DO SEMANTIC FEATURE SPACES MAKE BETTER WORLD ACTION MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

World action models (WAMs) offer a promising approach to robotic control by coupling future visual prediction with action generation. Many existing WAMs predict future states in reconstruction-oriented variational autoencoder (VAE) latent spaces, raising the question: do semantic feature spaces provide a better foundation for robot control? We investigate this question through a systematic comparison of the Stable Diffusion 3.5 VAE, DINOv2, DINOv3, and SigLIP 2 within a common cascaded WAM framework. We use flow matching to predict future features directly in each encoder's representation space and an inverse dynamics model to decode actions from these predictions, with training configurations tuned for each representation. We analyze action decoding, future dynamics prediction, and closed-loop task success across training data and compute budgets. Experiments on RoboCasa365 and transfer from real-world DROID data to high-fidelity simulation in RoboLab assess WAMs' performance and generalization. Across the evaluated setups, semantic-space WAMs achieve higher task success and scale more favorably with compute, data, and pretraining on additional tasks than VAE-based WAMs. These findings support semantic feature spaces as a stronger foundation for effective and scalable world action modeling.

1 INTRODUCTION

Vision-language-action (VLA) models have emerged as a powerful paradigm for robotic control (Brohan et al., 2023a; Kim et al., 2024; Black et al., 2026; Physical Intelligence et al., 2025b; Bjorck et al., 2025), benefiting from semantic knowledge acquired through large-scale vision-language training. Their training objective, however, typically focuses on predicting actions directly, without explicitly supervising the future scene changes needed to accomplish a task. World action models (WAMs) (Kim et al., 2026; Ye et al., 2026) address this by coupling future visual prediction with action generation. Their effectiveness raises a fundamental question: *which representation space should a WAM use to support reliable robot control?*

Most recent WAMs (Wang et al., 2025; Aditi et al., 2026; Gao et al., 2026; Kim et al., 2026) use the latent space of a variational autoencoder (VAE) (Kingma & Welling, 2022). The compactness of VAE latents makes them convenient targets for diffusion or flow matching (Rombach et al., 2022; Esser et al., 2024; Lipman et al., 2023), but their suitability for visual reconstruction does not establish their suitability for action prediction. Semantic visual features such as DINO and SigLIP (Oquab et al., 2024; Siméoni et al., 2025; Tschannen et al., 2025) offer an alternative that may make task-relevant information more accessible for robot control. However, their higher dimensionality and complicates future prediction, while their learned invariances may discard details needed for precise control. A useful WAM representation must satisfy two requirements: being predictable from the available context and preserving information needed to recover executable actions.

Recent work has begun to investigate the representation choice. Nilaksh et al. (2026) compare reconstruction and semantic representations in world models for planning and policy evaluation. A concurrent work OpenWAM (Wang et al., 2026b) extends this investigation to WAMs and demonstrates the effectiveness of semantic representations with dimensional compression. These studies motivate three complementary questions: *How effectively can WAMs operate directly in uncompressed semantic feature spaces? Which representations best support future prediction, action decoding,*

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.

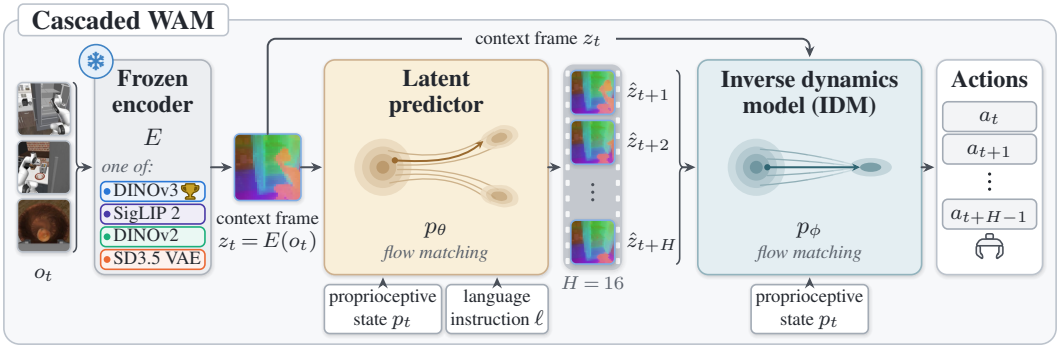

Figure 1: **Cascaded world action model (WAM).** A frozen encoder E encodes multi-view visual observations o_t into context features z_t. Conditioned on z_t, the language instruction ℓ, and proprioceptive state p_t, the latent predictor p_θ generates future features $\hat{z}_{t+1:t+H}$. The inverse dynamics model p_ϕ converts the current and predicted future features, together with p_t, into actions $a_{t:t+H-1}$. We compare four visual encoders within this common framework.

and ultimately closed-loop manipulation? How does each representation benefit from scaling the amount of training data and the diversity of pretraining tasks?

We systematically study these questions in a cascaded WAM framework, as illustrated in Figure 1. Cascaded WAMs separate future prediction from action decoding (Wang et al., 2026a), allowing us to examine both stages and their combined effect on control compared to Joint WAMs. Building on high-dimensional space flow matching (Porcher et al., 2026), we predict uncompressed visual features conditioned on current observations, language, and proprioception, avoiding an additional compression stage. A flow matching based inverse dynamics model (IDM) then converts these predicted futures into actions. We compare four representations: Stable Diffusion 3.5 VAE (Esser et al., 2024), DINOv2 (Oquab et al., 2024), DINOv3 (Siméoni et al., 2025), and SigLIP 2 (Tschannen et al., 2025), covering reconstruction, self-supervised visual learning, and vision-language pretraining. We optimize the WAM training configuration for each representation to isolate the effect of the representation from other design choices.

Our analysis connects IDM action error given ground-truth future features, future feature prediction quality, and closed-loop task success. On RoboCasa365 (Nasiriany et al., 2026), we evaluate 50 target tasks over 1,500 episodes, varying the compute budget, the amount of training data, and whether models are pretrained on additional tasks. We also test transfer from real-robot pretraining on DROID (Khazatsky et al., 2025) with generalization evaluations on a high-fidelity simulation benchmark RoboLab (Yang et al., 2026). Our results challenge the prevailing use of VAE latents in WAMs: all three semantic representations achieve higher task success on both RoboCasa365 and RoboLab and benefit more from pretraining, supporting semantic feature spaces as a stronger foundation for world action modeling.

