

World Model for Robot Learning: A Comprehensive Survey

Bohan Hou^{1,*,\dagger}, Gen Li^{1,*}, Jindou Jia^{1,*}, Tuo An^{1,*}, Xinying Guo^{1,*}, Sicong Leng¹, Haoran Geng², Yanjie Ze³, Tatsuya Harada⁴, Philip Torr⁵, Oier Mees⁶, Marc Pollefeys⁷, Zhuang Liu⁸, Jiajun Wu³, Pieter Abbeel², Jitendra Malik², Yilun Du⁹, Jianfei Yang^{1,\dagger}

¹Nanyang Technological University, ²University of California, Berkeley, ³Stanford University, ⁴The University of Tokyo, ⁵University of Oxford, ⁶Microsoft, ⁷ETH Zurich, ⁸Princeton University, ⁹Harvard University

*Equal Contribution (alphabetical order), †Corresponding Author

World models, which are predictive representations of how environments evolve under actions, have become a central component of robot learning. They support policy learning, planning, simulation, evaluation, data generation, and have advanced rapidly with the rise of foundation models and large-scale video generation. However, the literature remains fragmented across architectures, functional roles, and embodied application domains. To address this gap, we present a comprehensive review of world models from a robot-learning perspective. We examine how world models are coupled with robot policies, how they serve as learned simulators for reinforcement learning and evaluation, and how robotic video world models have progressed from imagination-based generation to controllable, structured, and foundation-scale formulations. We further connect these ideas to navigation and autonomous driving, and summarize representative datasets, benchmarks, and evaluation protocols. Overall, this survey systematically reviews the rapidly growing literature on world models for robot learning, clarifies key paradigms and applications, and highlights major challenges and future directions for predictive modeling in embodied agents. To facilitate continued access to newly emerging works, benchmarks, and resources, we will maintain and regularly update the accompanying GitHub repository alongside this survey.



<https://github.com/NTUMARS/Awesome-World-Model-for-Robotics-Policy>



<https://ntumars.github.io/wm-robot-survey/>

Correspondence: Jianfei Yang at jianfei.yang@ntu.edu.sg;

1 Introduction

Robotic policy learning is rapidly shifting from task-specific control pipelines toward foundation-model-driven embodied intelligence. Recent Vision-Language-Action (VLA) (Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2024; Intelligence et al., 2025b; Wu et al., 2024) policies aim to unify perception, language understanding, and control by mapping multimodal observations directly to robot actions, promising broad task generalization and flexible instruction following. Yet despite strong scaling trends (Xiao et al., 2025; Li et al., 2025b; Zhu et al., 2026), purely reactive VLA policies remain limited in complex physical environments, where they often struggle with long-horizon reasoning, temporal credit assignment, and robustness under compounding errors. A growing body of work argues that these limitations stem not only from insufficient action prediction capacity (Ye et al., 2026b; Dang et al., 2026), but also from the lack of explicit predictive structure for anticipating how the world may evolve under the agent’s behavior. This has renewed interest in world models (Craik, 1943; Bryson and Ho, 1975; Ha and Schmidhuber, 2018), predictive representations that capture environmental dynamics and enable reasoning about future states before acting.

The term *world model* (Craik, 1943; Bryson and Ho, 1975; Ha and Schmidhuber, 2018) has a long intellectual lineage. At its core, it describes how a system or environment evolves from its current state under intervention or action, and in its most standard form can be viewed as a state-transition model that predicts the next

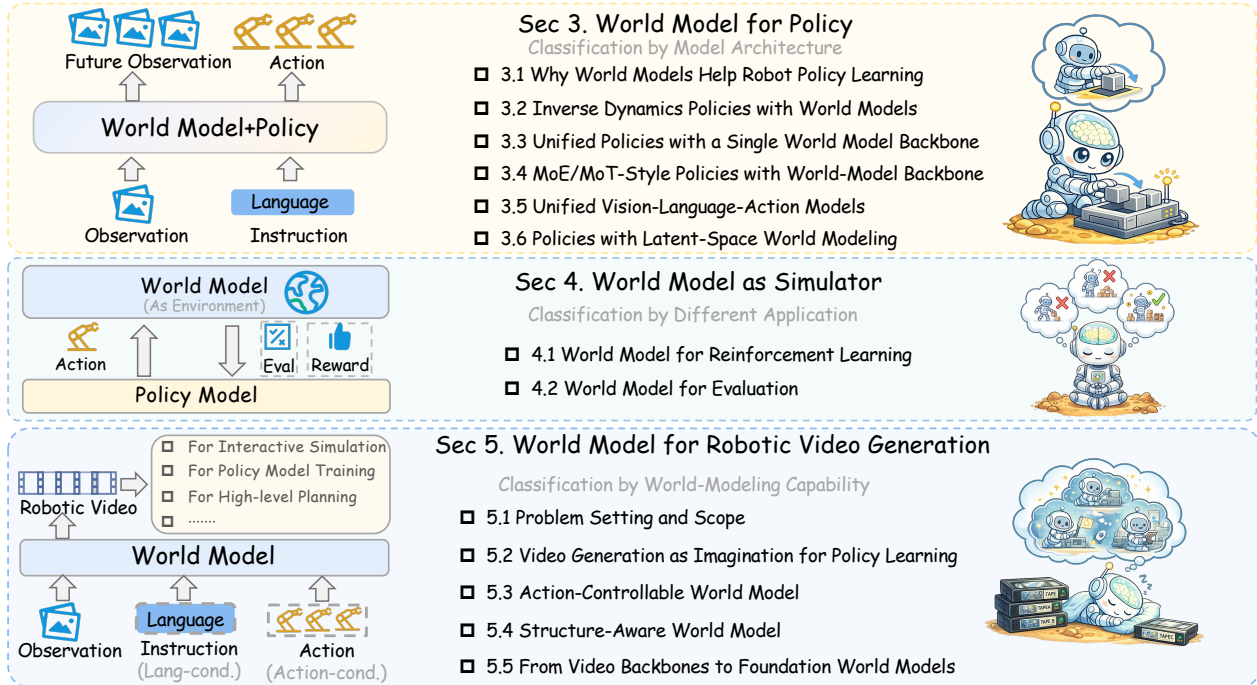


Figure 1 Overview of the organization of this survey. **Section 3** reviews how world models are coupled with robot policies from an architectural perspective. **Section 4** examines world models as simulators from an application perspective. **Section 5** focuses on robotic video world models and organizes the literature by world-modeling capability.

state or a sequence of future states from the current state and action. Early ideas emerged in cognitive science during the 1960s (Miller et al., 1960), where internal models were proposed to support mental simulation, prediction, and planning. Similar ideas also appeared in control theory and model-based decision-making (Conant and Ashby, 1970; Bryson and Ho, 1975; Richalet et al., 1978), and in classical robot planning, where internal models of geometry, constraints, and action consequences are used to support decision making before execution (Lozano-Perez, 1983). In modern machine learning, the resurgence of world models is driven mainly by two lines of progress (Ha and Schmidhuber, 2018): model-based reinforcement learning (Nguyen and Widrow, 1990; Jiang et al., 2026; Zhu et al., 2026), which uses learned dynamics for planning and policy improvement, and large-scale generative modeling (Ali et al., 2025; Guo et al., 2025; Jiang et al., 2025b; Jang et al., 2025b), especially video generation, which learns rich spatiotemporal regularities from large-scale visual or interaction data. Together, these developments make it increasingly plausible to learn predictive representations directly from pixels and reuse them for embodied decision making.

In this survey, rather than enforcing a single narrow formal definition, we take a robot-learning-centered view of world models. Our focus is on how predictive models of future world evolution support robotic policy learning, planning, simulation, evaluation, and data generation. Under this view, world models may support action selection through explicit rollout, future-conditioned action inference, or joint predictive-control modeling. What unifies them is not a single factorization, but their role as predictive structures that make robot decision-making more informed and physically grounded. We also use the notion of action in a broad predictive-control sense: *low-level motor commands* specify how the agent moves, while *high-level language instructions* specify what the future should be realized. This perspective also distinguishes robotic world models from generic perceptual predictors: in embodied AI, predictive quality matters only insofar as it is useful for action. Accordingly, an actionable world model should provide three core capabilities: **foresight** (Mi et al., 2026; Li et al., 2026b; Gu et al., 2026; Bi et al., 2025), i.e., anticipating future states or action consequences before execution; **imagination-driven planning** (Kim et al., 2026), i.e., using imagined rollouts to compare and select candidate behaviors; and **data amplification** (Jang et al., 2025b; Ali et al., 2025), i.e., synthesizing additional demonstrations or interaction trajectories to improve learning. These capabilities are especially important for embodied tasks such as manipulation, navigation, and driving, where success depends

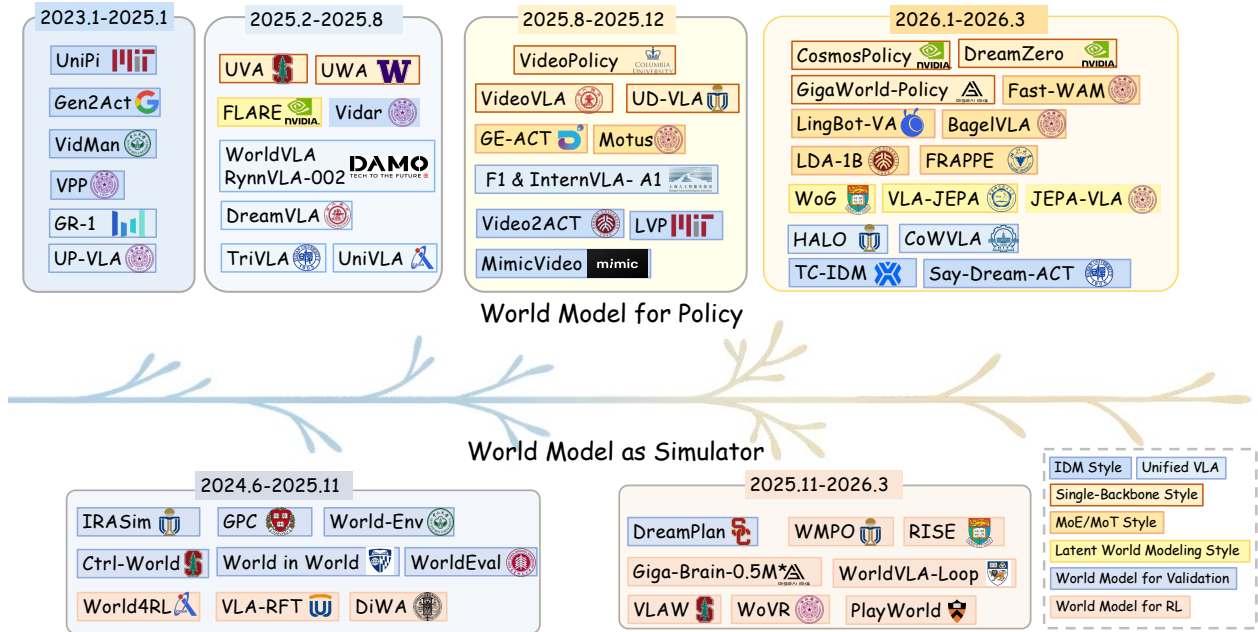


Figure 2 Temporal evolution of representative works on world models for robotic policy learning. The upper branch summarizes the progression of **world model for policy methods**, showing a trend from early decoupled video-generation-plus-IDM pipelines toward tighter integration through single-backbone, MoE/MoT, unified VLA, and latent world-modeling designs. The lower branch summarizes the progression of **world model as simulator methods**, where world models evolve from rollout-based validation and candidate evaluation into learned simulators for policy reinforcement learning, post-training, and co-evolving optimization. These trends should be understood as dominant directions rather than strictly sequential replacements.

on reasoning about contact, dynamics, and other physical regularities that language-centric pretraining alone does not capture. In this sense, world models are not merely a generative enhancement, but a predictive bridge from semantic intent to physically realizable behavior.

Historically, the integration of world models into robotic policies has evolved along two directions: tighter coupling between predictive modeling and action generation (Du et al., 2023; Li et al., 2025c; Zhu et al., 2025a), and broader use of learned world models as simulators for validation, post-training, and reinforcement learning (Xiao et al., 2025; Li et al., 2025b; Chandra et al., 2025). With the rise of foundation-scale video models (Wan, 2025; Ali et al., 2025), recent methods explore adapting large video generators into robot policies (Li et al., 2025c; Zhu et al., 2025a), aiming to improve generalization and sample efficiency through future prediction (Jang et al., 2025b), while later systems move toward unified training and closed-loop co-optimization with VLA policies (Cen et al., 2025). In parallel, world models are increasingly used as controllable simulators for post-training and evaluation (Zhu et al., 2026; Xiao et al., 2025), highlighting that the key objective is not only to generate plausible futures, but to generate control-consistent futures that support decision-making. Motivated by these trends, our survey differs from prior surveys (Zhang et al., 2025d) in three main respects: it offers a more fine-grained view of major world-model paradigms, a more comprehensive analysis of their roles across policy learning, planning, simulation, evaluation, and video generation, and a clearer robotics-centered definition of world models in relation to VLA policies and robot learning. By emphasizing action-conditioned consistency, long-horizon reliability, and practical deployability, this survey aims to clarify when and why world models translate into measurable gains in real robotic behavior.

We first introduce background on world models, video generation, and VLA/policy models in Sec. 2. As summarized in Fig. 1, we then review world models for policy in Sec. 3, world models as simulators in Sec. 4, and robotic video world models in Sec. 5. We further discuss broader embodied domains including navigation and autonomous driving in Sec. 6, and present benchmarks, datasets and results in Sec. 7, before concluding with open challenges and future directions in Sec. 8. In particular, Sec. 3 first introduces a probabilistic lens

that connects policy models, passive and controllable world models, and inverse-dynamics models as related queries of a shared predictive-control distribution.

Figure 2 highlights two closely related trends in the recent literature. On the policy side, early decoupled pipelines (Hu et al., 2025; Du et al., 2023) remain an important line, while the design space has progressively expanded toward single-backbone (Kim et al., 2026), unified VLA (Cen et al., 2025), and latent world-modeling (Su et al., 2026) approaches with tighter integration between prediction and action generation. On the simulator side, their roles have expanded from validating or ranking candidate actions based on imagined futures to serving as learned environments for reinforcement learning, post-training, and even co-evolution with policies (Li et al., 2025b; Guo et al., 2026a; Liu et al., 2026b). Taken together, these two trends indicate that world models are no longer used only as auxiliary predictors, but are increasingly integrated into the core learning and decision-making loop of robotic systems. To complement this survey, we will also continuously maintain and update the accompanying GitHub repository so that it remains aligned with the fast-moving progress of the field.

In summary, our main contributions are as follows:

- We present a policy-centric survey of world models for robot learning, with a particular focus on how predictive models are coupled with VLA policies to support action generation, planning, simulation, evaluation, and data generation.
- We provide a more fine-grained taxonomy of the field by distinguishing major architectural paradigms and functional roles of world models, revealing important differences that are often overlooked in broader discussions.
- We offer a more comprehensive and clearly defined treatment of robotic world models by clarifying their relationship to robot learning, VLA policies, video generation, and simulator-style usage, and by summarizing representative benchmarks, datasets, and open challenges.

2 Background

2.1 World Model and Video Generation Model

To establish a precise vocabulary for the remainder of this survey, we first clarify two closely related concepts used throughout the paper. In recent embodied AI literature, the term world model has been used rather broadly, referring to latent dynamics models, future state predictors, video predictors, and even implicit predictive structures inside large policies. Since our focus is policy-centric rather than purely generative, we use these terms in a more precise and functional sense.

2.1.1 World Model

In this survey, we use the term world model in a robotics- and embodiment-centered sense, rather than in the broadest possible generative sense. Concretely, a world model refers to a predictive model of agent-environment dynamics that captures how a robotic or embodied system evolves under actions. In its most standard form, it models a state-transition process: given the current state or observation together with an action, it predicts the next state or a sequence of future states as illustrated in Fig. 1 bottom.

Here we use the notion of action in a broad predictive-control sense. That is, both low-level motor commands and high-level language instructions are treated as *actions*: the former are concrete physical actions executed by the agent, while the latter are high-level semantic actions that specify what the future should be realized. For notational consistency with the rest of this survey, we keep these two forms of action separate, denoting low-level physical actions by a and high-level language or task actions by l . Under this convention, a general formulation can be written as

$$p(x_{t+1:t+H} \mid x_t, a_{t:t+H-1}, l), \quad (1)$$

where x_t denotes the modeled state at time t , $a_{t:t+H-1}$ denotes an action sequence over a horizon H , and l denotes the high-level action specification, such as a language instruction or goal description. This formulation

is intentionally agnostic to the choice of state space. What matters in our setting is whether the predicted futures are actionable for downstream embodied decision making.

Under this formulation, we use *world model* in a functional sense to refer to predictive models whose outputs support policy-related computation, including control, planning, simulation, evaluation, and data generation. Its defining property is not merely to predict a plausible future, but to predict how the future changes under robot-relevant actions in a way that supports embodied decision making. This definition is therefore narrower than generic future prediction in computer vision: a model does not qualify as a world model in our sense simply because it generates plausible future images or videos. Rather, it must capture environment evolution in a form relevant to robot interaction and useful for downstream policy-related computation. In embodied control, the most important subclass is the action-conditioned world model, since visually plausible but action-inconsistent futures offer limited value for closed-loop decision making. Depending on the method, the modeled variable x_t may be a visual observation, latent state, structured physical state, or even an abstract symbolic state used for planning (Liang et al., 2026, 2025c; Athalye et al., 2026; Liang et al., 2026), covering both classical latent dynamics models and newer generative predictive models for robot learning. In the symbolic case, the world model predicts transitions over predicates, object relations, affordances, or causal processes rather than over pixels (Liang et al., 2025c; Athalye et al., 2026; Liang et al., 2026).

In current embodied systems, however, the most common and scalable realization of state is precisely an observation stream, especially a visual observation sequence. For this reason, many practical world models in robotics are instantiated directly in visual observation space. Accordingly, although *world model* is the more general concept, the concrete models of primary interest in this survey are predominantly visual world models, i.e., video generation models defined over future observations.

