Reconstruction or Semantics? What Makes a Latent Space Useful for Robotic World Models

Nilaksh[*1,2,3] Saurav Jha[*1,2,3] Artem Zholus[*1,2,3] Sarath Chandar[1,2,3,4]
[1]Chandar Research Lab [2]Mila – Quebec AI Institute
[3]Polytechnique Montréal [4]Canada CIFAR AI Chair
{nilaksh.nilaksh, saurav.jha}@mila.quebec
hskalin.github.io/semantic-wm

Abstract

World model-based policy evaluation is a practical proxy for testing real-world robot control by rolling out candidate actions in action-conditioned video diffusion models. As these models increasingly adopt latent diffusion modeling (LDM), choosing the right latent space becomes critical. While the status quo uses autoencoding latent spaces like VAEs that are primarily trained for pixel reconstruction, recent work suggests benefits from pretrained encoders with representation-aligned semantic latent spaces. We systematically evaluate these latent spaces for action-conditioned LDM by comparing six reconstruction and semantic encoders to train world model variants under a fixed protocol on BridgeV2 dataset, and show effective world model training in high-dimensional representation spaces with and without dimension compression. We then propose three axes to assess robotic world model performance: visual fidelity, planning and downstream policy performance, and latent representation quality. Our results show visual fidelity alone is insufficient for world model selection. While reconstruction encoders like VAE and Cosmos achieve strong pixel-level scores, semantic encoders such as V-JEPA 2.1 (strongest overall on policy), Web-DINO, and SigLIP 2 generally excel across the other two axes at all model scales. Our study advocates semantic latent space as stronger foundation for policy-relevant robotics diffusion world models.

1. Introduction

Action-conditioned video world models are emerging as a practical interface between generative modeling and robotics [6, 15, 45]. Given observation and action histories, they predict future observations and serve as learned proxies for robot-environment interaction when handcrafted simulators are difficult to build [10, 38]. Recent works show that such models can support policy evaluation with good correlation to real-world outcomes [40], and policy improvement [35, 49, 55]. Yet current evaluations say little about which representation makes a world model faithful to robotic dynamics.

This question is increasingly important because many video world models are latent diffusion models (LDMs) [32, 41] that learn dynamics in an encoder-defined latent space. The standard choice is a reconstruction-aligned autoencoder, such as a VAE [20] or recent variants [1, 11, 46], whose latents are optimized for pixel fidelity and stable decoding. But robotic world models are more than video generators, where planning and evaluations require predictions that preserve physical, spatial, and task dynamics. This motivates using the semantic spaces of self-supervised and vision-language encoders as latents for robot world modeling [3, 7, 16, 17, 27, 31, 39]. These spaces expose object layout and task structure more directly than pixel-trained autoencoders [36]. However, they are hard to use for diffusion due to their higher dimensionality yielding off-manifold latent generation with poor object structures [51]. RAE [52] makes them more tractable with a dimension-dependent noise-schedule shift and a wide DDT head [43], while S-VAE [51] learns a compact, KL-regularized latent space using a semantic autoencoder as an adapter over the frozen semantic features.

Still, the effect of semantic latents on action-conditioned LDM for robotics remains open. DINO-WM [53] and V-JEPA 2-AC [3] show that pretrained feature spaces support planning, but they are not diffusion models: DINO-WM is an autoregressive feature-prediction world model, while V-JEPA 2-AC is a JEPA predictor [2]. RAE-NWM [50] shows that DINOv2 [27] spaces support diffusion-based navigation world modeling. Yet navigation differs from contact-rich manipulation, where gripper motion, object state, geometry, and policy rollouts all matter. This leads to our question: **what effects does latent space choice have for LDM-based robotic world modeling?**

