ISO-WAM: RETHINKING REPRESENTATION ALIGNMENT IN WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World-Action Models (WAMs) adapt pretrained video generators for robot manipulation, but pixel-level future prediction entangles action-induced visual changes with scene appearances unrelated to the action, making control brittle under visual distribution shifts. Aligning policies with future static representation, such as semantic or geometric features, offers an alternative source of supervision, but these representations may still retain irrelevant background information. We propose **Iso-WAM**, which disentangles physical dynamics from visual appearance through dynamic representation alignment. Iso-WAM uses learnable action queries to predict latent action targets provided by a pretrained Unified Latent Action Model (Uni-LAM). Uni-LAM integrates semantic and geometric transitions into a single latent action space for alignment, providing the policy with comprehensive and robust dynamic guidance. Experiments show that Iso-WAM achieves strong in-distribution performance on the standard LIBERO benchmark and robust zero-shot OOD generalization on LIBERO-Plus and real-world robot deployments. In particular, Iso-WAM attains a 75.1% average success rate on LIBERO-Plus, improving over the Fast-WAM-Joint baseline by 7.5%. Code will be released.

1 INTRODUCTION

World-Action Models (WAMs) (Yuan et al., 2026a;b; Ye et al., 2026b; Wang et al., 2026c; Liao et al., 2026) leverage pretrained video generators for robot manipulation, transferring their learned visual and motion priors to action generation. However, their reliance on pixel-level future prediction entangles action-induced visual changes with scene appearance unrelated to the action (Yuan et al., 2026a; Chen et al., 2026). This entanglement can make policies sensitive to changes in lighting, background textures, and other visual conditions, even when the task and required behavior remain unchanged. As a result, visual distribution shifts can disrupt action generation and undermine control robustness (Wang et al., 2026b; Yan et al., 2026).

To address this limitation, recent methods (Yuan et al., 2026a; Yan et al., 2026) attempt to improve robustness by aligning the policy with future depth maps or high-level semantic features. We refer to these approaches as *static representation alignment*: their targets describe the scene at a given time, rather than the transition between observations. Although more structured than pixels, these representations still encode task-irrelevant background information and may fail to eliminate the errors induced by appearance change. In contrast, we propose *dynamic representation alignment*, which supervises policies with latent representations of the transition between current and future observations (*i.e.*, latent actions). By emphasizing how the scene changes rather than its full future appearance, these transition targets can reduce the influence of static visual context and isolate the most useful action cues, providing more focused guidance for robust action generation.

Motivated by this insight, we introduce **Iso-WAM**, a framework that isolates physical dynamics from visual appearance through dynamic representation alignment. Specifically, Iso-WAM prepends learnable query tokens to the action sequence and aligns their output representations with latent action targets extracted from observation pairs. To identify suitable targets, we systematically explore semantic and geometric transition representations. We find that semantic transitions offer greater robustness to lighting, background, and noise variations, while geometric transitions provide better spatial understanding and resilience under camera viewpoint shifts.

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.

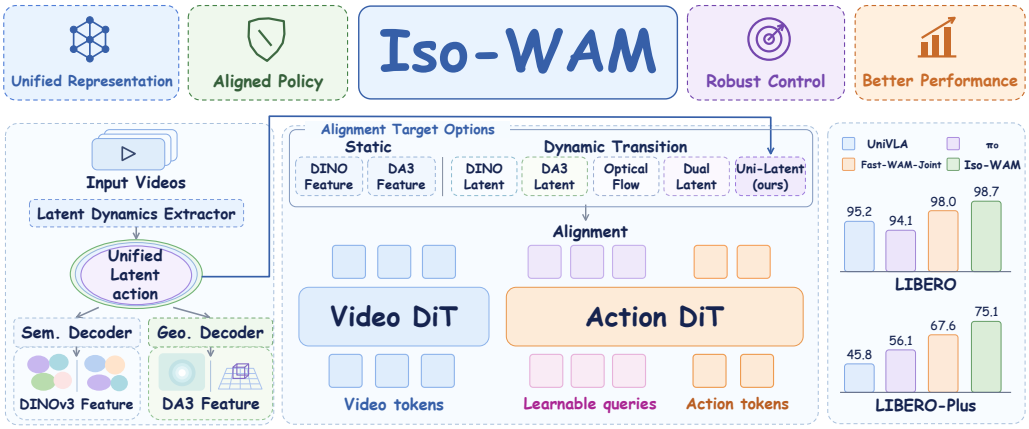


Figure 1: **Overview of Iso-WAM. (Left):** We extract a unified latent action by fusing semantic and geometric transitions. **(Middle):** The Action DiT aligns learnable queries with this unified target, decoupling physical control from task-irrelevant visual generation. **(Right):** Iso-WAM achieves state-of-the-art robustness on both ID (LIBERO) and OOD (LIBERO-Plus) benchmarks.