2.1.2 Video Generation Model

A video generation model predicts the future directly in image or video space. In the embodied setting, it can be written as

$$p(v_{t+1:t+H} \mid o_t, a_{t:t+H-1}, l), \quad (2)$$

where o_t denotes the current observation, it can represent observations from multiple perspectives, and $v_{t+1:t+H}$ denotes future frames or video segments. Compared with latent-state world models, this formulation preserves richer spatial, temporal, and interaction details, since the future is represented explicitly as visual evidence rather than abstract state variables. From the perspective above, such a model can be understood as a world model instantiated in visual observation space. Because visual observation is the most common form of state available to embodied agents, this visual instantiation is also the dominant one considered throughout this survey. This focus should not be read as assuming that pixel-level prediction is the optimal abstraction for control; rather, it reflects the prominence of video-based world models in the recent robot-learning literature.

This visual explicitness, however, also makes the modeling problem substantially more demanding. Beyond perceptual realism, an embodied video generation model must maintain temporal coherence, action consistency, physical plausibility, and long-horizon stability. Recent advances in large-scale video generative backbones have made such modeling increasingly viable in robotics (Yang et al., 2024b). As a result, video generation models are no longer used only for passive visual continuation. They are increasingly adapted into action-conditioned predictive modules that support imagination-based supervision, controllable rollout, simulator construction, and synthetic data generation for robot learning (Liang et al., 2024; Zhou et al., 2024; Pai et al., 2025; Zhu et al., 2025b; Guo et al., 2026b; Huang et al., 2026; Liao et al., 2026).

Among them, action-conditioned video generation models occupy a particularly important place in embodied AI. Here, the notion of action should be understood broadly: conditioning may come from low-level continuous controls, but also from higher-level task or language descriptions that specify what the future should be realized. Under both forms, these models inherit the expressive power of video prediction while modeling how the visual future changes as a consequence of candidate actions. This makes them especially suitable for the policy-centric setting of this survey: they can serve not only as generators of plausible futures, but also as predictive substrates for control, planning, and policy improvement. Therefore, unless otherwise specified, the world models discussed in the remainder of this survey are predominantly video-based world models, with special emphasis on the action-conditioned case.

2.2 Robot Policy

State-of-the-art robot control methods have shifted from analytical controllers to end-to-end learning models (Ai et al., 2025). Formally, the robot policy is a decision-making model that frames physical control as an action prediction task, mapping current environmental observations to future action trajectories. Here, we specifically focus on the imitation learning paradigm, where policies are trained to synthesize behaviors directly from expert demonstrations.

Given the current observation o_t (including visual and proprioceptive states) and an *optional* language instruction l , the policy predicts future action sequences $a_{t+1:t+k}$. This process is typically modeled as the following conditional probability distribution:

$$p(a_{t+1:t+k}|o_t, l). \quad (3)$$

In practice, structuring predicted actions as temporal chunks with length k has emerged as a predominant strategy to ensure temporal coherence and mitigate compounding errors (Chi et al., 2023; Zhao et al., 2023; Wu et al., 2026b). From an architectural perspective, contemporary robot policies are primarily bifurcating into two paradigms: specialized visuomotor policies and generalist Vision-Language-Action (VLA) models. The former, represented by frameworks like Diffusion Policy (Chi et al., 2023, 2025a; Dasari et al., 2025), focuses on training task-specific, often lightweight, end-to-end networks that leverage generative modeling to capture complex action distributions with high precision and low latency. Conversely, VLA models, pioneered by RT-2 (Zitkovich et al., 2023), OpenVLA (Kim et al., 2025), and π_0 (Black et al., 2024), are developed by fine-tuning large-scale Vision-Language Models (VLMs) on large scale robotic trajectory data (Open X-Embodiment Collaboration, 2024), thereby inheriting the vast semantic knowledge and open-vocabulary reasoning capabilities of foundational models to achieve superior cross-task (Octo Model Team et al., 2024) and cross-embodiment generalization (Doshi et al., 2024).

2.2.1 Visuomotor Policy

Visuomotor policies establish a direct mapping from raw states to the action space, resulting in a generally lightweight yet generalization-bounded architecture. The most straightforward approach formulates this mapping as a *regression* task (Bain and Sammut, 1995; Osa et al., 2018; Zhao et al., 2023). In this paradigm, neural networks encode the current observation and directly regress the continuous physical action values deterministically.

To address the inherent multi-modality of human demonstrations, recent visuomotor policies have increasingly adopted *generative* models. These approaches capture the full action distribution using generative techniques, such as Diffusion Policy (Chi et al., 2023, 2025a) based on diffusion models (Ho et al., 2020; Song et al., 2021), and flow matching (Zhang and Gienger, 2024; Lipman et al., 2023; Liu, 2022). By framing action prediction as a conditional generation process, these models can synthesize high-fidelity, multimodal action sequences starting from initial Gaussian noise. Furthermore, to enhance sampling efficiency and accelerate the generation process, recent advancements have explored replacing the standard Gaussian noise with more informative base distributions, such as visual representations (Gao et al., 2025a) and action history (Jia et al., 2026b). Pan et al. (2026) found that generative policies outperform regression by improving manifold adherence through stochasticity injection and supervised iterative computation.

2.2.2 Vision-Language-Action Policy

To leverage the powerful reasoning capabilities of VLM, VLA models typically equip the pre-trained backbone with a dedicated action head, co-fine-tuning the entire framework on robotic trajectory data. In VLA models, action prediction mainly integrates discrete and continuous representation paradigms.

On one hand, *discrete action tokenization* quantizes continuous actions into tokens that reside within the same vocabulary space as the language model, directly utilizing the next-token prediction ability of VLM (Liu et al., 2025a), successfully exemplified by RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2025). While standard binning-based discretization can struggle with high-frequency control, FAST (Pertsch et al., 2025) introduces Frequency-space Action Sequence Tokenization (FAST), which uses the discrete cosine transform (DCT) to compress action chunks into a dense token sequence. This enables autoregressive VLAs to handle

highly dexterous tasks with the precision of generative models while significantly reducing training time. Its imitation-learning objective is identical to the standard negative log-likelihood loss.

On the other hand, to overcome quantization errors and maintain precision in high-frequency control, *continuous action representation* has emerged as a promising alternative. This approach usually treats the action head as a conditional generator, learning a probabilistic generative model, such as diffusion models or flow matching, successfully exemplified by the π family (Black et al., 2024; Intelligence et al., 2025a,b). Instead of predicting deterministic values, these generative formulations model the full multimodal distribution of human demonstrations.

3 World Model for Policy

The upper branch of Fig. 2 places existing policy-coupling paradigms in a broader temporal context. The field has gradually moved from decoupled predict-then-act pipelines to more unified and internalized forms of predictive control. Importantly, this progression should not be read as implying that video-pretrained backbones are inherently superior to VLM, latent, structured, or symbolic alternatives for control. Which predictive substrate is most effective remains an open empirical question; our focus here is to organize the rapidly growing family of methods that couple visual or video-based world-model priors with robot policies, while also highlighting latent and structured variants where they arise.

3.1 Why World Models Help Robot Policy Learning

Recent robotics policies increasingly incorporate world models, often instantiated as video generative models, because large-scale video pretraining may provide useful priors over temporal dynamics and physical regularities. Their benefit is not limited to predicting the future, but also lies in providing structured predictive representations that make action generation less ambiguous. By conditioning on anticipated outcomes rather than only the current observation, the policy gains longer-horizon foresight and a more informative basis for control.

This trend can be interpreted from a probabilistic perspective. Suppose the objective is to model the joint conditional distribution of future observations and future actions. Let o_t denote the current observation, a_t denote an action, and l denote the instruction (e.g., a language instruction or task specification). The distribution is expressed as

$$p(o_{t+1:t+k}, a_{t+1:t+k} \mid o_t, l). \quad (4)$$

Then several seemingly distinct paradigms can be viewed as different marginals or conditionals of the same underlying predictive-control model:

$$\text{Policy Model: } p(a_{t+1:t+k} \mid o_t, l) = \int p(o_{t+1:t+k}, a_{t+1:t+k} \mid o_t, l) d_o, \quad (5)$$

$$\text{Passive World Model: } p(o_{t+1:t+k} \mid o_t, l) = \int p(o_{t+1:t+k}, a_{t+1:t+k} \mid o_t, l) d_a, \quad (6)$$

$$\text{Controllable World Model: } p(o_{t+1:t+k} \mid o_t, a_{t+1:t+k}), \quad (7)$$

$$\text{Inverse Dynamics Model: } p(a_{t+1:t+k} \mid o_{t:t+k}). \quad (8)$$

In this sense, policy model, passive world model (video generation model), controllable world model and inverse dynamics model are not entirely separate abstractions; rather, they correspond to different ways of querying or factorizing the same idealized joint distribution. This also explains why world models and policies can be naturally coupled: a policy may use future observations generated by a world model as an intermediate latent variable, while an inverse-dynamics-style decoder can recover executable actions from such predicted futures.

Therefore, integrating a world model into policy learning can be viewed more generally as introducing predictive structure into action generation. Instead of learning a direct, monolithic mapping from the current observation to actions, the model reasons about future observations as auxiliary predictive variables that inform or constrain action selection. In some formulations, the model first predicts a plausible future and

Table 1 Comparison of architectural paradigms for world-model-based policies in Sec. 3.

Paradigm	Representative Work	Future Generation at Inference	Backbone	Coupling Style
IDM-style	UniPi (Du et al., 2023)	Explicit video rollout	VGM	Decoupled
	VidMan (Wen et al., 2024)	Explicit video rollout	VGM	Decoupled
	Vidar (Feng et al., 2025)	Explicit video rollout	VGM	Decoupled
	Gen2Act (Bharadhwaj et al., 2025)	Explicit human-video rollout	VGM	Decoupled
	VPP (Hu et al., 2025)	Latent predictive features	VGM	Decoupled
	Video2Act (Jia et al., 2025b)	Latent predictive features	VGM	Decoupled
	MimicVideo (Pai et al., 2025)	Latent visual plan	VGM	Decoupled
	TC-IDM (Mi et al., 2026)	Structured execution plan	VGM	Decoupled
	LVP (Chen et al., 2025a)	Visual plan	VGM	Decoupled
Say-Dream-ACT (Gu et al., 2026)	Video prompt	VGM	Decoupled	
Single-backbone	UVA (Li et al., 2025c)	Joint latent prediction	VGM	Shared backbone
	UWA (Zhu et al., 2025a)	Joint diffusion process	VGM	Shared backbone
	VideoVLA (Shen et al., 2025)	Joint video rollout	VGM	Shared backbone
	VideoPolicy (Liang et al., 2025a)	Video policy substrate	VGM	Shared backbone
	Cosmos Policy (Kim et al., 2026)	Parallel action/state/value outputs	VGM	Shared backbone
	DreamZero (Ye et al., 2026b)	Chunk-wise joint rollout	VGM	Shared backbone
	UD-VLA (Chen et al., 2026b)	Synchronous denoising	VGM	Shared backbone
	GigaWorld-Policy (Ye et al., 2026a)	Optional visual branch	VGM	Shared backbone
MoE/MoT	GE-Act (Bharadhwaj et al., 2025)	Latent visual guidance	VGM	Expert fusion
	Motus (Bi et al., 2025)	Expert rollout	VGM	MoT fusion
	LingBot-VA (Li et al., 2026b)	Visual predictive context	VGM	MoT fusion
	BagelVLA (Hu et al., 2026)	Single-step visual foresight	VGM	MoT fusion
	Fast-WAM (Yuan et al., 2026)	Train-time video, test-time skipped	VGM	MoT fusion
	LDA-1B (Lyu et al., 2026)	Latent dynamics only	VGM	Expert fusion
	FRAPPE (Zhao et al., 2026)	Latent representation alignment	VGM	Parallel experts
	DiT4DiT (Ma et al., 2026)	Latent video guidance	VGM	Expert fusion
Unified VLA	GR-1 (Wu et al., 2024)	Future image prediction	UMM	Joint co-training
	UP-VLA (Zhang et al., 2025c)	Future image prediction	UMM	Joint co-training
	WorldVLA (Cen et al., 2025)	Future image (mainly train-time)	UMM	Joint co-training
	DreamVLA (Zhang et al., 2025e)	Structured world knowledge	UMM	Joint co-training
	UniVLA (Wang et al., 2025)	Latent world modeling	UMM	Joint co-training
	CoWVLA (Yang et al., 2026a)	Latent dynamics	UMM	Joint co-training
	F1 (Lv et al., 2025)	Visual foresight	UMM	Unified MoT
	InternVLA-A1 (Cai et al., 2026)	Latent foresight	UMM	Unified MoT
	HALO (Shou et al., 2026)	Visual subgoal prediction	UMM	Unified multi-expert
TriVLA (Liu et al., 2025d)	Episodic dynamics	UMM	Multi-system	
Latent-space WM	FLARE (Zheng et al., 2025)	Latent alignment	MLLM	Latent internalization
	VLA-JEPA (Sun et al., 2026)	Latent target prediction	MLLM	Latent internalization
	JEPA-VLA (Miao et al., 2026)	Predictive embeddings	MLLM	Latent internalization
	WoG (Su et al., 2026)	Future condition only	MLLM	Latent internalization
	DIAL (Chen et al., 2026c)	Latent visual foresight	MLLM	Latent internalization

then decodes actions conditioned on that future. In others, candidate actions are generated first and then evaluated or regularized through predicted future outcomes. More unified approaches model observations and actions jointly within a shared generative process.

In practice, incorporating such predictive structure can provide a useful inductive bias for control, especially when robot action data are limited and large-scale predictive pretraining is available (Jang et al., 2025b). Motivated by this perspective, we organize this section from an architectural viewpoint, categorizing world-model-based policy methods according to how predictive generation interacts with action production, ranging from decoupled pipelines to tightly integrated end-to-end formulations. Table 1 summarizes these paradigms from a comparative architectural perspective, highlighting their representative methods, whether future generation remains active at inference time, the underlying backbone family, and the style of coupling between world modeling and action generation. This architectural progression also echoes broader trends in foundation models, where the design space has expanded from modular pipelines to include shared backbones, expert-coupled architectures, and latent forms of capability internalization.

3.2 Inverse Dynamics Policies with World Models

A representative line of work incorporates world models into robot control through a decoupled design, in which future prediction and action generation are realized by two distinct modules. The central idea is to first use a world model, most commonly an image or video generative model, to predict a task-conditioned future observation sequence (or its latent representation), and then train a separate policy module to infer executable actions from the current observation together with the predicted future. Unlike unified end-to-end policies that jointly model perception, prediction, and control within a single backbone, this paradigm preserves an explicit functional separation: the world model provides a structured hypothesis of “what should happen next,” while the policy translates such anticipated futures into low-level actions. As illustrated in Fig. 3(a), this family adopts a decoupled predict-then-act pipeline: a world model first produces future observations or their predictive representations, and a separate inverse-dynamics-style policy then maps these anticipated futures to executable actions.

This family of methods first constructs a predicted future trajectory:

$$\hat{\mathbf{o}}_{t+1:t+H} = \mathcal{W}(o_t, l), \quad (9)$$

or, more generally, a future latent representation:

$$\hat{\mathbf{z}}_{t+1:t+H} = \mathcal{W}(E_{\text{img}}(o_t), E_{\text{text}}(l)), \quad (10)$$

where \mathcal{W} denotes the world model and H is the prediction horizon. The policy is then conditioned on both the current observation and the generated future

$$\pi(a_{t+1,t+H'} \mid o_t, l) = P(a_t \mid E_{\text{img}}(o_t), E_{\text{text}}(l), \Phi(\hat{\mathbf{o}}_{t+1:t+H})), \quad (11)$$

or equivalently in latent form

$$\pi(a_{t+1,t+H'} \mid o_t, l) = P(a_t \mid E_{\text{img}}(o_t), E_{\text{text}}(l), \hat{\mathbf{z}}_{t+1:t+H}), \quad (12)$$

where $\Phi(\cdot)$ denotes a feature extractor over predicted future observations, and H' denotes the action chunk size. From a control perspective, this formulation is inverse-dynamics-style: rather than inferring actions solely from the present state, the policy leverages an anticipated state transition or future evolution signal, thereby reducing ambiguity in action generation.

Historically, early works in this line established the basic decoupled paradigm of predict first, then act. UniPi (Du et al., 2023) is a representative early example: it uses a task-conditioned world model to generate future video trajectories, and then trains a separate inverse dynamics model to derive an action representation by comparing adjacent frames. Subsequent methods mainly advance this paradigm by progressively redesigning what form of future representation is exposed to the policy. Early visual extensions, such as VidMan (Wen et al., 2024) and Vidar (Feng et al., 2025), retain the canonical two-stage recipe while introducing masked inverse dynamics to emphasize action-relevant regions. Gen2Act (Bharadhwaj et al., 2025) follows a similar decoupled design, but conditions execution on generated human videos rather than robot-centric future rollouts. Besides, works such as VPP (Hu et al., 2025) and Video2Act (Jia et al., 2025b) represent a closely related variant that moves away from explicit pixel-space rollout and instead treats the video world model as a source of compact predictive representations. Rather than decoding actions from fully rendered future frames, these methods extract control-relevant features from the latent space of a pretrained video diffusion model and inject them into a separate action head, yielding a more compact and stable interface between prediction and control. A related direction is explored by V2A (Luo and Du, 2025), which further grounds generated video states into action through goal-conditioned exploration: instead of learning a direct inverse-dynamics decoder from predicted futures, it treats synthesized video states as visual goals and learns a goal-conditioned policy through hindsight-style self-exploration.

Moving toward tighter representations, MimicVideo (Pai et al., 2025) replaces explicit video prediction with partially denoised latent visual plans, yielding a more compact and control-aligned interface. TC-IDM (Mi et al., 2026) and LVP (Chen et al., 2025a) push this abstraction further by translating generated futures into execution-oriented intermediates, such as tool-centric geometric trajectories or retargetable visual plans.

