtask Analysis

The analysis of success rates across different downstream tasks, particularly on the Meta-World and Push-T benchmarks, reveals significant insights into the performance of various models. The best-performing methods in these tasks are notably *DINO-WM* and *DDP-WM*, both utilizing Self-Supervised Learning (SSL) paradigms. *DINO-WM* achieves an impressive 98% success rate on Push-T, while *DDP-WM* follows closely with 90%. These high performance rates can be attributed to the models' ability to leverage pre-trained visual features and self-supervised learning for efficient world modeling and long-term planning, allowing them to generalize well on unseen tasks without requiring additional labeled data. Methods such as *DreamerV3* and Sparse Imagination, both under the Reinforcement Learning (RL) paradigm, show strong performance, with *DreamerV3* reaching 78.3% on *Push-T*. While these models benefit from end-to-end training in the environment, they generally rely on a higher computational cost and more task-specific training data compared to SSL-based approaches, which may explain their relatively lower success rates. Methods based on Imitation Learning (IL), like *DAML*, *T-OSVI*, and *VILP*, exhibit varied success across benchmarks. *OSVI-WM* stands out with 83.8% success rate on *Meta-World*, demonstrating the effectiveness of imitation learning in tasks that require high-level behavior replication, especially when expert demonstrations are available. However, these methods' performance can be highly sensitive to the quality of the demonstration data, which may limit their generalizability compared to self-supervised methods.

doi:10.20944/preprints202605.0435.v1

Table 2. Performance comparison of different paradigm video world models on nuScenes and VBench datasheet, including FID, FVD, Image Quality, Motion Smoothness, Subject Consistency, and Human Action.

Model	Resolution	Paradigm	nuScenes		VBench			
			FID ↓	FVD ↓	Image Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Human Action ↑
GAIA-1 [79]	–	Diffusion	2.15	12.8	–	–	–	–
Sora [80]	–	Diffusion	–	–	–	–	–	–
DIAMOND [81]	–	Diffusion	–	–	–	–	–	–
Astra [95]	–	Diffusion	–	–	–	–	–	–
AVID [82]	–	Diffusion	–	–	–	–	–	–
Vid2World [83]	–	Diffusion	–	–	–	–	–	–
Vidar [84]	–	Diffusion	1.40	385.2	–	–	–	–
PA-VDM [85]	–	Diffusion	–	98.5	–	–	–	–
Geometry Forcing [87]	–	Diffusion	–	–	–	–	–	–
VerseCrafter [88]	–	Diffusion	9.80	165.4	–	–	–	–
DAWM [93]	–	Diffusion	–	–	–	–	–	–
World4RL [92]	–	Diffusion	–	–	–	–	–	–
CWMDT [96]	–	Diffusion	4.20	210.3	–	–	–	–
DriveDreamer [23]	128×192	Diffusion	14.90	340.80	–	–	–	–
GenAD [115]	256×448	Diffusion	15.40	184.00	–	–	–	–
ProphetDWM [116]	256×448	Diffusion	6.90	190.50	–	–	–	–
Epona [86]	512×1024	Diffusion	7.50	82.80	–	–	–	–
MaskGWM [117]	288×512	Diffusion	4.00	59.40	–	–	–	–
LongDWM [118]	480×720	Diffusion	12.30	102.90	–	–	–	–
GeoDrive [119]	480×720	Diffusion	4.10	61.60	–	–	–	–
Vista [97]	576×1024	Diffusion	6.90	89.40	–	–	–	–
VC2 + VideoDPO [120]	–	Diffusion+RL	–	–	–	92.18	95.69	99.00
Turbo + VideoDPO [120]	–	Diffusion+RL	–	–	–	88.85	96.10	94.00
CogVid + VideoDPO [120]	–	Diffusion+RL	–	–	–	88.64	94.67	81.00
Turbo-v1 + [121]	–	Diffusion+RL	–	–	72.07	–	–	95.00
Turbo-v2 + HALO [121]	–	Diffusion+RL	–	–	69.11	–	–	97.60
CogVidX-2B + HALO [121]	–	Diffusion+RL	–	–	61.90	–	–	98.00
GAPO [122]	–	Diffusion+RL	–	–	68.98	99.13	95.20	–
OnlineVPO [123]	–	Diffusion+RL	–	–	67.36	99.36	97.58	–
InstructVideo [124]	–	Diffusion+RL	–	–	70.09	96.76	96.45	–
BEVGen [125]	–	Autoregressive	41.20	–	–	–	–	–
VaViM-s [126]	–	Autoregressive	–	–	–	–	–	–
DrivingWorld [127]	–	Autoregressive	7.40	90.90	–	–	–	–
DriveArena [128]	–	Autoregressive	–	–	–	–	–	–
DreamForge [129]	–	Autoregressive	28.77	197.9	–	–	–	–
Drive-WM [24]	192×384	Autoregressive	15.80	122.70	–	–	–	–
Panacea [130]	256×512	Autoregressive	16.96	139.00	–	–	–	–
CogDriving [131]	480×720	Autoregressive	15.30	37.80	–	–	–	–
DrivePhysica [132]	256×448	Autoregressive	3.96	38.06	–	–	–	–
UniScene [133]	256×512	Autoregressive	6.45	71.94	–	–	–	–
DiST-4D [134]	424×800	Autoregressive	6.83	22.67	–	–	–	–
VisionReward [135]	–	RL	–	–	–	–	–	98.40
RDPO [136]	–	RL	–	–	65.11	99.27	97.04	–
CogVidX-2B + VPO [137]	–	RL	–	–	–	–	–	99.00
CogVidX-5B + VPO [137]	–	RL	–	–	–	–	–	99.60
VADER [138]	–	RL	–	–	66.08	98.89	95.53	–
CogVidX-2B + IPO [139]	–	RL	–	–	62.87	98.17	96.79	–

 doi:10.20944/preprints202605.0435.v1

6. Application Areas

6.1. Autonomous Driving Systems

Autonomous Driving Systems (ADS) enable vehicles to perceive their surroundings, plan trajectories, and execute control actions through onboard sensors (e.g., cameras, LiDAR, and radar), artificial intelligence algorithms, and computing units, without continuous human intervention [140,141]. According to the SAE International standard, automation levels range from L0 (no automation) to L5 (full automation) [142]. Currently, the industry has deployed L4 Robotaxi services commercially (e.g., Waymo operates driverless ride-hailing services in multiple cities), while L2/L2+ advanced driver-assistance systems, such as Tesla Autopilot and Ford BlueCruise, have been widely adopted. However, fully autonomous driving at the L5 level remains unrealized [143]. This limitation primarily arises from three key challenges: data scarcity and long-tail scenario coverage, information loss in modular architectures, and the simulation-to-reality gap.