To integrate these complementary properties into a single dynamic representation, we propose **Uni-LAM**, a unified latent action model designed to combine appearance robustness with spatial awareness. Uni-LAM learns a shared latent action through joint reconstruction of future semantic and geometric features, conditioned on current-frame features. During policy training, Iso-WAM aligns its learnable queries with the shared latent actions provided by the frozen Uni-LAM, giving the action stream a unified supervision signal that captures both semantic and geometric changes.

Extensive experiments demonstrate the effectiveness of Iso-WAM. On the standard LIBERO benchmark (Liu et al., 2023), Iso-WAM achieves an average success rate of 98.7%, attaining state-of-the-art in-distribution (ID) performance among the evaluated methods. On the challenging OOD benchmark LIBERO-Plus (Fei et al., 2025), it achieves the highest average success rate among the evaluated baselines and representation alignment variants at 75.1%, outperforming Fast-WAM-Joint by 7.5 percentage points. Real-world robot experiments further demonstrate strong zero-shot robustness to unseen lighting and background texture variations.

In summary, our main contributions are as follows:

- We propose **Iso-WAM**, a framework that aligns learnable queries in the action stream with dynamic transition representations to improve robot control under visual distribution shifts.
- We systematically investigate the representation design space for WAMs and introduce **Uni-LAM**, which integrates complementary semantic and geometric transitions into a shared latent action for unified policy alignment.
- We demonstrate strong ID performance on LIBERO and improved zero-shot OOD robustness on LIBERO-Plus and real-world tasks under lighting and background texture shifts.

2 Related Works

World Action Models. World Action Models (WAMs) (Yuan et al., 2026a;b; Wang et al., 2026b;a; Ye et al., 2026b; Yang et al., 2026; Zhang et al., 2026) combine pretrained video generation models with action-generation modules, using video prediction as an intermediate reasoning medium to guide robot manipulation. Representative approaches include DreamZero (Ye et al., 2026b) and LingBot-VA (Li et al., 2026a), which jointly model future video and robot actions, while Fast-WAM (Yuan et al., 2026b) omits explicit future-video generation during deployment to improve inference efficiency. However, beyond the differences in their inference paradigms, most of these methods still optimize representations primarily toward pixel-level generation fidelity. This design may cause action generation to be affected by task-irrelevant visual factors, such as illumination

and sensor noise. Iso-WAM explores this question through auxiliary supervision derived from latent representations of observation transitions.

Representation Alignment in Video Models and WAMs. Researchers have increasingly begun to explore the distillation of auxiliary world knowledge (Assran et al., 2025; Mur-Labadia et al., 2026; Oquab et al., 2023; Siméoni et al., 2025; Lin et al., 2025) into video generation models (Xie et al., 2026; Zhang et al., 2025; Wu et al., 2026; Yu et al., 2025) and world action models (Wang et al., 2026a; Ma et al., 2026). In the field of video generation, works such as VideoREPA (Zhang et al., 2025) and Geometry Forcing (Wu et al., 2026) align generative models with features from foundation models that capture semantic, geometric, and other forms of structured information. In the domain of action world models, WAM4D (Li et al., 2026b) and X-WAM (Guo et al., 2026) have begun incorporating geometric information for representation alignment, while ST-WAM (Wang et al., 2026b) seeks to obtain stable latent-state representations through semantic alignment. Different from these approaches, in this paper, we provide a systematic discussion on the impact of various representation alignment methods, and ultimately employ dynamic representation alignment for the World-Action Model (WAM) as a robust control solution.

3 METHOD

Iso-WAM trains a world-action model to predict compact representations of observation transitions alongside its action-generation objective. The key idea is to provide a shared target that captures changes in both semantic content and spatial structure. We first learn this target with Uni-LAM, which encodes an observation pair into one latent code used to reconstruct future semantic and geometric features. We then freeze Uni-LAM and align learnable queries in the policy with its outputs. Figure 2 shows how this supervision enters the action stream. We describe the policy alignment first, followed by the construction of its target.

3.1 PRELIMINARIES

At time t, a WAM receives an observation o_t and a language instruction l, and generates a future video sequence and an action chunk $A_t = a_{t+1:t+H}$ of length H. Its video and action experts operate on visual latents from a pretrained video VAE and action tokens, respectively. We build on the Fast-WAM framework, which supports action-only denoising and joint denoising of actions and future visual latents.

Each expert is trained with flow matching. Let x_1 denote a clean training target and $x_0 \sim \mathcal{N}(0, I)$ a noise sample. At noise time $\tau \in [0, 1]$, the model receives $x_\tau = (1 - \tau)x_0 + \tau x_1$ and predicts the velocity toward the clean target:

$$\ell_{FM} = \mathbb{E}_{\tau, x_0, x_1}\left[\|v_\theta(x_\tau, o_t, l, \tau) - (x_1 - x_0)\|_2^2\right]. \tag{1}$$

Here, x denotes the action or video target for the corresponding expert. We denote the base model's weighted combination of the two flow-matching losses by $\mathcal{L}_{FM}$. Iso-WAM retains this objective and adds supervision for observation transitions through the action stream.