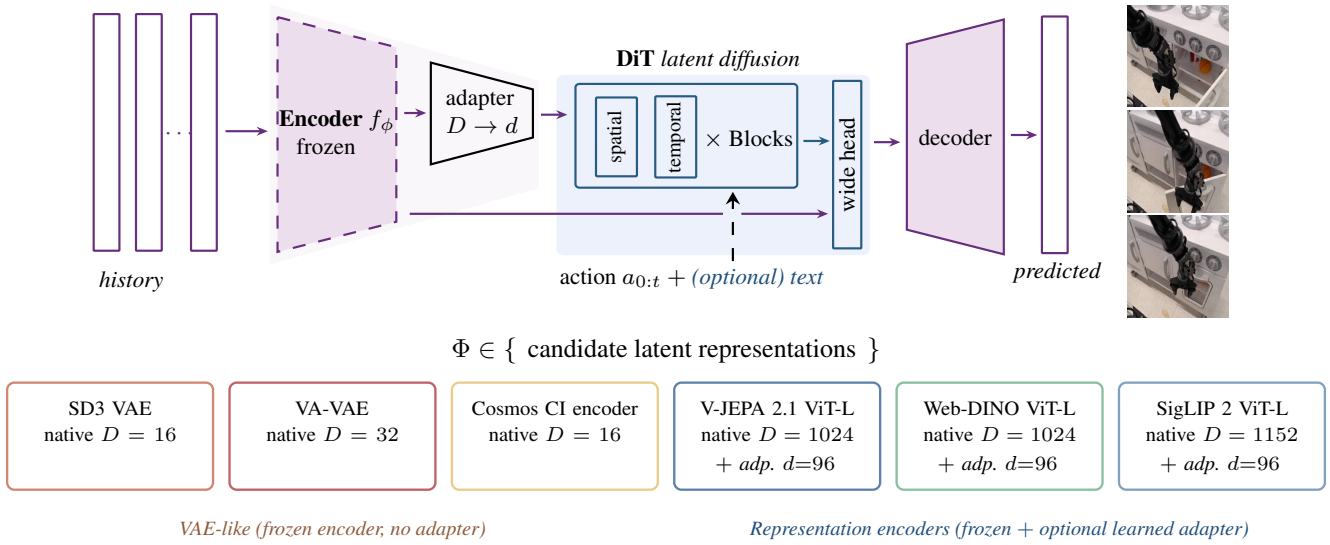


Figure 1. **Which latent space makes a better robotic world model?** For a latent diffusion model, we fix the Diffusion Transformer (DiT) transition model, action conditioning, and training data. We vary only the encoder f_ϕ defined latent interface: encoder, optional compression adapter, and the associated decoder path. This isolates how reconstruction-aligned and semantic representations affect action-faithful dynamics, generated rollouts, and downstream policy performance for robot control. We show the encoder families compared in the bottom panels.

We answer this with a controlled evaluation study that varies only the representation space in which the transition model operates (see Fig. 1). For effective semantic space LDM training, we adapt RAE's wide-head and schedule-shift recipe [52] alongside the compact S-VAE adapter [51], and train on the Bridge V2 dataset [42] with the same DiT transition model [28] and action-conditioning scheme. We then propose an evaluation suite spanning three axes: visual fidelity, planning and downstream policy performance, and latent quality. Our findings show that semantic latents improve action recoverability, task-success classification, CEM planning, and policy-in-the-loop success, while reconstruction latents mainly retain photometric advantages. Our key contributions are three-fold:

1. Our primary contribution is the *evaluation* of representation spaces for latent diffusion world modeling. We do controlled analyses of how latent space choice affects not only visual generation, but also robotics tasks and robustness through our proposed three evaluation axes.
2. We propose an effective recipe for *training diffusion world models in high dimensional semantic spaces*, by leveraging the recent advances in semantic space diffusion and extending them to action-conditioned world modeling. We also study the effects of different design choices.
3. We show that semantic latent spaces are consistently more useful for policy evaluation and planning, even when reconstruction latents match or exceed them on low-level pixel fidelity, establishing that the best robotic world model latent space is the one that preserves action-relevant structure, not merely the one that reconstructs images the best.

2. Problem Formulation

We consider multi-task robot manipulation from partial observations. The offline dataset is $\mathcal{D} = \{(o_{0:T}, a_{0:T-1}, \ell, y)\}$, where $o_t \in \mathcal{O}$ is an RGB observation, $a_t \in \mathbb{R}^{d_a}$ is a continuous robot action, ℓ is an optional language instruction, and $y \in \{0, 1\}$ denotes episode success. Tasks vary in object configurations and instructions, but share a robot embodiment; we therefore view the data as samples from related partially observed Markov Decision Processes with shared dynamics and task-dependent goals. Because a single observation does not generally determine the next observation under an action, we condition on a finite visual-action history of length H and model the action-conditioned predictive distribution over a rollout horizon K: $p(o_{t+1:t+K} \mid o_{t-H:t}, a_{t-H:t+K-1})$.

2.1. Latent Space World Models

Rather than predicting future frames directly in pixel space, latent world models learn predictive dynamics in a representation space. Each model consists of a frozen encoder, an optional frozen adapter, an action-conditioned transition model, and a decoder.

Encoder and adapter. A pretrained image encoder maps each observation to a spatial latent $z_t = f_\phi(o_t) \in \mathbb{R}^{N \times D}$, where $N = h \times w$ is the number of patches and D is the encoder's native channel dimension. The encoder is frozen,

so f_ϕ fixes the representation space in which dynamics are learned. For high-dimensional semantic representation encoders, we optionally use a frozen adapter α_ψ to obtain compact diffusion-friendly latents $\tilde{z}_t = \alpha_\psi(z_t) \in \mathbb{R}^{N\times d}$ [51]. For compressed reconstruction-aligned latent spaces, the adapter is simply the identity map.

Transition model. An action-conditioned DiT [28] predicts future latent trajectories: $\tilde{z}_{t+1:t+K} \sim p_\theta(\cdot \mid \tilde{z}_{t-H:t}, a_{t-H:t+K-1})$. Only the transition model is updated during world model training; the encoder, adapter, and decoder remain fixed. For semantic encoders without adapters, we add a lightweight wide DDT head [43], which adds few parameters but addresses the width bottleneck of DiT for high-dimensional latent spaces [52]. Otherwise, variants share the same transition backbone and differ only in representation and decoding path. Because all variants keep the same patch-token count, the DiT backbone with adapter does not incur an increase in transition-model parameter count or GFLOPs.

Decoder. Predicted latents are mapped back to pixels as $\hat{o}_{t+1:t+K} = \mathrm{Dec}(\tilde{z}_{t+1:t+K})$. The decoder is needed for visual rollouts and pixel-level evaluation, but decoded image quality alone does not determine world model quality: a model may render plausible frames while missing action-relevant dynamics, or preserve control-relevant structure despite minor photometric errors.

2.2. The Role of the Latent Space in Robotics

The encoder-defined latent space determines the state representation on which the transition model p_θ learns dynamics. In LDM, reconstruction-aligned latents $z_t^{\mathrm{pix}} = f_\phi^{\mathrm{PIX}}(o_t) \in \mathbb{R}^{N\times D_{\mathrm{pix}}}$ are commonly used because they preserve pixel-level information and provide reliable decoders [9]. For robotic world models, however, the relevant state is not only what an image looks like, but how it changes under actions and whether those changes preserve task progress, object state, contact, and geometry. This creates a multi-objective problem where useful latents should be action-controllable, task-informative, visually decodable, and useful for planning or policy evaluation.