To address these challenges, video world models offer enhanced generative modeling capabilities and introduce a paradigm shift from data-driven reactive perception to implicit-dynamics-driven predictive simulation. Video world models significantly expand the support for training data distributions through generative modeling, alleviating the scarcity of long-tail scenarios. Representative methods, such as the *DriveDreamer* series [144,145], leverage large language models to provide high-level scene constraints and conditional control, enabling world models to synthesize counterfactual and extreme 4D driving scenarios from real-world videos for systematic data augmentation. *GAIA-1* further generates multi-agent driving trajectories conditioned on textual descriptions and actions, reducing reliance on high risk real-world road testing. Video world models also reshape ADS architecture by learning a unified latent world representation that integrates perception, prediction, and planning into a single dynamic model, mitigating information loss and cascading errors inherent in modular pipelines. *LAW* [146] introduces a self-supervised latent world model, achieving impressive end-to-end trajectory prediction performance on the challenging *nuScenes*, *NAVSIM*, and *CARLA* benchmarks.

Furthermore, video world models support neural closed-loop simulation and counterfactual safety validation, bridging the simulation-to-reality gap. For example, Vista supports multimodal action control and long-horizon prediction (over 40 seconds), simulating millions of kilometers of driving in a neural closed-loop setting and achieving over 27% improvements in FID and FVD metrics. *TrafficBots* [147] and *GenAD* utilize multi-agent simulation and structured latent spaces to evaluate counterfactual branches (e.g., lane-change collision risk), while *OmniNWM* [148] extends this paradigm to panoramic multimodal world modeling for comprehensive safety validation. In summary, video world models present innovative approaches to addressing data scarcity, information loss, and the simulation-to-reality gap in autonomous driving through generative modeling, large-scale data augmentation, unified integration, and neural closed-loop simulation.

6.2. Robotics

Video world models have emerged as powerful embodied world models in robotics, enabling agents to predict environmental dynamics and plan actions effectively. These models learn spatiotemporal mappings that capture the evolution of environments over space and time, providing robots with predictive capabilities crucial for autonomous operation [149]. Recent advances demonstrate that video generation models can serve as cost-effective alternatives to traditional simulators for robotic training, generating synthetic data that captures complex robot environment interactions [150].

One prominent application area is robotic navigation, where *Navigation World Models (NWMs)* predict future visual representations based on past frames and agent actions, enabling long-horizon planning in complex environments [151]. For instance, Lu et al. [152] introduced *GWM* Robotics, which is a specialized world model trained on real-world robotics data that predicts video rollouts conditioned on robot actions, supporting both offline training and online deployment. In manipulation tasks, video world models have shown success in generating realistic robot video data after fine-tuning on target platforms, significantly improving policy generalization across different robotic embodiments.

 doi:10.20944/preprints202605.0435.v1

The integration of foundation models with video world models has unlocked new capabilities in robot learning. Jang et al. [153] proposed *DreamGen*, which leverages video world models to enhance generalization by creating diverse training scenarios through predictive simulation. Moreover, video world models facilitate the development of cloud-based simulation environments that can be deployed on devices for real-time decision making. This capability addresses the sim-to-real gap by enabling robots to simulate action outcomes before execution, reducing the need for extensive real-world trial and error learning. The ability to generate counterfactual scenarios, which involves predicting how environments would evolve under different action sequences, further enhances robotic safety and adaptability in uncertain conditions.

6.3. Embodied Intelligence

Embodied intelligence represents a paradigm where agents perceive complex multimodal environments, act within them, and anticipate how their actions will alter future states [154]. Video world models serve as foundational components for embodied intelligence by functioning as internal simulators that capture environmental dynamics, enabling both forward and counterfactual reasoning capabilities. These models bridge the gap between language-based reasoning and physical world understanding, moving beyond the limitations of pure language intelligence toward comprehensive spatial awareness [155].

Recent papers have systematized world models for embodied AI through a three-axis taxonomy that clarifies their functionality across representation learning [156], predictive modeling, and action-conditioned generation. This unified framework formalizes the problem setting and learning objectives for embodied world models, providing a structured approach to evaluating their performance across robotics, autonomous driving, and general interactive environments. A key advancement in embodied intelligence is the tight coupling between morphology, action, and environment [157], whereby world models enable agents to develop intelligence through continuous interaction with their surroundings. Modern approaches leverage world models to create digital twins that simulate physical interactions, social dynamics, and environmental constraints, providing agents with a cognitive framework for reasoning about complex scenarios [158]. Video world models with long-term spatial memory [159] demonstrated how long-term spatial memory, which integrates geometric constraints across spatial, working, and situational memory types, can significantly enhance embodied agents' ability to navigate and interact with persistent environments. The synergy between large language models (LLMs) and world models represents a crucial frontier in embodied intelligence research. While LLMs excel at abstract reasoning and language understanding, world models provide the spatial and physical grounding necessary for real-world interaction [160]. This integration enables agents to reason about action sequences, predict environmental changes, and adapt strategies based on internal simulations, representing capabilities essential for tasks ranging from household assistance to industrial automation. According to Feng et al. [155], world models are envisioned as the counterpart to LLMs in language, aiming to establish a general-purpose foundation for embodied intelligence that generalizes across domains and tasks.

However, significant challenges remain in scaling world models to handle the complexity of real-world embodied scenarios. Issues such as long-term consistency, multi-agent interaction modeling, and safety guarantees require continued research attention. The establishment of standardized datasets, evaluation metrics, and benchmark environments is essential for advancing the field and enabling embodied world models to move from laboratory demonstrations to real-world applications in smart manufacturing, healthcare, and service robotics.