3.2 ISO-WAM POLICY ALIGNMENT

Transition targets. A future-frame feature describes the scene at one endpoint. Our alignment target instead represents the transition between two observations. Given an observation pair separated by Δ steps, the frozen latent action model Uni-LAM (Section 3.3) produces a deterministic target:

$$\bar{z} = f_{\mathrm{LAM}}(o_t, o_{t+\Delta}) = \mu_\phi(o_t, o_{t+\Delta}). \tag{2}$$

Here, μ_ϕ is the posterior mean predicted by Uni-LAM, and $\bar{z} \in \mathbb{R}^{32}$ is the policy-alignment target. Uni-LAM learns this representation through conditional future-feature reconstruction, as described in Section 3.3. We apply the same encoder separately to each adjacent sampled-frame pair from each camera. In the LIBERO configuration, a 32-step action chunk with $\Delta = 4$ yields eight transitions per camera and 16 targets across two cameras.

Queries in the action stream. We prepend a set of learnable queries to the action tokens, with one query assigned to each camera–interval pair. At noise time τ, the queries pass through the Action

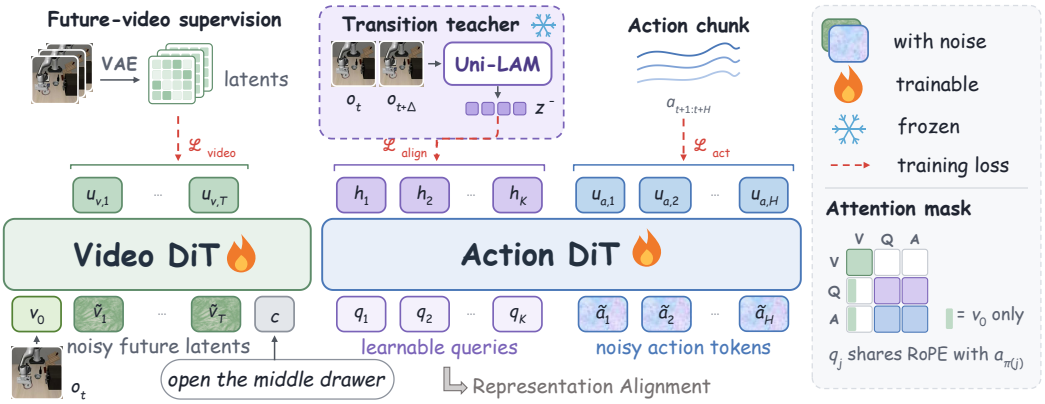


Figure 2: **Iso-WAM policy training.** Video and Action DiTs predict their respective flow velocities. Learnable queries in the action stream are projected and aligned with the frozen Uni-LAM target $\bar{z}$ extracted from observation pairs. Red dashed arrows, flames, and snowflakes denote losses, trainable, and frozen modules, respectively. The attention mask restricts queries to current-frame visuals (v_0) while enabling full bidirectional interaction between queries and action tokens.

DiT together with the noisy action chunk. They directly attend to current-frame visual tokens and interact bidirectionally with action tokens. For each query, a shared linear head projects its final hidden state h^τ into a latent action prediction, giving $\hat{z}^\tau = Wh^\tau + b$. Thus, each query predicts the transition assigned to its camera and interval, while its interaction with action tokens couples this prediction task to action generation.

Alignment supervision. We match query predictions to the frozen targets with a cosine loss:

$$\mathcal{L}_{align} = \mathbb{E}\left[1 - \cos\left(\hat{z}^\tau, \bar{z}\right)\right]. \tag{3}$$

The loss is averaged over valid query–target pairs, training examples, sampled noise, and flow-matching noise times. The targets are computed from recorded observation pairs during training, with no gradients passed to Uni-LAM. At deployment, the queries remain in the policy, but neither Uni-LAM nor recorded future observations are required. The same alignment design supports the action-only and joint video–action inference variants.

3.3 Unified Latent Action Pre-training

As we will detail in our subsequent ablation studies, semantic and geometric latent action spaces exhibit distinct strengths and weaknesses in handling out-of-distribution environments. Motivated by the desire to synergize their complementary advantages, we introduce the Unified Latent Action Model (Uni-LAM). Uni-LAM constructs a single, cohesive transition target from two complementary feature spaces: DINOv3 supplies semantic features, and Depth Anything 3 supplies geometric features. We integrate them through a shared latent bottleneck. By constraining both future-feature decoders to reconstruct from the exact same latent code, the synergistic fusion of representations is explicitly learned before the target is ever passed to the downstream policy. Figure 3 summarizes this teacher pre-training stage.

Encoding an observation pair. For each endpoint $s \in \{t, t+\Delta\}$, frozen encoders produce semantic and geometric patch features $F_s^m = \mathcal{E}_m(o_s)$, where $m \in \{\text{sem}, \text{geo}\}$. We use DINOv3 (Siméoni et al., 2025) and DA3 (Lin et al., 2025), align their spatial grids, and project their features to a common width. A QFormer-style encoder processes the features from both frames with a single learnable aggregation query. Its attention mask restricts direct interactions between semantic and geometric context tokens, while allowing the query to read both modalities across time. This teacher query aggregates an observed transition; it is distinct from the policy queries that predict transition targets during action denoising.