2 Related Work

2.1 Vision-Language-Action Models (VLA)

Vision-language-action (VLA) models directly map visual observations and language instructions to robot actions. RT-1 (Brohan et al., 2023b) showed that transformer policies can scale with robot demonstration data, while RT-2 (Brohan et al., 2023a) transferred knowledge from vision-language pretraining to robotic control. RT-X (O'Neill et al., 2025) and OpenVLA (Kim et al., 2024) extended this direction with diverse cross-robot data and open models. More recent models pair a vision-language backbone with a continuous action generator. The π series introduces a flow-matching action model (Black et al., 2026) and extends it with heterogeneous co-training for open-world generalization (Physical Intelligence et al., 2025b), reinforcement learning from real-world experience (Physical Intelligence et al., 2025a), and context conditioning for steerable behavior (Physical Intelligence et al., 2026). GR00T models (Bjorck et al., 2025) use a similar design with pretraining on

a data pyramid spanning Internet data, synthetic robot data, and real robot demonstrations. Unlike WAMs, these methods primarily predict actions without explicitly predicting future visual states.

2.2 World Action Models (WAM)

In contrast to VLAs, world action models also model the distribution of future states. They are divided into two types depending on whether they factorize or jointly model future states and actions (Wang et al., 2026a). **Cascaded WAMs** first predict future visual states and then map the predicted trajectory to actions with an inverse dynamics model (Du et al., 2023; Wu et al., 2023; Tian et al., 2024; Hu et al., 2025; Pai et al., 2025). **Joint WAMs** predict future states and actions within one model (Li et al., 2025; Zhu et al., 2025). To reduce inference cost, another line of work uses future prediction as an auxiliary training objective to shape the policy representation, then discards the prediction head at inference (Li et al., 2026).

Most WAMs build on pretrained video generators rather than training a world model from scratch. Models like Wan 2.2 (Wang et al., 2025) provide a strong generative initialization that can be adapted to robot data through action fine-tuning, and recent large-scale systems such as Cosmos 3 (Aditi et al., 2026) and DreamDojo (Gao et al., 2026) follow a similar paradigm. This strategy leverages the rich visual and temporal priors learned by large-scale video models for robotics, but it also inherits the VAE latent space on which these generators operate.

Recent work explores semantic representations as alternatives to VAE latents. DINO-WM (Zhou et al., 2025) predicts DINOv2 features for goal-conditioned planning. EgoWAM (Li et al., 2026) finds that DINO features and 3D flow outperform VAE latents as auxiliary training targets. Nilaksh et al. (2026) compare reconstruction and semantic spaces in world models for planning and policy evaluation. Concurrent work OpenWAM (Wang et al., 2026b) studies representation choice within a broader investigation of WAM design, demonstrating the effectiveness of dimensionally compressed semantic features. We complement these efforts by comparing reconstruction, self-supervised, and vision-language representations in their native spaces within a cascaded WAM, examining how representation space affects task success and the benefits of additional data and pretraining tasks.

2.3 Flow Matching

Flow matching (Lipman et al., 2023) learns a vector field that transports noise to a target distribution, making it well suited to modeling complex and uncertain outputs. It has been widely used in image generation (Esser et al., 2024), video generation (Wang et al., 2025), and robot action generation (Physical Intelligence et al., 2025b).

Latent visual generators often separate representation learning from generation. Latent Diffusion Models (Rombach et al., 2022) first train a low-dimensional VAE with reconstruction, perceptual, adversarial, and KL losses, then freeze it and train a generative model in the resulting latent space. Diffusion Transformers (DiTs) (Peebles & Xie, 2023) and rectified-flow transformers (Esser et al., 2024) follow this design.

Recent work has shown that high-dimensional semantic spaces can nevertheless support strong generative modeling. Representation Autoencoders (RAEs) (Zheng et al., 2025) pair frozen DINOv2, SigLIP 2, or MAE encoders with learned decoders and adapt the DiT architecture and noise schedule to wider latent spaces. FlowWM (Porcher et al., 2026) applies flow matching directly to DINOv3 features for stochastic world modeling. JiT (Li & He, 2026) further studies generative learning in high-dimensional, low-rank spaces and shows that clean-sample prediction is more effective as the ambient dimensionality grows. However, these works do not systematically compare representation spaces in terms of both future prediction and robot action prediction. We address this gap by comparing VAE, DINOv2, DINOv3, and SigLIP 2 within the same cascaded WAM framework.

3 Method

3.1 Problem Formulation

Let o_t denote all RGB camera observations at time t, E a visual encoder, $z_t = E(o_t)$ the corresponding visual features, p_t the robot's proprioceptive state, and ℓ a language instruction. Given a

prediction horizon H, a world action model represents the joint distribution

$$p(z_{t+1:t+H}, a_{t:t+H-1} \mid z_t, p_t, \ell)$$

over future visual features and actions.

In this work, we adopt the cascaded WAM (Wang et al., 2026a), which factorizes this distribution as

$$\underbrace{p_\theta(z_{t+1:t+H} \mid z_t, p_t, \ell)}_{\text{latent predictor}} \underbrace{p_\phi(a_{t:t+H-1} \mid z_t, z_{t+1:t+H}, p_t, \ell)}_{\text{inverse dynamics model}}.$$

The latent predictor generates future visual features consistent with the current observation and task instruction. The inverse dynamics model (IDM) predicts the actions required to follow this feature sequence. At inference, the latent predictor first generates $\hat{z}_{t+1:t+H}$, and the IDM then decodes the actions from these predicted features. Figure 1 illustrates the overall framework.

3.2 Representation Spaces

A useful representation for cascaded WAMs must support future prediction while preserving the information needed to decode actions. Reconstruction-oriented VAE latents provide compact prediction targets, whereas semantic features from pretrained visual encoders may make task-relevant information more accessible. We focus on image encoders to compare visual representations without introducing differences in temporal encoding.

We therefore compare four encoders spanning reconstruction-based pretraining (Stable Diffusion 3.5 VAE (Esser et al., 2024)), self-supervised visual learning (DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025)), and vision-language learning (SigLIP 2 Tschannen et al. (2025). SigLIP 2 also serves as the visual encoder in the VLM backbone of GR00T VLA (Bjorck et al., 2025), providing a shared representation family for comparisons between WAMs and VLAs.

Table 1 summarizes the visual encoders used in our comparison. The Stable Diffusion 3.5 VAE produces 16-channel latents. We concatenate each 2×2 spatial block into a single token, yielding 256 tokens of dimension $d = 64$. For DINOv2, DINOv3, and SigLIP 2, we use ViT-L variants of comparable size, each with approximately 300 million parameters and feature dimension $d = 1024$.

Table 1: Visual encoders used in our comparison. Token counts include spatial tokens only. †The VAE encodes each 8×8 pixel patch into a 16-channel latent. We merge 2×2 neighboring latents into one token, so each token covers 16×16 pixels and has dimension 64.

