EVEWORLD: PHYSICAL EVOLUTION SUPERVISION FOR EMBODIED WORLD MODELS

Anonymous authors
Paper under double-blind review

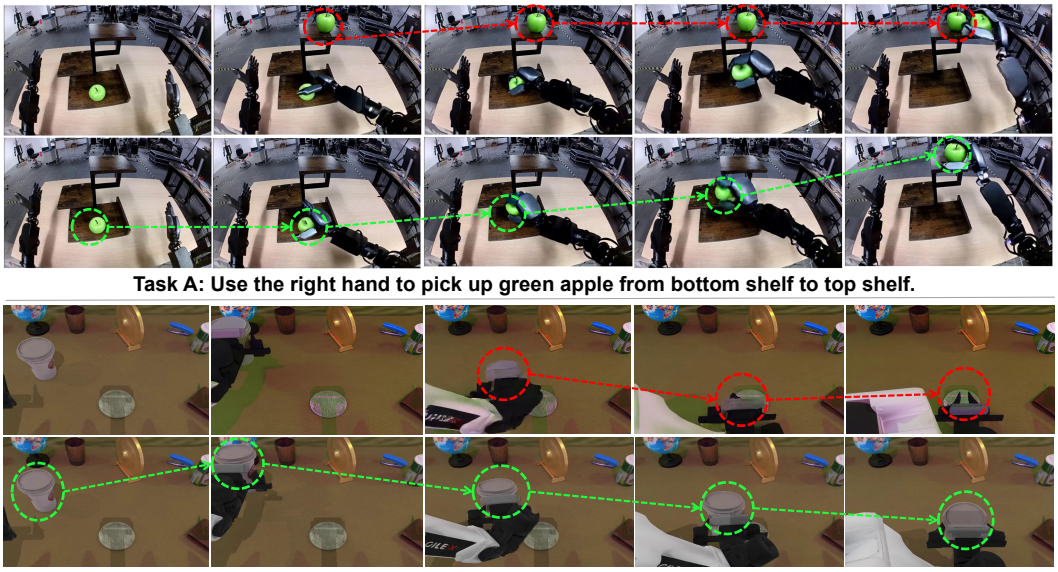


Figure 1: Embodied world models face the challenge of **Model Laziness**, where instance consistency is violated by spurious duplication or disappearance during interaction. **Red** denotes Standard SFT, which exhibits such failures, while **Green** denotes EVEWORLD, which better maintains physical consistency throughout the interaction. Dashed arrows trace target evolution over time.

ABSTRACT

Embodied world models enable scalable simulation of embodied interactions for robot learning. However, existing models are prone to **Model Laziness**, as they focus on visual fidelity at the expense of physical reasoning and lack process-level supervision over the temporal dynamics of manipulated objects. In this work, we propose EVEWORLD, a physical evolution-supervision framework for physically consistent target evolution. EVEWORLD consists of two components: Instance-Guided Restoration (IGR) and Temporal Instance Alignment (TIA). First, IGR promotes instance consistency through restoration supervision. Second, TIA promotes cross-frame consistency by aligning target instances across adjacent frames. We further introduce the Model Laziness Rate (MLR), a metric that measures persistent violations of instance consistency in generated trajectories. Extensive experiments on DreamGenBench, EWMBench, and PBench demonstrate the effectiveness of EVEWORLD, notably achieving an 87.5% reduction in MLR compared with GigaWorld-0. On the WorldArena 2.0 Track 1 leaderboard, our model ranks 6th in JEPA Similarity and 17th overall, which further validates the performance of our evolution supervision strategy. Source code is available at an anonymous link: `https://anonymous.4open.science/r/EVEWorld-release-C8B4`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

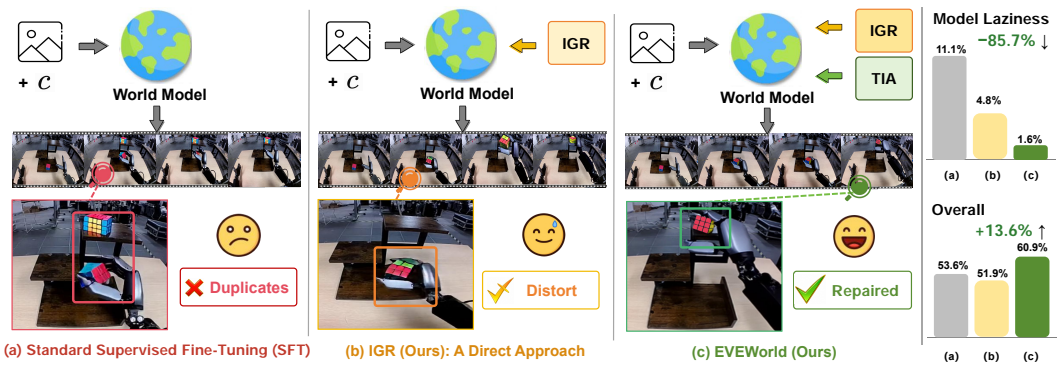


Figure 2: **Comparison of Standard SFT, IGR, and EVEWORLD. (a)** Existing video world models (GigaWorld Team et al., 2025; Yang et al., 2025; Team Wan et al., 2025; NVIDIA et al., 2025), including our Standard SFT baseline, may exhibit instance-consistency failures, such as target duplication at the goal. **(b)** Our IGR variant improves instance consistency, but cross-frame consistency failures such as target distortion may remain. **(c)** Our full EVEWORLD model combines IGR and TIA to maintain physically consistent target evolution throughout the interaction. Relative to Standard SFT, EVEWORLD reduces MLR by 85.7% and improves overall Gemini-IF by 13.6%. Additionally, the inputs consist of an initial image and an instruction c.