A shared variational code. The encoder output parameterizes a diagonal Gaussian posterior $q_\phi(z \mid o_t, o_{t+\Delta})$ with mean μ_ϕ and standard deviation σ_ϕ. During Uni-LAM pretraining, we sample a 32-dimensional code z from this posterior using the reparameterization trick and use it to condition

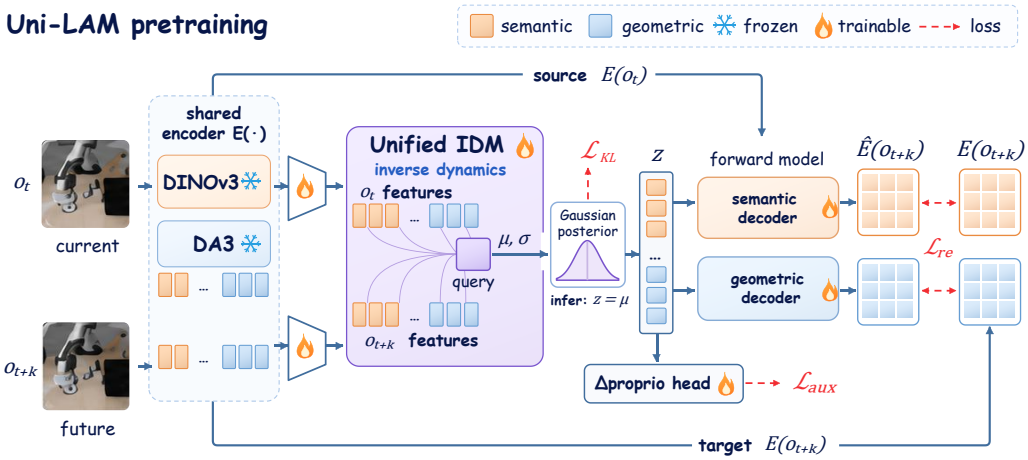


Figure 3: Uni-LAM pretraining. Frozen DINOv3 and DA3 encoders provide semantic and geometric features, which a query-based unified IDM encodes into a variational transition code z. Conditioned on z and current-frame features, modality-specific decoders reconstruct future teacher features, while an auxiliary head predicts proprioceptive change. Training combines reconstruction, KL, and auxiliary losses. The posterior mean $\bar{z} = \mu$ serves as the frozen policy-alignment target. Here, $k = \Delta$ in the text.

both future-feature decoders. After pretraining, we freeze Uni-LAM and use its posterior mean as the policy-alignment target defined in Eq. (2). Thus, z denotes a sampled code for teacher training, whereas $\bar{z}$ denotes the deterministic target for policy alignment.

Learning transitions through conditional reconstruction. Two modality-specific decoders predict future features from current features and the shared code: $\widehat{F}^m_{t+\Delta} = D_m(F^m_t, z)$. The current-frame input provides scene context directly to each decoder, while the bottleneck is encouraged to encode the additional information needed to predict the change. Both decoders receive the same z, requiring it to support semantic and geometric prediction jointly. We implement the decoders as separate AdaLN Transformers, with z modulating their hidden states. An auxiliary head receives z and the current proprioceptive state to predict the end-effector state change, adding explicit motion supervision to the feature-reconstruction task.

3.4 Training Objective

The training process of our framework is strictly decoupled into two stages: (1) pre-training the Unified Latent Action Model (Uni-LAM) and (2) optimizing the Iso-WAM policy for joint generation and synergistic representation alignment.

Stage 1: Uni-LAM Objective. The unified LAM is designed to capture both semantic and geometric dynamics within a single, cohesive latent space. It is optimized via a joint feature reconstruction loss (reconstructing future representations), a KL-divergence loss $\mathcal{L}_{KL}$ to regularize the posterior q_ϕ toward a standard normal prior, and an auxiliary proprioceptive state-change prediction loss $\mathcal{L}_{aux}$ to encourage motion-related information in z

$$\mathcal{L}_{LAM} = \mathcal{L}^{sem}_{recon} + \mathcal{L}^{geo}_{recon} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{aux}\mathcal{L}_{aux} \tag{4}$$

Stage 2: Iso-WAM Policy Objective. Once trained, Uni-LAM is frozen, and its posterior mean provides the deterministic alignment target $\bar{z}$ defined in Eq. equation 2. The Iso-WAM policy is then optimized using a joint objective. It is supervised by the standard Flow-Matching loss $\mathcal{L}_{FM}$ for joint action and video generation. Simultaneously, the alignment loss $\mathcal{L}_{align}$ explicitly enforces dynamic transition consistency. The final objective is:

$$\mathcal{L}_{policy} = \mathcal{L}_{FM} + \lambda_{align}\mathcal{L}_{align} \tag{5}$$

This design provides the policy with a single alignment target that integrates semantic and geometric transitions, avoiding the need for separate modality-specific alignment objectives.