Representation	Pretraining	Architecture	Parameters	d	Patch size	Resolution	Tokens/frame
Stable Diffusion 3.5 VAE	reconstruction	VAE	84M	$64^\dagger$	$16^\dagger$	256×256	256
DINOv2	self-supervised	ViT-L	300M	1024	14	252×252	324
DINOv3	self-supervised	ViT-L	300M	1024	16	256×256	256
SigLIP 2	language-supervised	ViT-L	316M	1024	16	256×256	256

3.3 Flow Matching Training

We train the latent predictor and the IDM separately with the standard conditional flow matching objective (Lipman et al., 2023). Let $x_0 \sim p_{\text{noise}}$, $x_1 \sim p_{\text{data}}$, and $\tau \sim \mathcal{U}(0, 1)$. We define

$$x_\tau = (1 - \tau)x_0 + \tau x_1$$

and minimize

$$\mathcal{L}_{\text{FM}}(\psi) = \mathbb{E}\left[\|u_\psi(x_\tau, \tau; c) - (x_1 - x_0)\|_2^2\right],$$

where c denotes the conditioning inputs. For the latent predictor, $x_1 = z_{t+1:t+H}$ and $c = (z_t, p_t, \ell)$. For the IDM, $x_1 = a_{t:t+H-1}$ and $c = (z_t, z_{t+1:t+H}, p_t, \ell)$, where $z_{t+1:t+H}$ contains ground-truth future features during training.

Figure 2 illustrates the latent predictor and IDM architectures, with model configurations provided in Appendix A. We tune the training settings for each representation to account for differences in optimization requirements. Implementation details, including the time shift selected for each representation, are provided in Appendix B.

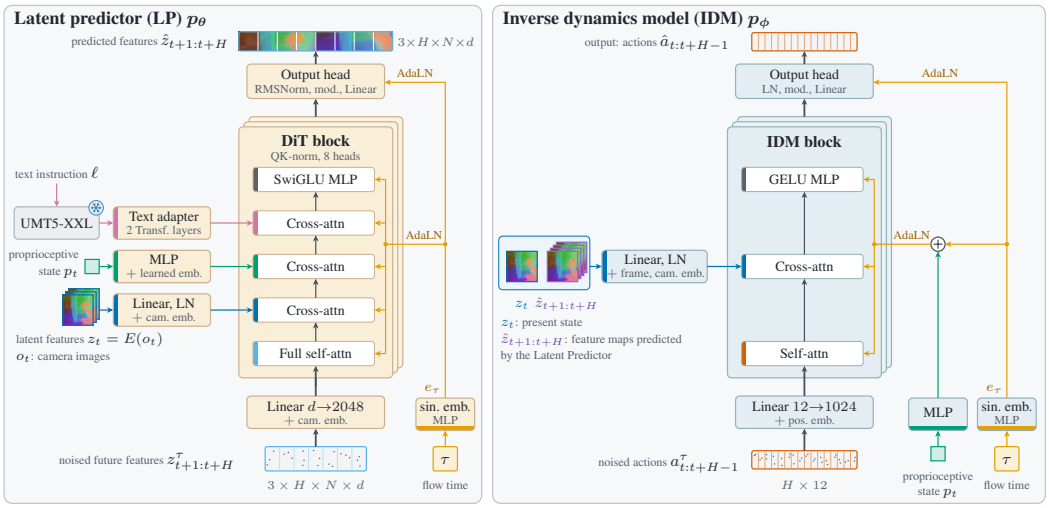

Figure 2: Latent predictor (left) and inverse dynamics model (right). The latent predictor generates future feature maps from z_t, p_t, and ℓ, which serve as conditioning for the IDM.

4 Experiments

4.1 Experimental Setup

RoboCasa365. RoboCasa365 (Nasiriany et al., 2026) is a large-scale simulation benchmark for generalist robot policies, comprising 365 household manipulation tasks across diverse kitchen environments. Three cameras are used including left shoulder, right shoulder and wrist. We consider two training settings. In the *target-only* setting, we train on directly on the Target50 human demonstration set, which contains 50 tasks with about 500 demonstrations per task, totaling about 25,000 episodes (193 hours). In the *pretrain-and-finetune* setting, we first pretrain on the Pretrain300 human demonstrations and then fine-tune on Target50. Pretrain300 contains 300 tasks with about 100 demonstrations per task, totaling about 30,000 episodes (482 hours). Comparing these two settings allows us to assess how broad multi-task pretraining affects downstream control performance.

We evaluate all 50 Target50 tasks with 30 episodes per task and seed, using three random seeds. This corresponds to 1,500 episodes per seed and 4,500 episodes per model. We report the mean success rate and standard deviation across the three seeds.

DROID and RoboLab. DROID (Khazatsky et al., 2025) is a large-scale real-robot dataset containing approximately 76,000 trajectories (350 hours), collected across diverse scenes, tasks, and operators. It provides multi-view RGB observations, robot proprioception, actions, and language annotations. For DROID training, we use all successful episodes and no held-out split. The model observes two exterior views and one wrist-camera view. Both the proprioceptive state and actions consist of the seven arm joint positions and the gripper position, recorded at 15 Hz. We randomly sample one of the available language annotations for the training data. The latent predictor and IDM use the same architecture and training recipe as on RoboCasa365 (Appendix A).

RoboLab (Yang et al., 2026) is a high-fidelity simulation benchmark to evaluate the generalization of task-generalist robot policies. It has been shown to correlate with real-robot performance, making it a useful proxy for large-scale real-world evaluation. We follow the RoboLab protocol: the models are trained exclusively on DROID and evaluated zero-shot in RoboLab, without using any simulation data for fine-tuning. RoboLab-120 comprises 120 tabletop manipulation tasks spanning visual, relational, and procedural competencies across three difficulty levels. We use the DROID embodiment, consisting of a Franka Panda arm with a Robotiq gripper, an exterior camera and a wrist camera, and joint-position control. We run 10 episodes per task, for a total of 1,200 episodes per model. We report both success rate and the RoboLab subtask-progress score, which assigns partial credit for subtasks a policy completes and reaches 1 for successful episodes.

4.2 Inverse Dynamics Model

We first evaluate how well each representation supports action prediction given ground-truth future observations, isolating action decoding from errors in future prediction. We train the IDM with action horizon $H = 16$ on 80% of the demonstrations in RoboCasa365 Target50 and evaluate on the remaining 20%. At evaluation, we perform full iterative denoising with eight steps and report the denormalized action mean absolute error (MAE).

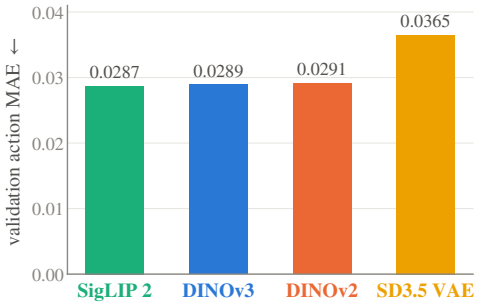

Figure 3: **IDM action error given ground-truth future features.** Lower MAE is better. Semantic representations perform similarly and outperform VAE latents.