As an initial diagnostic, we use an inverse dynamics model (IDM) to probe whether an encoder makes action-relevant change explicit in latent space. Different encoders induce markedly different action-aligned trajectory geometries, suggesting that encoder choice changes which aspects of robot dynamics are easy for a transition model to learn. This motivates us to treat the latent space f_ϕ as the experimental variable, and evaluate its effect beyond visual fidelity on axes spanning controllability, task semantics, and policy performance.

We thus compare reconstruction-aligned latents with semantic latents from pretrained vision foundation models [3, 27, 39], denoted as $z_t^{\mathrm{rep}} = f_\phi^{\mathrm{REP}}(o_t) \in \mathbb{R}^{N\times D_{\mathrm{rep}}}$. Since D_{rep} is typically large, we evaluate both native features and compact adapter latents $\tilde{z}_t = \alpha_\psi(z_t^{\mathrm{rep}})$. We train one world model per candidate in $\Phi = \{f_\phi^{(1)}, \ldots, f_\phi^{(m)}\}$ while fixing the data, history, action conditioning, optimizer, and transition backbone, so that each model learns a different latent transition $p_\theta^{(\phi)}(\tilde{z}_{t+1:t+K} \mid \tilde{z}_{t-H:t}, a_{t-H:t+K-1})$. The decoder differences are controlled through reconstruction gap metrics, latent-space metrics, and planning metrics.

3. Experiments

3.1. Dataset and Training

Benchmark protocol. We isolate the effect of the encoder-defined latent space by fixing the dataset, history length, action conditioning, transition architecture, optimizer, and training schedule, and varying only the encoder f_ϕ, optional adapter α_ψ, and decoder path. For each encoder–adapter pair, we train an LDM from scratch and evaluate the resulting world model for visual fidelity, representation quality, and downstream policy performance.

Dataset. We train and evaluate on Bridge V2 [42], a real-robot manipulation dataset with ≈60K WidowX 250 demonstrations across 13 task families. Each episode includes RGB observations, 7 Degrees of Freedom (DoF) end-effector actions covering position, rotation, and gripper state, and a language instruction. For trajectory success classification, we use SOAR [54] which contains roughly 30.5K success/failure class episodes for WidowX 250 with a 1:2 class split.

Encoder variants. We compare two encoder families. reconstruction-aligned encoders f_ϕ^{PIX} include: Stable Diffusion 3 (SD3) VAE [11] with D=16, VA-VAE [46] with D=32, and Cosmos [1] with D=16; for these, $\alpha_\psi \equiv \mathbb{I}$. Semantics-aligned encoders f_ϕ^{REP} include: V-JEPA 2.1 [26] with D=1024, Web-DINO [13], adapted from DINOv2 [27], with D=1024, and SigLIP 2 [39] with D=1152. For semantic encoders, we evaluate both native latents and compact latents from a pretrained semantic VAE adapter [51], which maps $D\to d$ with d=96.

Adapter, decoder, and transition model. The S-VAE adapter [51] is pretrained to reconstruct frozen encoder features with a KL-regularized loss, and is paired with a lightweight pixel decoder. All transition models are DiTs trained on Bridge V2 [42] with flow matching [25]. Each DiT layer factorizes attention into a spatial block within each frame and a causal temporal block across frames. We sample every second frame, condition on H=2 history frames,

and predict 8 future frames. We do not make use of language instruction conditioning while training the DiT. For all non-VAE encoders, we apply a dimension-dependent noise-schedule shift [11]. At inference, models roll out autoregressively one frame at a time using a 10-frame sliding context; VAE variants use their native pixel decoders, while semantic variants use the learned adapter decoder.

3.2. Evaluation Metrics

To study how the choice of latent representation propagates through to downstream tasks, we propose an evaluation suite that segregates this effect across three axes.

1. **Planning and downstream policy performance.** For robotics applications, a latent world model should enable planning, *i.e.*, searching for the optimal action sequence given a goal state [3, 53]. Evaluating planning helps separate the latent world modeling performance from the pixel decoder performance, which visual metrics conflate together. Given a real k-step transition, we use the cross-entropy method (CEM) [33] to recover the action sequence whose predicted latent best matches the target, and report CEM error at single-step $(k = 1)$ and multi-step $(k = 4)$ horizons.
   We also test whether the world model can serve as a policy-evaluation environment. We roll out OpenVLA-7B [19] inside each world model on 20 Bridge V2 test episodes with 8 trials per episode, and a subset of 10 of these were used for Out-Of-Distribution (OOD) evaluations. We use two Vision-Language Models (VLMs): InternVL 3.5 [44] and Qwen 3.6 [30], to judge the tasks' success. We report consensus success rate, Borda rank, and robustness under distractor-object and OOD-instruction perturbations.
2. **Pixel fidelity and scene geometry.** Decoded rollouts must remain visually coherent to support visual policies. We report image/video metrics: FID, SSIM, LPIPS, FVD, temporal LPIPS, and point-track consistency, together with perceptual and geometric scores from WorldArena [34]. This family measures generation and motion quality, temporal consistency, and scene geometry.
3. **Latent representation quality.** Because the transition model operates in latent space, we directly probe whether generated latents preserve action and task-relevant structure. We train an inverse dynamics model (IDM) [37] on frozen encoder latents to recover action chunks for horizon $k\in\{1, 4\}$, and apply the IDM to world model latents to measure generation-induced degradation. We train a classifier on latent trajectories of SOAR [54], a language and success label annotated dataset of trajectories, to classify whether a trajectory was a success given the text instruction. We again measure the degradation in accuracy induced by evaluating on generated latents.