Table 1: Success rates (%) on LIBERO with 50 rollouts per task. The best overall result is bold.

Method	Params	Spatial	Object	Goal	Long	Overall
OpenVLA	7B	84.7	88.4	79.2	53.7	76.5
π_0	3.3B	96.8	98.8	95.8	85.2	94.1
$\pi_{0.5}$	3.3B	98.8	98.2	98.0	92.4	96.9
π_0-Fast	3.3B	96.4	96.8	88.6	60.2	85.5
UniVLA	7B	96.5	96.8	95.6	92.0	95.2
WorldVLA	7B	87.6	96.2	83.4	60.0	81.8
Motus	8B	96.8	99.8	96.6	97.6	97.7
Fast-WAM	6B	98.2	100.0	97.0	95.2	97.6
Iso-WAM-uncond	6B	98.6	99.8	97.6	94.6	97.7
Fast-WAM-Joint	6B	99.4	98.2	98.8	95.6	98.0
Iso-WAM	6B	98.8	99.6	97.8	98.6	**98.7**

4 EXPERIMENTS

We organize our experiments to address three core questions: (1) *Does Iso-WAM significantly improve out-of-distribution (OOD) generalization while preserving competitive in-distribution (ID) performance?* (2) *Do these robustness benefits effectively transfer to real-world physical manipulation?* (3) *Which design choices are critical for optimal representation alignment?* Accordingly, we first detail the experimental setup, evaluate the method on simulation benchmarks, systematically ablate our core designs, and finally validate the policy in unconstrained real-world scenarios.

4.1 EXPERIMENTAL SETUP

4.1.1 BENCHMARKS.

We evaluate Iso-WAM on standard LIBERO, LIBERO-Plus, and real-world manipulation tasks. LIBERO (Liu et al., 2023) measures in-distribution performance across four suites with 40 tasks, using 50 rollouts per task. LIBERO-Plus (Fei et al., 2025) measures zero-shot generalization across seven perturbation categories, using 10,030 episodes per seed and averaging results over two seeds, without training on LIBERO-Plus data. For real-world evaluation, we compare Iso-WAM with Fast-WAM-Joint on two UR5 pick-and-place tasks. Each task uses 80 demonstrations collected under standard conditions and 20 evaluation rollouts per condition, covering normal conditions, lighting shifts, and texture shifts. We additionally evaluate clean-to-random generalization on RoboTwin. Detailed benchmark descriptions and RoboTwin results are provided in the appendix.

4.1.2 IMPLEMENTATION DETAILS.

We evaluate Iso-WAM under two distinct inference paradigms. The first paradigm, akin to Fast-WAM (Yuan et al., 2026b), conditions solely on the initial video frame and directly denoises the action sequence. The second paradigm, following Fast-WAM-Joint, jointly denoises both future visual latents and robot actions. All model variants are built upon the pre-trained Wan2.2-5B (Wan et al., 2025) VideoDiT foundation model. For the LIBERO benchmarks (Liu et al., 2023), the models are trained across 8 NVIDIA H100 GPUs with an effective batch size of 64. We strictly follow the standard evaluation protocols for each benchmark and report the task success rates. Furthermore, all representation alignment variants also adhere to the Fast-WAM-Joint paradigm, are trained using the aforementioned settings, and are evaluated on both LIBERO and LIBERO-Plus.

4.2 GENERALIZATION IN SIMULATION

Standard LIBERO. We first evaluate the in-distribution (ID) performance on the standard LIBERO task suites. As shown in Table 1, Iso-WAM-uncond maintains highly competitive performance across all suites, while the full Iso-WAM further pushes the overall success rate of Fast-WAM-Joint from 98.0% to 98.7%, delivering a notable +3.0% improvement in the most challenging LIBERO-

Long suite (98.6% vs. 95.6%). Furthermore, Iso-WAM exhibits remarkable absolute strength, establishing a new state-of-the-art (SOTA) overall success rate among all evaluated VLAs and WAMs. Strikingly, despite not utilizing any massive embodied pretraining data, our model confidently outperforms heavily pretrained powerhouses such as π_0 (Black et al., 2024), $\pi_{0.5}$ (Physical Intelligence et al., 2025), and Motus (Bi et al., 2026). These results compellingly prove that explicitly isolating and aligning with dynamic latent actions does not disrupt the rich dynamics priors originally present in the pretrained video model; rather, it provides a more precise and stable control guidance for complex, long-horizon manipulation tasks.

Table 2: LIBERO-Plus success rates (%). Bold and underlined values denote the best and second-best results, respectively. Iso-WAM demonstrates significantly greater robustness against camera viewpoint shifts, background variations, and sensor noise.