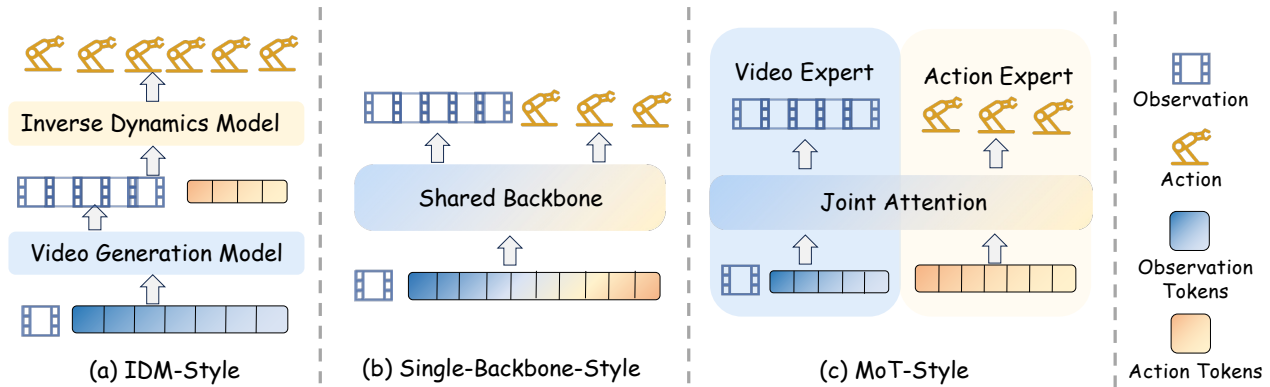


Figure 3 Representative architectural paradigms for using world models as policies. **(a) IDM-style.** A video generation model first predicts future observations, and an inverse dynamics model then recovers actions from the predicted visual trajectory, yielding a decoupled predict-then-act pipeline. **(b) Single-backbone-style.** Observation and action tokens are processed within a unified shared backbone, so future prediction and action generation are modeled jointly in a common latent space. **(c) MoT-style.** Video and action experts remain partially specialized, while cross-modal interaction is achieved through shared joint attention, enabling deeper coupling between world modeling and policy generation.

From a prompting perspective, Say-Dream-ACT (Gu et al., 2026) uses generated video plans not as explicit inverse-dynamics targets, but as in-context visual guidance for a separate action policy. Taken together, these methods reveal a common trend within this family: the predicted future gradually shifts from raw pixel-space rollout to increasingly structured, compact, and execution-friendly representations, while the world model and policy remain architecturally decoupled.

A related and complementary direction introduces more structured geometric intermediates into this decoupled pipeline. Rather than using generated or demonstrated videos only as raw visual futures, these methods further extract 3D-aware motion structure from video and use it as a more action-relevant predictive prior. In this sense, the key interface is still visually grounded, but the future is represented in a more structured form, such as dense correspondences, 3D trajectories, motion fields, or actionable 3D flow. Representative examples include AVDC (Ko et al., 2024), which recovers actions from synthesized videos through dense correspondences; VidBot (Chen et al., 2025b), which extracts 3D hand trajectories and interaction cues from human videos; Object-centric 3D Motion Field (Yin et al., 2025), which represents actions through object-centric 3D motion structure similar to Hind4sight-Net (Nematollahi et al., 2020); and NovaFlow (Li et al., 2026a), which distills generated videos into actionable 3D object flow for downstream execution. These works can be viewed as a structured extension of the decoupled paradigm: the world model remains visually grounded, but 3D representations serve as an intermediate structured prior that makes downstream action recovery more direct and robust.

A defining feature of this family is its architectural decoupling: the predictive model is typically pretrained first and then frozen, lightly adapted, or connected to a separate policy head, rather than optimized jointly with action generation. This separation brings modularity, reusable video priors, and interpretable future prediction, but also limits performance by the fidelity and controllability of the generated future, and can accumulate error when visually plausible predictions are not action-consistent. Even so, this paradigm remains one of the earliest and most influential routes by which world models became directly useful for robot policy learning, and it naturally motivates the tighter couplings introduced in later video-action architectures.

3.3 Unified Policies with a Single World Model Backbone

Different from the decoupled inverse-dynamics-style pipeline above, a more tightly coupled line of work uses a single generative backbone to jointly model future visual evolution and future actions. Figure 3(b) summarizes this shift visually: instead of passing predictions from a world model to a downstream policy module, observation and action tokens are processed within one shared backbone, so future modeling and action generation are coupled inside the same generative process. A more fundamental motivation behind this

design is not merely that video models can “imagine” future observations, but that pretrained video-generative backbones are optimized for temporally predictive modeling. In contrast to many VLM backbones (Kim et al., 2025; Black et al., 2024), which are primarily pretrained through image–text or vision–language alignment objectives and therefore emphasize semantic correspondence, video generation models are trained to model temporally ordered observations and may encode useful priors over motion continuity, temporal causality, and approximate physical dynamics. When action generation is embedded into the same denoising or generative process that models future world evolution, the policy may therefore benefit from a backbone already biased toward propagating constraints across time. However, whether video-pretrained backbones are consistently superior to matched-scale VLM backbones for robotic control remains an open empirical question; current results should be viewed as suggestive evidence for a promising inductive bias rather than a definitive architectural conclusion.

At a high level, this family replaces the two-stage factorization of “predict first, then act” with a unified multimodal generative objective. Let $\mathbf{x} = [z^v; z^a]$ denote the concatenation of future visual and action representations. A shared backbone f_θ is trained on corrupted inputs $\tilde{\mathbf{x}}_\tau$ under conditioning (o_t, l)

$$\hat{y} = f_\theta(\tilde{\mathbf{x}}_\tau, o_t, l, \tau), \quad \mathbf{x} = [z^v; z^a], \quad (13)$$

where τ means denoising steps, with a generic unified objective

$$\mathcal{L}_{\text{unified}} = \mathbb{E}[\ell(\hat{y}, y)], \quad (14)$$

where the exact target depends on the specific instantiation: y may correspond to diffusion noise in continuous denoising models, a velocity field in flow-matching variants, or masked tokens in discrete denoising formulations.

Representative early unified designs such as UVA (Li et al., 2025c), UWA (Zhu et al., 2025a), and later VideoVLA (Shen et al., 2025) already make this perspective explicit. Rather than preserving a modular “world model + policy head” decomposition, they treat control as a direct interface to a unified predictive generator. UVA learns a joint video–action latent space and supervises both modalities jointly, while retaining efficient deployment through lightweight modality-specific decoding heads that allow policy inference to bypass explicit video generation. UWA pushes this coupling further into the diffusion process itself by integrating video and action diffusion within a single transformer under modality-specific timesteps, and can be queried as a policy by marginalizing out the visual future through timestep control. Building on this unified view, VideoVLA shifts the emphasis toward directly converting a pretrained video generator into a robotic control model: it extends a Video Diffusion Transformer into a Video-Action Diffusion Transformer that jointly predicts future visual outcomes and action sequences, thereby making the pretrained video model itself the backbone of the policy. A closely related perspective is taken by VideoPolicy (Liang et al., 2025a), which treats video generation as the primary policy substrate and reduces action prediction to a lightweight interface layered on top of the generated behavior.

Subsequent methods tighten this coupling further by minimizing the representational gap between visual prediction and control. Cosmos Policy (Kim et al., 2026) is a particularly direct realization of this idea: it keeps the pretrained video diffusion architecture essentially unchanged and encodes robot actions, future states, and values as additional latent “frames” within the original diffusion sequence. At inference time, these outputs need not all be used symmetrically: in direct policy mode, only the action output is required for execution, whereas in planning mode the future-state and value predictions can be used to rank candidate trajectories. DreamZero (Ye et al., 2026b) follows the same end-to-end philosophy with an autoregressive flow-matching video-action DiT, but performs closed-loop chunk-wise joint denoising rather than free-running long-horizon rollout, thereby limiting compounding error while preserving tight video–action alignment. UD-VLA (Chen et al., 2026b) extends the same principle to a discrete multimodal setting, coupling future-image tokens and action tokens within a single synchronous denoising trajectory while introducing dedicated test-time efficiency techniques. GigaWorld-Policy (Ye et al., 2026a) provides a more explicitly action-centered variant: it jointly optimizes future action prediction and action-conditioned future video generation within a single shared transformer stack, while using a causal design that makes the visual branch optional at inference time.

Ultimately, the key difference across these unified methods is not whether they all render full future videos online, but how much of the visual branch remains active during control. Some preserve explicit future prediction for consistency or planning, whereas others retain the benefits of joint training while marginalizing,

truncating, or partially discarding the visual branch for efficiency. In all cases, unlike the decoupled methods above, the world model is not treated as a separate upstream module consumed by a downstream policy. Instead, world modeling and policy learning are collapsed into a single generative process, providing one possible route for injecting spatiotemporal priors from large-scale video pretraining into control.

3.4 MoE/MoT-Style Policies with Expert World-Model Backbones

Compared with the single-backbone generators above, a related but architecturally distinct line of work preserves explicit specialization by maintaining separate expert streams for video prediction, action generation, and sometimes language or scene understanding. Rather than collapsing all modalities into one shared diffusion backbone, these methods adopt MoE/MoT-style (Liang et al., 2025b) or multi-branch designs, where modality-specific experts interact through shared attention, cross-attention, or interleaved autoregressive sequences. As illustrated in Fig. 3(c), unlike single-backbone models, they retain separate video and action experts while coupling them through repeated interaction. The motivation remains to transfer the spatiotemporal and physical priors of pretrained video diffusion models (Wan, 2025; Ali et al., 2025) into control, but under a different architectural assumption: full parameter sharing is not always optimal, since video prediction and action generation have different temporal frequencies, representational scales, and optimization requirements. In this sense, these models resemble expertized VLA designs such as π_0 (Black et al., 2024) and $\pi_{0.5}$ (Intelligence et al., 2025b), except that their backbone is not primarily a static semantic encoder, but a temporally predictive video generator whose representations may contain useful cues about motion continuity, temporal causality, and approximate physical dynamics.

At a high level, these approaches can be viewed as learning a coupled predictive-control mapping with specialized experts:

$$(\mathbf{h}_{\ell+1}^v, \mathbf{h}_{\ell+1}^a) = \mathcal{F}_\ell^{\text{mix}}(\mathbf{h}_\ell^v, \mathbf{h}_\ell^a; o_t, l), \quad (15)$$

where ℓ indexes the layer, and $\mathcal{F}_\ell^{\text{mix}}$ denotes a layerwise interaction operator, such as joint attention, cross-attention, or shared-attention fusion (Bao et al., 2023), that couples a video expert and an action expert while preserving their distinct parameterization. Under this view, the video branch serves as a temporally predictive latent stream, and the policy is obtained by repeatedly injecting this foresight into the action branch, rather than by decoding actions from an entirely separate downstream head.

Within this family, one common pattern is parallel expert coupling, where a pretrained video diffusion backbone is paired with a lighter action branch. GE-Act (Liao et al., 2026) follows this pattern by introducing a parallel flow-matching action pathway alongside a pretrained video diffusion world model, using deep cross-attention to inject visual latent features into action generation. Here, the video branch provides predictive world-state structure, while the action branch translates it into executable control without requiring full video rendering online. Earlier instantiations of this paradigm leveraged image-editing diffusion models to predict subgoals for a goal-conditioned policy to follow (Black et al., 2023; Hatch et al., 2025).

A second and more explicit pattern is Mixture-of-Transformers-based deep interaction, in which multiple experts are retained throughout the network and fused repeatedly via shared attention. Motus (Bi et al., 2025), LingBot-VA (Li et al., 2026b), BagelVLA (Hu et al., 2026) and more recently DiT4DiT (Ma et al., 2026) are representative examples. Motus formulates the design most directly as a Mixture-of-Transformers with dedicated experts for understanding, video generation, and action. LingBot-VA pushes this idea toward causal world modeling by interleaving video and action tokens into a shared autoregressive sequence and using a dual-stream MoT with shared attention, turning imagined future states into a context for action refinement. BagelVLA extends the same intuition to longer-horizon manipulation, interleaving linguistic planning, visual forecasting, and action generation within one execution loop; its Residual Flow Guidance further makes visual foresight practical through single-step denoising (Liu et al., 2023b) rather than full video rollout. DiT4DiT (Ma et al., 2026) follows the same intuition by coupled architecture, using intermediate denoising features from the video branch to guide action prediction. Fast-WAM (Yuan et al., 2026) can be viewed as a hybrid point in this family: it adopts a shared-attention Mixture-of-Transformers backbone with coupled video and action branches, yet concludes that the main benefit may come more from video co-training during training than from explicit future imagination at inference time. Across these variants, the video branch is increasingly treated not as an output to be faithfully rendered, but as a predictive latent process whose hidden states guide action generation.

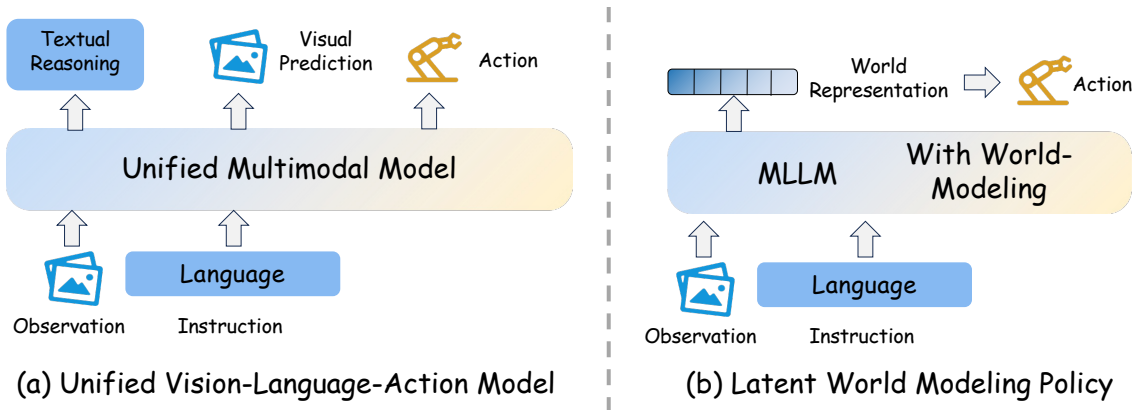


Figure 4 Two MLLM-based routes for internalizing world modeling into policy learning. **(a) Unified Vision-Language-Action model.** A unified multimodal model jointly processes observation and language, and is trained to produce action together with auxiliary future-oriented outputs such as textual reasoning or visual prediction. **(b) Latent-space world modeling for VLA.** Instead of explicitly predicting future images, the model internalizes future dynamics as a compact world representation or world modeling inside the same MLLM backbone, and then maps this latent predictive knowledge to action.

A third pattern is latent-space expertization, which shifts world modeling from pixel space to structured latent dynamics while retaining specialized multimodal branches. LDA-1B (Lyu et al., 2026) represents this direction by moving visual forecasting into a DINO (Caron et al., 2021) latent space and coupling visual and action experts through shared self-attention inside a multimodal diffusion transformer. FRAPPE (Zhao et al., 2026) follows a related philosophy from the perspective of future-representation alignment: instead of reconstructing future observations, it trains multiple parallel expert streams with separate adapters and aligns them to visual foundation models in latent space. Although more training-oriented than the explicitly architected MoT models above, it reflects the same underlying idea that deeply coupled specialized predictive streams can improve world-aware action generation.

Taken together, these methods bridge the gap between detached modular pipelines and fully unified single-backbone generators. They embed world modeling directly into the policy while preserving architectural specialization. Specifically, video diffusion models provide predictive foresight, while MoE/MoT mechanisms translate this into action without losing modality-specific structures. Compared to single-backbone approaches, the key distinction is architectural: both aim to couple future prediction with action generation, but these methods achieve it through deeply interacting specialized experts rather than full parameter sharing.

3.5 Unified Vision-Language-Action Models

Unified VLA models provide another route to world model as policy. Although they do not always employ an explicit video world model, they still learn future-oriented predictive structure within the same multimodal policy backbone, for example through future-image prediction, visual foresight, or structured world knowledge. As shown in Fig. 4(a), this family differs from the preceding video-backbone paradigms in that future modeling is internalized inside a unified VLA architecture rather than introduced through a separate predictive module.

One important subclass performs explicit future-state prediction. These methods directly predict future images, either a single frame or a short sequence, as part of the unified training objective. GR-1 (Wu et al., 2024) is an early representative that jointly predicts actions and future images within a single GPT-style transformer. UP-VLA (Zhang et al., 2025c) follows a similar strategy, using future-image prediction to improve both action generation and visual generalization. WorldVLA (Cen et al., 2025) further unifies action and image understanding and generation in one autoregressive framework, while using future-image prediction mainly as a joint training signal rather than a mandatory inference-time output.

A second subclass replaces pixel-level prediction with implicit or latent future modeling. Instead of forecasting future frames directly, these methods predict compact future-aware representations that are more tightly aligned with action. DreamVLA (Zhang et al., 2025e) predicts structured world knowledge, including dynamic,

spatial, and semantic cues, to support inverse dynamics modeling. UniVLA (Wang et al., 2025) incorporates world modeling during post-training over a native multimodal tokenization framework, allowing the model to absorb causal dynamics from large-scale video data without introducing a separate external world model. CoWVLA (Yang et al., 2026a) pushes this direction further by modeling latent motion and compact future visual targets instead of reconstructing redundant future frames.

A third subclass consists of *multi-expert or multi-system unified models*, which remain unified at the training and task level but preserve explicit functional specialization inside the architecture. This category includes F1 (Lv et al., 2025), InternVLA-A1 (Cai et al., 2026), HALO (Shou et al., 2026), and TriVLA (Liu et al., 2025d). Although these methods also adopt expertized or MoT-style designs, their predictive branch is better understood as visual foresight or subgoal generation within a unified VLA framework, rather than as a native video-backbone world model. F1 predicts future visual states as planning targets within a Mixture-of-Transformers architecture. InternVLA-A1 extends this design with lightweight latent visual foresight and joint optimization of foresight prediction and action generation. HALO pushes the predictive branch toward visual subgoal prediction and embodied reasoning, while TriVLA organizes grounding, episodic dynamics perception, and control as coordinated subsystems.

Taken together, unified VLA models extend the notion of world model as policy beyond explicit video generation. Some do so through direct future-image prediction, others through compact latent or semantic world knowledge, and still others through unified multi-expert systems with explicit foresight modules. Across these variations, the shared principle is that action generation is no longer treated as a purely reactive mapping from the current observation, but is jointly trained with an internal predictive objective that captures future state evolution or its compact surrogate. Relative to the preceding subsections, the key distinction is therefore not whether the model contains an explicit standalone world model, but whether future-oriented predictive modeling is internalized within the same multimodal policy backbone.