4. Findings

4.1. Does the choice of latent space affect planning and policy performance?

Semantic latents offer better policy-in-the-loop performance. Table 1 shows that encoder choice strongly affects downstream VLA policy rollouts at DiT-S. Reconstruction-aligned spaces perform worst: VAE and VA-VAE have the lowest consensus success rates and weakest Borda ranks, while semantic encoders improve policy success, interaction quality, and robustness. V-JEPA 2.1 and SigLIP 2 variants give the strongest DiT-S results. Semantic-family VLA SR and CEM outperform reconstruction-family under paired bootstrap over tasks.

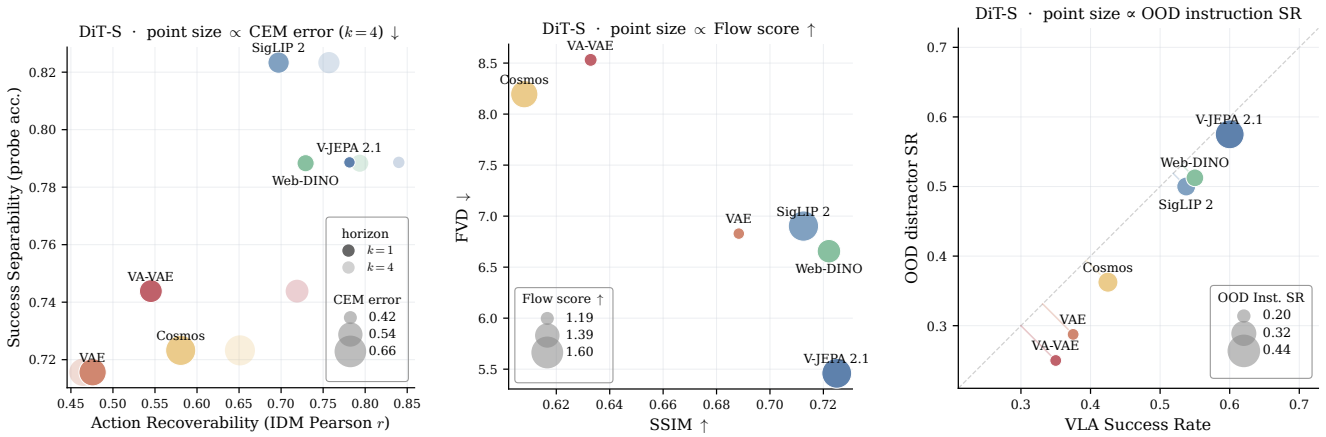
Native semantic spaces preserve action geometry for planning. Representation aligned spaces have the lowest DiT-S action-recovery errors (Table 1). For example, V-JEPA 2.1 is best at k=4 and SigLIP 2 is best at k=1. Fig. 2c likewise shows semantic encoders closer to the upper-right diagonal in the VLA–OOD plane, while VAE-family models fall lower and suffer larger distractor-induced drops.

Scaling narrows policy gaps but not action-centric gaps. For DiT-L, the gaps in VLA success and OOD robustness for VAE and Cosmos narrow relative to semantic encoders. We attribute this to improved visual fidelity at larger model size, which benefits the VLA policy. However, both still lag on CEM action recovery, which depends directly on latent transition structure rather than rendered visual quality; at DiT-L, VAE and Cosmos have larger k=1 CEM errors than all semantic encoders. They also lag on IDM r and classifier accuracy.

4.2. Does the latent space affect action recoverability and preservation of task semantics?

Semantic latents make action-relevant changes more recoverable. Table 2 shows that semantic encoders retain substantially more action information than reconstruction-aligned ones. On encoder latents, V-JEPA 2.1 and WebDINO achieve the strongest IDM Pearson r across both horizons, and this advantage largely persists after world model (WM) generation.

Semantic latents better preserve task-success information. From Table 2, we also see that success classifiers trained on frozen encoder latents achieve higher accuracy for semantic encoders, and their performance degrades less when evaluated on generated WM latents, with SigLIP 2 having best WM latent accuracy. This indicates that semantic spaces not only encode local action effects, but also retain higher-level task progress signals useful for policy evaluation.



(a) Latent space utility (b) Visual utility (c) Policy performance

Figure 2. **Latent space effect overview:** each point is a DiT-S world model trained by varying only the encoder and the associated decoder path. (a) **Upper-right is favorable.** Latent space metrics show that semantic encoders improve action recoverability, task-success separability, and action planning error (CEM) relative to reconstruction-aligned encoders. (b) **Lower-right is favorable.** Visual utility metrics show that pixel fidelity alone does not explain downstream performance: reconstruction-aligned spaces remain competitive on low-level image quality, while semantic spaces often improve video and motion quality. (c) **Upper-right is favorable.** Closed-loop evaluations show that semantic spaces generally yield higher VLA success and stronger robustness to OOD distractor objects and instructions. Details about all metrics are in Sec. 3.2.