1 INTRODUCTION

Embodied world models and related policy-learning methods are emerging as scalable predictive engines for embodied intelligence, modeling diverse interaction dynamics to support robot learning, planning, and simulation (Yang et al., 2024a; Bruce et al., 2024; Zhou et al., 2024; Jang et al., 2025; Li et al., 2025b; Du et al., 2023; Schmidt & Jiang, 2024; Ye et al., 2025; Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2025; Physical Intelligence et al., 2025). Recent work (Huang et al., 2025; 2026; Guo et al., 2026; GigaWorld Team et al., 2026; Xiao et al., 2025; Pang et al., 2025; Dou et al., 2026; Duan et al., 2026; Zheng et al., 2026; Bardhan et al., 2026; Li et al., 2025c; 2026; Quevedo et al., 2026) has further advanced embodied world modeling and its evaluation through high-fidelity generation, multiview and long-context modeling, action conditioning, persistent memory, and task-oriented evaluation.

Recent embodied world models further strengthen structural faithfulness through explicit 4D RGB–depth–normal prediction, geometry-aware action conditioning, and disentangled dynamics and visual synthesis, with related physical-evaluation benchmarks (Zhen et al., 2025; DreamX Team et al., 2026; Shan et al., 2026; Liu et al., 2026; Li et al., 2025a; Bansal et al., 2026; Meng et al., 2025; Qin et al., 2025; Shang et al., 2026a; Motamed et al., 2026; Garrido et al., 2025). Nevertheless, stronger visual and structural modeling does not necessarily ensure physically consistent target evolution: a rollout may exhibit target-instance consistency failures, such as duplication or disappearance, and cross-frame consistency failures, such as deformation and identity drift (Yuan et al., 2026; Foo et al., 2026). Figure 1 shows a representative instance consistency failure after standard supervised fine-tuning (SFT). We refer to such instance consistency failures as **Model Laziness**: frame-level reconstruction can admit a shortcut that satisfies appearance or endpoint cues without maintaining instance consistency throughout the interaction.

In this work, we propose EVEWORLD, an evolution-supervision framework that promotes physically consistent target evolution and mitigates Model Laziness in embodied world models, as shown in Figure 2. EVEWORLD comprises two complementary components: **Instance-Guided Restoration (IGR)**, which promotes instance consistency by restoring count-perturbed videos, and **Temporal Instance Alignment (TIA)**, which promotes cross-frame consistency by aligning target instances across adjacent frames. To quantify Model Laziness, we introduce the **Model Laziness Rate (MLR)**, a metric that measures persistent violations of target-instance consistency over a rollout. Extensive evaluations on DreamGenBench, WorldArena 2.0, EWMBench, and PBench demonstrate the effectiveness of EVEWORLD in promoting physically consistent target evolution. Notably, on the WorldArena 2.0 Track 1 leaderboard, EVEWORLD achieves an EWMScore-P of 70.17, ranking 17th overall and 6th in JEPA Similarity. These results further support the effectiveness of evolution supervision under large-scale world-model evaluation.

2 Related Work

Video Generation Models for Robotics. Large-scale video models provide strong priors for embodied prediction and scalable robot-data generation (Blattmann et al., 2023; Brooks et al., 2024; Team Wan et al., 2025; Kong et al., 2024), enabling world models for robot learning, planning, and simulation (Yang et al., 2024a; Zhou et al., 2024; Jang et al., 2025; Lu et al., 2025; Gao et al., 2025; GigaWorld Team et al., 2025). Recent work (Zhen et al., 2025; DreamX Team et al., 2026; Shan et al., 2026; Fu et al., 2026) further strengthens structural modeling through richer representations and conditioning: TesserAct jointly predicts RGB, depth, and normals, DreamX-Phi incorporates geometric action conditioning and depth supervision, and DVG-WM disentangles dynamics from visual synthesis, and RoboMaster improves multi-object interaction modeling. In contrast, our EVEWorld retains the host model's native representation and introduces evolution supervision during post-training to promote physically consistent target evolution.

Physical Reasoning in Embodied World Models. Prior works (Yang et al., 2024b; Traub et al., 2024; Jeong et al., 2025; Chen et al., 2025; Wu et al., 2026; Xiao et al., 2025; Dou et al., 2026) improve object-centric and temporally consistent modeling through correspondence, tracking, object permanence, and persistent memory, while other studies (Deng et al., 2026; Sun et al., 2025; Khose et al., 2026) examine object dynamics, quantity behavior, and broader aspects of physical reasoning from different perspectives. However, despite these advances and evaluations (Yang et al., 2024b; Traub et al., 2024; Jeong et al., 2025; Xiao et al., 2025; Dou et al., 2026; Deng et al., 2026; Sun et al., 2025; Khose et al., 2026), physically consistent target evolution remains challenging, with manipulated objects exhibiting failures in both instance consistency and cross-frame consistency. In contrast, our EVEWorld introduces evolution supervision to explicitly promote physically consistent target evolution.

3 Why Embodied World Models Become Lazy

3.1 Problem Statement

Given an initial observation I_0 and an instruction c, an embodied video world model predicts future frames $x_{1:T}$. We focus on manipulation scenarios that require instance consistency, where the target-instance count is expected to remain unchanged throughout the interaction.

Let $N_t(I_0, x_t, c)$ denote the number of instances of the target specified by c, with $N_0(I_0, c)$ denoting its initial count. We define Model Laziness as

$$\mathrm{ML}(I_0, x_{1:T}, c) = \mathbb{I}[\exists t \in \{1, \dots, T\} : N_t(I_0, x_t, c) \neq N_0(I_0, c)]\,. \tag{1}$$

This captures instance consistency failures caused by unsupported target duplication or disappearance. Interactions that intentionally change the target-instance count, such as cutting or assembly, are outside the scope of the current formulation and MLR protocol. Robot-induced occlusion is handled separately through an occlusion-aware check during MLR evaluation.