3.6 Policies with Latent-Space World Modeling

A further route to world model as policy is defined by methods that internalize future prediction entirely in representation space, without relying on explicit image or video generation. Rather than synthesizing future observations, these approaches construct predictive latent targets, future-aware embeddings, or compact control conditions, and couple them with action generation within the same policy framework. In this context, world modeling is realized not as visual reconstruction, but as learning a future-aware representation that captures how the environment may evolve in a form directly useful for control. Such methods therefore retain the core benefit of world modeling by injecting predictive structure into action generation, while avoiding the computational overhead and redundancy of explicit generative decoding. Conceptually, this direction is related to the JEPa family (Assran et al., 2023, 2025), which models prediction in embedding space rather than pixels, but the focus here is not JEPa itself; rather, it is the emergence of VLA methods that turn this representation-space predictive principle into a practical mechanism for policy learning. Figure 4(b) illustrates this more internalized variant. Here, the backbone is again typically MLLM-based rather than video-DiT-based, but future modeling is absorbed more deeply into latent world representations or parameterized world knowledge, so that action generation is guided by internal predictive structure without requiring explicit future-image decoding.

Chronologically, FLARE (Zheng et al., 2025) is an early representative of this direction. It introduces “Future Latent Representation Alignment”, aligning hidden features of the action denoising network with latent embeddings of future observations, so that the policy can implicitly anticipate future states without explicitly generating them. VLA-JEPa (Sun et al., 2026) makes this more explicit by adopting a JEPa-style pretraining objective for VLA: its key design is leakage-free state prediction, where future frames are used only to produce latent targets for supervision, encouraging the model to learn action-relevant state transitions in latent space rather than shortcutting through pixel variation. JEPa-VLA (Miao et al., 2026) takes a complementary route: instead of adding an explicit latent prediction head, it argues that predictive embeddings learned by video JEPa models, especially V-JEPa 2 (Assran et al., 2025), already provide stronger policy priors than static visual representations, and therefore adapts these predictive embeddings as a better backbone for existing VLA models. Most recently, WoG (Su et al., 2026) moves world modeling into the condition space of action generation: rather than predicting future images or generic future latents, it learns to predict

compact future-oriented conditions together with actions, so that the model directly forecasts the part of future information that is most useful for precise control. DIAL (Chen et al., 2026c) provides a closely related recent example by decoupling high-level intent from low-level action through latent world modeling, using latent visual foresight in the VLM feature space as a structured bottleneck for downstream action generation.

Beyond neural latent representations, a related but more classical non-pixel abstraction appears in symbolic or planner-facing world models. Unlike the neural policy backbones reviewed above, these methods usually externalize world modeling as an abstract transition model over predicates, object relations, affordances, operators, or causal processes, which is then queried by a symbolic or task-and-motion planner to produce high-level skill sequences (Silver et al., 2021; Shah et al., 2025; Liang et al., 2025c; Athalye et al., 2026; Liang et al., 2026). We include this line as a complementary perspective to emphasize that useful world models may not depend on predicting pixels; they can also capture abstract logic, object relations, causal regularities, and symbolic dynamics for planning and control.

Taken together, this subsection highlights a non-pixel route for world-model-based policy learning. The main line is latent-space world modeling, where the policy avoids explicit future-image or video decoding while still internalizing future dynamics for action generation. The symbolic planner examples above further reinforce the same broader point: control-relevant prediction can be expressed through compact latent or abstract variables when they provide a more direct interface to action.

4 World Model as Simulator

Beyond serving as a predictive module for conditioning, planning, or internal supervision, world models can also be used more directly as interactive simulators. In this paradigm, the value of a world model lies not only in its ability to model future evolution, but in its ability to stand in for the environment itself (Xiao et al., 2025; Li et al., 2025b; Zhu et al., 2026; Gemini Robotics Team et al., 2025): given the current observation, task instruction, and candidate actions, the model can roll out future states, provide feedback signals, and support downstream decision making through imagined interaction. This makes world model as simulator a particularly direct and practical extension of world modeling for embodied intelligence.

This direction is especially appealing for visuomotor policies because conventional reinforcement learning on physical robots is often slow, expensive, difficult to reset, and potentially unsafe, while pure imitation learning remains limited by demonstration quality and cannot easily learn from failures. Recent work therefore replaces costly real-world interaction with learned simulators built from world models, enabling policy improvement through imagined rollouts rather than repeated physical trial-and-error (Wu et al., 2023). In frameworks such as World-Env (Xiao et al., 2025), VLA-RFT (Li et al., 2025b), and WMPO (Zhu et al., 2026), the world model is explicitly used as a low-cost, controllable virtual environment for post-training, substantially reducing the dependence on real interaction while improving data efficiency and robustness.

At the same time, the simulator view provides benefits beyond reinforcement training. Because a world model can roll out action-conditioned future states, it can also expose verifiable signals from predicted trajectories, such as reward-like feedback (Xiao et al., 2025; Li et al., 2025b), task completion cues, or rollout consistency, which are useful not only for policy optimization but also for evaluation, ranking, and test-time decision making. This is already reflected in systems that augment the learned simulator with reward or termination feedback, and it naturally extends to rollout-based assessment of candidate behaviors. Here, the same predictive ability that supports imagined training also becomes a mechanism for judging whether a candidate action sequence is likely to succeed before it is executed.

For this reason, we organize this section around two complementary uses of the simulator paradigm. As summarized in Fig. 5, world models can support policy learning in at least two distinct functional roles: as learned simulators for reinforcement learning, and as evaluators for decision-time validation. We first discuss world model for reinforcement training, where the learned simulator is used to replace physical interaction and support policy improvement through imagined rollouts. We then discuss world model for evaluation, where the same simulator capability is used to verify, rank, or assess candidate behaviors through predictive rollout and future-state feedback.

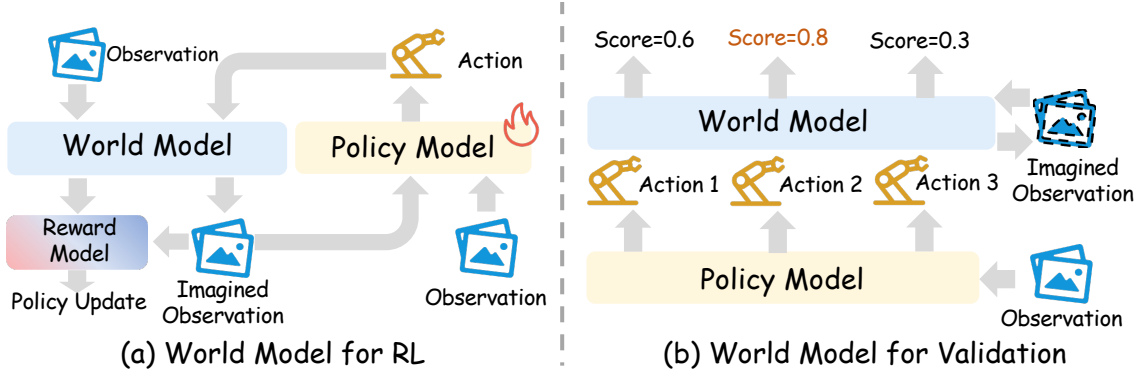


Figure 5 Two uses of world models for policy learning. **(a)** In the reinforcement-learning setting, the world model serves as a learned simulator that produces imagined transitions for policy improvement. **(b)** In the validation setting, the world model scores imagined consequences of candidate actions to support decision-time selection.

4.1 World Model for Reinforcement Learning

A further role of world models in embodied learning is to serve as interactive simulators for reinforcement post-training. Different from the preceding paradigms, where the world model mainly provides predictive conditioning, planning cues, or internal supervision, the methods in this subsection directly use the world model as a learned environment in which a Vision-Language-Action (VLA) policy can roll out trajectories, receive rewards, and improve through imagined interaction. In this setting, the world model is no longer merely a predictor of plausible futures; it becomes the medium through which reinforcement learning is carried out. At a high level, these methods optimize a policy inside a learned simulator

$$(\hat{o}_{t+1}, \hat{r}_t, \hat{d}_t) \sim p_\phi(\cdot \mid o_{\leq t}, a_{\leq t}, l), \quad (16)$$

where the world model p_ϕ provides imagined transitions, and optionally rewards and termination signals. The policy is then improved from imagined rollouts by maximizing expected return,

$$J(\theta) = \mathbb{E}_{\hat{\tau} \sim (\pi_\theta, p_\phi)} \left[\sum_t \gamma^t \hat{r}_t \right], \quad (17)$$

or, in practice, through a GRPO-style policy optimization objective,

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_t \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right], \quad r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}, \quad (18)$$

together with task-specific variants adapted to chunked or flow-based action heads. Under this shared view, the common goal is to replace expensive physical interaction with reinforcement learning inside a learned world simulator.

Within this first-level paradigm, early works such as UniSim (Yang et al., 2024a), World-Env (Xiao et al., 2025) and VLA-RFT (Li et al., 2025b) establish the basic recipe of reinforcement learning in a learned simulator by coupling action-conditioned world simulation with reward generation. DiWA (Chandra et al., 2025) shows that a frozen world model learned from large-scale play data can already support fully offline adaptation of diffusion policies, while World4RL (Jiang et al., 2025d) extends this idea to higher-fidelity manipulation refinement through a diffusion world model for end-to-end imagined policy optimization.

Subsequent works make this paradigm increasingly compatible with modern VLA architectures and larger embodied datasets. World-Gymnast (Quevedo et al., 2025) demonstrates that RL inside a video world model can outperform both supervised finetuning and software simulators. PlayWorld (Yin et al., 2026) learns robot world models from autonomous play and shows that reinforcement learning in the learned simulator improves downstream real-world performance. RehearseVLA (Xiao et al., 2026) adapts the same principle to VLA post-training with a physically consistent world simulator and an instant reflector for reward and termination feedback. In parallel, WMPO (Zhu et al., 2026) emphasizes pixel-space imagination and

on-policy GRPO, ProphRL (Zhang et al., 2025b) adapts RL updates to flow-based action heads through FA-GRPO and FlowScale, RISE (Yang et al., 2026b) augments the simulator with compositional dynamics and progress-value estimation, and GigaBrain-0.5M* (Team et al., 2026) scales world-model-based RL to pretrained VLA adaptation. Despite these differences, all of these methods treat the world model primarily as the environment in which policy optimization takes place.

A second-level development explicitly recognizes that the learned simulator is itself imperfect and must be improved together with the policy. World-VLA-Loop (Liu et al., 2026b), VLAW (Guo et al., 2026a), and WoVR (Jiang et al., 2026) exemplify this shift in different ways. World-VLA-Loop (Liu et al., 2026b) jointly predicts future observations and rewards, and closes the loop by using policy failure rollouts to refine the simulator. VLAW (Guo et al., 2026a) follows an iterative repair-and-improve strategy, alternating between real-world data for simulator refinement and synthetic data for VLA improvement. WoVR (Jiang et al., 2026) pushes this direction further by treating simulator reliability as the central bottleneck, introducing controllable action-conditioned video modeling, Keyframe-Initialized Rollouts, and explicit world-model-policy co-evolution:

$$\phi^{k+1} \leftarrow \text{UpdateWM}(\phi^k, D_{\text{real}} \cup D_{\text{policy}}(\pi_{\theta^k})), \quad \theta^{k+1} \leftarrow \text{UpdatePolicy}(\theta^k, \hat{D}(\phi^{k+1})), \quad (19)$$

where policy rollouts refine the world model, and the improved world model in turn produces better imagined data for subsequent policy updates. In this sense, the focus moves beyond reinforcement learning in a world model toward reinforcement learning with a world model whose fidelity, action-following precision, and rollout reliability must themselves be continuously improved.

Taken together, this subsection reveals a clear progression in the role of world models for policy improvement. The first-level paradigm treats world models as learned simulators for reinforcement training, differing mainly in reward design, rollout representation, and optimization compatibility. The second-level paradigm further recognizes that imagined reinforcement learning is only as effective as the simulator is reliable, and therefore introduces simulator refinement, rollout regulation, and policy-world-model co-evolution as integral parts of the loop.

4.2 World Model for Evaluation

Beyond serving as a learned simulator for reinforcement post-training, a world model can also evaluate candidate behaviors before execution. Here, the goal is not to improve a policy through repeated imagined interaction, but to estimate which candidate action sequence, policy, or checkpoint is most likely to succeed in the real world. As illustrated in Fig. 5(b), the world model supports decision-time selection by scoring or verifying imagined consequences of candidate actions. Given the current observation, task instruction, and one or more candidate actions, it rolls out predicted futures and uses them for ranking, rejection, or safety filtering. In this sense, the evaluator role is a natural extension of the simulator view: once a world model can stand in for the environment, it can be used not only for training in imagination, but also for judging what the policy should do next.

One direct form of evaluation is rollout-based candidate assessment. Here, the policy proposes multiple action sequences, the world model predicts their future outcomes, and the system selects the candidate with the most favorable imagined consequence. GPC (Qi et al., 2026) is a particularly clean example: rather than retraining the policy, it augments a frozen generative robot policy at deployment with an action-conditioned world model and uses predictive look-ahead to rank and refine candidate actions online. IRASim (Zhu et al., 2025b) similarly demonstrates model-based planning by simulating multiple candidate trajectories and selecting the one with the highest predicted value. World-in-World (Zhang et al., 2025a) extends this idea to closed-loop planning, where candidate plans are rolled out in imagination, evaluated by a revision policy, and revised before execution. DreamPlan (Jia et al., 2026a) turns the same evaluator logic into a training signal by constructing preference pairs over candidate actions from world-model rollouts. Across these methods, the world model acts as a decision-time or near-decision-time selector that converts imagined futures into action choice.

Beyond simple ranking of discrete candidates, a more active paradigm treats the world model as the transition dynamics for Model Predictive Control (MPC). In this setting, the system does not merely select from a few

pre-defined actions but actively optimizes an action sequence within the world model’s imagined trajectories to minimize a cost function. Works such as TD-MPC2 (Hansen et al., 2024) and LeWorldModel (Maes et al., 2026) demonstrate that latent-space MPC can significantly enhance the long-horizon reasoning of embodied agents. By performing gradient-based planning through the world model, the agent can discover complex strategies that are not explicitly present in the training demonstrations. This synergy effectively transforms the world model from a passive judge of actions into an active navigational map for continuous control optimization.

A second, more explicit form is to use the world model itself as a policy evaluator. A recent large-scale example is Evaluating Gemini Robotics Policies in a Veo World Simulator Team et al. (2025a), which uses a video world simulator for offline policy evaluation, OOD testing, and safety probing. WorldEval (Li et al., 2025e) is the clearest example: it studies whether a world model can serve as a scalable proxy for real-world policy evaluation, ranking different robot policies and even different checkpoints of the same policy entirely in imagination, while also functioning as a safety detector. The same role appears at the benchmark level in WorldArena (Shang et al., 2026), which explicitly identifies policy evaluation as a core downstream use of embodied world models.

A third form arises when the simulator is equipped with explicit feedback heads that convert imagined rollouts into assessment signals. World-Env (Xiao et al., 2025) augments the simulator with continuous reward prediction and action termination prediction. VLA-RFT (Li et al., 2025b) uses verified rewards computed on imagined trajectories inside a controllable world simulator. World-VLA-Loop (Liu et al., 2026b) jointly predicts future observations and reward signals in a state-aware video world model. RISE (Yang et al., 2026b) makes this evaluator role even more explicit by introducing a progress value model that scores imagined outcomes according to task advancement. In these systems, imagined rollouts are not only a source of synthetic training data, but also a basis for deciding whether a behavior is promising, complete, or worth executing.

A related but lighter-weight perspective appears in latent-space predictive world models, especially the JEPA line. Rather than generating explicit pixel-space futures for ranking candidate actions, these methods perform prediction and planning in embedding space. V-JEPA 2 (Assran et al., 2025) and V-JEPA 2.1 (Mur-Labadia et al., 2026) are representative examples, with the latter further showing that a latent action-conditioned world model can support zero-shot robot planning with image goals. More recently, LeWorldModel (Maes et al., 2026) pushes this direction toward a simpler and faster end-to-end JEPA formulation, while also showing that latent predictive models can detect physically implausible events. At present, however, these methods are better viewed as an adjacent direction for predictive planning and plausibility checking than as fully developed policy evaluators in the broader embodied-control sense considered here.

This evaluator perspective also clarifies why action faithfulness and rollout reliability matter so much. An evaluator is useful only if its imagined future preserves the causal consequences of candidate actions. Ctrl-World (Guo et al., 2026b) makes this connection explicit by showing that action-faithful rollouts can support policy evaluation in imagination. At the same time, WoVR (Jiang et al., 2026) highlights an important caveat: hallucination and long-horizon error do not merely reduce visual quality, but can directly corrupt the assessment signal itself. For evaluation, realism alone is therefore insufficient; what matters is whether the rollout remains reliable enough to support ranking, selection, and rejection in a way that tracks real-world execution.

Taken together, these works reveal a clear broadening of the simulator paradigm. In embodied robot learning, a world model is no longer only a low-cost environment for reinforcement training; it is increasingly also used as an evaluator that can compare policies, score candidate behaviors, detect likely failures, and support both decision-time action selection and offline policy assessment (Li et al., 2025e; Team et al., 2025a). This shift is conceptually important for the remainder of the survey, because it shows that the value of a world model lies not only in generating future trajectories, but in generating trajectories that are trustworthy enough to support policy evaluation and action choice.