Table 1. **DiT-S policy and behavioral metrics.** **Best** and runner-up per column. In-distribution (ID) SR and Out-of-Distribution (OOD) SR are calculated on a subset of 10 episodes with InternVL 3.5. Consenus SR and Borda rank aggregate InternVL3.5-14B and Qwen3.6-27B rankings. Interaction quality measures the plausibility of robot-object contact. PCK coverage measures point tracking recall. Muted ± terms show one standard deviation error averaged over episodes.

Encoder	VLA SR		Interaction quality		PCK	OOD robustness			CEM error	
	Consensus SR ↑	Borda rank ↓	IQ score ↑	Instr. follow ↑	PCK coverage ↑	ID SR ↑	OOD SR distractor ↑	OOD SR instruction ↑	k=1 ↓	k=4 ↓
VAE	0.169 ±0.030	25	3.26	3.48	0.719	0.375 ±0.054	0.287 ±0.051	0.200 ±0.045	0.111 ±0.009	0.612 ±0.023
VA-VAE	0.175 ±0.030	23	3.22	3.42	0.715	0.350 ±0.053	0.250 ±0.048	0.200 ±0.045	0.097 ±0.005	0.543 ±0.023
Cosmos	0.244 ±0.034	16	3.32	3.51	0.707	0.425 ±0.055	0.362 ±0.054	0.275 ±0.050	0.112 ±0.009	0.661 ±0.033
V-JEPA 2.1	0.344 ±0.038	**6**	3.43	3.78	**0.735**	0.600 ±0.055	0.575 ±0.055	**0.400** ±0.055	0.084 ±0.008	**0.424** ±0.014
V-JEPA 2.1$_{96}$	**0.362** ±0.038	8	**3.52**	**3.84**	0.735	0.600 ±0.055	0.537 ±0.056	0.250 ±0.048	0.089 ±0.007	0.548 ±0.017
Web-DINO	0.212 ±0.032	21	3.34	3.58	0.735	0.550 ±0.056	0.512 ±0.056	0.250 ±0.048	0.090 ±0.007	0.474 ±0.026
Web-DINO$_{96}$	0.300 ±0.036	11	3.44	3.77	0.732	0.600 ±0.055	0.512 ±0.056	0.275 ±0.050	0.090 ±0.007	0.531 ±0.025
SigLIP 2	0.325 ±0.037	9	3.43	3.58	0.730	0.537 ±0.056	0.500 ±0.056	0.263 ±0.049	**0.082** ±0.006	0.523 ±0.030
SigLIP 2$_{96}$	0.331 ±0.037	15	3.42	3.71	0.731	**0.625** ±0.054	**0.588** ±0.055	0.312 ±0.052	0.086 ±0.005	0.537 ±0.026

4.3. How does the latent space affect visual fidelity?

Semantic latent spaces remain visually competitive. Table 3 shows that the policy gains from semantic encoders do not come at the cost of decoded visual quality. At DiT-S scale, these encoders dominate most perceptual, structural, and video-level metrics, particularly when used with adapters d_{96}: SigLIP 2$_{96}$ gives the best SSIM, V-JEPA 2.1$_{96}$ gives the best FVD, and Web-DINO variants are strongest on JEPA similarity, subject consistency, depth error, and temporal LPIPS. VAE-style spaces remain competitive on image quality, and qualitatively tend to preserve sharper local appearance details, but they lag behind semantic spaces on global structure and temporal generation quality. Fig. 3 shows semantic space models have lower gap for pixel reconstruction, particularly while extrapolating beyond the 10-frame horizon length seen during training.

Scaling details. The main table focuses on DiT-S to emphasize controlled latent-space comparisons. The key pattern remains that stronger visual metrics do not reliably predict policy-facing metrics.

Table 2. **IDM Pearson** r (horizons $k \in \{1, 4\}$) and Success classifier for DiT-S, reported on encoder (Enc.) and world model (WM) latents.

Encoder	Pearson r				Classifier Acc.	
	Enc.↑		WM↑		Whole-video	
	k=1	k=4	k=1	k=4	Enc.↑	WM↑
• VAE	0.507	0.478	0.476	0.464	0.835	0.716
• VA-VAE	0.549	0.744	0.545	0.719	0.868	0.744
• Cosmos	0.626	0.673	0.581	0.651	0.851	0.723
• V-JEPA 2.1	**0.829**	**0.865**	**0.781**	**0.840**	0.905	0.789
• Web-DINO	0.820	0.845	0.729	0.794	**0.906**	0.788
• SigLIP 2	0.772	0.793	0.697	0.757	0.903	**0.823**

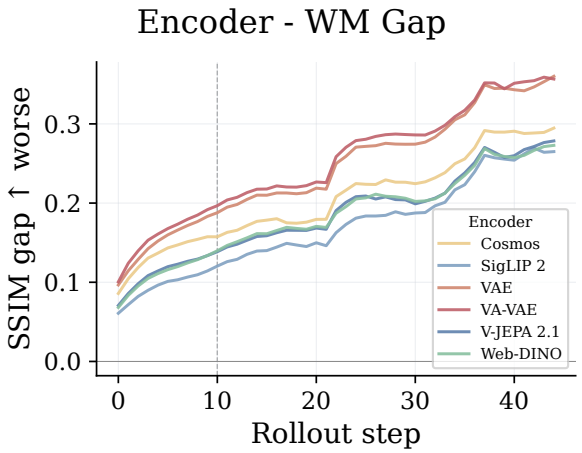


Figure 3. **SSIM gap** over steps.