As shown in Figure 3, SigLIP 2, DINOv3, and DINOv2 achieve similar MAEs, while the error of the Stable Diffusion 3.5 VAE is about 25% higher. Thus, all three semantic feature spaces support more accurate action decoding than VAE latents, but there is no clear advantage among the semantic encoders. This comparison provides a diagnostic of each representation's suitability for inverse dynamics; subsequent closed-loop evaluations (Section 4.3 and 4.4) examine whether these advantages translate into improved task success.

4.3 Cascaded WAM on RoboCasa365

We evaluate closed-loop task success on RoboCasa365 under the two training settings described in Section 4.1, with results reported in Table 2.

Training on Target50 only. DINOv3 achieves the highest success rate, while all three semantic representations outperform the VAE. The VAE's lower success rate is consistent with its higher IDM error.

Table 2: Success rate (%) on RoboCasa Target50, averaged over three seeds. Δ is the gain from Pretrain300 pretraining. The four WAMs differ only in latent representation, while GR00T is a VLA baseline without predicting future states. All visual feature encoders remain frozen.

Model	Type	Target50 ↑	Pretrain300 → Target50 ↑	Δ
DINOv3	WAM	**66.1** ±1.9	**73.5** ±2.0	**+7.4**
SigLIP 2	WAM	62.4 ±2.1	68.7 ±2.2	+6.3
DINOv2	WAM	61.7 ±2.0	66.5 ±2.1	+4.8
Stable Diffusion 3.5 VAE	WAM	49.1 ±2.4	51.5 ±2.3	+2.4
GR00T 1.7	VLA	50.0 ±2.3	52.0 ±2.2	+2.0

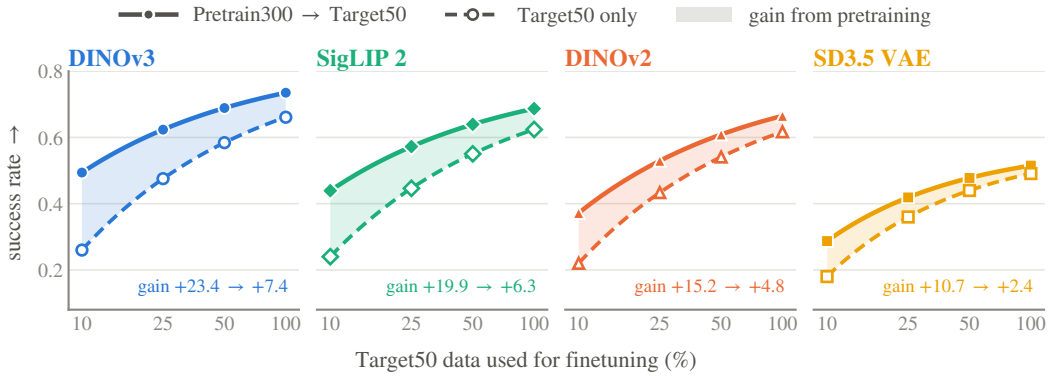

Figure 4: Success rate versus Target50 data fraction. Solid/dashed curves denote training with/without Pretrain300 pretraining; shading indicates pretraining gains.

Pretraining on broad tasks. Pretraining on Pretrain300 improves all four representations without changing their ranking. However, the semantic encoders gain 4.8 - 7.4 percentage points, compared with only 2.4 points for the VAE. These results suggest that semantic-space WAMs benefit more from pretraining across diverse tasks.

Scaling with target task data. Figure 4 shows success rates as we vary the amount of Target50 training data, with one panel per representation. The shaded gap between curves indicates the gain from Pretrain300 pretraining. Pretraining provides the largest gains when target data is scarce, with semantic representations benefiting more than VAE latents.

Scaling with training compute. Figure 5 reports success on the full Target50 split as a function of training FLOPs. We use FLOPs rather than training steps so that representations with different per-step costs line up. For example, DINOv2 uses 324 tokens per frame, whereas the other representations use 256. The VAE performs better at low compute budgets, potentially benefiting from its compact 64-dimensional latents. With additional compute, however, its performance plateaus while semantic-space WAMs continue to improve and reach higher success rates.

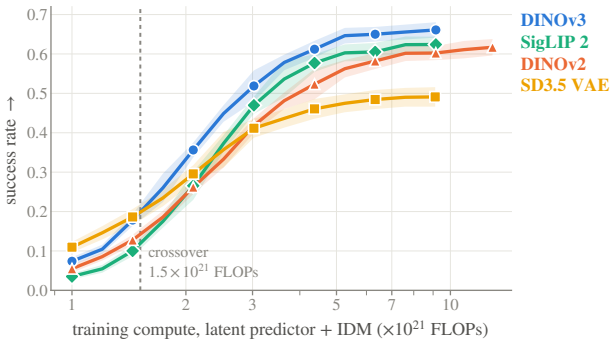

Figure 5: Target50 success rate versus training FLOPs. Shaded bands show the spread across three seeds.

4.4 Generalization: DROID Pretraining, RoboLab Evaluation

We next examine whether the representation ranking extends to training on real robot data. We train models on DROID and then evaluate on RoboLab. This setting complements the RoboCasa365 experiments by placing greater emphasis on generalization to new tasks and environments. Table 3 reports the results.

The representation ranking remains consistent with Table 2. DINOv3 achieves the highest success rate and progress score, while the VAE performs worst, reaching approximately half the progress score of DINOv3. The three semantic encoders are close, between 13.8% and 16.7%. Success rates are significantly lower than on RoboCasa365, which is expected as the evaluation is out-of-domain and much more challenging. Nevertheless, these results suggest that the advantage of semantic representations extends beyond training on simulated demonstrations to generalization from real-robot data.

Table 3 also includes GR00T and FLUX 3 Action (Black Forest Labs, 2026) as reference baselines. Our best-performing WAM outperforms GR00T, consistent with the RoboCasa evaluation, while FLUX 3 Action achieves substantially higher performance. However, differences in model scale and pretraining limit direct comparison with FLUX 3 Action. Our WAMs contain 1.3B parameters and are only trained on DROID using frozen pretrained visual encoders, without initializing the predictor from a pretrained generative model.

Table 3: Success rate (%) and progress score on RoboLab-120 after DROID pretraining, with 10 episodes per task. The four WAM rows are ours and differ only in the representation space; bold marks the best of them. Parameter counts exclude frozen encoders.