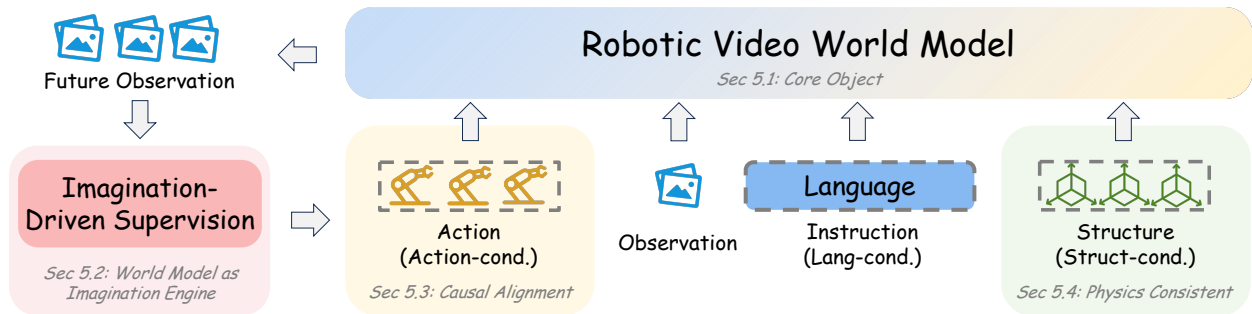


Figure 6 Unified view of robotic video world models. **Section 5.1** defines the core object as a robotic video world model that predicts future observations in visual space. Building on this core, **Section 5.2** uses the predicted future as an imagination engine for supervision, **Section 5.3** introduces action conditioning to improve causal alignment and controllability, and **Section 5.4** further incorporates structure priors to enhance physical and interaction consistency. The resulting future observation is therefore not only visually plausible, but increasingly actionable for downstream robot learning and decision making. Finally, **Section 5.5** highlights the broader transition from task-specific video prediction to a scalable and reusable world-model interface built on strong video priors.

5 World Model for Robotic Video Generation

5.1 Problem Setting and Scope

An important route to embodied world modeling represents the future directly in image or video space. In this setting, the model predicts the visual evolution of robot–environment interaction from the current observation, task specification, and often a sequence of candidate actions. Unlike generic video synthesis, robotic video generation is subject to substantially stronger requirements: the predicted future should be not only visually plausible, but also temporally coherent, action-consistent, physically credible, and useful for downstream decision making. For this reason, robotic video generation should not be understood merely as a perceptual generation problem, but as a concrete mechanism for constructing visually explicit world models that support robotics policy learning, planning, evaluation, simulation, and data generation. Recent progress in large-scale video generation backbones, such as CogVideoX, has played an important enabling role by showing that long-horizon and high-fidelity spatiotemporal generation can be learned at scale and later adapted to embodied settings (Yang et al., 2024b). We focus on video-based world models in this section because they form a rapidly growing and practically influential branch of recent work, not because pixel-level prediction is assumed to be the most compact or universally optimal representation for embodied control.

In this survey, we also treat task or language conditioning as a form of high-level action. Under this view, robotic video generation includes not only low-level action-conditioned rollout, but also text- or task-guided visual prediction that specifies what future should be realized before low-level control is grounded. Accordingly, the key question is not simply whether a model can generate visually convincing future videos, but whether those futures are actionable: whether they remain faithful to the conditioning actions, preserve physically plausible interaction dynamics, and can be translated into executable robot behavior. In robotics, the value of a video world model depends on whether it preserves action consequences, interaction structure, and physical regularities in a way that improves policy behavior. Following this viewpoint, as shown in Fig. 6, we organize the literature into four stages of progression: video generation as imagination for policy learning, action-controllable rollout models, structure-aware generation with richer interaction priors, and foundation-scale video backbones adapted into reusable world models. Table 2 summarizes representative methods under this capability-oriented taxonomy.

5.2 Video Generation as Imagination for Policy Learning

A first class of methods uses video generation primarily as an imagination engine for policy learning. The central idea is to exploit strong generative priors to synthesize future task executions and then convert these imagined futures into supervision for robot control. In this line, the video model is valuable not because it produces visually impressive clips, but because it expands supervision beyond the narrow support of collected

Table 2 Comparison of representative methods in Sec. 5, grouped by the four capability regimes discussed in Sec. 5. Checkboxes indicate whether a feature is explicitly supported or emphasized by the original paper. Here, *Foundation-scale* is reserved for methods that explicitly build on or present large-scale pretrained/foundation video or world models.

Group	Method	Task-cond.	Action-cond.	Structure-aware	Foundation-scale	Main use
Imagination-Based	UniPi (Du et al., 2023)	✓	–	–	✓	Plan.
	Video Language Planning (Du et al., 2024)	✓	–	–	✓	Plan.
	Dreamitate (Liang et al., 2024)	–	–	–	✓	Plan.
	RoboDreamer (Zhou et al., 2024)	✓	–	–	–	Plan.
	ManipDreamer (Li et al., 2025f)	✓	–	✓	–	Plan.
	DreMa (Barcellona et al., 2025)	–	✓	✓	–	Data
	PhysWorld (Mao et al., 2025)	✓	–	✓	–	Plan.
	DreamGen (Jang et al., 2025a)	✓	–	–	✓	Data
Action-Controllable	IRASim (Zhu et al., 2025b)	–	✓	–	–	Plan.
	RoboEnvision (Yang et al., 2025)	✓	–	–	–	Plan.
	RoboMaster (Fu et al., 2026)	–	✓	✓	–	Data
	Ctrl-World (Guo et al., 2026b)	–	✓	✓	–	Eval.
	EnerVerse-AC (Jiang et al., 2025c)	–	✓	✓	–	Eval.
	Interactive World Simulator (Wang et al., 2026c)	–	✓	–	–	Sim.
	EVA (Wang et al., 2026b)	✓	–	–	–	Eval.
Structure-Aware	Mask2IV (Li et al., 2025a)	✓	–	✓	–	Data
	Tesseract (Zhen et al., 2025)	✓	–	✓	–	Sup.
	RoboVIP (Wang et al., 2026a)	✓	–	✓	–	Data
Foundation Video WM	Vid2World (Huang et al., 2026)	–	✓	–	✓	Sim.
	Genie Envisioner (Liao et al., 2026)	✓	✓	–	✓	Sim.
	DreamDojo (Gao et al., 2026a)	–	✓	–	✓	Sim.
	WoW (Chi et al., 2025c)	–	–	–	✓	Plan.
	UnifLM-WMA-0 (Unitree, 2025)	✓	✓	–	–	Sim.
	Cosmos Predict 2.5 (Ali et al., 2025)	✓	–	–	✓	Sim.
	GigaWorld-0 (Team et al., 2025b)	–	–	✓	✓	Data
ABot-PhysWorld (Chen et al., 2026d)	–	✓	–	✓	Sim.	

robot trajectories.

A closely related branch is text- or task-guided robotic video generation. In our definition, language can be viewed as a high-level action that specifies what future should be realized. From this perspective, methods such as UniPi (Du et al., 2023) and Video Language Planning (Du et al., 2024) are not merely performing perceptual synthesis, but robotic world modeling in which semantic actions are translated into predictive visual trajectories. This branch is mainly useful for data amplification and visual planning, as it provides task-relevant demonstrations and future-oriented supervision before low-level action grounding.

Dreamitate (Liang et al., 2024) is an early and representative example. It fine-tunes a video diffusion model on task-specific human demonstrations and, at test time, uses synthesized executions in a novel scene directly as action-guiding visual plans for real-world robot control. RoboDreamer (Zhou et al., 2024) extends this direction through compositional world modeling, where instructions are decomposed into reusable primitives and generation is conditioned on these structured components, improving generalization to unseen combinations of objects and actions. ManipDreamer (Li et al., 2025f) further strengthens this line by introducing an action tree representation together with depth and semantic visual guidance, improving instruction following as well as temporal and physical consistency.

A related but more explicit perspective is to reinterpret imagination as a learnable digital twin. DreMa (Barcellona et al., 2025) combines Gaussian Splatting with a physics simulator to reconstruct an explicit and manipulable scene representation, allowing the model to generate additional demonstrations for imitation learning. PhysWorld (Mao et al., 2025) addresses the gap between photorealistic motion and physically executable behavior by reconstructing a physical world model from generated videos and grounding predicted

motion into robot actions through object-centric residual reinforcement learning. These methods move beyond pure visual imagination and begin to connect generated futures to physically meaningful execution.

This imagination paradigm also scales naturally toward synthetic data generation and high-level planning. When future generation is conditioned on task or language descriptions, the resulting videos can serve as high-level demonstration surrogates or task-relevant synthetic supervision. They can also function as visual plans for long-horizon decision making (Du et al., 2024; Chen et al., 2025a). DreamGen (Jang et al., 2025a) adapts strong video generators to a target embodiment, synthesizes neural trajectories, and recovers executable actions through latent action modeling or inverse dynamics. Its central message is that stronger video world models can be used not only to regularize policies, but also to produce synthetic experience that improves downstream generalization. Taken together, these works establish the first major role of robotic video generation, namely to serve as an imagination engine that broadens the supervisory and planning signals available for policy learning.

5.3 Toward Action-Controllable Video World Models

A second class of methods shifts the emphasis from imagined supervision to explicit controllability. Here the central question is no longer whether a model can produce plausible future videos, but whether the generated future follows the commanded action sequence with sufficient precision to support manipulation reasoning and downstream control. This shift is important because, in embodied settings, a visually convincing rollout is of limited value if it does not respond faithfully to action intervention.

IRASim (Zhu et al., 2025b) is representative of this transition. It formulates robot manipulation as a trajectory-to-video problem and introduces frame-level action conditioning within each transformer block to strengthen alignment between individual actions and corresponding future frames. RoboEnvision (Yang et al., 2025) focuses on long-horizon multi-task manipulation and emphasizes the difficulty of preserving semantic and temporal consistency over extended task evolution. RoboMaster (Fu et al., 2026) addresses more complex robot-object interactions through collaborative trajectory control. By decomposing manipulation into multiple phases and modeling the coupled motion of the robot arm and the manipulated object, it improves faithfulness under rich contact dynamics. Ctrl-World (Guo et al., 2026b) pushes controllability further toward policy-in-the-loop rollout. It combines joint multi-view prediction, frame-level action control, and memory-based long-horizon generation so that predicted futures can support both policy evaluation and targeted policy improvement. EnerVerse-AC (Jiang et al., 2025c) follows a related direction and formulates the world model as an action-conditional multi-view generator that can act both as a data engine and as an evaluator for robotic inference. Interactive World Simulator (Wang et al., 2026c) pushes this line from controllable generation toward genuinely interactive simulation, emphasizing high-frequency, long-horizon, and stable policy-conditioned interaction for closed-loop rollout, demonstration collection, and policy evaluation. EVA (Wang et al., 2026b) complements this direction from the perspective of post-training alignment by targeting the executability gap between visually plausible rollouts and physically executable robot behavior, using inverse-dynamics rewards to align video world models with smooth, embodiment-consistent action sequences.

Collectively, these works mark a decisive conceptual shift. For robotic video world models, fidelity is increasingly measured not only by realism, but also by action faithfulness, controllable interaction, and usefulness for closed-loop decision making.

5.4 Structure-Aware Generation with Interaction and Geometry Priors

A closely related thread improves controllability by introducing richer intermediate structure for interaction. Instead of conditioning only on low-dimensional action sequences, these methods encode masks, geometry, viewpoints, or identity cues that better preserve contact relations and scene structure. The underlying intuition is that robotic video generation becomes substantially more useful when the model is required to preserve explicit interaction structure, rather than merely synthesize visually convincing motion.

Mask2IV (Li et al., 2025a) is illustrative of this idea. It adopts a two-stage design that first predicts interaction trajectories for the actor and the object, and then generates a video conditioned on these trajectories. This removes the need for dense user-provided masks while retaining flexible control over interaction outcomes.

Tesseract (Zhen et al., 2025) pushes structure further by extending the representation space from 2D video to a 4D embodied world model over RGB, depth, and normal signals, improving spatial consistency and enabling stronger inverse dynamics and policy learning. RoboVIP (Wang et al., 2026a) focuses on a practical requirement of modern policy learners, namely temporally coherent multi-view observations. It introduces visual identity prompting to guide multi-view video diffusion and uses the resulting videos as scalable augmentation for manipulation data.

This structure-aware view also connects robotic video world models to a broader line of structured and symbolic world modeling. While the methods above preserve structure inside generated visual futures, another family abstracts the world into predicates, object relations, affordances, or causal processes and predicts their transitions for planning (Liang et al., 2025c; Athalye et al., 2026; Liang et al., 2026). These approaches do not aim to improve visual fidelity; instead, they target more compact and compositional predictive variables that may be better aligned with long-horizon reasoning and executable control.

Although these methods differ in representation and objective, they share a common principle: richer structural priors can make generated futures more controllable, more consistent across views and contacts, and ultimately more useful for downstream embodied learning.

5.5 From Video Backbones to Foundation World Models

The most recent works reinterpret robotic video world models as general-purpose interactive predictors built by adapting large-scale video backbones. In this regime, video generation is no longer merely a downstream augmentation tool. It becomes a reusable substrate for simulation, planning, evaluation, and large-scale robot data production.

Vid2World (Huang et al., 2026) is a canonical example of this shift. Rather than training a robotic world model from scratch, it systematically transforms a pretrained video diffusion model into an interactive world model suitable for action-conditioned rollout. Genie Envisioner (Liao et al., 2026) extends this idea into a unified world foundation platform that integrates video world modeling with action decoding for robotic manipulation. DreamDojo (Gao et al., 2026a) pushes the foundation-model regime further by pretraining on large-scale human egocentric videos, introducing continuous latent actions to bridge unlabeled human interaction and robot control, and then post-training the model for target embodiments. It shows that such a model can support long-horizon real-time rollout, policy evaluation, and model-based planning.

A complementary argument is made by WoW (Chi et al., 2025c). It emphasizes that physical intuition cannot be acquired from passive video observation alone, and instead trains a large generative world model on extensive robot interaction trajectories. By coupling generative rollout with inverse dynamics and critique, it explicitly closes the imagination-to-action loop. At the platform level, UnifoLM-WMA-0 (Unitree, 2025) and Cosmos Predict 2.5 (Ali et al., 2025) further reflect the trend toward reusable world backbones, while GigaWorld-0 (Team et al., 2025b) makes the data-engine perspective fully explicit by combining a controllable video branch with a physically grounded 3D branch for large-scale embodied data synthesis. ABot-PhysWorld (Chen et al., 2026d) extends this trend toward physics-aligned world foundation models. It explicitly targets physically plausible and action-controllable manipulation video generation.

Taken together, these works indicate a broader transition in the field. Robotic video generation is increasingly treated not as an isolated generative task, but as a foundation layer for interactive world modeling.

5.6 Technical Progression and Open Challenges

Viewed together, the literature reveals a clear technical progression. Early methods such as Dreamitate, RoboDreamer, DreMa, ManipDreamer, DreamGen, and PhysWorld mainly treat video generation as imagination that supplies additional supervision for policy learning (Liang et al., 2024; Zhou et al., 2024; Barcellona et al., 2025; Li et al., 2025f; Jang et al., 2025a; Mao et al., 2025). The next wave, including IRASim, RoboEnvision, RoboMaster, Ctrl-World, EnerVerse-AC, Interactive World Simulator, and EVA, makes action alignment, controllable rollout, interactive usability, executability, and evaluation utility central objectives (Zhu et al., 2025b; Yang et al., 2025; Fu et al., 2026; Guo et al., 2026b; Jiang et al., 2025c; Wang et al., 2026c,b). A parallel line introduces richer interaction structure through masks, geometry, and multi-view identity cues, as

in Mask2IV, TesserAct, and RoboVIP (Li et al., 2025a; Zhen et al., 2025; Wang et al., 2026a). The newest systems, including Vid2World, Genie Envisioner, DreamDojo, WoW, ABot-PhysWorld, UnifoLM-WMA-0, Cosmos Predict 2.5, and GigaWorld-0, increasingly elevate robotic video generation into a reusable foundation layer for embodied world modeling (Huang et al., 2026; Liao et al., 2026; Gao et al., 2026a; Chi et al., 2025c; Chen et al., 2026d; Unitree, 2025; Ali et al., 2025; Team et al., 2025b).

This progression also clarifies the central bottleneck of the field. The key challenge is no longer simply to generate realistic futures. It is to generate futures that remain causally aligned with robot actions, physically and kinematically self-consistent over long horizons, coherent across views and embodiments, stable under interaction, and executable enough to support real policy improvement. For robotics, therefore, the true value of video generation lies in turning future prediction into a controllable, interactive, and actionable interface between perception and decision making.

6 World Model for Other Applications

6.1 World Model for Navigation

World models have become a useful abstraction for embodied navigation, where agents must act under severe partial observability and reason about spaces, objects, or paths that are not yet visible. Instead of treating navigation as a purely reactive next-step decision problem, this line of work uses world models to imagine action-conditioned future observations, construct future-aware planning states, or derive value-like signals from imagined trajectories before the agent physically moves (Koh et al., 2021; Bar et al., 2025; Huang et al., 2025b; Nie et al., 2025; Sharma et al.). In this sense, the world model turns unseen space into a predictive planning substrate for reasoning about future visibility, traversability, and goal progress (Koh et al., 2021; Bar et al., 2025).

Early works emphasize look-ahead prediction of unseen observations. Pathdreamer generates plausible future 360° RGB, depth, and semantic observations for unvisited indoor viewpoints, showing that planning with imagined observations can substantially reduce the gap to planning with true future observations (Koh et al., 2021). VISTA introduces an “imagine-and-align” strategy for instruction-conditioned visual imagination (Huang et al., 2025c), while VISTA_{v2} extends this into a navigation world model that rolls out egocentric futures under candidate actions and projects them into an online value map for planning (Huang et al., 2025b). A parallel trend scales this idea into controllable video predictors: NWM formulates controllable video generation explicitly as a navigation world model (Bar et al., 2025); SparseVideoNav replaces dense long-horizon rollout with sparse future generation for faster deployment (Zhang et al., 2026); and EgoWM adapts pretrained Internet-scale video diffusion models into action-conditioned egocentric world models through lightweight conditioning (Bagchi et al., 2026). Taken together, these methods show that the value of world models in navigation lies less in visual realism itself than in exposing hidden future structure in a form usable for planning, trajectory ranking, and value estimation.

6.2 World Model for Autonomous Driving

World models have likewise become an important paradigm in autonomous driving, where they increasingly unify perception, prediction, planning, and simulation. Compared with robotic manipulation, driving places stronger demands on long-horizon forecasting, multi-agent interaction, structured geometry, and safety-critical planning. Accordingly, driving world models often learn future-evolving scene representations—in image space, multi-view space, occupancy space, or compact latent space—and use them to support downstream planning or end-to-end driving decisions (Hu et al., 2022, 2023; Zheng et al., 2024; Wang et al., 2024b). Early and representative works already reveal two complementary routes. One emphasizes compact or structured predictive states: MILE learns a latent dynamics model with geometric inductive bias for urban driving (Hu et al., 2022), while OccWorld formulates world modeling in 3D occupancy space so that ego motion and scene evolution are represented in a planning-compatible form (Zheng et al., 2024). The other makes the generative world-model view more explicit: GAIA-1 casts driving world modeling as multimodal sequence modeling over video, text, and action tokens (Hu et al., 2023), and DriveDreamer uses diffusion-based modeling with structural constraints to capture complex traffic evolution from real driving data (Wang et al., 2024a).