Method	Cam.	Rob.	Lang.	Light	Back.	Noise	Lay.	Avg.
UniVLA	1.8	46.2	69.6	69.0	81.0	21.2	31.9	45.8
OpenVLA-OFT	<u>56.4</u>	31.9	79.5	88.7	**93.3**	75.8	74.2	<u>71.4</u>
π_0	13.8	6.0	58.8	85.0	<u>81.4</u>	**79.0**	68.9	56.1
π_0-FAST	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	62.5
WorldVLA	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.6
Fast-WAM	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.4
Iso-WAM-uncond	42.1 (+25.7)	52.2 (+7.7)	76.2 (+7.3)	92.4 (+14.2)	62.7 (+9.0)	50.4 (+12.7)	68.9 (+8.2)	63.6 (+12.2)
Fast-WAM-Joint	36.3	<u>60.6</u>	<u>91.0</u>	<u>92.9</u>	49.9	62.2	**80.3**	67.6
Iso-WAM	47.5 (+11.2)	**71.2** (+10.6)	**93.5** (+2.5)	**96.3** (+3.4)	62.9 (+13.0)	<u>76.5</u> (+14.3)	<u>77.8</u> (-2.5)	**75.1** (+7.5)

LIBERO-Plus. We next investigate the out-of-distribution (OOD) generalization capabilities on the LIBERO-Plus benchmark without any OOD fine-tuning. As presented in Table 2, both variants of our method exhibit remarkable robustness. Iso-WAM-uncond outperforms the standard Fast-WAM by a significant average margin of 12.2%. Furthermore, the full Iso-WAM achieves the highest average success rate of 75.1%, surpassing its direct baseline Fast-WAM-Joint and the strongest external baseline OpenVLA-OFT. Crucially, our method demonstrates exceptional resilience to severe visual shifts that typically confound generative models. For instance, under the Background and Noise variations, Iso-WAM improves upon its baseline by 13.0% and 14.3%, respectively. While baseline WAMs easily confuse static appearance shifts with physical dynamics, our dynamic bottleneck effectively filters out these visual distractors. This validates our core hypothesis: isolating actionable transitions (Δ) from visual entanglement significantly boosts zero-shot OOD robustness.

4.3 Real-World Deployment

Table 3: Success rates of real-world robotic manipulation under the Lighting Shift and Texture Shift settings. Iso-WAM significantly outperforms Fast-WAM-Joint in OOD environments.

(a) Lighting Shift	**Lighting Conditions**			**Avg.**
	Normal (ID)	Blue (OOD)	Purple (OOD)	
Fast-WAM-Joint	85.0	45.0	32.5	54.2
Iso-WAM (Ours)	**92.5**	**82.5**	**70.0**	**81.7**
Improvement (pp)	*+7.5*	*+37.5*	*+37.5*	*+27.5*
(b) Texture Shift	**Background Textures**			**Avg.**
	Normal (ID)	Cloth A (OOD)	Cloth B (OOD)	
Fast-WAM-Joint	85.0	32.5	37.5	51.7
Iso-WAM (Ours)	**92.5**	**75.0**	**77.5**	**81.7**
Improvement (pp)	*+7.5*	*+42.5*	*+40.0*	*+30.0*

Lighting Shifts. As shown in Table 3(a), lighting shifts substantially degrade the performance of Fast-WAM-Joint. Its success rate drops from 85.0% under normal lighting to 45.0% and 32.5% under blue and purple lighting, respectively. Iso-WAM achieves 92.5% under normal lighting and

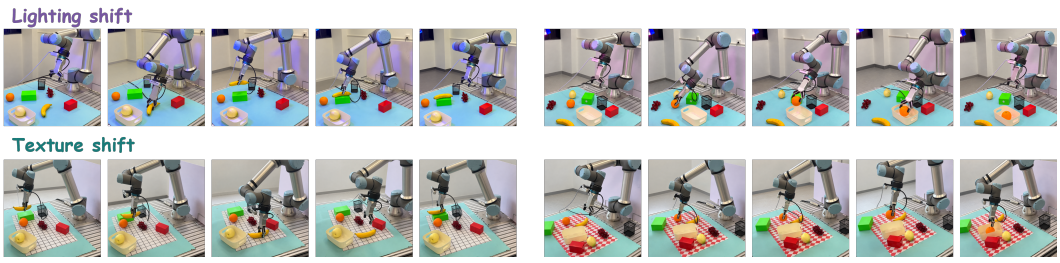


Figure 4: Real-world rollouts under lighting shifts and texture shifts.

retains success rates of 82.5% and 70.0% under the two OOD conditions, outperforming Fast-WAM-Joint by 37.5 percentage points in each case. Averaged across all three lighting conditions, including Normal (ID), Iso-WAM achieves an 81.7% success rate, compared with 54.2% for Fast-WAM-Joint, yielding a 27.5-percentage-point improvement. These results support the effectiveness of dynamic representation alignment for robust control under lighting variations.