3.2 Causes of Model Laziness

Standard training emphasizes visual reconstruction while providing limited process-level supervision for physically consistent target evolution (Denton & Fergus, 2018; Bengio et al., 2015; Jeong et al., 2025). This leaves two key gaps: instance consistency and cross-frame consistency.

Missing Instance-Consistency Supervision. Experimental results show that Standard SFT retains 92%–99% of the injected target area across different perturbation settings (Appendix Figure 6, Appendix C.2). Because the perturbed region is small relative to the full frame, its reconstruction error can be diluted by the surrounding scene, providing insufficient supervision for recovering the correct target-instance count. IGR (Section 4.1) addresses this gap by providing a direct restoration target and assigning greater spatial weight to the perturbed region.

Missing Cross-frame Consistency Supervision. IGR improves instance consistency, but a correct target-instance count alone does not ensure cross-frame consistency; the target may still become distorted, fade, or shift abruptly across adjacent frames, as shown in Figure 2(b). This stems from

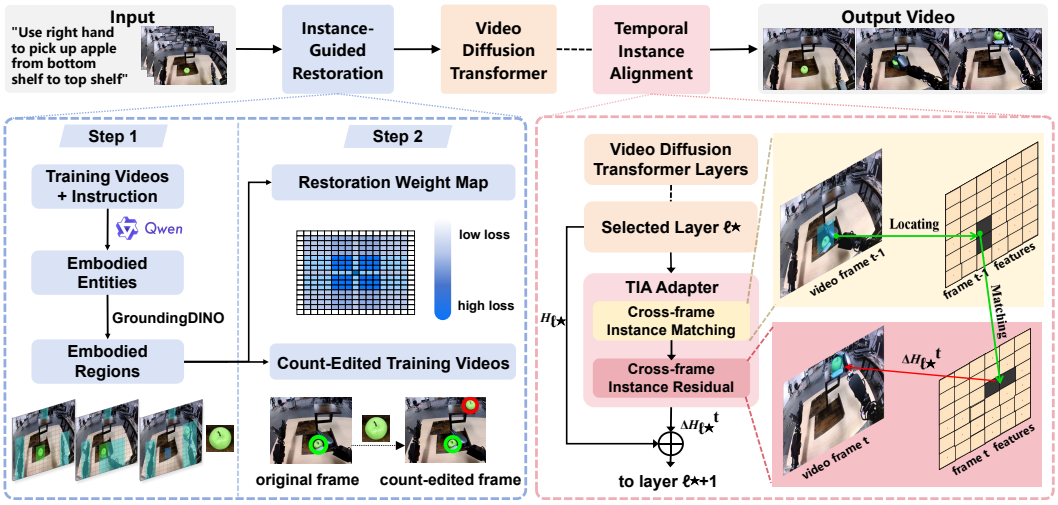


Figure 3: **Overall architecture of EVEWORLD.** Given a clean demonstration x and instruction c during training, IGR constructs a count-edited video $\tilde{x}$, which is processed by the video diffusion transformer to restore x. Within the transformer, TIA operates at the selected layer $\ell^\star$ to perform cross-frame instance matching and residual refinement, after which the refined features pass through the remaining layers to produce the final output.

the lack of explicit supervision linking the same target across frames, allowing locally plausible predictions to evolve inconsistently over time. TIA (Section 4.2) addresses this gap by aligning target instances across adjacent frames.

4 METHOD

Framework Overview. Given a clean demonstration video x and its instruction c, EVEWORLD uses IGR (Section 4.1) to construct a count-edited video $\tilde{x}$ and trains the model to restore x, assigning higher weights to restoration-related regions to promote instance consistency. TIA (Section 4.2) operates within the video diffusion transformer to align target features across adjacent frames and promote cross-frame consistency. Together, IGR and TIA enable physically consistent target evolution in the final video output. Notably, the demonstration video is used for training supervision; at inference, EVEWORLD takes only the initial image and instruction as input.

4.1 INSTANCE-GUIDED RESTORATION

To address failures in physically consistent target evolution, we introduce Instance-Guided Restoration (IGR) to promote instance consistency. IGR inserts an extra target instance into clean demonstrations to form count-edited samples, then trains the model to restore the corresponding clean latent representation encoded by a pretrained VAE. A language model and GroundingDINO (Liu et al., 2024a) localize the target trajectory to sample a target patch p and valid paste region Ω:

$$\tilde{x} = \mathrm{Paste}(x, p, \Omega), \qquad z = \mathrm{VAE}(x), \qquad \tilde{z} = \mathrm{VAE}(\tilde{x}). \tag{2}$$

Training the model to map the count-edited latent $\tilde{z}$ back to the clean latent z encourages recovery of the correct target-instance count. To prevent this local instance-consistency signal from being diluted by global reconstruction, we assign restoration-related regions three times the reconstruction weight of ordinary regions and normalize the resulting spatial weight map to unit mean. We denote the normalized map by $\hat{W}$; its exact construction is given in Appendix B.1. Using the EDM-style noise weighting $\lambda(\sigma)$ (Karras et al., 2022), the restoration objective is

$$\mathcal{L}_{\mathrm{IGR}} = \mathbb{E}\left[\lambda(\sigma)\frac{1}{CTHW}\left|\hat{W}^{1/2} \odot (D_\theta(\tilde{z} + \sigma\epsilon, c, \sigma) - z)\right|_2^2\right]. \tag{3}$$

Minimizing $\mathcal{L}_{\mathrm{IGR}}$ provides explicit instance-consistency supervision by recovering the correct target-instance state from count-perturbed inputs; details are provided in Appendix B.1.

4.2 Temporal Instance Alignment

IGR improves instance consistency, but a correct target-instance count alone does not guarantee cross-frame consistency. Failures such as identity drift, distortion, fading, or abrupt motion may still occur across adjacent frames. We therefore introduce Temporal Instance Alignment (TIA) to explicitly promote cross-frame consistency by aligning target instances across adjacent frames.