7. Open Problems and Future Directions

7.1. Towards Mobile Video World Models

Recent video world models often rely on large-scale architectures, such as transformers or diffusion-based models, with billions of parameters [161]. These models achieve high performance in

 doi:10.20944/preprints202605.0435.v1

long-horizon simulation and controllable generation. However, they demand significant computational resources and memory, limiting their deployment on mobile devices, edge platforms, and real-time robotic systems [162,163]. Therefore, developing efficient and lightweight video world models is a crucial research direction. The main challenge lies in the high computational cost of spatiotemporal modeling. Video world models process sequences of high-dimensional visual tokens and often require iterative generation steps. Both autoregressive rollouts and diffusion-based sampling introduce latency [164]. Reducing the model size alone is insufficient if inference remains slow. Efficient design must take into account parameter count, memory footprint, and sampling complexity together. Model compression, knowledge distillation, and structured pruning are promising techniques [165]. Knowledge distillation can transfer temporal reasoning ability from a large teacher model to a smaller student model, while pruning can remove redundant attention heads, temporal layers, or token channels to reduce computational cost, though care must be taken to avoid damaging long-term consistency. Quantization further reduces memory usage and inference latency, but applying it to video world models requires careful handling of temporal accumulation errors. To enable deployment in embodied agents, autonomous systems, and interactive applications, mobile video world models need to balance efficiency with long-horizon simulation fidelity. This challenge remains unresolved and warrants further investigation in future research.

7.2. Towards Long-Horizon Consistency Video World Models

Long-horizon consistency is a critical challenge for video world models. Current diffusion-based and autoregressive frameworks exhibit intrinsic limitations in maintaining coherent environment evolution over extended time horizons. Diffusion-based models generate future frames through iterative denoising, conditioned on short-term context. Although this process achieves high visual fidelity, it lacks mechanisms to maintain consistent state propagation over long rollouts. As the prediction horizon increases, reliance on local visual cues leads to drift in object identity, motion, and scene structure, making it difficult to maintain consistency or plan for the future. Autoregressive models, on the other hand, model temporal causality through step-by-step predictions. However, their sequential nature introduces compounding errors over time. Small inaccuracies early in the sequence accumulate, leading to semantic drift and loss of object consistency, along with implausible dynamics. Furthermore, the computational cost of autoregressive rollouts grows linearly with sequence length, limiting their scalability for long horizon simulations. These limitations suggest that achieving long-horizon consistency will require more than simply scaling diffusion or autoregressive models. Future video world models may require the integration of explicit state abstraction, hierarchical temporal modeling, or memory-augmented dynamics to disentangle long-term evolution from short-term synthesis [166,167]. Addressing this challenge at the latent state transition level, rather than frame generation, is likely essential for building reliable and scalable video world models.

7.3. Towards Physically Grounded Video World Models

Despite significant progress in visual fidelity, current video world models still lack robust physical grounding. Many models can generate visually plausible sequences while violating basic physical principles, such as object permanence, gravity, or collision consistency [168]. This gap reveals a fundamental limitation of purely data-driven video generation when used as a surrogate for environment modeling. Physically grounded video world models require stable internal representations that reflect the underlying structure of the physical world. Rather than learning pixel-level correlations alone, models should encode object-centric states, interaction dynamics, and physically consistent transitions. Incorporating physical priors, structured dynamics modules, or hybrid neural–physical components may help constrain long-term evolution and improve controllability in interactive settings [169,170]. Another open challenge lies in evaluation. Existing benchmarks primarily assess visual realism and short-term temporal coherence, but rarely measure physical correctness [171]. Developing physics-oriented evaluation protocols, including evaluations of object permanence, energy conservation, and

 doi:10.20944/preprints202605.0435.v1

multi-object interactions, is fundamental to the systematic advancement of physically grounded video world models.

8. Conclusion

Video world models have emerged as a unifying framework that bridges video generation, representation learning, and environment simulation. This paper provides a systematic review of the field, organizing existing approaches along learning paradigms and architectural designs, and highlighting the conceptual shift from frame-level generation to environment-level modeling. We reviewed both implicit state deduction and explicit visual modeling approaches, covering reinforcement learning, self-supervised learning, imitation learning, as well as diffusion-based and autoregressive architectures. While recent models demonstrate impressive progress in visual fidelity and short-term coherence, our analysis reveals persistent challenges in physical grounding and long-horizon consistency. In particular, diffusion-based models lack persistent state propagation, whereas autoregressive models suffer from compounding errors and scalability limitations over long rollouts. These limitations indicate that long-term world modeling cannot be achieved by scaling generative models alone. Video world models represent a foundational step toward visual environment understanding and simulation. Continued progress in this direction has the potential to enable more reliable planning, reasoning, and interaction in embodied intelligence, autonomous systems, and unified world models.

References

1. Ha, D.; Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122* **2018**, *2*.
2. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **2021**, *65*, 99–106.
3. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139–1.
4. Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; Wang, X. 4d gaussian splatting for real-time dynamic scene rendering. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 20310–20320.
5. Li, Z.; Niklaus, S.; Snavely, N.; Wang, O. Neural scene flow fields for space-time view synthesis of dynamic scenes. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6498–6508.
6. Zheng, P.; Jiang, J.; Zhang, Y.; Zeng, C.; Qin, C.; Li, Z. CGC-net: A context-guided constrained network for remote-sensing image super resolution. *Remote Sensing* **2023**, *15*, 3171.
7. Zhang, Y.; Zheng, P.; Zeng, C.; Xiao, B.; Li, Z.; Gao, X. Jointly rs image deblurring and super-resolution with adjustable-kernel and multi-domain attention. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *63*, 1–16.
8. Guo, J.; Chen, X.; Xia, Q.; Wang, Z.; Ou, J.; Qin, L.; Yao, S.; Tian, W. HASH-RAG: bridging deep hashing with retriever for efficient, fine retrieval and augmented generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 26847–26858.
9. Cao, S.; Zhang, J.; Zheng, P.; Yan, J.; Qin, C.; Ye, Y.; Dong, W.; Wang, P.; Yang, Y.; Zhang, C. Language-guided token compression with reinforcement learning in large vision-language models. *arXiv preprint arXiv:2603.13394* **2026**.
10. Zheng, P.; Pu, X.; Chen, K.; Huang, J.; Yang, M.; Feng, B.; Ren, Y.; Jiang, J.; Zhang, C.; Yang, Y.; et al. Joint lossless compression and steganography for medical images via large language models. *arXiv preprint arXiv:2508.01782* **2025**.
11. Tewari, A.; Thies, J.; Mildenhall, B.; Srinivasan, P.; Tretschk, E.; Yifan, W.; Lassner, C.; Sitzmann, V.; Martin-Brualla, R.; Lombardi, S.; et al. Advances in neural rendering. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2022, Vol. 41, pp. 703–735.
12. Zhang, K.; Shuang, K.; Yang, X.; Yao, X.; Guo, J. What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 393–409.