Model	Parameters	Pretraining before robot data	Success (%) ↑	Score ↑
WAM, DINOv3	1.3B	none	**16.7**	**28**
WAM, SigLIP 2	1.3B	none	14.2	25
WAM, DINOv2	1.3B	none	13.8	24
WAM, Stable Diffusion 3.5 VAE	1.3B	none	9.2	14
GR00T N1.6	3B	VLA pretraining	7.2	17.1
FLUX 3 Action	7B	FLUX 3 generative model	42.9	–

4.5 Does Latent Prediction Error Predict Control?

As the latent predictor is trained in a different space for each representation, the latent prediction error L_1 has a different scale and a different dimensionality in each case, and thus its values cannot be compared across representations. We therefore examine whether the latent error still correlates with closed-loop task success within each representation.

Figure 6 plots success rate against validation latent L_1 error across checkpoints for each of the four latent predictors. Within each representation, lower prediction error generally corresponds to higher success, with logistic fits illustrating this trend. However, these relationships do not provide a common ranking across representations: DINOv2 and SigLIP 2 exhibit more than three times the latent error of the VAE yet achieve higher success rates. We therefore only use latent L_1 only to select hyperparameters within a representation, and rely on closed-loop success for comparisons across representations.

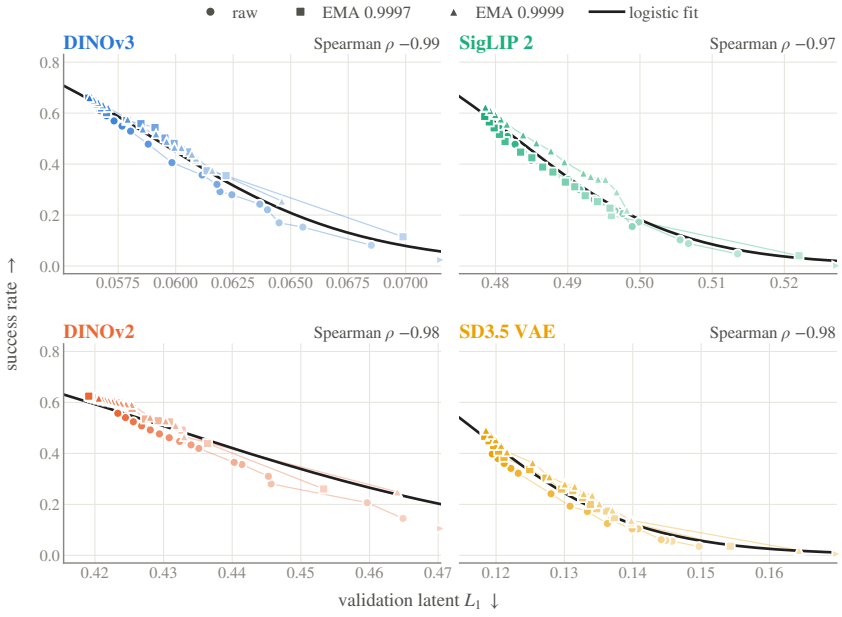

Figure 6: Success rate against validation latent L_1, one panel per representation. Each point is a checkpoint evaluated with the raw weights or with an exponential moving average (EMA) of the weights, and darker points are later checkpoints. The black curve is a logistic fit. The horizontal scales differ because latent L_1 cannot be compared across representations.

4.6 Where Do Semantic Representations Help?

Table 2 shows a clear gap between VAE and semantic encoders. To understand this gap, we examine task characteristics. We find no clear pattern based on task horizon, but observe a stronger distinction between articulated-part and free-object manipulation. Articulated parts are attached to fixtures and move along constrained paths, such as a refrigerator door rotating about a hinge or a dishwasher rack sliding along a rail (10 atomic tasks). Free objects are not attached to fixtures and must be grasped and transported, as when removing toast from a toaster (7 atomic tasks). While some articulated-part tasks permit a range of successful contact locations (Appendix Figure 9), free-object manipulation requires coordinated motion of the gripper and the grasped object.

With the full WAM, the VAE trails the semantic representations by 12.5% on articulated-part tasks and 38.1% on free-object tasks (Table 4). The VAE latents might be too poor for the IDM to recover the right actions, so we give each IDM the ground-truth future latents of recorded demonstrations. The gap then nearly vanishes, to about 2%. The VAE therefore falls behind because its latents are hard to predict: from a VAE context, the model struggles to anticipate how the arm and an object will interact, particularly for tasks requiring coordinated robot-object motion (Figure 7).

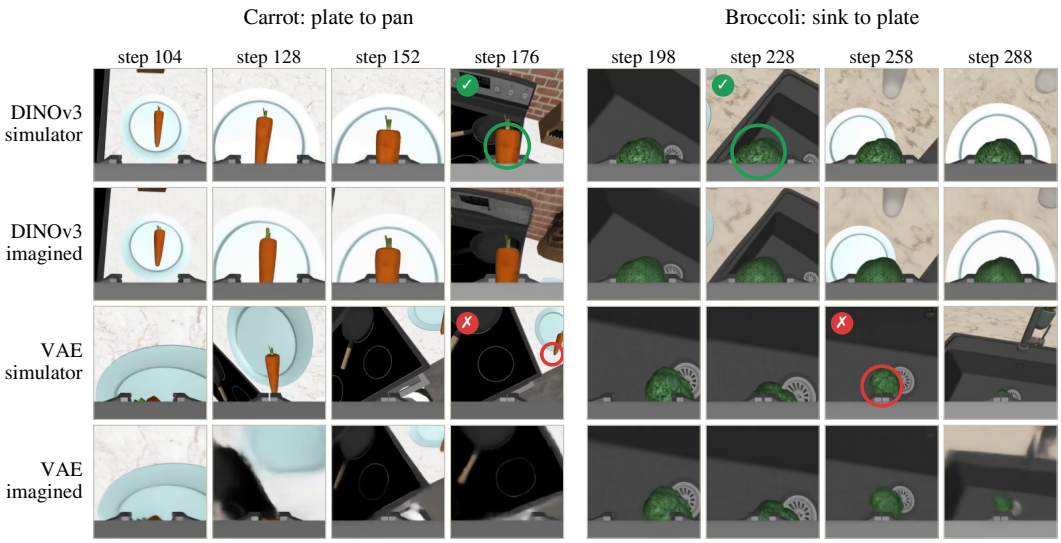

Figure 7: Two free-object tasks seen from the wrist camera. In each task, DINOv3 and the VAE start from the same state; rows show the simulator and the decoded prediction. ✅ DINOv3 holds the object. ❌ The grasp of the VAE fails: the carrot slips out of the gripper, and the gripper closes before reaching the broccoli.