We select the attachment layer using an offline layer-wise matching probe. For each candidate layer ℓ, the frame-$(t-1)$ target feature retrieves its best match over all token cells of frame t, and we select the layer with the lowest endpoint error (EPE):

$$\hat{u}_t^\ell = \arg\max_v f_{t-1}^\ell(u_{t-1})^\top f_t^\ell(v), \qquad \ell^\star = \arg\min_{\ell\in\mathcal{B}} \frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}} \left\|\hat{u}_t^\ell - u_t\right\|_2 . \tag{4}$$

where $\mathcal{V}$ contains adjacent-frame pairs with the target visible in both frames and detector confidence above a threshold. At layer $\ell^\star$, TIA has two stages (Figure 3, right). *Local instance locating* matches target features within a local neighborhood, while *cross-frame instance transport* converts pre-transport cosine similarities into softmax transport weights, splats projected previous-frame features into frame t, normalizes by transport mass, and writes them back through a zero-initialized bounded residual.

$$H_t' = H_t^{\ell^\star} + \gamma\, \tanh\Big(W_{\mathrm{out}}\big(\bar{H}_t - W_{\mathrm{in}} H_t^{\ell^\star}\big)\Big). \tag{5}$$

Here, $\bar{H}_t$ denotes the transported feature field. GroundingDINO and target locations u_t provide offline alignment supervision during training only; at inference, dense matching and transport operate directly on the model's hidden features and require neither target localization nor an external tracker. The transported features are further supervised to favor the true next-frame target location over nearby candidates:

$$\mathcal{L}_{\mathrm{TIA}} = -\frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}} \log \frac{\exp(s_{t,u_t}/\tau)}{\sum_{v\in\mathcal{C}_t}\exp(s_{t,v}/\tau)}, \qquad s_{t,v} = \bar{f}_{t-1}^{\ell^\star}(u_{t-1})^\top \bar{f}_t^{\ell^\star}(v), \tag{6}$$

where $\bar{f}^{\ell^\star}$ is the normalized feature after transport. The locating, transport, and alignment components jointly constrain target identity across frames. Further details are provided in Appendix B.2.

5 Experiments

5.1 Experimental Setup and Metrics

Benchmarks and Data Splits. We use the official DreamGen GR1 fine-tuning split, which contains 92 robot-manipulation videos paired with instructions, and primarily evaluate on DreamGenBench, comprising 126 manipulation tasks across unseen environments, objects, and behaviors (Jang et al., 2025). We further evaluate cross-domain generalization on WorldArena 1.0, cross-trial robustness on PBench, cross-distribution generalization on AgiBot–EWMBench, and cross-backbone generalization using FlowWAM on WorldArena 2.0 and held-out RoboTwin actions (Shang et al., 2026a;b; AgiBot-World-Contributors et al., 2025; Hu et al., 2025; Mu et al., 2025; NVIDIA, 2025b; Huang et al., 2024; Liu et al., 2024b; Unterthiner et al., 2018; Chen et al., 2026).

Evaluation Metrics. We evaluate instruction following with Qwen-IF and Gemini-IF (Bai et al., 2025; Gemini Team et al., 2023); WorldArena 1.0/2.0, EWMBench, PBench, and RoboTwin assess complementary aspects of perceptual quality, motion, semantics, consistency, and task fidelity.

Related evaluations (Deng et al., 2026; Khose et al., 2026; Sun et al., 2025), including RBench, WFM-Eval, and video numeracy benchmarks, provide indirect signals related to object quantity and persistence. In contrast, we propose the **Model Laziness Rate (MLR)**, a metric that explicitly measures target-instance consistency at the rollout level using the observed initial count N_0 as its reference, without requiring a prompt-specified count or reference future. We exclude under-counts attributable to robot-induced occlusion while retaining all over-counts, and denote the resulting

Table 1: **Overall comparison on DreamGenBench.** MLR denotes Model Laziness Rate, and Qwen-IF/Gemini-IF measure instruction-following accuracy. General video models are evaluated zero-shot, whereas Standard SFT and EVEWORLD are post-trained on DreamGen. MLR is computed on a shared set of 63 eligible items across all models.

Metric	CogVideoX	Wan-TI2V	Wan-A14B	Cosmos	GigaWorld-0	Standard SFT	EVEWorld
MLR (%) ↓	28.57	17.46	11.11	14.29	12.70	11.11	**1.59**
Qwen-IF (%) ↑	38.89	38.89	64.29	62.70	79.37	73.81	**80.16**
Gemini-IF (%) ↑	5.56	10.32	15.87	24.60	60.19	53.57	**60.85**

Table 2: **Fine-grained comparison on DreamGenBench.** GR1-Env, GR1-Object, and GR1-Behavior denote generalization to unseen environments, objects, and behaviors, respectively. Qwen-IF and Gemini-IF measure instruction-following accuracy. General video models are evaluated zero-shot, whereas Standard SFT and EVEWORLD share the same DreamGen post-training setting.