 doi:10.20944/preprints202605.0435.v1

13. Wang, J.; Li, Q.; Zheng, P.; Pu, X.; Ren, Y. Chronoselect: Robust learning with noisy labels via dynamics temporal memory. In Proceedings of the Proceedings of the 7th ACM International Conference on Multimedia in Asia, 2025, pp. 1–7.
14. Zhang, J.; Sun, Q.; Zhang, C.; Wang, X.; Huang, Z.; Zhou, Y.; Zheng, P.; Tai, C.l.A.; Bae, S.H.; Ma, Z.; et al. TDA-RC: Task-Driven Alignment for Knowledge-Based Reasoning Chains in Large Language Models. *arXiv preprint arXiv:2604.04942* **2026**.
15. Zheng, P.; Chen, K.; Huang, J.; Chen, B.; Liu, J.; Ren, Y.; Pu, X. Efficient medical image restoration via reliability guided learning in frequency domain. *arXiv e-prints* **2025**, pp. arXiv–2504.
16. Yin, F.; Nie, H.; Pu, X.; Zheng, P.; Zhu, Q.; Ren, Y.; Deng, L. Dual Ontology-enhanced Clinical Decision Learning for First-admission Mortality Prediction. *IEEE Journal of Biomedical and Health Informatics* **2026**.
17. Zhang, J.; Zhang, C.; Chen, S.; Wang, X.; Huang, Z.; Zheng, P.; Yuan, S.; Zheng, S.; Sun, Q.; Zou, J.; et al. Ghs-tda: A synergistic reasoning framework integrating global hypothesis space with topological data analysis. *arXiv e-prints* **2026**, pp. arXiv–2602.
18. Zhang, Y.; Zheng, P.; Jiang, J.; Xiao, P.; Gao, X. FCIR: Rethink aerial image super resolution with Fourier analysis. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
19. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* **2019**.
20. Hafner, D.; Lillicrap, T.; Norouzi, M.; Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193* **2020**.
21. Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; Davidson, J. Learning latent dynamics for planning from pixels. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 2555–2565.
22. Lee, A.X.; Nagabandi, A.; Abbeel, P.; Levine, S. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems* **2020**, *33*, 741–752.
23. Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; Lu, J. Drivedreamer: Towards real-world-drive world models for autonomous driving. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 55–72.
24. Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; Zhang, Z. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14749–14759.
25. Li, X.; He, X.; Zhang, L.; Wu, M.; Li, X.; Liu, Y. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732* **2025**.
26. Ding, J.; Zhang, Y.; Shang, Y.; Zhang, Y.; Zong, Z.; Feng, J.; Yuan, Y.; Su, H.; Li, N.; Sukiennik, N.; et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys* **2025**, *58*, 1–38.
27. Fu, A.; Zhou, Y.; Zhou, T.; Yang, Y.; Gao, B.; Li, Q.; Wu, G.; Shao, L. Exploring the interplay between video generation and world models in autonomous driving: A survey. *arXiv preprint arXiv:2411.02914* **2024**.
28. Zheng, P.; Zhang, C.; Mo, J.; Li, G.; Zhang, J.; Zhang, J.; Cao, S.; Zheng, S.; Qin, C.; Wang, G.; et al. LLaVA-FA: Learning Fourier Approximation for Compressing Large Multimodal Models. *arXiv preprint arXiv:2602.00135* **2026**.
29. Zheng, P.; Chen, K.; Huang, J.; Chen, B.; Liu, J.; Ren, Y.; Pu, X. Lightweight medical image restoration via integrating reliable lesion-semantic driven prior. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 2977–2986.
30. Guo, J.; Shuang, K.; Li, J.; Wang, Z.; Liu, Y. Beyond the granularity: Multi-perspective dialogue collaborative selection for dialogue state tracking. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2320–2332.
31. Zhang, J.; Zhang, C.; Chen, S.; Wang, X.; Huang, Z.; Zheng, P.; Yuan, S.; Zheng, S.; Sun, Q.; Zou, J.; et al. Learning global hypothesis space for enhancing synergistic reasoning chain. *arXiv preprint arXiv:2602.09794* **2026**.
32. Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.C.; et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* **2023**.