More recent work pushes these directions toward planning-oriented and unified driving intelligence. DriveWM generates controllable multiview future videos and uses imagined multi-future rollouts together with image-based rewards to select safer trajectories (Wang et al., 2024b). UniDWM further argues for a structure- and dynamics-aware latent world representation as a unified substrate for perception, prediction, and planning (Xiong et al., 2026). DriveWorld-VLA strengthens the connection between world modeling and action generation by using latent world states as the planner’s decision state, allowing action-conditioned imagination to guide control without expensive pixel rollout (Liu et al., 2026a). From the scaling perspective, DriveVLA-W0 shows that future-image prediction through world modeling provides dense self-supervision that improves end-to-end driving VLAs beyond low-dimensional action supervision alone (Li et al., 2026c). Furthering this hierarchical synergy, SteerVLA (Gao et al., 2026b) conceptually frames the high-level VLM as a semantic world model that generates fine-grained, common-sense reasoning to steer a low-level VLA policy through complex, long-tail driving maneuvers. Overall, world models in autonomous driving are best understood as a bridge from passive scene understanding to predictive driving intelligence (Hu et al., 2023; Wang et al., 2024b; Xiong et al., 2026): some methods use them as controllable simulators of future scenes, some as structured state spaces for planning, and some as dense predictive supervision for scaling end-to-end driving policies (Zheng et al., 2024; Li et al., 2026c). Across these variants, the shared intuition is the same: safe driving requires not only recognizing the current scene, but reasoning over how it may evolve under ego behavior and surrounding traffic dynamics (Hu et al., 2022; Wang et al., 2024a; Xiong et al., 2026).

7 Benchmarks, Datasets, and Results

7.1 Benchmarks for World Model Evaluation

In embodied intelligence, evaluating world models differs fundamentally from evaluating video generation models in conventional computer vision. In robotics, the value of a world model depends on whether it can generate action-conditioned future states that remain consistent with real physical dynamics. This requires the model to capture robot–environment interactions beyond surface-level realism, functioning instead as a faithful predictor of physically plausible and temporally coherent future observations.

More importantly, such a model should respond reliably to action interventions, remain coherent over long horizons, and support downstream tasks such as policy learning, planning, and evaluation. From this policy-centric perspective, visual realism alone is neither necessary nor sufficient (Shang et al., 2026), since rollouts may look convincing while still violating dynamics in ways that break closed-loop control (Qin et al., 2025; Li et al., 2025e). Accordingly, we organize existing embodied world model benchmarks into three complementary categories: (i) action-conditioned generation and open-loop predictive quality, (ii) closed-loop task utility and policy evaluation, and (iii) physical consistency, controllability, and executability diagnostics.

7.1.1 Action-conditioned generation and open-loop predictive quality

The first dimension evaluates embodied world models in an open-loop setting. Given the current observation together with an action sequence, language instruction, or task specification, the model is asked to autoregressively generate future observations without being embedded into a planner or control loop. The key question is whether the predicted future remains faithful to the commanded behavior over time, in terms of semantic correctness, temporal coherence, and action responsiveness, rather than visual plausibility alone. Open-loop benchmarks are attractive because they are relatively easy to scale and standardize, although their results should be interpreted with care.

Recent benchmarks in this direction have become increasingly embodied. Rather than treating robot video generation as generic video synthesis, RBench (Deng et al., 2026) and EWMBench (Yue et al., 2025) evaluate whether generated futures preserve the task-relevant structure of embodied interaction. RBench emphasizes structural consistency, physical plausibility, and action completeness across diverse robotic tasks and embodiments. EWMBench adopts a more factorized view, separating scene consistency, motion correctness, and semantic alignment. Together, they reflect a broader shift in open-loop evaluation from appearance-level realism to interaction-faithful prediction.

Other benchmarks further connect open-loop prediction to downstream utility. DreamGen Bench (Jang

et al., 2025a) evaluates instruction following and physics alignment, asking whether generated rollouts are useful as synthetic experience for policy learning rather than merely realistic. EVA-Bench (Chi et al., 2025b) complements this view by emphasizing long-horizon anticipation and out-of-domain robustness under variations in viewpoint, scene layout, and motion distribution. Overall, these benchmarks suggest that strong open-loop world models must do more than generate plausible futures: they must remain action-grounded, physically sensible, and robust enough to support downstream embodied use.

7.1.2 Closed-loop task utility and policy evaluation

While open-loop benchmarks evaluate whether a world model can generate action-conditioned futures, closed-loop benchmarks ask whether those predictions remain useful inside an interactive decision loop. In this setting, the world model is evaluated not as a passive predictor, but as an environment simulator, policy evaluator, or planning substrate that directly influences action selection over time. The focus therefore shifts from predictive plausibility to decision utility: whether the model preserves the task-relevant dynamics needed for policy ranking, value estimation, planning, and ultimately task success. This makes closed-loop evaluation more aligned with embodied intelligence, since small modeling errors can accumulate and break control once the agent repeatedly acts on model-generated futures.

Recent benchmarks in this category differ in protocol, but share the same principle: a useful embodied world model must support downstream decision making, not just realistic generation. WorldArena (Shang et al., 2026) makes this explicit by evaluating world models not only with perceptual criteria, but also through functional roles such as synthetic data generation, policy evaluation, and action planning, highlighting the gap between visual realism and embodied utility. WorldEval (Li et al., 2025e) operationalizes this idea through comparative policy assessment, asking whether rollouts in a learned world model preserve the relative ordering of robot policies and checkpoints. WorldGym (Quevedo et al., 2025) extends this setting by treating the learned model as an interactive environment for Monte Carlo evaluation, focusing on whether estimated policy values and success trends match those in the real world. Across these works, rank consistency, value fidelity, and decision reliability emerge as more informative criteria than pixel-level accuracy.

A stricter version of this evaluation places the world model directly inside a closed-loop planning pipeline and measures embodied task success. World-in-World (Zhang et al., 2025a) is representative of this setting: by providing a unified interface for integrating heterogeneous world models into online planning tasks, it tests whether the model can improve control under iterative prediction and replanning. This is a harder setting than open-loop rollout evaluation because it exposes compounding errors that arise when prediction and action interact over time. Overall, recent evidence suggests that visual plausibility is only a weak proxy for control utility, whereas action-grounded consistency and controllability are much more reliable indicators of downstream embodied usefulness.

7.1.3 Physical consistency, controllability, and executability diagnostics

While open-loop benchmarks evaluate predictive quality and closed-loop benchmarks evaluate downstream utility, diagnostic benchmarks ask a more targeted question: which properties of a generated rollout determine whether it is actually usable for embodied control? This dimension focuses on whether predicted futures preserve the physical and action-relevant structure required for execution, including consistency with dynamics, responsiveness to action interventions, and recoverability into valid control signals. Rather than measuring overall prediction quality or end-task success, it probes the specific failure modes that often explain why visually plausible rollouts still fail in planning, policy evaluation, or execution.

WorldSimBench (Qin et al., 2025) is representative of this direction. It combines perceptual evaluation with manipulative evaluation, asking not only whether generated videos look realistic, but also whether they remain sufficiently consistent with action and environment dynamics to support inverse-dynamics recovery and downstream control. WoW-World-Eval (Fan et al., 2026) provides a broader but closely related perspective. Although it spans perception, planning, prediction, execution, and generalization, it is especially relevant here because it introduces physical-law and execution-oriented criteria, including an IDM-based Turing Test for whether generated videos induce plausible and executable actions. Together, these benchmarks make clear that visual plausibility alone is insufficient: generated rollouts must also preserve physically grounded and operationally executable action consequences.

Table 3 Core attributes of representative datasets/resources for embodied world model training. Benchmark-only resources discussed in the evaluation section are excluded.

Name	Year	Source	X-Emb.	Act.	Obs./3D	Lang.	M/C
RoVid-X (Deng et al., 2026)	2026	Real/Robot video	–	–	–	✓	✗
Open X-Embodiment (OXE) (O’Neill et al., 2024)	2024	Real	✓	✓	–	–	–
DROID (Khazatsky et al., 2024)	2024	Real	✗	✓	–	✓	–
BridgeData V2 (Walke et al., 2023)	2023	Real	✗	✓	–	✓	–
AgiBot World (Bu et al., 2025a)	2025	Real	–	✓	–	✓	–
Galaxea Open-World Dataset (Jiang et al., 2025a)	2025	Real	–	✓	–	✓	–
Humanoid Everyday (Zhao et al., 2025)	2025	Real	–	✓	✓	✓	✓
RoboMIND 2.0 (Hou et al., 2025)	2025	Real+Sim	✓	✓	–	✓	✓
FastUMI-100K (Liu et al., 2025b)	2025	Real	–	✓	✓	✓	–
BRMData (Zhang et al., 2024)	2024	Real	–	✓	✓	–	–
UMI (Chi et al., 2024)	2024	Real	–	✓	–	–	✗
MV-UMI (Rayyan et al., 2025)	2025	Real	✓	✓	✓	–	✗
ActiveUMI (Zeng et al., 2025)	2025	Real	–	✓	✓	–	✗
TWIST2 (Ze et al., 2025)	2025	Real	–	✓	–	✗	✗
DexWild (Tao et al., 2025)	2025	Human+Robot	✓	–	–	✗	✗
EgoMimic (Kareer et al., 2025)	2025	Human+Robot	–	–	✓	✗	✗
PHSD / In-N-On (Cai et al., 2025)	2025	Human ego	✗	–	–	✗	✗
UniHand (Luo et al., 2025)	2025	Human video	–	–	–	✓	✗
UniHand 2.0 (Luo et al., 2026)	2026	Human+Robot+VLM	✓	✓	–	✓	✗
Hoi! (Engelbracht et al., 2025)	2025	Human+Robot	✓	✓	✓	✗	✓
FreeTacMan (Wu et al., 2025)	2025	Robot-free	–	✓	✓	✗	✓
Humanoid Visual-Tactile-Action (Kwon et al., 2025)	2025	Real	✗	✓	–	✗	✓
VTDexManip (Liu et al., 2025c)	2025	Human tactile	✗	–	–	✗	✓
RH20T (Fang et al., 2023)	2023	Real	–	✓	–	✓	✓
RH20T-P (Chen et al., 2025e)	2025	Real	–	✓	–	✓	✓
RoboTwin 2.0 (Chen et al., 2025d)	2025	Sim	✓	✓	–	✓	–
Action100M (Chen et al., 2026a)	2026	Web video	✗	–	✗	✓	✗

X-Emb.: cross-embodiment coverage. **Act.:** explicit action supervision or aligned action proxy. **Obs./3D:** strong observation support beyond basic monocular RGB, e.g., multi-view, depth, LiDAR, or 3D annotations. **Lang.:** language/task conditioning. **M/C:** multimodal or contact-rich signals such as force, tactile, audio, or dense proprioceptive/contact cues.

✓ denotes strong support, – denotes partial/moderate support, and ✗ denotes absent or not emphasized.

Related evidence appears in adjacent domains such as autonomous driving. DrivingGen (Zhou et al., 2026) evaluates generative driving world models not only by visual realism, but also by trajectory plausibility, temporal coherence, and controllability under ego conditioning. Its results reveal a trade-off between appearance quality and physically reliable motion generation, reinforcing the broader point that action-conditioned world models should be judged by control-relevant dynamics rather than visual appeal alone.

A complementary diagnostic direction examines the component abilities underlying world modeling itself. WM-ABench (Gao et al., 2025b) fits this role by decomposing evaluation into atomic capabilities such as spatial and temporal understanding, motion perception, mechanistic simulation, and controlled counterfactual reasoning. Although such benchmarks do not directly test rollout executability, they are useful for identifying which internal predictive or causal capacities are missing when a model fails in more integrated open-loop or closed-loop settings.

Overall, these three categories form a layered evaluation framework for embodied world models. Open-loop benchmarks test whether the model can generate coherent action-conditioned futures; closed-loop benchmarks test whether those predictions remain useful for planning and policy evaluation; diagnostic benchmarks test whether the generated futures are physically grounded, controllable, and executable. Together, they highlight a broader lesson from recent work: no single metric is sufficient for embodied world model evaluation. A strong model must not only predict plausible futures, but also preserve the action-relevant structure needed for reliable control.

Table 4 Relevance of representative datasets/resources to embodied world-modeling capabilities.

Name	General Traj.	Long-Horizon	X-Emb. Scaling	Human Prior	Contact / Physics	Synth. / Recipe
RoVid-X (Deng et al., 2026)	–	–	–	×	–	×
Open X-Embodiment (OXE) (O’Neill et al., 2024)	✓	–	✓	×	–	×
DROID (Khazatsky et al., 2024)	✓	–	×	×	–	×
BridgeData V2 (Walke et al., 2023)	✓	–	×	×	–	×
AgiBot World (Bu et al., 2025a)	✓	–	–	×	–	×
Galaxea Open-World Dataset (Jiang et al., 2025a)	✓	–	–	×	–	×
Humanoid Everyday (Zhao et al., 2025)	✓	✓	–	×	✓	×
RoboMIND 2.0 (Hou et al., 2025)	✓	✓	✓	×	✓	–
FastUMI-100K (Liu et al., 2025b)	✓	✓	–	×	–	×
BRMData (Zhang et al., 2024)	✓	✓	–	×	–	×
UMI (Chi et al., 2024)	–	–	–	✓	×	×
MV-UMI (Rayyan et al., 2025)	–	–	✓	✓	×	×
ActiveUMI (Zeng et al., 2025)	–	–	–	✓	×	×
TWIST2 (Ze et al., 2025)	–	–	–	×	–	×
DexWild (Tao et al., 2025)	–	–	✓	✓	×	×
EgoMimic (Kareer et al., 2025)	×	–	–	✓	×	×
PHSD / In-N-On (Cai et al., 2025)	×	–	×	✓	×	×
UniHand (Luo et al., 2025)	×	–	–	✓	×	×
UniHand 2.0 (Luo et al., 2026)	–	–	✓	✓	×	✓
Hoi! (Engelbracht et al., 2025)	–	–	✓	–	✓	×
FreeTacMan (Wu et al., 2025)	–	–	–	–	✓	×
Humanoid Visual-Tactile-Action (Kwon et al., 2025)	–	–	×	×	✓	×
VTDexManip (Liu et al., 2025c)	×	–	×	✓	✓	×
RH20T (Fang et al., 2023)	–	–	–	×	✓	×
RH20T-P (Chen et al., 2025e)	–	–	–	×	✓	×
RoboTwin 2.0 (Chen et al., 2025d)	–	–	✓	×	–	✓
Action100M (Chen et al., 2026a)	×	–	×	–	×	✓

✓ strong relevance; – partial relevance; × limited or no direct relevance.

7.2 Datasets for World Model Training

Complementary to benchmarks, which specify how embodied world models should be evaluated, training datasets determine what kinds of experience such models can learn from in the first place. For embodied intelligence, such data are not merely collections of videos, but samples of *agent–environment transitions* that may couple observations with actions, task progression, embodiment-specific constraints, and physical interaction dynamics. As a result, the value of a dataset is not defined by scale alone, but by whether it provides sufficiently rich action-conditioned transitions, long-horizon task structure, diversity across scenes and embodiments, and coverage of manipulation-relevant physical signals. These properties jointly determine whether a world model can acquire dynamics priors that are genuinely useful for prediction, planning, and control.

Existing resources relevant to embodied world model training are rarely well characterized by a single taxonomy. A given dataset may simultaneously serve as a general-purpose trajectory corpus, a cross-embodiment aggregation resource, a human-to-robot prior, and a multimodal interaction dataset. For this reason, rather than forcing datasets into mutually exclusive categories, we compare them along several complementary dimensions. Table 3 summarizes their core data attributes, including embodiment coverage, action supervision, observation and 3D support, language conditioning, and multimodal or contact-rich signals. Table 4 further organizes the same resources by the kinds of world-modeling capability they are most likely to support, including general trajectory pretraining, long-horizon modeling, cross-embodiment scaling, human-prior transfer, contact- and physics-aware modeling, and synthetic or recipe-driven data scaling.

Taken together, these comparisons suggest that current training resources are better understood as spanning several parallel axes rather than falling into disjoint groups. Large-scale robot trajectory corpora provide the basic transition coverage needed for action-conditioned prediction, while cross-embodiment datasets encourage more transferable dynamics priors across platforms. Human-video and human-to-robot resources offer an additional route for learning interaction regularities beyond robot-collected trajectories, whereas tactile-, force-, and contact-rich datasets are particularly important for grounding executability and physical consistency. In parallel, synthetic data and aggregated data recipes broaden the controllable variation available for training. This multi-axis view also clarifies a central limitation of the current landscape: despite the rapid growth of available resources, failure recovery, decision-sensitive variation, and dense physically grounded supervision remain much scarcer than large-scale successful demonstrations.

Table 5 Representative results on the LIBERO standard 4-suite benchmark, grouped by how world modeling is integrated with policy learning. Methods with no directly reported number under the standard Spatial/Object/Goal/Long protocol are omitted.