4.4. Does scaling along input views and model size help?

Multi-view training improves action recovery but can hurt video quality under limited data. We take the trained DiT-S models and finetune them for 20 epochs on the BridgeV2 episodes that contain three camera views. Fig. 4 (left) shows that while this does lead to superior CEM action prediction, it also degrades generation quality, possibly due to smaller number of training episodes. However, the semantic encoders are more robust to this degradation. **Model scaling improves both visual quality and policy success, with larger gains for reconstruction latents:** in Fig. 4 (right), we see that both generation (SSIM) and policy performance (VLA-SR) generally scale with the DiT size. Here, VAE scales notably well on visual metrics and approaches semantic encoders, which already perform strongly at DiT-S.

4.5. Do compressed adapter latents aid semantic encoders further for world modeling?

Adapters improve diffusion ease but can distort control geometry. Fig. 5, Table 1, and Table 3 show that the compressed space d_{96} of adapters helps the latent diffusion model, as also observed by Zhang et al. [51] and Bai et al. [4]. This leads to generally stronger performance than the native variants on most metrics except latent CEM action error, OOD robustness, and PCK coverage. These findings hint towards the adapter compressing the latent space in a way that is useful for high-level task completion such as diffusion denoising but hurtful for fine-grained tasks like trajectory optimization, where precise directional action information is needed.

4.6. Do reconstruction-aligned and semantic encoders fail differently?

The main failure modes differ: reconstruction latents hallucinate task semantics, while semantic latents miss geometry and contact. Our qualitative rollouts show that all encoder families share a common failure mode where static scene elements are faithfully preserved while manipulation-relevant details hallucinate. Beyond this universal pattern, encoder families show distinct hallucinations. Reconstruction encoders tend to fail at the object-semantic level: VAE and Cosmos can hallucinate task-relevant objects, producing coherent looking but task-incorrect states, and under OOD instructions, both can maintain the prior action pattern rather than updating to the new goal. Semantic encoders preserve task-level intent at the cost of geometric precision. We find the latter to better capture semantic distinctions even under instruction shift.

4.7. Do high-dimensional semantic latents and adapter add computational overhead?

High-dimensional semantic latents do not substantially increase DiT compute in our setup. The DiT always receives the same number of tokens per frame N=256, hence larger channel dimensions only affect the input/output projections. The main compute differences instead come from the frozen encoder and decoder architectures. In particular, ViT-based semantic encoders paired with the adapter pixel decoder remain competitive in total GFLOPs, while native high-dimensional semantic spaces require only a lightweight wide DDT head [43]. We report parameter counts and GFLOPs split by encoder, adapter, DiT, and decoder separately.

5. Practical Recipe, Related Work, and Limitations

A Recipe for Semantic Latent Diffusion Robotics World Modeling Our findings suggest a practical recipe for building robotic latent diffusion world models. **Do not begin by optimizing for visual realism alone**. Instead, choose a latent space that makes **action and task progress** explicit, make that space easy for diffusion to model, and evaluate the resulting world model with control- and policy-based metrics. Visual realism can often be improved through better decoder training, but transition quality and latent fidelity

Table 3. **Visual realism quality** for DiT-S. **Best** and runner-up.

Encoder	Visual quality						Content consistency		Motion quality			
	SSIM↑	LPIPS↓	FID↓	Image quality↑	Aesthetic quality↑	JEPA sim.↑	Subject consist.↑	Depth AbsRel↓	Dyn. degree↑	Flow score↑	FVD↓	t-LPIPS↓
VAE	0.688	0.218	17.428	**0.592**	0.467	0.871	0.810	0.390	0.767	1.186	6.829	0.0264
VA-VAE	0.633	0.226	15.488	0.585	0.464	0.783	0.817	0.455	0.765	1.204	8.531	0.0253
Cosmos	0.608	0.245	16.947	0.558	0.463	0.517	0.793	0.638	0.813	1.511	8.195	0.0223
V-JEPA 2.1	0.725	**0.176**	6.771	0.578	0.473	0.929	0.841	0.404	0.832	1.587	5.459	0.0197
V-JEPA 2.1_{96}	0.729	0.179	6.302	0.579	**0.474**	0.928	0.841	0.363	**0.843**	**1.653**	**5.224**	0.0212
Web-DINO	0.722	0.199	7.626	0.576	0.472	0.938	**0.849**	**0.350**	0.794	1.408	6.656	0.0234
Web-$DINO_{96}$	0.728	0.181	**5.998**	0.574	0.473	**0.944**	0.841	0.375	0.835	1.634	5.510	**0.0195**
SigLIP 2	0.713	0.205	7.858	0.566	0.471	0.931	0.839	0.394	0.827	1.602	6.902	0.0228
SigLIP 2_{96}	**0.738**	0.179	6.881	0.573	0.472	0.938	0.843	0.372	0.827	1.547	6.005	0.0223