 doi:10.20944/preprints202605.0435.v1

33. Parker-Holder, J.; Ball, P.; Bruce, J.; Dasagi, V.; Holsheimer, K.; Kaplanis, C.; Moufarek, A.; Scully, G.; Shar, J.; Shi, J.; et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model* **2024**.
34. Zheng, P.; Zhang, C.; Cui, M.; Chen, G.; Sun, Q.; Huang, J.; Zhang, J.; Kim, T.H.; Qin, C.; Ren, Y.; et al. Towards visual chain-of-thought reasoning: A comprehensive survey **2026**.
35. Guo, J.; Shuang, K.; Zhang, K.; Liu, Y.; Li, J.; Wang, Z. Learning to imagine: distillation-based interactive context exploitation for dialogue state tracking. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 12845–12853.
36. Zheng, P.; Zhang, C.; Wen, Y.; Liu, W.; Sun, Q.; Mo, J.; Zhang, J.; Lee, J.; Kim, T.H.; Liu, K.; et al. Topology-Aware Layer Pruning for Large Vision-Language Models. *arXiv preprint arXiv:2604.16502* **2026**.
37. Zhang, J.; Zhang, C.; Cao, S.; Liu, W.; Zheng, P.; Huang, J.; Qin, C.; Ye, Y.; Dong, W.; Yang, Y. RCP: Representation Consistency Pruner for Mitigating Distribution Shift in Large Vision-Language Models. *arXiv preprint arXiv:2604.04972* **2026**.
38. Zhang, J.; Cao, S.; Zhang, C.; Hong, Z.; Huang, J.; Zheng, P.; Qin, C.; Dong, W.; Yang, Y.; Liu, T. Immunizing 3D Gaussian Generative Models Against Unauthorized Fine-Tuning via Attribute-Space Traps. *arXiv preprint arXiv:2604.09688* **2026**.
39. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Mojtaba.; Komeili.; Muckley, M.; Rizvi, A.; Roberts, C.; et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, 2025, [arXiv:cs.AI/2506.09985].
40. Guan, Y.; Liao, H.; Li, Z.; Hu, J.; Yuan, R.; Zhang, G.; Xu, C. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* **2024**.
41. Feng, T.; Wang, W.; Yang, Y. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260* **2025**.
42. Zhang, P.F.; Cheng, Y.; Sun, X.; Wang, S.; Li, F.; Zhu, L.; Shen, H.T. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097* **2025**.
43. Guo, J.; Shuang, K.; Li, J.; Wang, Z. Dual slot selector via local reliability verification for dialogue state tracking. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 139–151.
44. Liu, Z.; Cao, S.; Zheng, P.; Liu, K.; Qin, C.; Qin, X.; Wei, J.; Zhang, C. Relaxing Anchor-Frame Dominance for Mitigating Hallucinations in Video Large Language Models. *arXiv preprint arXiv:2604.12582* **2026**.
45. Sun, Q.; Zhang, C.; Zhang, J.; Wang, X.; Xie, J.; Zheng, P.; Wang, H.; Lee, S.; Tai, C.l.A.; Yang, Y.; et al. GRASP: Guided Region-Aware Sparse Prompting for Adapting MLLMs to Remote Sensing. *arXiv preprint arXiv:2601.17089* **2026**.
46. Zhang, J.; Zhang, C.; Chen, S.; Huang, Z.; Zheng, P.; Wang, Z.; Guo, P.; Mo, F.; Bae, S.H.; Zou, J.; et al. Lightweight llm agent memory with small language models. *arXiv preprint arXiv:2604.07798* **2026**.
47. Lee, A.X.; Zhang, R.; Ebert, F.; Abbeel, P.; Finn, C.; Levine, S. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523* **2018**.
48. Franceschi, J.Y.; Delasalles, E.; Chen, M.; Lamprier, S.; Gallinari, P. Stochastic latent residual video prediction. In Proceedings of the International Conference on Machine Learning. PMLR, 2020, pp. 3233–3246.
49. Ou, J.; Guo, J.; Jiang, S.; Wang, Z.; Qin, L.; Yao, S.; Tian, W. Accelerating adaptive retrieval augmented generation via instruction-driven representation reduction of retrieval overlaps. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 26983–27000.
50. Wang, X.; Zhu, Z.; Huang, G.; Wang, B.; Chen, X.; Lu, J. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985* **2024**.
51. Hu, J.; Stone, P.; Martín-Martín, R. SLAC: Simulation-Pretrained Latent Action Space for Whole-Body Real-World RL. *arXiv preprint arXiv:2506.04147* **2025**.
52. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. *Advances in neural information processing systems* **2022**, *35*, 8633–8646.
53. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* **2022**.
54. Kaiser, L.; Babaeizadeh, M.; Milos, P.; Osinski, B.; Campbell, R.H.; Czechowski, K.; Erhan, D.; Finn, C.; Kozakowski, P.; Levine, S.; et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374* **2019**.
55. Oh, J.; Singh, S.; Lee, H. Value prediction network. *Advances in neural information processing systems* **2017**, *30*.

 doi:10.20944/preprints202605.0435.v1

56. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609.
57. Hansen, N.; Wang, X.; Su, H. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955* **2022**.
58. Wu, J.; Ma, H.; Deng, C.; Long, M. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *Advances in Neural Information Processing Systems* **2023**, *36*, 39719–39743.
59. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* **2018**.
60. Mendonca, R.; Bahl, S.; Pathak, D. Structured world models from human videos. *arXiv preprint arXiv:2308.10901* **2023**.
61. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15619–15629.
62. Wang, Z.; Wang, K.; Zhao, L.; Stone, P.; Bian, J. Dyn-O: Building Structured World Models with Object-Centric Representations. *arXiv preprint arXiv:2507.03298* **2025**.
63. Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; Kipf, T. Object-centric learning with slot attention. *Advances in neural information processing systems* **2020**, *33*, 11525–11538.
64. Xu, Z.; Qiu, Q.; She, Y. VILP: Imitation Learning with Latent Video Planning. *IEEE Robotics and Automation Letters* **2025**.
65. Goswami, R.G.; Krishnamurthy, P.; LeCun, Y.; Khorrami, F. Osvi-wm: One-shot visual imitation for unseen tasks using world-model-guided trajectory generation. *arXiv preprint arXiv:2505.20425* **2025**.
66. Robine, J.; Uelwer, T.; Harmeling, S. Smaller world models for reinforcement learning. *Neural Processing Letters* **2023**, *55*, 11397–11427.
67. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* **2023**.
68. Agarwal, P.; Andrews, S.; Kahou, S.E. Learning to play atari in a world of tokens. *arXiv preprint arXiv:2406.01361* **2024**.
69. Bruce, J.; Dennis, M.D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. Genie: Generative interactive environments. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
70. Lee, S.; Cho, D.; Park, J.; Kim, H.J. Cqm: Curriculum reinforcement learning with a quantized world model. *Advances in Neural Information Processing Systems* **2023**, *36*, 78824–78845.
71. Scannell, A.; Nakhaei, M.; Kujanpää, K.; Zhao, Y.; Luck, K.S.; Solin, A.; Pajarinen, J. Discrete Codebook World Models for Continuous Control. *arXiv preprint arXiv:2503.00653* **2025**.
72. Pan, M.; Zheng, Y.; Li, J.; Wang, Y.; Yang, X. Video-Enhanced Offline Reinforcement Learning: A Model-Based Approach. *arXiv preprint arXiv:2505.06482* **2025**.
73. Yan, W.; Zhang, Y.; Abbeel, P.; Srinivas, A. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157* **2021**.
74. Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A.G.; Yang, M.H.; Hao, Y.; Essa, I.; et al. Magvit: Masked generative video transformer. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10459–10469.
75. Yu, L.; Lezama, J.; Gundavarapu, N.B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Birodkar, V.; Gupta, A.; Gu, X.; et al. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737* **2023**.
76. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.
77. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **2019**, *32*.
78. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* **2020**.
79. Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* **2023**.