Method	Param.	GR1-Env IF (%)		GR1-Object IF (%)		GR1-Behavior IF (%)	
		Qwen ↑	Gemini ↑	Qwen ↑	Gemini ↑	Qwen ↑	Gemini ↑
General video models (zero-shot)							
CogVideoX1.5-5B-I2V	5B	48.28	15.52	24.00	1.00	48.94	4.26
Wan2.2-TI2V-5B	5B	51.72	17.24	24.00	5.00	46.81	11.70
Wan2.2-I2V-A14B	A14B	58.62	20.69	66.00	13.00	65.96	15.96
Cosmos-Predict2-2B	2B	68.97	31.03	60.00	28.00	61.70	17.02
GigaWorld-based models							
GigaWorld-0	2B	79.31	54.60	**82.00**	**54.33**	**76.60**	**69.86**
Standard SFT	2B	79.31	51.72	76.00	42.00	68.09	67.02
EVEWorld	2B	**82.76**	**72.41**	**82.00**	50.33	**76.60**	64.89

evaluation count by $\widetilde{N}_t(x)$. MLR implements Eq. 1 with a two-frame persistence criterion to suppress isolated detection errors. We define

$$\mathrm{MLR} = \frac{100}{|\mathcal{D}_+|} \sum_{x \in \mathcal{D}_+} \mathbb{I}\Big[\exists t \in \{1, \ldots, T-1\} : \widetilde{N}_t(x) \neq N_0(x) \wedge \widetilde{N}_{t+1}(x) \neq N_0(x)\Big], \tag{7}$$

where $\mathcal{D}_+ = \{x \in \mathcal{D} : N_0(x) > 0\}$. Further details are provided in Appendix A.

Model and Training Setup. For the main experiments, we initialize EVEWORLD from the 2B-parameter GigaWorld-0 Video-Pretrain-2B (GigaWorld Team et al., 2025) and compare against CogVideoX1.5, Wan2.2, and Cosmos-Predict2 as zero-shot baselines (Yang et al., 2025; Team Wan et al., 2025; NVIDIA, 2025a). Standard SFT, EVEWORLD, and all ablations share the same initialization, training data, and 250-step optimization budget. We fine-tune all transformer parameters on eight NVIDIA H20Z GPUs with an effective batch size of 64 and a learning rate of 4.32×10^{-5}. For cross-distribution and cross-backbone experiments, we apply the same evolution-supervision method to the corresponding data and backbone. Details are provided in the appendix.

5.2 MAIN RESULTS

Overall Results. As shown in Table 1, EVEWORLD substantially mitigates Model Laziness while improving instruction following. Compared with Standard SFT, EVEWORLD reduces MLR from 11.11% to 1.59% (85.7% relative reduction), while Qwen-IF increases from 73.81% to 80.16% and Gemini-IF from 53.57% to 60.85%. Appendix A (Appendix Figure 4) further shows that the remaining violations under EVEWORLD are transient, indicating that evolution supervision improves physical consistency without sacrificing instruction following.

Fine-grained Analysis. Table 2 reports instruction-following performance across the Environment, Object, and Behavior splits. Compared with Standard SFT, EVEWORLD improves Qwen-IF on all three splits, with the largest gain on Behavior (68.09% to 76.60%), while Gemini-IF improves notably on Environment (51.72% to 72.41%) and Object (42.00% to 50.33%). Overall, these results show that evolution supervision yields broad gains across different generalization dimensions.

Table 3: **Zero-shot comparison on WorldArena 1.0.** All metrics follow the released evaluation protocol. IQ, AQ, DD, Flow, MS, SC, BC, and PC denote Image Quality, Aesthetic Quality, Dynamic Degree, Flow Score, Motion Smoothness, Subject Consistency, Background Consistency, and Photometric Consistency; Overall is their equal-weight mean. MLR is computed on 157 shared eligible items, and bold marks the best GigaWorld-based result.

Method	Appearance		Motion			Consistency			Main Metrics	
	IQ ↑	AQ ↑	DD ↑	Flow ↑	MS ↑	SC ↑	BC ↑	PC ↑	Overall ↑	MLR (%) ↓
General I2V models (zero-shot)										
CogVideoX	24.51	34.80	20.71	7.24	50.78	54.96	66.35	10.20	33.69	2.55
Wan-TI2V	33.48	37.90	33.00	11.38	61.86	68.88	77.17	18.19	42.73	17.20
Wan-A14B	51.93	43.37	42.29	8.04	64.13	67.69	75.37	33.61	48.30	29.94
Cosmos	57.88	38.83	50.46	16.69	71.61	64.14	73.71	21.42	49.34	8.28
GigaWorld-based models										
GigaWorld-0	58.53	**38.79**	44.36	7.74	61.67	77.90	88.13	**48.32**	53.18	27.39
Standard SFT	**60.71**	38.36	48.49	7.82	64.04	**79.34**	88.67	44.17	53.95	27.39
EVEWorld	59.82	37.35	**62.90**	**10.98**	**67.82**	78.72	**88.91**	47.58	**56.76**	**13.38**

Table 4: **Comparison on EWMBench after AgiBot post-training.** Standard SFT and EVEWORLD use the same AgiBot data and optimization budget. Motion, Semantics, and Overall are the official EWMBench aggregate scores, and DYN, HSD, nDTW, and Scene are component metrics. Complete results are provided in Appendix E.1.

Method	Motion ↑	Semantics ↑	DYN ↑	HSD ↑	nDTW ↑	Scene ↑	Overall ↑
GigaWorld-0	50.30	2.2497	7.19	23.29	19.83	**89.58**	3.6486
Standard SFT	61.51	2.2477	14.88	**24.35**	**22.28**	84.38	3.7066
EVEWorld	**63.65**	**2.2524**	**17.55**	24.12	21.97	86.37	**3.7525**

Table 5: **Cross-backbone evaluation and ablation on RoboTwin.** PSNR, SSIM, and LPIPS measure frame-level visual fidelity against aligned RGB ground truth, and Flow-EPE denotes optical-flow endpoint error.

Variant	IGR	TIA	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	Flow-EPE ↓	MLR (%) ↓
FlowWAM			12.218	0.748	0.383	3.033	52.05
Standard SFT			11.687	0.735	0.414	2.828	47.89
+ IGR	✓		12.544	0.770	0.380	2.487	40.03
+ TIA		✓	**13.445**	**0.776**	**0.333**	**1.850**	**22.54**
EVEWorld	✓	✓	12.765	0.769	0.365	2.207	35.21

5.3 GENERALIZATION EVALUATION

Zero-Shot Domain Generalization. On WorldArena 1.0 (Shang et al., 2026a), which includes unseen scenes and robot embodiments, EVEWORLD improves EWMScore-local-8 from 53.95 to 56.76 (+5.21%) over Standard SFT while reducing MLR from 27.39% to 13.38% (−51.15%) (Table 3). These results demonstrate improved cross-domain generalization in both motion quality and target-instance consistency.