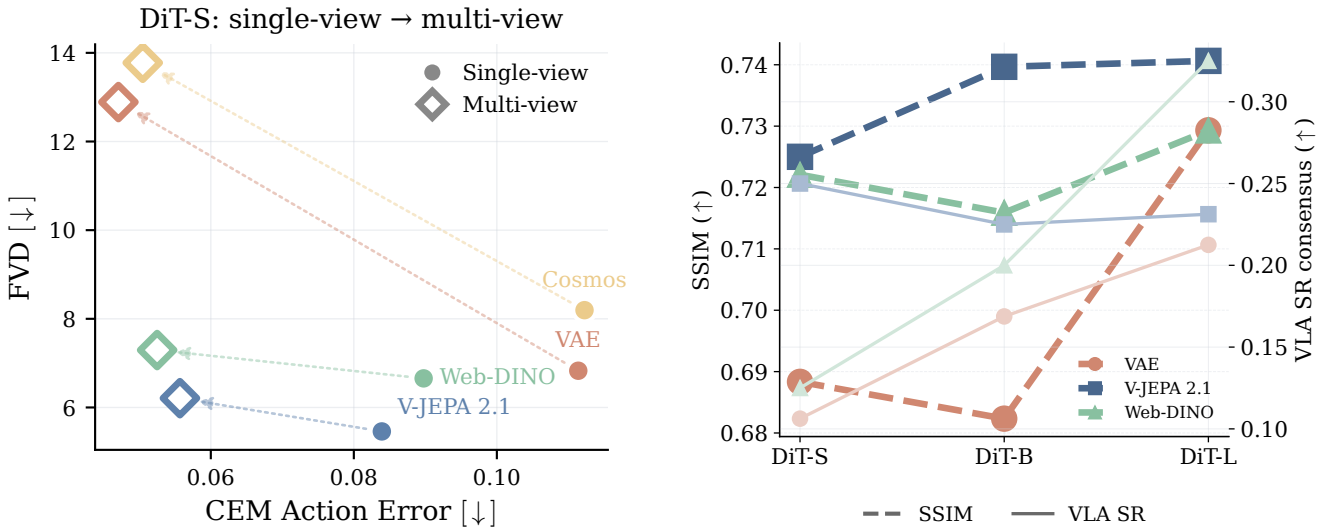


Figure 4. **Scaling** camera views (left) and DiT sizes (right).

remain important. Use robot demonstration datasets with preferably **multi-view trajectories** and, when available, success/failure labels to unlock diverse evaluations. **Choose pretrained semantic encoders** as the default latent state space, since they preserve action geometry and task progress better than reconstruction latents. **Pair them with adapter compression** when decoded rollout quality or VLA-in-the-loop evaluation matters. For transition model, a robust default for high-dimensional semantic spaces is: a **spatial-temporal DiT** with causal temporal blocks, a shallow-wide DDT head [43], and a dimension-aware noise shifting [52]. The spatial blocks stay non-causal since per-frame patches are denoised jointly. For training, diffusion forcing [8] can be used for autoregressive next-frame rollout. Finally, evaluate world models on **multiple axes** covering both visual, latent, and downstream task performance.

Related work. **Robotic world models** can be seen to span three related objectives. One line treats world models as policy-evaluation environments: WorldGym [29] and WorldEval [24] roll out policies in learned video models; [40] studies how pretraining, data diversity, and failure modes affect evaluation. A second line adapts pretrained generators into interactive simulators: UniSim [45] learns interactive real-world simulators from broad data; Vid2World [18] causalizes video diffusion with action guidance; Ctrl-World [14] studies multi-view, long-horizon, policy-in-the-loop manipulation. A third line moves prediction and planning into semantic feature space: DINO-WM [53], DINO-world [5], and V-JEPA 2-AC [3] show that pretrained representations can support latent space forecasting and zero-shot or few-shot planning. These works establish the utility of both video generation and semantic representations, but do not isolate the encoder-defined latent space within a unified action-conditioned framework.

World model evaluation has moved beyond rollout plausibility and policy ranking toward physics, semantics, and embodied utility [23, 26]. RBench [21] measures task correctness and structural realism. WorldModelBench [22] highlights instruction-following and physics-adherence failures

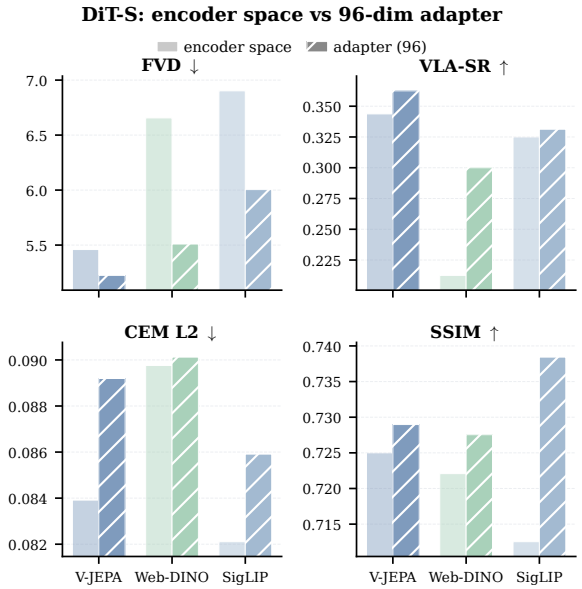


Figure 5. **Adapter** ablation results.

Key Empirical Takeaways

- **Visual fidelity does not always imply downstream performance.** Reconstruction latents can match or exceed semantic latents on pixel-level metrics, especially at larger DiT scale, yet lag on action recovery, task-success probes, CEM planning, and policy-in-the-loop evaluation.
- **Semantic latents scale better with multiple views.** Under limited data, adding multiple views improves planning but can hurt visual rollouts; semantic encoders retain the action recoverability benefit with substantially less degradation than reconstruction latents.
- **Adapters trade control geometry for diffusion ease.** Adapters ease diffusion and decoding, but can distort fine-grained action geometry compared with native semantic features.
- **World models in semantic spaces lower reconstruction and generation ceiling gap.** Training decoders with the same budget for semantic world models is more effective.
- **High-dimensional semantic latents are practical in DiTs.** With a fixed patch-token count, semantic width adds little to the transition-model cost.

missed by generic video metrics. EWMBench [47] evaluates scene consistency, motion correctness, and semantic alignment. World-in-World [48] prioritizes closed-loop task success, WoW-World-Eval [12] adds inverse-dynamics-based action plausibility, and WorldArena [34] exposes the gap between perceptual quality and downstream functionality. These benchmarks evaluate world models at system-level while we seek to evaluate them at model-level.