 doi:10.20944/preprints202605.0435.v1

80. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Video generation models as world simulators. *OpenAI Blog* **2024**, *1*, 1.
81. Alonso, E.; Jelley, A.; Micheli, V.; Kanervisto, A.; Storkey, A.J.; Pearce, T.; Fleuret, F. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems* **2024**, *37*, 58757–58791.
82. Rigter, M.; Gupta, T.; Hilmkil, A.; Ma, C. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822* **2024**.
83. Huang, S.; Wu, J.; Zhou, Q.; Miao, S.; Long, M. Vid2World: Crafting Video Diffusion Models to Interactive World Models. *arXiv preprint arXiv:2505.14357* **2025**.
84. Feng, Y.; Tan, H.; Mao, X.; Xiang, C.; Liu, G.; Huang, S.; Su, H.; Zhu, J. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898* **2025**.
85. Xie, D.; Xu, Z.; Hong, Y.; Tan, H.; Liu, D.; Liu, F.; Kaufman, A.; Zhou, Y. Progressive autoregressive video diffusion models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6322–6332.
86. Zhang, K.; Tang, Z.; Hu, X.; Pan, X.; Guo, X.; Liu, Y.; Huang, J.; Yuan, L.; Zhang, Q.; Long, X.X.; et al. Epona: Autoregressive Diffusion World Model for Autonomous Driving. *arXiv preprint arXiv:2506.24113* **2025**.
87. Wu, H.; Wu, D.; He, T.; Guo, J.; Ye, Y.; Duan, Y.; Bian, J. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982* **2025**.
88. Zheng, S.; Yin, M.; Hu, W.; Li, X.; Shan, Y.; Fu, Y. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. *arXiv preprint arXiv:2601.05138* **2026**.
89. Kim, B.; Kim, T.; Lee, J.; Joo, H. Dexterous World Models. *arXiv preprint arXiv:2512.17907* **2025**.
90. Huang, Y.; Zhang, J.; Zou, S.; Liu, X.; Hu, R.; Xu, K. LaDi-WM: A Latent Diffusion-based World Model for Predictive Manipulation. *arXiv preprint arXiv:2505.11528* **2025**.
91. Zhu, C.; Yu, R.; Feng, S.; Burchfiel, B.; Shah, P.; Gupta, A. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792* **2025**.
92. Jiang, Z.; Liu, K.; Qin, Y.; Tian, S.; Zheng, Y.; Zhou, M.; Yu, C.; Li, H.; Zhao, D. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *arXiv preprint arXiv:2509.19080* **2025**.
93. Li, Z.; Han, X.; Li, Y.; Strauss, N.; Schubert, M. DAWM: Diffusion Action World Models for Offline Reinforcement Learning via Action-Inferred Transitions. *arXiv preprint arXiv:2509.19538* **2025**.
94. Won, J.; Lee, K.; Jang, H.; Kim, D.; Shin, J. Dual-stream diffusion for world-model augmented vision-language-action model. *arXiv preprint arXiv:2510.27607* **2025**.
95. Zhu, Y.; Feng, J.; Zheng, W.; Gao, Y.; Tao, X.; Wan, P.; Zhou, J.; Lu, J. Astra: General Interactive World Model with Autoregressive Denoising. *arXiv preprint arXiv:2512.08931* **2025**.
96. Shen, Y.; Maksutova, A.; Li, C.; Unberath, M. Counterfactual World Models via Digital Twin-conditioned Video Diffusion. *arXiv preprint arXiv:2511.17481* **2025**.
97. Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; Li, H. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems* **2024**, *37*, 91560–91596.
98. Po, R.; Nitzan, Y.; Zhang, R.; Chen, B.; Dao, T.; Shechtman, E.; Wetzstein, G.; Huang, X. Long-context state-space video world models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8733–8744.
99. Ren, Z.; Wei, Y.; Guo, X.; Zhao, Y.; Kang, B.; Feng, J.; Jin, X. Videoworld: Exploring knowledge learning from unlabeled videos. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29029–29039.
100. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The" something something" video database for learning and evaluating visual common sense. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 5842–5850.
101. Yuan, L.; Gundavarapu, N.B.; Zhao, L.; Zhou, H.; Cui, Y.; Jiang, L.; Yang, X.; Jia, M.; Weyand, T.; Friedman, L.; et al. Videoglue: Video general understanding evaluation of foundation models. *arXiv preprint arXiv:2307.03166* **2023**.
102. Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21807–21818.