Table 4: Gap in success rate between the VAE and the semantic encoders: the success rate of the VAE model minus the mean success rate of the DINOv3, DINOv2 and SigLIP 2 models. *IDM on ground-truth latents*: each IDM is given the ground-truth future latents of a recorded demonstration, encoded by its own encoder. The two columns use different protocols, so only gaps within a column are comparable.

	Δ = VAE $-$ semantic	
Tasks	Δ full WAM	Δ IDM on ground-truth latents
All tasks	-14.3%	-2.1%
Articulated part	-12.5%	-1.9%
Free object	$\mathbf{-38.1\%}$	-2.3%

5 Conclusion

Most current world action models operate in VAE latent spaces inherited from video generators adapted for control. We investigate whether semantic feature spaces provide a better alternative. Within a common cascaded framework, we compare VAE, DINOv2, DINOv3, and SigLIP 2 representations using flow matching in each encoder's native space, connecting future prediction and action decoding to closed-loop performance. Across RoboCasa365 and DROID-to-RoboLab transfer, all three semantic representations achieve higher task success. They also benefit more from pretraining on diverse tasks and training data. These findings encourage the use of semantic representations for more effective and scalable world action models.

Limitations

World action models are slow at inference because they also denoise the future state. We use only ten denoising steps, and distillation could reduce this cost further. Semantic features are also 16 times larger per token than VAE latents (1024 against 64 dimensions), which increases the memory needed to store and load them.

AI USE STATEMENT

Generative AI tools were used to assist with drafting, language editing, and LaTeX formatting. All AI-assisted content was reviewed by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The final paper will document the model formulation, representation spaces, training protocol, evaluation procedure, implementation details, and hyperparameters needed to reproduce the main results. Additional ablations will be provided in the appendix and supplementary material.

REFERENCES

Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sieniawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiaxiang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang, Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical AI, 2026. URL https://arxiv.org/abs/2606.02800.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael 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, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-

language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Black Forest Labs. FLUX 3 Action: A 7b world action model for robot control. https://bfl.ai/models/flux-3-action, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale, 2023b. URL https://arxiv.org/abs/2212.06817.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. DreamDojo: A generalist robot world model from large-scale human videos, 2026. URL https://arxiv.org/abs/2602.06949.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations, 2025. URL https://arxiv.org/abs/2412.14803.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle

Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset, 2025. URL https://arxiv.org/abs/2403.12945.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning, 2026. URL https://arxiv.org/abs/2601.16163.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. Egowam: World action models beyond pixels with in-the-wild egocentric human data, 2026. URL https://arxiv.org/abs/2607.08436.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model, 2025. URL https://arxiv.org/abs/2503.00200.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise, 2026. URL https://arxiv.org/abs/2511.13720.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots, 2026. URL https://arxiv.org/abs/2603.04356.

Nilaksh, Saurav Jha, Artem Zholus, and Sarath Chandar. Reconstruction or semantics? what makes a latent space useful for robotic world models. *arXiv preprint arXiv:2605.06388*, 2026.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu,

Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models, 2025. URL https://arxiv.org/abs/2310.08864.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond VLAs, 2025. URL https://arxiv.org/abs/2512.15692.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025a. URL https://arxiv.org/abs/2511.14759.

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, 2025b. URL https://arxiv.org/abs/2504.16054.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

Francois Porcher, Nicolas Carion, Karteek Alahari, and Shizhe Chen. Flow matching in feature space for stochastic world modeling, 2026. URL https://arxiv.org/abs/2606.29059.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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, 2024. URL https://arxiv.org/abs/2412.15109.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, Mike Zheng Shou, Xuanjing Huang, Xipeng Qiu, and

Yu-Gang Jiang. World action models: The next frontier in embodied ai, 2026a. URL https://arxiv.org/abs/2605.12090.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026b.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation, 2023. URL https://arxiv.org/abs/2312.13139.

Jenai Xuning Yang, Rishit Dagli, Alex Zook, Hugo Hadfield, Ankit Goyal, Stan Birchfield, Fabio Ramos, and Jonathan Tremblay. Robolab: A high-fidelity simulation benchmark for analysis of task generalist policies, 2026. URL https://arxiv.org/abs/2604.09860.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. URL https://arxiv.org/abs/2510.11690.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning, 2025. URL https://arxiv.org/abs/2411.04983.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets, 2025. URL https://arxiv.org/abs/2504.02792.

A Model Configurations

Table 5 gives the configuration of the two components of the cascade. Both are trained separately on the Target50 split of RoboCasa365. The configuration is the same for every representation space: only the encoder changes, except for the time shift, which is tuned per representation: $\alpha = 1$ for the VAE predictor and $\alpha = 64$ for the semantic predictors (Appendix B). The parameter counts exclude the encoder and differ by less than 1% across representations.

Table 5: Configuration of the latent predictor and the inverse dynamics model.

	Latent predictor	Inverse dynamics model
Data		
Dataset	RoboCasa365, Target50	RoboCasa365, Target50
Training transitions	14.93M	14.93M
Architecture		
Encoder	frozen	frozen
Width	2048	1024
Depth	8	12
Self-attention heads	8	64
Cross-attention heads (visual / proprio)	8 / 8	8 / –
MLP ratio	4.0	4.0
QK normalization	yes	–
Dropout	–	0.1
Positional encoding	RoPE-3D	factorized
Multi-camera handling	encode, then concatenate	encode, then concatenate
Context conditioning	separate cross-attention	separate cross-attention
Proprioceptive conditioning	separate cross-attention	AdaLN
Text conditioning	separate cross-attention (UMT5-XXL)	none
Horizon H	16	16
Context frames	2 (z_t and $z_t - z_{t-1}$)	1
Parameters (excluding encoder)	1.01B	321M
Flow matching		
Prediction / loss space	x / v	v / v
Time sampling	uniform	uniform
Time shift (train / eval)	64 / 64 (semantic), 1 / 1 (VAE)	1 / 1
Inference steps	10	8
Solver	Euler	Euler
Optimization		
Global batch size	256	512
Gradient accumulation	2	1
Training steps	400K	400K
Learning rate	2×10^{-4}	5×10^{-4}
Minimum learning rate	10^{-6}	10^{-6}
Schedule	warmup–cosine, 10% warmup	warmup–cosine, 10% warmup
Gradient clipping	1.0	1.0
Weight decay	0.1	0
Adam (β_1, β_2)	(0.9, 0.95)	(0.9, 0.95)
EMA decays	0.9997, 0.9999	0.9997, 0.9999
Hardware	4 nodes $\times$ 8 H200, DDP	4 nodes $\times$ 8 H200, DDP