Texture Shifts. We observe a similar trend under texture shifts (Table 3(b)). When the plain tabletop is replaced with unseen patterned cloths, the success rate of Fast-WAM-Joint drops from 85.0% under normal conditions to 32.5% on Cloth A and 37.5% on Cloth B. Iso-WAM achieves 75.0% and 77.5% on the two OOD textures, improving over Fast-WAM-Joint by 42.5 and 40.0 percentage points, respectively. Averaged across all three conditions, including Normal (ID), Iso-WAM achieves an 81.7% success rate, compared with 51.7% for Fast-WAM-Joint, yielding a 30.0-percentage-point improvement. Together, these real-world results demonstrate improved robustness to unseen lighting and background texture variations in the evaluated manipulation tasks.

4.4 Ablation Study

Table 4: Ablation study on different alignment targets. "ID" denotes the in-distribution LIBERO benchmark (average success rate), while "OOD" details the out-of-distribution robustness across various perturbations in LIBERO-Plus.

Alignment Target	**ID**	**OOD (LIBERO-Plus)**							
	LIBERO	Cam.	Rob.	Lang.	Light	Back.	Noise	Lay.	**Avg.**
Baseline (No Alignment)	98.0	36.3	60.6	91.0	92.9	49.9	62.2	80.3	67.6
Static Feature Alignment									
DINO Feature	98.4	43.5	69.8	93.6	94.0	62.4	67.2	82.6	73.3
DA3 Feature	98.4	51.4	70.1	92.8	95.1	58.6	54.3	78.2	71.5
Dynamic Latent Action Alignment									
DINO Latent Action	**98.7**	41.5	71.7	94.8	95.3	65.1	72.5	79.4	74.3
DA3 Latent Action	98.5	49.2	72.3	94.5	94.2	61.8	58.7	80.5	73.0
Optical Flow	98.5	46.5	72.0	94.0	93.5	61.0	63.8	80.2	73.0
Dual Latent Action	98.0	45.8	60.6	93.5	92.5	53.6	67.6	76.4	70.0
Uni-Latent Action (Ours)	**98.7**	47.5	71.2	93.5	96.3	62.9	76.5	77.8	**75.1**

To systematically deconstruct the representation design space of WAMs and isolate the contributions of our design choices, we conduct an ablation study on the alignment targets across the ID and OOD benchmarks. As shown in Table 4, we analyze the empirical results from three critical dimensions, revealing the core insights behind our method:

Dynamic Latent Actions vs. Static Future Features. We first validate our core hypothesis. As shown in Table 4, replacing static future features with dynamic latent actions improves the average OOD success rate from 73.3% to 74.3% for DINO and from 71.5% to 73.0% for DA3, while maintaining or improving ID performance. These results support the advantage of modeling dynamic transitions over predicting static future features in the evaluated settings. High-dimensional static representations are intrinsically harder to predict and force models to entangle control with task-irrelevant visual rendering. In contrast, dynamic latent actions reside in a low-dimensional

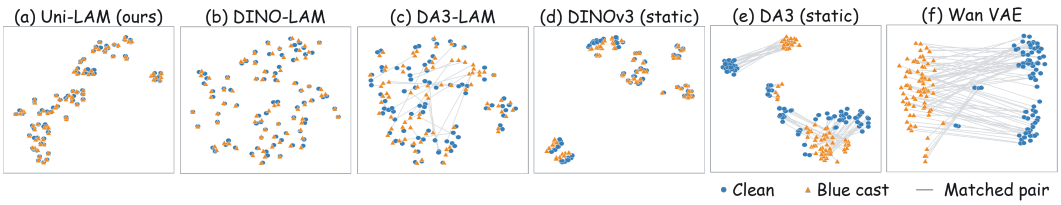


Figure 5: Representation stability under a blue color cast. Uni-LAM achieves the highest representation stability.

manifold, making them much easier to predict. They serve as a physics-informed information bottleneck, discarding static appearance biases and compelling the policy to focus entirely on actionable core kinematics.

Comparison of Different Representations. Next, we investigate the properties of semantic, geometric, and optical flow representations. We find that, as shown in the third row of Table 4, the DINO latent action model explicitly captures high-level, object-centric correspondences that are inherently invariant to superficial appearance shifts (e.g., lighting or backgrounds). Conversely, dense geometric features like DA3 encode crucial 3D spatial relationships, making them more robust to camera viewpoint shifts, achieving an excellent result of 49.2% under Camera Viewpoints variations. However, their reliance on low-level visual cues (e.g., edges and shadows) makes pure geometric alignment fragile under extreme domain perturbations. Furthermore, while optical flow explicitly encodes 2D motion, it essentially remains a pixel-centric feature highly susceptible to texture and illumination disturbances. Its lack of high-level physical abstraction therefore yields only minor OOD improvements on LIBERO-Plus.