 doi:10.20944/preprints202605.0435.v1

103. Yue, H.; Huang, S.; Liao, Y.; Chen, S.; Zhou, P.; Chen, L.; Yao, M.; Ren, G. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694* **2025**.
104. Chi, X.; Fan, C.K.; Zhang, H.; Qi, X.; Zhang, R.; Chen, A.; Chan, C.m.; Xue, W.; Liu, Q.; Zhang, S.; et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461* **2024**.
105. Liu, J.; Qu, Y.; Yan, Q.; Zeng, X.; Wang, L.; Liao, R. Fr\'echet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos. *arXiv preprint arXiv:2407.16124* **2024**.
106. Qin, Y.; Shi, Z.; Yu, J.; Wang, X.; Zhou, E.; Li, L.; Yin, Z.; Liu, X.; Sheng, L.; Shao, J.; et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072* **2024**.
107. Duan, H.; Yu, H.X.; Chen, S.; Fei-Fei, L.; Wu, J. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983* **2025**.
108. Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Proceedings of the Conference on robot learning. PMLR, 2020, pp. 1094–1100.
109. Yu, T.; Finn, C.; Xie, A.; Dasari, S.; Zhang, T.; Abbeel, P.; Levine, S. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557* **2018**.
110. Chang, M.; Gupta, S. One-shot visual imitation via attributed waypoints and demonstration augmentation. *arXiv preprint arXiv:2302.04856* **2023**.
111. Mandlekar, A.; Ramos, F.; Boots, B.; Savarese, S.; Fei-Fei, L.; Garg, A.; Fox, D. Iris: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 4414–4420.
112. Chun, J.; Jeong, Y.; Kim, T. Sparse Imagination for Efficient Visual World Model Planning. *arXiv preprint arXiv:2506.01392* **2025**.
113. Zhou, G.; Pan, H.; LeCun, Y.; Pinto, L. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983* **2024**.
114. Yin, S.; Yin, K.; Chen, W.; Liu, Y.; Li, G.; Lin, L. DDP-WM: Disentangled Dynamics Prediction for Efficient World Models. *arXiv preprint arXiv:2602.01780* **2026**.
115. Zheng, W.; Song, R.; Guo, X.; Zhang, C.; Chen, L. Genad: Generative end-to-end autonomous driving. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 87–104.
116. Wang, X.; Peng, P. Prophetdwm: A driving world model for rolling out future actions and videos. *arXiv preprint arXiv:2505.18650* **2025**.
117. Ni, J.; Guo, Y.; Liu, Y.; Chen, R.; Lu, L.; Wu, Z. Maskgwm: A generalizable driving world model with video mask reconstruction. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 22381–22391.
118. Wang, X.; Wu, Z.; Peng, P. LongDWM: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546* **2025**.
119. Chen, A.; Zheng, W.; Wang, Y.; Zhang, X.; Zhan, K.; Jia, P.; Keutzer, K.; Zhang, S. Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421* **2025**.
120. Liu, R.; Wu, H.; Zheng, Z.; Wei, C.; He, Y.; Pi, R.; Chen, Q. Videodpo: Omni-preference alignment for video diffusion generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 8009–8019.
121. Wang, S.; Tang, H.; Dou, Z.; Xiong, C. Harness Local Rewards for Global Benefits: Effective Text-to-Video Generation Alignment with Patch-level Reward Models. *arXiv preprint arXiv:2502.06812* **2025**.
122. Zhu, B.; Jiang, Y.; Xu, B.; Yang, S.; Yin, M.; Wu, Y.; Sun, H.; Wu, Z. Aligning anime video generation with human feedback. *arXiv preprint arXiv:2504.10044* **2025**.
123. Zhang, J.; Wu, J.; Chen, W.; Ji, Y.; Xiao, X.; Huang, W.; Han, K. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159* **2024**.
124. Yuan, H.; Zhang, S.; Wang, X.; Wei, Y.; Feng, T.; Pan, Y.; Zhang, Y.; Liu, Z.; Albanie, S.; Ni, D. Instructvideo: Instructing video diffusion models with human feedback. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6463–6474.
125. Swerdlow, A.; Xu, R.; Zhou, B. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters* **2024**, *9*, 3578–3585.
126. Bartoccioni, F.; Ramzi, E.; Besnier, V.; Venkataramanan, S.; Vu, T.H.; Xu, Y.; Chambon, L.; Gidaris, S.; Odabas, S.; Hurych, D.; et al. Vavim and vavam: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672* **2025**.

 doi:10.20944/preprints202605.0435.v1

127. Hu, X.; Yin, W.; Jia, M.; Deng, J.; Guo, X.; Zhang, Q.; Long, X.; Tan, P. DrivingWorld: Constructing world model for autonomous driving via video GPT. *arXiv preprint arXiv:2412.19505* **2024**.
128. Yang, X.; Wen, L.; Wei, T.; Ma, Y.; Mei, J.; Li, X.; Lei, W.; Fu, D.; Cai, P.; Dou, M.; et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 26933–26943.
129. Mei, J.; Hu, T.; Yang, X.; Wen, L.; Yang, Y.; Wei, T.; Ma, Y.; Dou, M.; Shi, B.; Liu, Y. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003* **2024**.
130. Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; Zhang, X. Panacea: Panoramic and controllable video generation for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6902–6912.
131. Lu, H.; Wu, X.; Wang, S.; Qin, X.; Zhang, X.; Han, J.; Zuo, W.; Tao, J. Seeing beyond views: Multi-view driving scene video generation with holistic attention. *arXiv preprint arXiv:2412.03520* **2024**.
132. Yang, Z.; Guo, X.; Ding, C.; Wang, C.; Wu, W. Physical informed driving world model. *arXiv preprint arXiv:2412.08410* **2024**.
133. Li, B.; Guo, J.; Liu, H.; Zou, Y.; Ding, Y.; Chen, X.; Zhu, H.; Tan, F.; Zhang, C.; Wang, T.; et al. Uniscene: Unified occupancy-centric driving scene generation. In Proceedings of the Proceedings of the computer vision and pattern recognition conference, 2025, pp. 11971–11981.
134. Guo, J.; Ding, Y.; Chen, X.; Chen, S.; Li, B.; Zou, Y.; Lyu, X.; Tan, F.; Qi, X.; Li, Z.; et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27231–27241.
135. Xu, J.; Huang, Y.; Cheng, J.; Yang, Y.; Xu, J.; Wang, Y.; Duan, W.; Yang, S.; Jin, Q.; Li, S.; et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059* **2024**.
136. Qian, W.; Wang, C.; Peng, H.; Tan, Z.; Li, H.; Zeng, A. Rdpo: Real data preference optimization for physics consistency video generation. *arXiv preprint arXiv:2506.18655* **2025**.
137. Cheng, J.; Lyu, R.; Gu, X.; Liu, X.; Xu, J.; Lu, Y.; Teng, J.; Yang, Z.; Dong, Y.; Tang, J.; et al. Vpo: Aligning text-to-video generation models with prompt optimization. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 15636–15645.
138. Prabhudesai, M.; Mendonca, R.; Qin, Z.; Fragkiadaki, K.; Pathak, D. VADER: Video Diffusion Alignment via Reward Gradients.
139. Yang, X.; Tan, Z.; Li, H. Ipo: Iterative preference optimization for text-to-video generation. *arXiv preprint arXiv:2502.02088* **2025**.
140. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access* **2020**, *8*, 58443–58469.
141. Badue, C.; Guidolini, R.; Carneiro, R.V.; Azevedo, P.; Cardoso, V.B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T.M.; Mutz, F.; et al. Self-driving cars: A survey. *Expert systems with applications* **2021**, *165*, 113816.
142. Committee, O.R.A.D.O. *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*; SAE international, 2021.
143. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.
144. Zhao, G.; Wang, X.; Zhu, Z.; Chen, X.; Huang, G.; Bao, X.; Wang, X. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 10412–10420.
145. Zhao, G.; Ni, C.; Wang, X.; Zhu, Z.; Zhang, X.; Wang, Y.; Huang, G.; Chen, X.; Wang, B.; Zhang, Y.; et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 12015–12026.
146. Li, Y.; Fan, L.; He, J.; Wang, Y.; Chen, Y.; Zhang, Z.; Tan, T. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481* **2024**.
147. Zhang, Z.; Liniger, A.; Dai, D.; Yu, F.; Van Gool, L. Trafficbots: Towards world models for autonomous driving simulation and motion prediction. *arXiv preprint arXiv:2303.04116* **2023**.
148. Li, B.; Ma, Z.; Du, D.; Peng, B.; Liang, Z.; Liu, Z.; Ma, C.; Jin, Y.; Zhao, H.; Zeng, W.; et al. OmniNWM: Omniscient Driving Navigation World Models. *arXiv preprint arXiv:2510.18313* **2025**.