B Flow-Matching Details

FlowWM (Porcher et al., 2026) shows that DiTs require critical modifications to converge in high-dimensional semantic feature spaces. Following this work, every latent predictor uses a 2048-wide projection head, larger than the feature dimension of every representation. We also shift the sampled timestep toward the noisy endpoint. Esser et al. (2024) apply a constant log-SNR offset in the convention where the clean sample is at time zero. Under our convention, in which $\tau = 0$ is noise and $\tau = 1$ is data, the same offset is

$$\tau' = \frac{\tau}{\alpha - (\alpha - 1)\,\tau}, \qquad \alpha > 0.$$

The endpoints are fixed and $\alpha = 1$ is the identity. Values $\alpha > 1$ move probability mass toward $\tau = 0$, so training emphasizes noisier timesteps. All latent predictors follow JiT (Li & He, 2026): the network predicts the clean sample $\hat{x}_1$, which is converted to a velocity $u_\psi = (\hat{x}_1 - x_\tau)/(1 - \tau)$ before applying the loss above. We tune α for each representation, which gives $\alpha = 64$ for the semantic predictors and $\alpha = 1$ for the VAE predictor. The IDM predicts the velocity directly, with $\alpha = 1$.

C Feature Map Visualizations

Figure 8 encodes the observed frames of one episode with each representation, projects the feature map onto its first three principal components, and maps them to RGB. The two DINO encoders produce smooth regions that follow objects and surfaces, and those regions stay stable as the arm moves. SigLIP 2 is much noisier at the patch level. The VAE latents follow edges and texture rather than objects, which is what a representation trained for reconstruction should do. This is the difference the paper tests: whether a target that groups by object is easier to predict, and more useful for control, than one that preserves appearance.

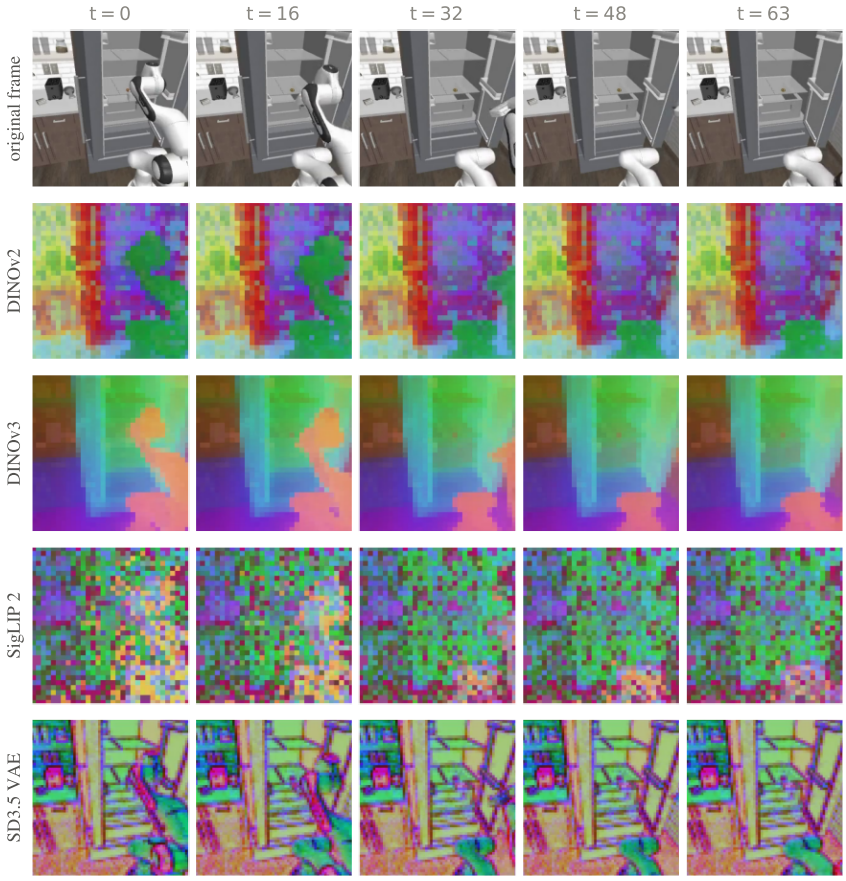

Figure 8: Observed frames of one episode encoded by each representation, with the feature map projected onto its first three principal components and mapped to RGB. Rows are the frame and the four representation spaces; columns are five frames of the episode. These are encodings of ground-truth frames, not predictions.

D ROLLOUT EXAMPLES

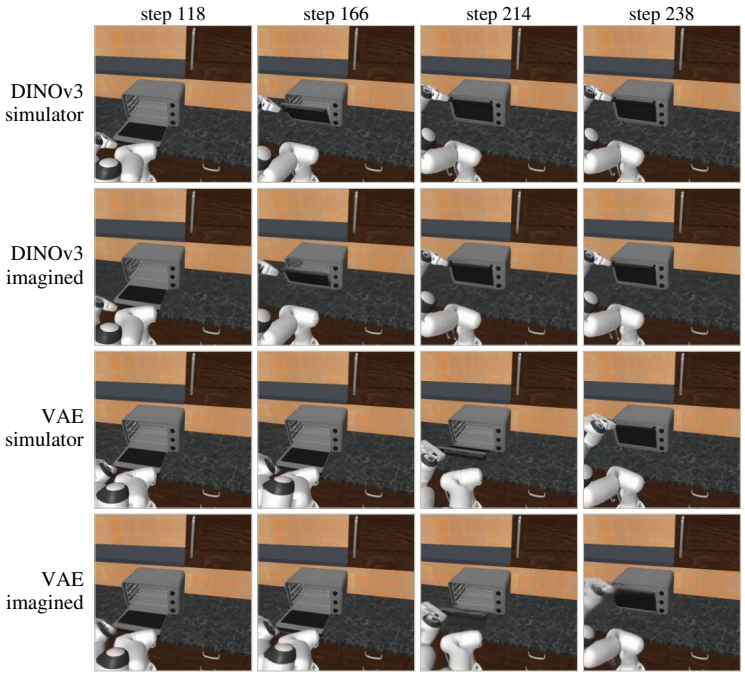

Figure 9: Closing a toaster oven door, an articulated-part task, seen from the right shoulder camera. DINOv3 and the VAE start from the same state; rows show the simulator and the decoded prediction. Both models push the door shut.