Group	Method	Spatial	Object	Goal	Long	Avg
Decoupled	UniPi (Du et al., 2023)	–	–	–	0.0	–
	MimicVideo (Pai et al., 2025)	94.2	96.8	90.6	94.0	93.9
	Say-Dream-ACT (Gu et al., 2026)	99.4	99.2	98.6	95.4	98.1
Single-backbone	UVA (Li et al., 2025c)	–	–	–	90.0	–
	VideoPolicy (Liang et al., 2025a)	–	–	–	94.0	–
	Cosmos Policy (Kim et al., 2026)	98.1	100.0	98.2	97.6	98.5
	UD-VLA (Chen et al., 2026b)	94.1	95.7	91.2	89.6	92.7
MoE / MoT	Motus (Bi et al., 2025)	96.8	99.8	96.6	97.6	97.7
	LingBot-VA (Li et al., 2026b)	98.5	99.6	97.2	98.5	98.5
Unified VLA	RynnVLA-002 (Cen et al., 2025)	99.0	99.8	96.4	94.4	97.4
	DreamVLA (Zhang et al., 2025e)	97.5	94.0	89.5	89.5	92.6
	UniVLA (Bu et al., 2025b)	96.5	96.8	95.6	92.0	95.2
	Unified VLA (Wang et al., 2025)	95.4	98.8	93.6	94.0	95.5
	CoWVLA (Yang et al., 2026a)	97.2	97.8	94.6	92.8	95.6
	F1 (Lv et al., 2025)	98.2	97.8	95.4	91.3	95.7
	TriVLA (Liu et al., 2025d)	91.2	93.8	89.8	73.2	87.0
Latent-space WM	VLA-JEPA (Sun et al., 2026)	96.2	99.6	97.2	95.8	97.2
	JEPA-VLA (Miao et al., 2026)	97.2	98.0	95.6	94.8	96.4

Methods are grouped by the manner in which world modeling is incorporated into policy learning, rather than by publication year alone.

Avg denotes the average success rate over the four LIBERO suites when directly reported by the original paper. “–” indicates that the corresponding suite-level result was not directly reported under the standard Spatial/Object/Goal/Long protocol.

7.3 Representative Results on Common Benchmarks

Building on the previous discussions of evaluation protocols and training data, we briefly summarize representative results on common downstream manipulation benchmarks. Because evaluation criteria for embodied world models are often benchmark-specific, we focus on task success rate and closely related completion metrics, which are the most widely reported and directly comparable indicators of downstream performance.

Tables 5 and 6 collect representative results from recent embodied world-model and world-model-related methods. For clarity, we group methods by how world modeling is integrated with policy learning, including decoupled pipelines, shared-backbone designs, mixture-based architectures, unified VLA-style formulations, and latent-space world-model variants. Although these categories are not strictly exclusive, they provide a useful high-level view of design differences.

Table 5 focuses on the LIBERO (Liu et al., 2023a) standard 4-suite setting. We retain the breakdown into Spatial, Object, Goal, and Long suites, since methods with similar averages can still differ substantially across subsets. Table 6 complements this with results on RoboTwin 2.0 (Chen et al., 2025d), CALVIN (Mees et al., 2022), and SIMPLER-style (Li et al., 2025d) benchmarks. These evaluations are more heterogeneous

Table 6 Representative results on RoboTwin, CALVIN, and SIMPLER-style benchmarks, grouped by how world modeling is integrated with policy learning. Methods with no directly reported number under the corresponding protocols are omitted.

Group	Method	RT-A	RT-B	C-A	C-D	S-G	S-W	S-O
Decoupled	UniPi (Du et al., 2023)	–	–	0.92	–	–	–	–
	VidMan (Wen et al., 2024)	–	–	3.42	–	–	–	–
	Vidar (Feng et al., 2025)	65.8	17.5	–	–	–	–	–
	VPP (Hu et al., 2025)	–	–	4.33	–	–	–	–
	Video2Act (Jia et al., 2025b)	54.6	54.1	–	–	–	–	–
	MimicVideo (Pai et al., 2025)	–	–	–	–	–	–	46.9/56.3
Single-backbone	VideoVLA (Shen et al., 2025)	–	–	–	–	73.1/62.8	53.1	63.0
	UD-VLA (Chen et al., 2026b)	–	–	–	4.64	–	62.5	–
MoE / MoT	Motus (Bi et al., 2025)	88.7	87.0	–	–	–	–	–
	LingBot-VA (Li et al., 2026b)	92.9	91.6	–	–	–	–	–
	LingBot-VLA (Wu et al., 2026a)	88.6	86.7	–	–	–	–	–
	BagelVLA (Hu et al., 2026)	75.3	20.9	4.41	–	–	–	–
	FRAPPE (Zhao et al., 2026)	57.5	25.5	–	–	–	–	–
Unified VLA	GR-1 (Wu et al., 2024)	–	–	3.06	4.21	–	–	–
	UP-VLA (Zhang et al., 2025c)	–	–	4.08	4.42	–	–	–
	DreamVLA (Zhang et al., 2025e)	–	–	4.44	–	–	–	–
	Unified VLA (Wang et al., 2025)	–	–	4.41	4.63	–	69.8	–
	CoWVLA (Yang et al., 2026a)	–	–	4.21	4.47	60.9	76.0	–
	F1 (Lv et al., 2025)	–	–	–	–	–	–	72.9
	InternVLA-A1 (Cai et al., 2026)	89.4	87.0	–	–	–	–	–
	HALO (Shou et al., 2026)	80.5	26.4	–	–	–	–	–
Latent-space WM	TriVLA (Liu et al., 2025d)	–	–	4.37	–	–	–	–
	VLA-JEPA (Sun et al., 2026)	–	–	–	–	65.2	57.3	–
	JEPA-VLA (Miao et al., 2026)	73.5	17.7	–	–	–	–	–
	WoG (Su et al., 2026)	–	–	–	–	69.4	63.5	–

Methods are grouped by the manner in which world modeling is incorporated into policy learning, rather than by publication year alone.

RT-A and **RT-B** denote the two RoboTwin evaluation settings, corresponding to the simple setting with non-randomized testing environments and the harder setting with randomized environments, respectively.

C-A and **C-D** denote the CALVIN ABCD and ABCDD protocols, respectively.

S-G, **S-W**, and **S-O** denote SIMPLER-style results on Google Robot, WidowX, and other reported setups, respectively.

“–” indicates that no directly reported number was found under the corresponding protocol. Entries of the form “a/b” denote two reported protocol variants in the original source.

in embodiment and protocol, so they are less suitable for strict ranking but still useful for revealing cross-benchmark variation.

Several patterns are worth noting. First, strong results are not limited to a single architectural paradigm: competitive performance appears across decoupled, shared-backbone, unified, mixture-based, and latent predictive designs. This suggests that the utility of world modeling for embodied control is not tied to one specific implementation. Second, the LIBERO breakdown shows that long-horizon manipulation remains a key differentiator. While many methods already perform strongly on Spatial and Object suites, larger drops are more common on Goal and especially Long suites, where success depends more on sustained, action-grounded consistency over extended trajectories.

Results on RoboTwin, CALVIN, and SIMPLER-style benchmarks further support this point, while also highlighting stronger benchmark dependence. Compared with LIBERO, these settings are more fragmented, and strong performance on one benchmark does not necessarily transfer to another. This suggests that current embodied world models are still sensitive to differences in embodiment, action space, task composition, and evaluation protocol.

Overall, these results suggest three conclusions. First, embodied world models already show strong practical utility on standard downstream manipulation benchmarks. Second, high performance can emerge from multiple design paradigms, indicating that photorealistic video generation is not necessary for effective embodied

control. Third, the main remaining challenges lie in long-horizon robustness, cross-benchmark generalization, and the lack of standardized reporting across platforms.

8 Challenges and Future Directions

Despite the promise of world models for robot learning, reliable deployment in complex embodied tasks remains limited by several core challenges beyond simple scaling. Current systems must address causal conditioning gaps in action-dependent dynamics, efficiency bottlenecks in training and inference, limited integration of non-visual sensory feedback, and the lack of standardized evaluation focused on functional utility rather than visual realism. Another important frontier is symbolic and structured abstraction: while pixel- or latent-space prediction is powerful, long-horizon reasoning may require object-centric, relational, or rule-like structure that provides a more compact interface for planning and control. In this section, we discuss these challenges and outline future directions toward reliable, efficient, and actionable world models for embodied agents.

8.1 Causal Conditioning Gaps

Current VLA frameworks often couple world models with inverse dynamics (Li et al., 2026b; Ye et al., 2026b; Team et al., 2026), using future-state prediction to regularize policy learning. However, a causal misalignment can arise when the predicted future is conditioned more strongly on historical context or task intent than on the specific pending robot action. In such cases, the world model may generate futures that are semantically plausible or intention-consistent, but not necessarily faithful to the physical consequences of the candidate action. This limits its usefulness for precise closed-loop control, where the key requirement is not only to predict a likely future, but to predict how the future changes under the robot’s own intervention.

The technical bottleneck is weak action conditioning: many predictive world-model objectives are trained mainly from observation history and task intent, so their futures can be plausible without being causally tied to the robot action to be executed. To reduce this mismatch, WorldVLA (Cen et al., 2025) adopts implicit unified training strategies that couple future-state prediction with action generation, encouraging more policy-aligned predictive dynamics.

8.2 Efficiency Bottlenecks

World-model-based policies are far more computationally intensive than VLA models, especially during both training and inference. This overhead arises because models either jointly predict future videos and actions or require fine-tuning before policy learning, making adaptation costly due to large model sizes and complex environment dynamics. Parameter-efficient strategies, like lightweight adapters, can mitigate this by keeping the base model largely frozen. Efficiency issues also appear at inference, particularly for diffusion-based video prediction, where iterative denoising causes high latency. Recent approaches like Mimic Video (Pai et al., 2025) and LingBot-VA (Li et al., 2026b) mitigate this via partial denoising. These methods prioritize motion dynamics over fine-grained visual details, capturing essential cues for decision-making without the cost of full reconstruction.

More fundamentally, recent approaches rethink world models entirely. Latent-space models, such as LeWorld-Model (Maes et al., 2026), reduce training and inference costs by focusing on predictive representations rather than full high-dimensional generation. Emerging paradigms, like Fast-WAM (Yuan et al., 2026), further decouple world modeling from deployment, using it only for training to enhance representations while eliminating it at inference.

8.3 Multi-Modal Perception Bottlenecks

Current world models excel in visual synthesis but remain decoupled from the physical dynamics of real-world interaction. Relying predominantly on vision and proprioception fails to capture unobservable properties such as friction, stiffness, and contact stability. To resolve these, integrating haptic sensing and force feedback (Tang et al., 2026; Huang et al., 2025a) is indispensable for providing ground-truth interaction signals. Recent visuo-tactile models (Higuera et al., 2026; Zheng et al., 2026) have begun addressing this by learning joint latent representations to enhance robustness in contact-rich tasks.

A significant architectural challenge lies in aligning asynchronous signals with divergent frequencies and dimensions. While tactile sensors capture high-frequency transient events, their low-dimensional signals are often diluted or overwhelmed by high-dimensional visual features during joint latent optimization (Chen et al., 2025c). Effectively balancing these heterogeneous inputs is essential for preventing visual dominance and ensuring that sparse visual semantics are fused with dense physical feedback, a critical step toward physics-aware robotic intelligence.

8.4 Classical Control Integration

World models serve as forward dynamics for proactive planning via MPC (Hansen et al., 2022, 2024; Maes et al., 2026). By optimizing action sequences to minimize cumulative costs, agents use imagined rollouts to bridge reactive execution with strategic reasoning. However, a major bottleneck is the massive computational overhead. MPC requires iterative world model rollouts for action optimization, which significantly limits the real-time deployment of high-capacity models in dynamic environments.

Unlike analytical kinematics, world models capture the joint stochastic evolution of both the agent and its environment. A critical frontier lies in reconciling this neural expressivity with formal control guarantees, such as Lyapunov stability or robust control (Jia et al., 2025a). Fusing learned dynamics with existing mature control principles, not just MPC, presents a potential pathway toward self-adaptive robotic systems capable of operating in non-stationary, open-world settings.

8.5 Symbolic Structure Integration

While this survey has primarily focused on visual and latent world models, symbolic world models provide an important complementary direction. Instead of predicting pixels, they operate over structured states, such as objects, relations, predicates, or occupancy maps, enabling more stable and compositional predictions. A key limitation of pixel-based rollouts is long-horizon error accumulation, which can degrade planning reliability. Symbolic representations mitigate this by abstracting away low-level details and modeling discrete or rule-based transitions, allowing more reliable reasoning over extended horizons. However, they often require suitable abstractions and perception grounding, and can struggle when high-dimensional observations cannot be cleanly mapped into predefined symbols. A promising direction is therefore to build hybrid world models that combine learned perceptual representations with symbolic structure (Liang et al., 2026, 2025c; Shah et al., 2025). This is compelling because much of the real world is inherently structured: object-centric or relational abstractions learned from data, together with symbolic constraints in generative models, may offer a principled path toward scalable and reliable long-horizon world modeling.

8.6 Open Challenges in Evaluation Metrics

Another challenge for embodied world models is the absence of a widely accepted evaluation metric. Unlike conventional video generation, where perceptual fidelity is often central, embodied world models are ultimately judged by their functional value for decision making (Shang et al., 2026; Zhang et al., 2025a). A model may produce visually plausible futures yet still fail to preserve action-conditioned dynamics, causal consistency, or controllability, all of which are critical for policy learning and closed-loop execution. Conversely, limited visual realism does not necessarily preclude utility for planning or policy evaluation (Quevedo et al., 2025). As a result, evaluation remains inherently multi-dimensional, spanning predictive quality, downstream control utility, and physical executability (Fan et al., 2026), and current comparisons are still fragmented across benchmarks and protocols.

A key direction is therefore to develop function-aware evaluation frameworks that better reflect the intended role of the world model. Rather than relying on appearance-driven scores alone, future metrics should jointly assess predictive realism, action sensitivity, long-horizon consistency, and control utility. A practical goal is to establish a compact set of standardized metrics, such as task success, policy-ranking fidelity, and executability-oriented diagnostics, to enable more consistent comparison across tasks and embodiments and to distinguish visually plausible models from truly actionable ones.

References

- Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Tobias Pfaff, Cheston Tan, Henrik I Christensen, Hao Su, Jiajun Wu, and Yunzhu Li. A review of learning-based dynamics models for robotic manipulation. *Science Robotics*, 10(106):eadt1497, 2025.
- Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiabin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.
- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.
- Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Yanpin Tao, Pascal Vincent, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- Ashay Athalye, Nishanth Kumar, Tom Silver, Yichao Liang, Jiuguang Wang, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. From pixels to predicates: Learning symbolic world models via pretrained vlms. *IEEE Robotics and Automation Letters*, 11(4):4002–4009, 2026. doi: 10.1109/LRA.2026.3662533.
- Anurag Bagchi, Zhipeng Bao, Homanga Bharadhwaj, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Walk through paintings: Egocentric world models from internet priors. *arXiv preprint arXiv:2601.15284*, 2026.
- Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence*, pages 103–129. Oxford University Press, 1995.
- Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023.
- Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15791–15801, 2025.
- Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. In *International Conference on Learning Representations*, 2025.
- Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Conference on Robot Learning*, pages 3936–3951. PMLR, 2025.
- Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.
- Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π 0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- Arthur E. Bryson, Jr. and Yu-Chi Ho. *Applied optimal control: Optimization, estimation, and control*. Hemisphere Publishing Corporation, Washington, D.C., 1975. Revised printing.
- Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025a.
- Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025b.
- Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. InternVLA-A1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

- Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-N-On: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.
- Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, et al. Rynnvla-002: A unified vision-language-action and world model. *arXiv preprint arXiv:2511.17502*, 2025.
- Akshay L Chandra, Iman Nematollahi, Chenguang Huang, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. DiWA: Diffusion policy adaptation with world models. In *Conference on Robot Learning*, pages 3378–3400. PMLR, 2025.
- Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025a.
- Delong Chen, Tejaswi Kasarla, Yejin Bang, Mustafa Shukor, Willy Chung, Jade Yu, Allen Bolourchi, Theo Moutakanni, and Pascale Fung. Action100M: A large-scale video action dataset. *arXiv preprint arXiv:2601.10592*, 2026a.
- Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. VidBot: Learning generalizable 3D actions from In-the-Wild 2D human videos for zero-shot robotic manipulation. 2025b.
- Haonan Chen, Jiaming Xu, Hongyu Chen, Kaiwen Hong, Binghao Huang, Chaoqi Liu, Jiayuan Mao, Yunzhu Li, Yilun Du, and Katherine Driggs-Campbell. Multi-modal manipulation via multi-modal policy consensus. *arXiv preprint arXiv:2509.23468*, 2025c.
- Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Feilong Tang, Donglin Wang, and Haoang Li. Unified diffusion VLA: Vision-language-action model via joint discrete diffusion process. In *International Conference on Learning Representations*, 2026b.
- Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025d.
- Yi Chen, Yuying Ge, Hui Zhou, Mingyu Ding, Yixiao Ge, and Xihui Liu. Dial: Decoupling intent and action via latent world modeling for end-to-end vla. *arXiv preprint arXiv:2603.29844*, 2026c.
- Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. ABot-PhysWorld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026d.
- Zeren Chen, Zhelun Shi, Xiaoya Lu, Lehan He, Sucheng Qian, Enshen Zhou, Zhenfei Yin, Wanli Ouyang, Jing Shao, Yu Qiao, et al. RH20T-P: A primitive-level robotic manipulation dataset towards composable generalization agents in real-world scenarios. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 20532–20539. IEEE, 2025e.
- Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.
- Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *Robotics: Science and Systems*, 2024.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025a.
- Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-Min Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, et al. Empowering world models with reflection for embodied video prediction. In *International Conference on Machine Learning*, pages 10383–10410. PMLR, 2025b.
- Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025c.

- Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970. doi: 10.1080/00207727008920220.
- K.J.W. Craik. *The nature of explanation*. Cambridge University Press, 1943. ISBN 9780521047555.
- Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. RynnBrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979v1*, 2026.
- Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. The ingredients for robotic diffusion transformers. In *IEEE International Conference on Robotics and Automation*. IEEE, 2025.
- Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world. *arXiv preprint arXiv:2601.15282*, 2026.
- Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Conference on Robot Learning*. PMLR, 2024.
- Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36:9156–9172, 2023.
- Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Pack Kaelbling, et al. Video language planning. In *International Conference on Learning Representations*, 2024.
- Tim Engelbracht, René Zurbrügg, Matteo Wohlapp, Martin Büchner, Abhinav Valada, Marc Pollefeys, Hermann Blum, and Zuria Bauer. Hoil: A multimodal dataset for force-grounded, cross-view articulated manipulation. *arXiv preprint arXiv:2512.04884*, 2025.
- Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, et al. Wow, wo, val! a comprehensive embodied world model evaluation turing test. *arXiv preprint arXiv:2601.04137*, 2026.
- Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.
- Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.
- Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di ZHANG, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. In *International Conference on Learning Representations*, 2026.
- Dechen Gao, Boqi Zhao, Andrew Lee, Ian Chuang, Hanchu Zhou, Hang Wang, Zhe Zhao, Junshan Zhang, and Iman Soltani. VITA: Vision-to-action flow matching policy. *arXiv preprint arXiv:2507.13231*, 2025a.
- Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26170–26195, 2025b.
- Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K.R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026a.
- Tian Gao, Celine Tan, Catherine Glossop, Timothy Gao, Jiankai Sun, Kyle Stachowicz, Shirley Wu, Oier Mees, Dorsa Sadigh, Sergey Levine, and Chelsea Finn. Steervla: Steering vision-language-action models in long-tail driving scenarios. *arXiv preprint arXiv:2602.08440*, 2026b.
- Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

- Songen Gu, Yunuo Cai, Tianyu Wang, Simo Wu, and Yanwei Fu. Say, dream, and act: Learning video world models for instruction-driven robot manipulation. *arXiv preprint arXiv:2602.10717*, 2026.
- Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.
- Yanjiang Guo, Tony Lee, Lucy Xiaoyang Shi, Jianyu Chen, Percy Liang, and Chelsea Finn. VLAW: Iterative co-improvement of vision-language-action policy and world model. *arXiv preprint arXiv:2602.12063*, 2026a.
- Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026b.
- David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pages 2451–2463. 2018.
- Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control, 2024.
- Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, pages 8387–8406. PMLR, 2022.
- Kyle Beltran Hatch, Ashwin Balakrishna, Oier Mees, Suraj Nair, Seohong Park, Blake Wulfe, Masha Itkina, Benjamin Eysenbach, Sergey Levine, Thomas Kollar, and Benjamin Burchfiel. Ghil-glu: Hierarchical control with filtered subgoal images. In *IEEE International Conference on Robotics and Automation*. IEEE, 2025.
- Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.
- Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems*, 35:20703–20716, 2022.
- Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning*, pages 24328–24346. PMLR, 2025.
- Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, et al. BagelVLA: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*, 2026.
- Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-VLA: Unlocking vision-language-action model’s physical knowledge for tactile generalization, 2025a.
- Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026.
- Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. VISTA v2: World imagination for indoor vision-and-language navigation. *arXiv preprint arXiv:2512.00041*, 2025b.
- Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. Vista: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868*, 2025c.
- Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi_{0.6}^*$: A VLA that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025a.
- Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg,

- Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025b.
- Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzheng Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through video world models. In *Conference on Robot Learning*. PMLR, 2025a.
- Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv preprint*, pages arXiv–2505, 2025b.
- Emily Yue-Ting Jia, Weiduo Yuan, Tianheng Shi, Vitor Guizilini, Jiageng Mao, and Yue Wang. DreamPlan: Efficient reinforcement fine-tuning of vision-language planners via video world models. *arXiv preprint arXiv:2603.16860*, 2026a.
- Jindou Jia, Zihan Yang, Meng Wang, Kexin Guo, Jianfei Yang, Xiang Yu, and Lei Guo. Feedback favors the generalization of neural ODEs. In *International Conference on Learning Representations*, 2025a.
- Jindou Jia, Gen Li, Xiangyu Chen, Tuo An, Yuxuan Hu, Jingliang Li, Xinying Guo, and Jianfei Yang. Action-to-Action flow matching. *arXiv preprint arXiv:2602.07322*, 2026b.
- Yueru Jia, Jiaming Liu, Shengbang Liu, Rui Zhou, Wanhe Yu, Yuyang Yan, Xiaowei Chi, Yandong Guo, Boxin Shi, and Shanghang Zhang. Video2Act: A dual-system video diffusion policy with robotic spatio-motional modeling. *arXiv preprint arXiv:2512.03044*, 2025b.
- Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system VLA model. *arXiv preprint arXiv:2509.00576*, 2025a.
- Yuming Jiang, Siteng Huang, Shengke Xue, Yaxi Zhao, Jun Cen, Sicong Leng, Kehan Li, Jiayan Guo, Kexiang Wang, Mingxiu Chen, Fan Wang, Deli Zhao, and Xin Li. RynnVLA-001: Using human demonstrations to improve robot manipulation. *arXiv preprint arXiv:2509.15212*, 2025b.
- Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, and Guanghui Ren. EnerVerse-AC: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025c.
- Zhennan Jiang, Kai Liu, Yuxin Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Chao Yu, Haoran Li, and Dongbin Zhao. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *arXiv preprint arXiv:2509.19080*, 2025d.
- Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, et al. WoVR: World models as reliable simulators for post-training VLA policies with RL. *arXiv preprint arXiv:2602.13977*, 2026.
- Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation*, pages 13226–13233. IEEE, 2025.
- Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.
- Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.
- Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations*, 2024.

- Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021.
- Eunju Kwon, Seungwon Oh, In-Chang Baek, Yucheon Park, Gyungbo Kim, JaeYoung Moon, Yunho Choi, and Kyung-Joong Kim. A humanoid visual-tactile-action dataset for contact-rich manipulation. *arXiv preprint arXiv:2510.25725*, 2025.
- Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories. *arXiv preprint arXiv:2510.03135*, 2025a.
- Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025b.
- Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Konidaris, and Jiahui Fu. Novaflow: Zero-shot manipulation via actionable flow from generated videos. 2026a.
- Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026b.
- Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025c.
- Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, pages 3705–3728. PMLR, 2025d.
- Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025e.
- Ying Li, Xiaobao Wei, Xiaowei Chi, Yuming Li, Zhongyu Zhao, Hao Wang, Ningning Ma, Ming Lu, and Shanghang Zhang. Manipdreamer: Boosting robotic manipulation world model with action tree and visual guidance. *arXiv preprint arXiv:2504.16464*, 2025f.
- Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, AnYasong, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations*, 2026c.
- Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *Conference on Robot Learning*. PMLR, 2024.
- Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Parth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025a.
- Weixin Liang, LILI YU, Liang Luo, Srinu Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856.
- Yichao Liang, Nishanth Kumar, Hao Tang, Adrian Weller, Joshua B. Tenenbaum, Tom Silver, João F. Henriques, and Kevin Ellis. Visualpredicator: Learning abstract world models with neuro-symbolic predicates for robot planning. In *International Conference on Learning Representations (ICLR)*, 2025c.
- Yichao Liang, Thanh Dat Nguyen, Cambridge Yang, Tianyang Li, Joshua B. Tenenbaum, Carl Edward Rasmussen, Adrian Weller, Zenna Tavares, Tom Silver, and Kevin Ellis. Exopredicator: Learning abstract models of dynamic worlds for robot planning. In *International Conference on Learning Representations*, 2026.
- Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng YAN, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, 2026.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.
- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023a.

- Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025a.
- Kehui Liu, Zhongjie Jia, Yang Li, Pengan Chen, Song Liu, Xin Liu, Pingrui Zhang, Haoming Song, Xinyi Ye, Nieqing Cao, et al. FastUMI-100K: Advancing data-driven robotic manipulation with a large-scale UMI-style dataset. *arXiv preprint arXiv:2510.08022*, 2025b.
- Lin Liu, Ziyang Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, Long Chen, et al. DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. *arXiv preprint arXiv:2602.06521*, 2026a.
- Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.
- Qingtao Liu, Yu Cui, Zhengnan Sun, Gaofeng Li, Jiming Chen, and Qi Ye. VTDexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *International Conference on Learning Representations*, 2025c.
- Xiaokang Liu, Zechen Bai, Hai Ci, Kevin Yuchen Ma, and Mike Zheng Shou. World-VLA-Loop: Closed-loop learning of video world model and VLA policy. *arXiv preprint arXiv:2602.06508*, 2026b.
- Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. InstafLOW: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2023b.
- Zhenyang Liu, Yongchong Gu, Sixiao Zheng, Xiangyang Xue, and Yanwei Fu. Trivla: A triple-system-based unified vision-language-action model for general robot control. *arXiv preprint*, pages arXiv-2507, 2025d.
- T Lozano-Perez. Robot programming. In *IEEE Proceedings*, volume 71, pages 821–841, 1983.
- Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.
- Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-H0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.
- Yunhao Luo and Yilun Du. Grounding video models to actions through goal conditioned exploration. In *International Conference on Learning Representations*, 2025.
- Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.
- Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. LDA-1B: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*, 2026.
- Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.
- Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable End-to-End joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.
- Jiageng Mao, Sicheng He, Hao-Ning Wu, Yang You, Shuyang Sun, Zhicheng Wang, Yanan Bao, Huizhong Chen, Leonidas Guibas, Vitor Guizilini, et al. Robot learning from a physical world model. *arXiv preprint arXiv:2511.07416*, 2025.
- Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, pages 7327–7334, 2022.
- Weishi Mi, Yong Bao, Xiaowei Chi, Xiaozhu Ju, Zhiyuan Qin, Kuangzhi Ge, Kai Tang, Peidong Jia, Shanghang Zhang, and Jian Tang. TC-IDM: Grounding video generation for executable zero-shot robot motion. *arXiv preprint arXiv:2601.18323*, 2026.
- Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. JEPA-VLA: Video predictive embedding is needed for VLA models. *arXiv preprint arXiv:2602.11832*, 2026.

- George A. Miller, Eugene Galanter, and Karl H. Pribram. *Plans and the structure of behavior*. Holt, Rinehart and Winston, New York, 1960. doi: 10.1037/10039-000.
- Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.
- Iman Nematollahi, Oier Mees, Lukas Hermann, and Wolfram Burgard. Hindsight for foresight: Unsupervised structured dynamics models from physical interaction. pages 5319–5326. IEEE, 2020.
- Derrick Nguyen and Bernard Widrow. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *International Joint Conference on Neural Networks*, pages 21–26. IEEE, 1990.
- Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2392–2399. IEEE, 2025.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.
- Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*. IEEE, 2024.
- Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018. doi: 10.1561/23000000053.
- Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *IEEE International Conference on Robotics and Automation*, pages 6892–6903. IEEE, 2024.
- Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. Mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.
- Chaoyi Pan, Giri Anantharaman, Nai-Chieh Huang, Claire Jin, Daniel Pfrommer, Chenyang Yuan, Frank Permenter, Guannan Qu, Nicholas Boffi, Guanya Shi, and Max Simchowitz. Much ado about noising: Dispelling the myths of generative robotic control. In *International Conference on Learning Representations*, 2026.
- Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025.
- Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Inference-time enhancement of generative robot policies via predictive world modeling. *IEEE Robotics and Automation Letters*, 2026.
- Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. WorldSimBench: Towards video generation models as world simulators. In *International Conference on Machine Learning*, pages 50338–50362. PMLR, 2025.
- Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.
- Omar Rayyan, John Abanes, Mahmoud Hafez, Anthony Tzes, and Fares Abu-Dakka. MV-UMI: A scalable multi-view interface for cross-embodiment learning. *arXiv preprint arXiv:2509.18757*, 2025.
- Jacques Richalet, André Rault, Jean-Louis Testud, and Jean Papon. Model predictive heuristic control. *Automatica*, 14(5):413–428, 1978.
- Naman Shah, Jayesh Nagpal, and Siddharth Srivastava. From real world to logic and back: Learning generalizable relational concepts for long horizon robot planning. In *Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 5362–5434. PMLR, 2025.
- Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyong Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

- Rishabh Sharma, Gijs Hogervorst, Wayne Mackey, David Heeger, and Stefano Martiniani. Cross-view world models. In *ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling*.
- Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. VideoVLA: Video generators can be generalizable robot manipulators. 2025.
- Quanxin Shou, Fangqi Zhu, Shawn Chen, Puxin Yan, Zhengyang Yan, Yikun Miao, Xiaoyi Pang, Zicong Hong, Ruikai Shi, Hao Huang, et al. HALO: A unified vision-language-action model for embodied multimodal chain-of-thought reasoning. *arXiv preprint arXiv:2602.21157*, 2026.
- Tom Silver, Rohan Chitnis, Joshua B. Tenenbaum, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Learning symbolic operators for task and motion planning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3182–3189, 2021. doi: 10.1109/IROS51168.2021.9635941.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- Yue Su, Sijin Chen, Haixin Shi, Mingyu Liu, Zhengshen Zhang, Ningyuan Huang, Weiheng Zhong, Zhengbang Zhu, Yuxiao Liu, and Xihui Liu. World guidance: World modeling in condition space for action generation. *arXiv preprint arXiv:2602.22010*, 2026.
- Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.
- Tutian Tang, Xingyu Ji, Wanli Xing, Ce Hao, Wenqiang Xu, Lin Shao, Cewu Lu, Qiaojun Yu, Jiangmiao Pang, and Kaifeng Zhang. Towards human-like manipulation through RL-Augmented teleoperation and Mixture-of-Expertise VLA, 2026.
- Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. DexWild: Dexterous human interactions for In-the-Wild robot policies. *Robotics: Science and Systems*, 2025.
- Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, Fangchen Liu, Anirudha Majumdar, Andrew Marmon, Carolina Parada, Yulia Rubanova, Dhruv Shah, Vikas Sindhwani, Jie Tan, Fei Xia, Ted Xiao, Sherry Yang, Wenhao Yu, and Allan Zhou. Evaluating gemini robotics policies in a veo world simulator, 2025a. <https://arxiv.org/abs/2512.10675>.
- GigaBrain Team, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Jie Li, Jindi Lv, Jingyu Liu, Lv Feng, et al. GigaBrain-0.5 M*: A VLA that learns from world model-based reinforcement learning. *arXiv preprint arXiv:2602.12099*, 2026.
- GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied AI. *arXiv preprint arXiv:2511.19861*, 2025b.
- Unitree. UnifoLM-WMA-0: A world-model-action (WMA) framework under UnifoLM family, 2025.
- Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Boyang Wang, Haoran Zhang, Shujie Zhang, Jinkun Hao, Mingda Jia, Qi Lv, Yucheng Mao, Zhaoyang Lyu, Jia Zeng, Xudong Xu, et al. RoboVIP: Multi-view video generation with visual identity prompting augments robot manipulation. *arXiv preprint arXiv:2601.05241*, 2026a.
- Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. EVA: Aligning video world models with executable robot actions via inverse dynamics rewards. *arXiv preprint arXiv:2603.17808*, 2026b.
- Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024a.
- Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026c.

- Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024b.
- Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*, 2025.
- Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37:41051–41075, 2024.
- Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.
- Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Yufei Jiang, Ran Huang, Guoying Gu, and Hongyang Li. Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation. *arXiv preprint arXiv:2506.01941*, 2025.
- Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.
- Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic VLA foundation model. *arXiv preprint arXiv:2601.18692*, 2026a.
- Xiao-Ming Wu, Bin Fan, Kang Liao, Jian-Jian Jiang, Runze Yang, Yihang Luo, Zhonghua Wu, Wei-Shi Zheng, and Chen Change Loy. VLANeXt: Recipes for building strong VLA models. *arXiv preprint arXiv:2602.18532*, 2026b.
- Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-Env: Leveraging world model as a virtual environment for VLA post-training. *arXiv preprint arXiv:2509.24948*, 2025.
- Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. RehearseVLA: Simulated post-training for VLAs with physically-consistent world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.
- Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. UniDrive-WM: Unified understanding, planning and generation world model for autonomous driving. *arXiv preprint arXiv:2601.04453*, 2026.
- Fuxiang Yang, Donglin Di, Lulu Tang, Xuancheng Zhang, Lei Fan, Hao Li, Wei Chen, Tonghua Su, and Baorui Ma. Chain of world: World model thinking in latent motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.
- Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, et al. RISE: Self-improving robot policy with compositional world model. *arXiv preprint arXiv:2602.11075*, 2026b.
- Liudi Yang, Yang Bai, George Eskandar, Fengyi Shen, Mohammad Altilawi, Dong Chen, Soumajit Majumder, Ziyuan Liu, Gitta Kutyniok, and Abhinav Valada. Roboenvision: A long-horizon video generation model for multi-task robot manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 21281–21288. IEEE, 2025.
- Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024a.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-Video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2024b.
- Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. GigaWorld-Policy: An efficient action-centered world-action model. *arXiv preprint arXiv:2603.17240*, 2026a.
- Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

- Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, et al. PlayWorld: Learning robot world models from autonomous play. *arXiv preprint arXiv:2603.09030*, 2026.
- Zhao-Heng Yin, Sherry Yang, and Pieter Abbeel. Object-centric 3D motion field for robot learning from human videos. In *Advances in Neural Information Processing Systems*, 2025.
- Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.
- Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.
- Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.
- Qiyuan Zeng, Chengmeng Li, Jude St John, Zhongyi Zhou, Junjie Wen, Guorui Feng, Yichen Zhu, and Yi Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025.
- Fan Zhang and Michael Gienger. Affordance-based robot manipulation with flow matching. *arXiv preprint arXiv:2409.01083*, 2024.
- Hai Zhang, Siqi Liang, Li Chen, Yuxian Li, Yukuan Xu, Yichao Zhong, Fu Zhang, and Hongyang Li. Sparse video generation propels real-world beyond-the-view vision-language navigation. *arXiv preprint arXiv:2602.05827*, 2026.
- Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, et al. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025a.
- Jiahui Zhang, Ze Huang, Chun Gu, Zipei Ma, and Li Zhang. Reinforcing action policies by prophesying. *arXiv preprint arXiv:2511.20633*, 2025b.
- Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. UP-VLA: A unified understanding and prediction model for embodied agent. In *International Conference on Machine Learning*. PMLR, 2025c.
- Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097*, 2025d.
- Tianle Zhang, Dongjiang Li, Yihang Li, Zecui Zeng, Lin Zhao, Lei Sun, Yue Chen, Xuelong Wei, Yibing Zhan, Lusong Li, et al. Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks. *arXiv preprint arXiv:2405.18860*, 2024.
- Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. 2025e.
- Han Zhao, Jingbo Wang, Wenxuan Song, Shuai Chen, Yang Liu, Yan Wang, Haoang Li, and Donglin Wang. FRAPPE: Infusing world modeling into generalist policies via multiple future representation alignment. *arXiv preprint arXiv:2602.17259*, 2026.
- Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.
- Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.
- Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.
- Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, et al. FLARE: Robot learning with implicit world modeling. In *Conference on Robot Learning*, pages 3952–3971. PMLR, 2025.

- Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3D occupancy world model for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.
- Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.
- Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, pages 61885–61896. PMLR, 2024.
- Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L Waslander. DrivingGen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.
- Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025a.
- Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025b.
- Fangqi Zhu, YAN Zhengyang, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World model-based policy optimization for vision-language-action models. In *International Conference on Learning Representations*, 2026.
- Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.