 doi:10.20944/preprints202605.0435.v1

149. Mei, Z.; Yin, T.; Shorinwa, O.; Badithela, A.; Zheng, Z.; Bruno, J.; Bland, M.; Zha, L.; Hancock, A.; Fisac, J.F.; et al. Video Generation Models in Robotics-Applications, Research Challenges, Future Directions. *arXiv preprint arXiv:2601.07823* **2026**.
150. Kawaharazuka, K.; Matsushima, T.; Gambardella, A.; Guo, J.; Paxton, C.; Zeng, A. Real-world robot applications of foundation models: A review. *Advanced Robotics* **2024**, *38*, 1232–1254.
151. Bar, A.; Zhou, G.; Tran, D.; Darrell, T.; LeCun, Y. Navigation world models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15791–15801.
152. Lu, G.; Jia, B.; Li, P.; Chen, Y.; Wang, Z.; Tang, Y.; Huang, S. Gwm: Towards scalable gaussian world models for robotic manipulation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 9263–9274.
153. Jang, J.; Ye, S.; Lin, Z.; Xiang, J.; Bjorck, J.; Fang, Y.; Hu, F.; Huang, S.; Kundalia, K.; Lin, Y.C.; et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705* **2025**.
154. Sun, F.; Chen, R.; Ji, T.; Luo, Y.; Zhou, H.; Liu, H. A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. *CAAI Artificial Intelligence Research* **2024**, *3*, 1.
155. Feng, T.; Wang, X.; Jiang, Y.G.; Zhu, W. Embodied ai: From llms to world models. *arXiv preprint arXiv:2509.20021* **2025**.
156. Zhang, Y.; Tian, J.; Xiong, Q. A review of embodied intelligence systems: a three-layer framework integrating multimodal perception, world modeling, and structured strategies. *Frontiers in Robotics and AI* **2025**, *12*, 1668910.
157. Liu, H.; Guo, D.; Cangelosi, A. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys* **2025**, *57*, 1–36.
158. Fung, P.; Bachrach, Y.; Celikyilmaz, A.; Chaudhuri, K.; Chen, D.; Chung, W.; Dupoux, E.; Gong, H.; Jégou, H.; Lazaric, A.; et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355* **2025**.
159. Wu, T.; Yang, S.; Po, R.; Xu, Y.; Liu, Z.; Lin, D.; Wetzstein, G. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284* **2025**.
160. Long, X.; Zhao, Q.; Zhang, K.; Zhang, Z.; Wang, D.; Liu, Y.; Shu, Z.; Lu, Y.; Wang, S.; Wei, X.; et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917* **2025**.
161. Karnewar, A.; Korzhenkov, D.; Lelekas, I.; Karjauv, A.; Fathima, N.; Xiong, H.; Vaidyanathan, V.; Zeng, W.; Esteves, R.; Singhal, T.; et al. Neodragon: Mobile Video Generation using Diffusion Transformer. *arXiv preprint arXiv:2511.06055* **2025**.
162. Yahia, H.B.; Korzhenkov, D.; Lelekas, I.; Ghodrati, A.; Habibian, A. Mobile video diffusion. *arXiv preprint arXiv:2412.07583* **2024**.
163. Ahmad, S.; Hafeez, M.; Zaidi, S.A.R. Vision-Language Models on the Edge for Real-Time Robotic Perception. *arXiv preprint arXiv:2601.14921* **2026**.
164. Zhang, J.; Zheng, K.; Jiang, K.; Wang, H.; Stoica, I.; Gonzalez, J.E.; Chen, J.; Zhu, J. TurboDiffusion: Accelerating Video Diffusion Models by 100-200 Times. *arXiv preprint arXiv:2512.16093* **2025**.
165. Wu, Y.; Zhang, Z.; Li, Y.; Xu, Y.; Kag, A.; Sui, Y.; Coskun, H.; Ma, K.; Lebedev, A.; Hu, J.; et al. Snapgen-v: Generating a five-second video within five seconds on a mobile device. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 2479–2490.
166. Yu, J.; Bai, J.; Qin, Y.; Liu, Q.; Wang, X.; Wan, P.; Zhang, D.; Liu, X. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In Proceedings of the Proceedings of the SIGGRAPH Asia 2025 Conference Papers, 2025, pp. 1–11.
167. Liu, Z.; Deng, X.; Chen, S.; Wang, A.; Guo, Q.; Han, M.; Xue, Z.; Chen, M.; Luo, P.; Yang, L. Worldweaver: Generating long-horizon video worlds via rich perception. *arXiv preprint arXiv:2508.15720* **2025**.
168. Gao, Z.; Li, Z.; Wang, X.; Huang, J.; Ren, Z.; Shao, M.; Zhang, H.; Huang, T.; Cheng, Y.; Guo, Y.; et al. Mirage2Matter: A Physically Grounded Gaussian World Model from Video. *arXiv preprint arXiv:2602.00096* **2026**.
169. Wang, C.; Chen, C.; Huang, Y.; Dou, Z.; Liu, Y.; Gu, J.; Liu, L. Physctrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358* **2025**.
170. Yang, Y.; Zhang, Z.; Zhang, X.; Zeng, Y.; Li, H.; Zuo, W. Physworld: From real videos to world models of deformable objects via physics-aware demonstration synthesis. *arXiv preprint arXiv:2510.21447* **2025**.
171. He, X.; Feng, W.; Zheng, K.; Lu, Y.; Zhu, W.; Li, J.; Fan, Y.; Wang, J.; Li, L.; Yang, Z.; et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407* **2024**.

 doi:10.20944/preprints202605.0435.v1