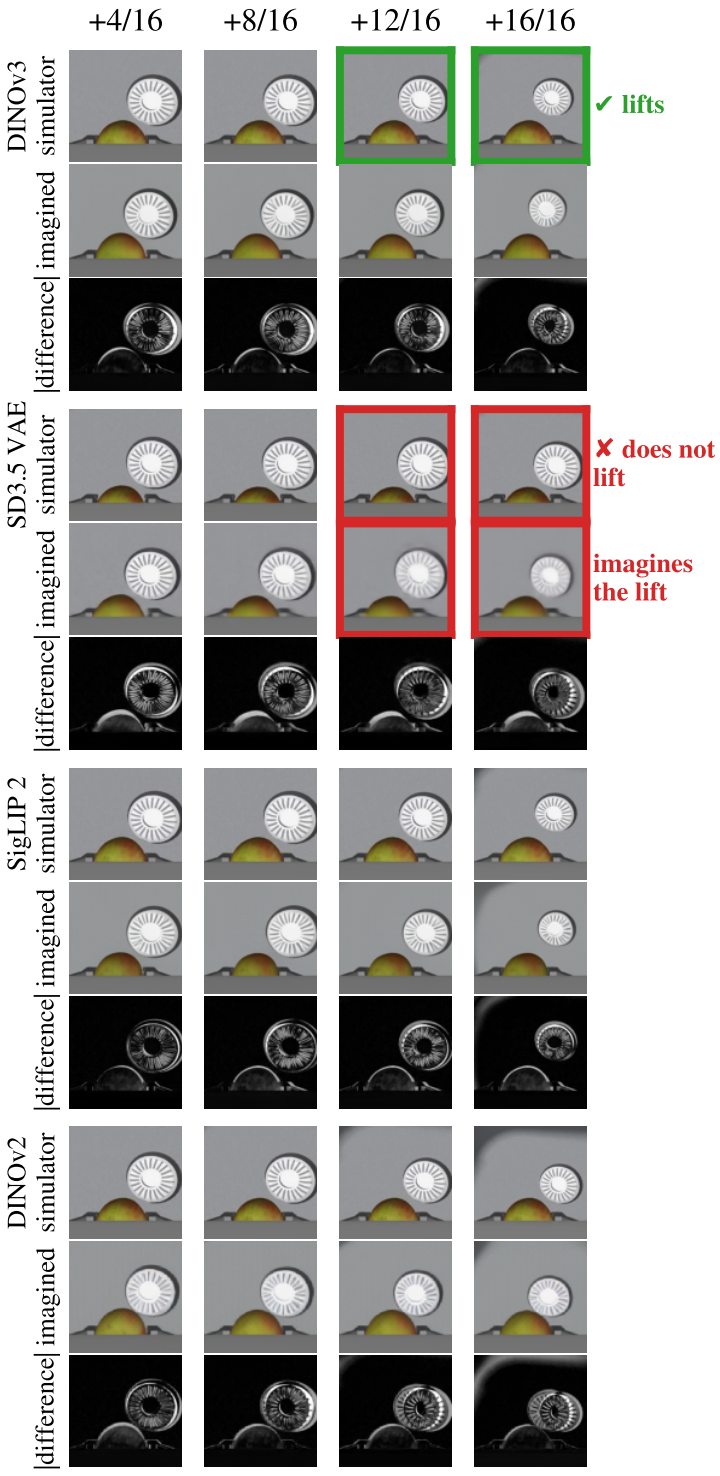

Figure 10: DINOv3 lifts the fruit, while the VAE imagines lifting it but does not lift it in the simulator: the sink drain shrinks only when the wrist camera rises. All four WAMs plan from the same simulator state; rows are the simulator, the decoded prediction, and their absolute difference.

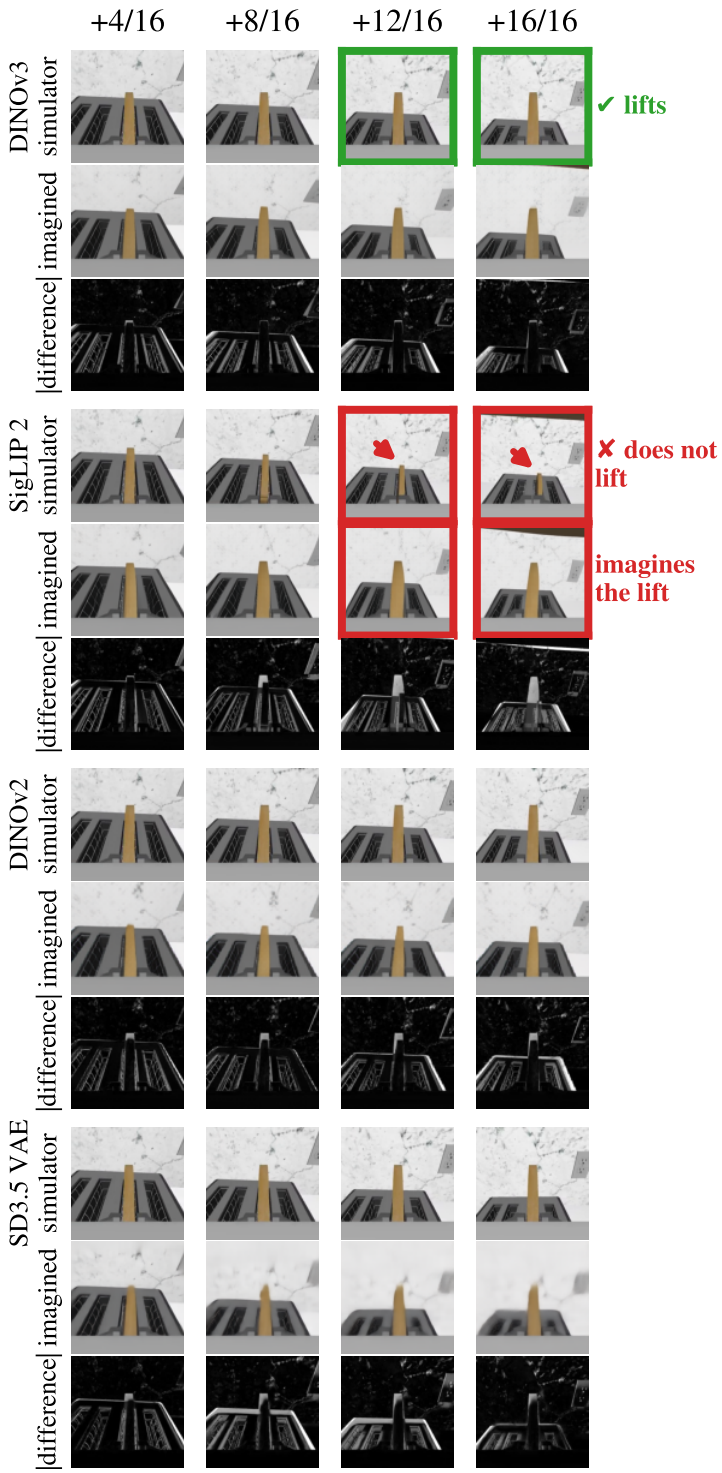

Figure 11: DINOv3 lifts the toast, while SigLIP 2 imagines lifting it but in the simulator the toast stays in the toaster. Rows as in Figure 10.