Generalization Across Training Distributions. On EWMBench (Hu et al., 2025), EVEWORLD trained on the same AgiBot data (AgiBot-World-Contributors et al., 2025) and optimization budget as Standard SFT improves Motion and Overall by 3.48% and 1.24%, respectively (Table 4). These results show that the gains of EVEWORLD extend beyond the DreamGen training distribution, particularly in motion modeling.

Cross-Backbone Generalization. We replace the GigaWorld image-text-conditioned IT2V backbone with FlowWAM's flow-conditioned WAM architecture (Table 5). On FlowWAM, IGR improves all reported metrics over Standard SFT on held-out RoboTwin actions, but TIA yields consistently larger gains. The full model continues to outperform Standard SFT, suggesting that the contributions of IGR and TIA are backbone-dependent. The resulting FlowWAM-based EVEWorld model, Super-

Table 6: **Component ablation on DreamGenBench.** All variants share the same initialization, training data, and optimization budget. Gemini-IF measures instruction-following accuracy.

Variant	Component		Gemini-IF (%)				MLR (%) ↓
	IGR	TIA	Env ↑	Object ↑	Behavior ↑	Overall ↑	
Standard SFT			51.72	42.00	67.02	53.57	11.11
IGR only	✓		49.43	36.67	**69.50**	51.85	4.76
TIA only		✓	55.17	42.67	68.79	55.29	7.94
EVEWorld	✓	✓	**72.41**	**50.33**	64.89	**60.85**	**1.59**

Table 7: **Exploration of Alternative Approaches.** PhysicsLatent uses weakly supervised process tokens; EAG uses energy-guided denoising; LAD-LoRA uses energy-regularized LoRA; Causal Frontier uses causal block-wise generation; ICH-D uses learned duplicate removal; and EVEWORLD combines instance restoration with temporal alignment. Each route is evaluated against its matched control; for EVEWORLD, the matched control is Standard SFT. Dyn., Train, Test, Causal, and Repair denote explicit dynamics/evolution modeling, training-time intervention, test-time intervention, causal generation, and learned repair, respectively. Δ denotes relative change from the matched control. Further details are provided in Appendix D.

Route	Design elements					Relative change	
	Dyn.	Train	Test	Causal	Repair	Δ MLR (%) ↓	Δ Gemini-IF (%) ↑
PhysicsLatent	✓	✓				↓ 7.0	↓ 11.9
EAG sampling guidance	✓		✓			↑ 150.0	↑ 10.0
LAD-LoRA	✓	✓				0.0	↓ 6.1
Causal Frontier		✓		✓		↑ 35.4	↓ 53.8
ICH-D	✓	✓	✓		✓	↑ 16.6	↓ 30.6
EVEWORLD (ours)	✓	✓			✓	↓ **85.7**	↑ **13.6**

Table 8: **Cross-trial robustness on PBench.** Each SFT_{250}–EVEWORLD pair uses the same inference seed, and the three trials use different seeds under otherwise identical settings. Scores denote the overall PBench Robot score, judged by Gemini-3.8-Flash. Δ denotes the gain over SFT_{250}, and Mean Δ averages the three paired gains.

Horizon	Trial 1			Trial 2			Trial 3			Mean Δ
	SFT_{250}	EVEWORLD	Δ	SFT_{250}	EVEWORLD	Δ	SFT_{250}	EVEWORLD	Δ	
3.8 s	57.37	**59.27**	+1.90	53.40	**55.07**	+1.67	59.57	**60.61**	+1.04	+1.54
5.8 s	59.42	**60.02**	+0.60	60.36	**63.02**	+2.66	60.79	**62.21**	+1.42	+1.56

visionWM, achieves an EWMScore-P of 70.17 on the official WorldArena 2.0 Track 1 leaderboard, ranking 17th overall and 6th in JEPA Similarity. These results support the cross-backbone generalization of our evolution-supervision framework, while the optimal component configuration may depend on the model architecture.

5.4 Ablation Studies and Analysis

Component Contributions of IGR and TIA. Table 6 shows that IGR primarily improves instance consistency, reducing MLR from 11.11% to 4.76%; TIA improves cross-frame consistency and Overall Gemini-IF. Combining both yields 1.59% MLR and a Gemini-IF score of 60.85, demonstrating complementary effects under the GigaWorld-0 image-text-conditioned IT2V architecture.

Cross-Trial Robustness. We further evaluate EVEWORLD on PBench Robot across multiple stochastic inference trials under a matched-budget setting, with evolution supervision as the only methodological difference from SFT_{250}. As shown in Table 8, EVEWORLD improves all six paired trials across 3.8 s and 5.8 s rollouts, with mean gains of $+1.54$ and $+1.56$, respectively, demonstrating consistent robustness across inference seeds and rollout horizons.