Limitations and future work. Our study isolates the effect of encoder-defined latent spaces within a controlled action-conditioned LDM protocol. The conclusions are therefore scoped to the Bridge V2 manipulation setting and a shared robot embodiment. Evaluating broader embodiments, domains, and data regimes is an important next step. Our policy-in-the-loop experiments also focus on evaluating a fixed VLA policy inside generated rollouts, while policy improvement and sim-to-real transfer would test a complementary use of the same world models. Lastly, our evaluation partially relies on VLM-based success judgments, which may introduce evaluator bias. We reduce this dependence by aggregating multiple VLMs and pairing them with non-VLM diagnostics, including CEM planning, inverse dynamics, latent success classification, and visual/geometric metrics.

Evaluation details. For CEM action recovery, we optimize a diagonal Gaussian over the searched action coordinates, use 400 candidate sequences and 5 update iterations per transition, and reuse one sampled rollout-noise tensor within each transition so all candidates see the same stochastic objective. For VLA-in-the-loop scoring, each rollout is judged from 16 sampled frames by InternVL 3.5 and Qwen 3.6, and success is counted only when both raters agree.

6. Conclusion

Our study shows that the encoder-defined latent space is a central design choice for action-conditioned latent diffusion world models in robotics. Across visual, latent, planning, and policy-in-the-loop evaluations, semantic representation spaces such as that of V-JEPA 2.1, Web-DINO, and SigLIP 2 generally provide stronger action recoverability, task-success classification accuracy, robustness, and downstream policy performance than reconstruction-aligned VAE-style latents, even when the latter remains competitive or superior on low-level photometric metrics. These results support the view that robotic world models should not be selected solely by visual realism, but by whether their latent dynamics preserve action-relevant structure and policy evaluation accuracy.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 3

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023. 1

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. 1, 3, 4, 7

[4] Jianhong Bai, Xiaoshi Wu, Xintao Wang, Xiao Fu, Yuanxing Zhang, Qinghe Wang, Xiaoyu Shi, Menghan Xia, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Kun Gai. SemanticGen: Video Generation in Semantic Space, December 2025. 6

[5] Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025. 7

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3(1):3, 2024. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[8] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. 7

[9] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ. 3

[10] Tom Erez, Yuval Tassa, and Emanuel Todorov. Simulation tools for model-based robotics: Comparison of bullet, havok, mujoco, ode and physx. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 4397–4404. IEEE, 2015. 1

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 3, 4

[12] 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. 8

[13] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 370–382, 2025. 3

[14] 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. 7

[15] David Ha and Jürgen Schmidhuber. World models. *eprint arXiv: 1803.10122*, 2018. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[18] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025. 7

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025. 4

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[21] Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. R-bench: Are your large multimodal model robust to real-world corruptions?, 2024. 7

[22] 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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=a3hafrDzuA. 7

[23] 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, 2025. 7

[24] 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*, 2025. 7

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 3

[26] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike 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. 3, 7

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3

[29] 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. 7

[30] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b. 4

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[33] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004. 4

[34] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou 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. 4, 8

[35] Ansh Kumar Sharma, Yixiang Sun, Ninghao Lu, Yunzhe Zhang, Jiarao Liu, and Sherry Yang. Worldgymnast: Training robots with reinforcement learning in a world model. *arXiv preprint arXiv:2602.02454*, 2026. 1

[36] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kdpeJNbFyf. 1

[37] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=meRCKuUpmc. 4

[38] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012. 1

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1, 3

[40] Wei-Cheng Tseng, Jinwei Gu, Qinsheng Zhang, Hanzi Mao, Ming-Yu Liu, Florian Shkurti, and Lin Yen-Chen. Scalable policy evaluation with video world models. *arXiv preprint arXiv:2511.11520*, 2025. 1, 7

[41] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021. 1

[42] 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, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 1723–1736. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/walke23a.html. 2, 3

[43] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 1, 3, 6, 7

[44] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 4

[45] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *NeurIPS Workshop on Generalization in Planning*, 2023. 1, 7

[46] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3

[47] 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. 8

[48] 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*, 2025. 8

[49] Jiahui Zhang, Ze Huang, Chun Gu, Zipei Ma, and Li Zhang. Reinforcing action policies by prophesying. *arXiv preprint arXiv:2511.20633*, 2025. 1

[50] Mingkun Zhang, Wangtian Shen, Fan Zhang, Haijian Qin, Zihao Pei, and Ziyang Meng. Rae-nwm: Navigation world model in dense visual representation space. *arXiv preprint arXiv:2603.09241*, 2026. 1

[51] Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, et al. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. *arXiv preprint arXiv:2512.17909*, 2025. 1, 2, 3, 6

[52] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 1, 2, 3, 7

[53] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In *Forty-Second International Conference on Machine Learning*, June 2025. 1, 4, 7

[54] Zhiyuan Zhou, Pranav Atreya, Abraham Lee, Homer Walke, Oier Mees, and Sergey Levine. Autonomous improvement of instruction following skills via foundation models. *arXiv preprint arXiv:407.20635*, 2024. 3, 4

[55] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. *arXiv preprint arXiv:2511.09515*, 2025. 1