Exploration of Alternative Approaches. We explored five alternative approaches to mitigate Model Laziness while improving instruction fidelity: PhysicsLatent, EAG, LAD-LoRA, Causal

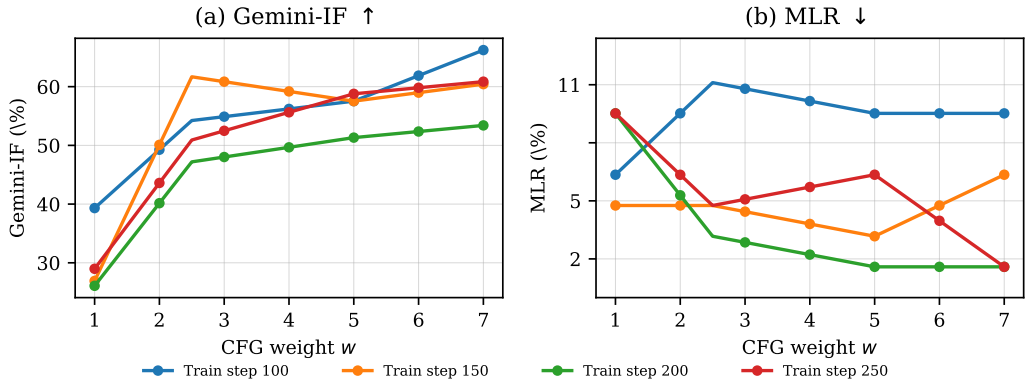


Figure 4: **CFG sensitivity across training stages on DreamGenBench.** Gemini-IF and MLR are evaluated under different classifier-free guidance weights w, with each curve denoting a training checkpoint.

Table 9: **Sequence-length scaling and inference cost.** We report latency (p50/p95: median/95th percentile), throughput, memory, and GPU cost across output durations.

Output		Latency (s) ↓		Throughput	Memory	Cost
Dur. (s)	Frames	p50	p95	Frames/min ↑	Peak/GPU (GiB) ↓	GPU-h/1k ↓
3.8	61	8.166	8.403	442.1	38.19	18.40
5.8	93	10.932	11.352	502.2	38.33	24.69
9.8	157	18.323	19.249	504.8	38.99	41.46
15.8	253	31.503	33.523	469.5	41.47	71.85
19.8	317	41.737	45.007	443.3	40.70	95.34

Frontier, and ICH-D (Table 7). However, none achieved both objectives simultaneously. For example, PhysicsLatent reduces Model Laziness at an 11.9% cost in instruction fidelity, whereas EAG gains 10.0% in instruction fidelity at a 150.0% increase in Model Laziness. We share these trade-offs to provide practical insights for future research, with further details in Appendix D.

CFG Sensitivity Analysis. Figure 4 shows that Gemini-IF generally increases with higher CFG weights, but MLR varies non-monotonically across guidance weights and training stages. Thus, CFG does not reliably improve instruction fidelity and reduce Model Laziness simultaneously, indicating the importance of explicit supervision for target evolution.

Sequence-Length Scaling and Inference Cost. Table 9 shows that increasing output duration from 3.8 to 19.8 s raises p50/p95 latency from 8.17/8.40 to 41.74/45.01 s and GPU-hour cost per 1,000 videos from 18.40 to 95.34, with relatively stable throughput and peak memory. Thus, longer rollouts mainly increase computation with limited additional memory overhead.

6 Conclusion

We presented EVEWorld, a physical evolution supervision framework for embodied world models, and proposed the Model Laziness Rate (MLR), a metric to quantify target-instance consistency failures, which we term Model Laziness. EVEWorld combines IGR for instance consistency and TIA for cross-frame consistency to promote physically consistent target evolution. Experiments show an 85.7% MLR reduction on DreamGenBench with improved instruction following, 17th overall and 6th in JEPA Similarity on the WorldArena 2.0 Track 1 leaderboard, and generalization gains across training distributions and model backbones.

Limitations and Future Works. The limitations of our current formulation mainly arise from its local, pairwise supervision, which does not explicitly capture long-horizon state dependencies or complex multi-object interactions. Future work will further expand the EVEWorld framework with broader forms of evolution supervision for longer-horizon and relational dynamics.

Reproducibility Statement

The appendices document the training and evaluation settings. The anonymized supplement provides code, data indices, configurations, evaluation prompts, and item-level results. The anonymized repository is available at `https://anonymous.4open.science/r/EVEWorld-release-C8B4`.

AI use statement

During the preparation of this manuscript, the authors used generative AI tools solely for writing assistance, including language polishing, grammar checking, and improving clarity of presentation. The AI tools were not used for developing the method, conducting experiments, generating results, or making scientific decisions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this manuscript.

References

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, et al. AgiBot World colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. doi: 10.48550/arXiv.2503.06669. URL https://arxiv.org/abs/2503.06669.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026.

Jai Bardhan, Patrik Drozdik, Josef Sivic, and Vladimir Petrik. Persistent robot world models: Stabilizing multi-step rollouts via reinforcement learning. *arXiv preprint arXiv:2603.25685*, 2026. doi: 10.48550/arXiv.2603.25685. URL https://arxiv.org/abs/2603.25685.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, volume 235, pp. 4603–4623, 2024.

Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using diffusion priors for video amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, pp. 22890–22900, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Using_Diffusion_Priors_for_Video_Amodal_Segmentation_CVPR_2025_paper.html.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, Jianlou Si, Jun Huang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. Flowwam: Optical flow as a unified action representation for world action models, 2026. URL https://arxiv.org/abs/2607.13017.

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.15282.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, 2018.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. SlotMemory: Object-centric KV memory for streaming long-video generation. *arXiv preprint arXiv:2605.31033*, 2026. doi: 10.48550/arXiv.2605.31033. URL https://arxiv.org/abs/2605.31033.

DreamX Team, Rui Chen, Xiangxiang Chu, Geng Li, Jifan Li, Qingfeng Shi, Datao Tang, Jing Tang, Jun Wang, and Pengfei Zhang. DreamX-Phi 1.0: Action-conditioned video world model for robotic manipulation. *arXiv preprint arXiv:2608.13489*, 2026. doi: 10.48550/arXiv.2608.13489. URL https://arxiv.org/abs/2608.13489.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. LiveWorld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026. doi: 10.48550/arXiv.2603.07145. URL https://arxiv.org/abs/2603.07145.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4301–4311, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Foo_Physical_Simulator_In-the-Loop_Video_Generation_CVPR_2026_paper.html.

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. URL https://openreview.net/forum?id=OeDwYtp8n1.

Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Zhehao Cai, and Lin Shao. FLIP: Flow-centric generative planning as general-purpose manipulation world model. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/37bf751554eb05465ed5dc60dced3f38-Abstract-Conference.html.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, et al. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805. URL https://arxiv.org/abs/2312.11805.

GigaWorld Team, Angen Ye, Boyuan Wang, et al. GigaWorld-0: World models as data engine to empower embodied AI. *arXiv preprint arXiv:2511.19861*, 2025.

GigaWorld Team, Angyuan Ma, Boyuan Wang, Bohan Li, et al. GigaWorld-1: A roadmap to build world models for robot policy evaluation. *arXiv preprint arXiv:2607.02642*, 2026.

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*, 2026.

Yue Hu, 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. doi: 10.48550/arXiv.2505.09694. URL https://arxiv.org/abs/2505.09694.

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.

Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Yue Liao, Zhengkai Jiang, Yue Hu, Peng Gao, Hongsheng Li, Maoqing Yao, and Guanghui Ren. EnerVerse: Envisioning embodied future space for robotics manipulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1263.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, et al. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Joel Jang, Seonghyeon Ye, Zongyu Lin, et al. DreamGen: Unlocking generalization in robot learning through video world models. In *Conference on Robot Learning*, 2025.

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, and Duygu Ceylan. Track4Gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7276–7287, 2025. doi: 10.1109/CVPR52734.2025.00682.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Sahil Khose, Mengqi Zhang, Prithvijit Chattopadhyay, and Judy Hoffman. Wfm-eval: Interpretable error diagnostics for video world models in robotics. In *World Models and FMEA Workshops at CVPR*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. 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*, volume 270, pp. 2679–2713, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 61064–61088, 2025a. doi: 10.52202/085713-1834.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied AI. *arXiv preprint arXiv:2510.16732*, 2025b. doi: 10.48550/arXiv.2510.16732. URL https://arxiv.org/abs/2510.16732.

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*, 2025c.

Yaxuan Li, Zhongyi Zhou, Yefei Chen, Yaokai Xue, and Yichen Zhu. dWorldEval: Scalable robotic policy evaluation via discrete diffusion world model. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.22152.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024a. doi: 10.1007/978-3-031-72970-6_3.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, et al. EvalCrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4d video generation for robot manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KoXY5e7bVW.

Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. GWM: Towards scalable gaussian world models for robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9263–9274, 2025. doi: 10.1109/ICCV51701.2025.00865. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lu_GWM_Towards_Scalable_Gaussian_World_Models_for_Robotic_Manipulation_ICCV_2025_paper.html.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, et al. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. doi: 10.1109/WACV61042.2026.00099.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27649–27660, 2025.

NVIDIA. Cosmos-Predict2: World simulation model for physical AI, 2025a. URL https://github.com/nvidia-cosmos/cosmos-predict2.

NVIDIA. PBench: A physical AI benchmark for world models, 2025b. URL https://huggingface.co/datasets/nvidia/PBench.

NVIDIA, Niket Agarwal, Arslan Ali, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Jing-Cheng Pang, Nan Tang, Kaiyuan Li, Yuting Tang, Xin-Qiang Cai, Zhen-Yu Zhang, Gang Niu, Masashi Sugiyama, and Yang Yu. Learning view-invariant world models for visual robotic manipulation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8882d370cdafec9885b918a8cfac642e-Abstract-Conference.html.

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*, 2025. doi: 10.48550/arXiv.2504.16054. URL https://arxiv.org/abs/2504.16054.

Yiran Qin, Zhelun Shi, Jiwen Yu, et al. WorldSimBench: Towards video generation models as world simulators. In *International Conference on Machine Learning*, 2025.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024.

Ziyu Shan, Zhenyu Wu, Xiaofeng Wang, Zheng Zhu, and Ziwei Wang. DVG-WM: Disentangled video generation enables efficient embodied world model for robotic manipulation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.32028.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7: 208–226, 2026a. doi: 10.1016/j.aiopen.2026.07.002.

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, Haisheng Su, Weizhen He, Wei Wu, Haoyi Duan, Gordon Wetzstein, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, Chen Gao, and Yong Li. WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform, 2026b. URL https://arxiv.org/abs/2605.17912.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. doi: 10.1109/CVPR52734.2025.00787.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. doi: 10.48550/arXiv.2503.20314. URL https://arxiv.org/abs/2503.20314.

Manuel Traub, Frederic Becker, Sebastian Otte, and Martin V. Butz. Learning object permanence from videos via latent imaginations. In *International Conference on Artificial Neural Networks*, pp. 223–240, 2024. doi: 10.1007/978-3-031-72338-4_16.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, et al. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Junyi Wu, Van Nguyen Nguyen, Benjamin Planche, Jiachen Tao, Changchang Sun, Zhongpai Gao, Zhenghao Zhao, Anwesa Choudhuri, Gengyu Zhang, Meng Zheng, Feiran Wang, Terrence Chen, Yan Yan, and Ziyan Wu. Consistent instance field for dynamic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3793, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Consistent_Instance_Field_for_Dynamic_Scene_Understanding_CVPR_2026_paper.html.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1659.

Sherry Yang, Yilun Du, Seyed Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024a.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. FRESCO: Spatial-temporal correspondence for zero-shot video translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. NewtonGen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/15a321fbfc19fce9b325ec6e77dfb597-Abstract-Conference.html.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Learning 4D embodied world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5337–5347, 2025.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Jun Shi, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, and Xu Jia. Mem-World: Memory-augmented action-conditioned world models for persistent robot manipulation. In *Conference on Robot Learning*, 2026. URL https://arxiv.org/abs/2606.18960.

Siyuan Zhou, Yilun Du, Jiaben Chen, et al. RoboDreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, 2024.

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*, volume 229, pp. 2165–2183, 2023.