Dual vs. Unified Transition Targets. Although semantic and geometric transitions offer complementary benefits, directly aligning the policy with both targets does not necessarily improve performance. As shown in Table 4, dual latent action alignment achieves an OOD success rate of 70.0%, below both semantic-only (74.3%) and geometry-only (73.0%) alignment. One possible explanation is that DINO (Siméoni et al., 2025) and DA3 (Lin et al., 2025) operate in distinct latent spaces, and requiring a single downstream policy to simultaneously minimize two different transition objectives may increase optimization difficulty and instability, hindering the integration of complementary information. Uni-LAM instead combines semantic and geometric transition information into a shared latent code before policy alignment, reducing the need for the policy to reconcile separate supervision signals. This design achieves an OOD success rate of 75.1%, outperforming both separate-target supervision and either single-modality variant.

Representation Analysis. We examine representation stability under a controlled blue color cast. As shown in Figure 5, Uni-LAM preserves close correspondence between clean and perturbed observations. In the original feature space, it achieves a mean paired cosine similarity of 0.993 and normalized drift of 0.115, compared with 0.982/0.154 for DINO-LAM and 0.814/0.508 for DA3-LAM. These results support the stability of the unified latent representation under the tested appearance perturbation. More detailed representation analyses are provided in the supplementary material.

5 Conclusion

In this paper, we presented **Iso-WAM**, a framework that improves the robustness of World-Action Models through dynamic representation alignment. Our study of the representation design space supports the effectiveness of latent transition targets for OOD generalization and highlights the complementary benefits of semantic and geometric information. To integrate these benefits, we introduced **Uni-LAM**, which learns a shared latent action from semantic and geometric transitions and provides a unified alignment target for learnable queries in the policy's action stream. Experiments on LIBERO and LIBERO-Plus demonstrate strong in-distribution performance and improved zero-shot OOD robustness, while real-world evaluations further show robustness to unseen lighting and background texture variations. These results support dynamic representation alignment as a promising approach to robust robotic manipulation under visual distribution shifts.

AI USE STATEMENT

We used a large language model to assist with translation, drafting and revising prose, condensing text, and LaTeX formatting. The authors reviewed the AI-assisted content for accuracy and consistency with the underlying experiments and sources and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the Iso-WAM framework, Uni-LAM pre-training, and training objectives in Section 3. Appendices B.1 and B.2 document the model architectures, training data, preprocessing, optimization hyperparameters, and computational settings for Uni-LAM and Iso-WAM, respectively. Benchmark descriptions and evaluation protocols, including zero-shot out-of-distribution evaluation, are provided in Appendix C, while baseline methods and representation alignment variants are described in Appendix D. Appendix G further specifies the real-world robot and camera setup, demonstration collection, evaluation rollouts, and visual perturbations. Code will be released to support reproduction of our results.

REFERENCES

Mahmoud 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. *arXiv preprint arXiv:2506.09985*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

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. *arXiv preprint arXiv:2410.24164*, 2024. URL https://arxiv.org/abs/2410.24164.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, et al. Flowwam: Optical flow as a unified action representation for world action models. *arXiv preprint arXiv:2607.13017*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, et al. Wam4d: Fast 4d world action model via spatial register tokens. *arXiv preprint arXiv:2606.14048*, 2026b.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 88446–88463, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*, 2026.

Fulong Ma, Daojie Peng, Wenjun Yue, Jiahang Cao, Bintao Wang, Qiang Zhang, and Jun Ma. Geosem-wam: Geometry-and semantic-aware world action models. *arXiv preprint arXiv:2606.03188*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

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.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.09747.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junke Wang, Qihang Zhang, Shuai Yang, Yiming Luo, Yujun Shen, Zuxuan Wu, Yu-Gang Jiang, and Yinghao Xu. Repwam: World action modeling with representation visual-action tokenizers. *arXiv preprint arXiv:2606.13674*, 2026a.

Mingxin Wang, Bin Hu, Bin Qian, Kaitao Jiang, Haoning Wu, Feng Yan, Bowen Jing, Ruiyang Hao, Enyi Wang, Kangning Niu, et al. St-wam: Semantic-temporal world action model for robust manipulation under visual distribution shifts. *arXiv preprint arXiv:2607.28993*, 2026b.

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*, 2026c.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Zhihao Xie, Junfeng Wu, Xinting Hu, Junchao Huang, and Li Jiang. Videorae: Taming video foundation models for generative modeling via representation autoencoders. *arXiv preprint arXiv:2607.14088*, 2026.

Haodong Yan, Junfeng Li, Junjie He, Zhide Zhong, MingMing Yu, Wenxuan Song, Jiaguan Zhu, Yangyang Zheng, Yuqiao Du, Jiadi You, et al. Robust-wam: Bridging generative pretraining and semantic foresight in world-action models. *arXiv preprint arXiv:2608.05903*, 2026.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems, 2026a. URL https://arxiv.org/abs/2604.11757.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. Dreamwam: Beyond rgb future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 2026a.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-robotworld technical report: Unifying embodied world modeling through language-conditioned video generation. *arXiv preprint arXiv:2606.17030*, 2026.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025.

A LIMITATIONS

Our study has two limitations. First, directly aligning the policy with separate semantic and geometric targets underperforms single-modality alignment variants in our experiments. We provide only a preliminary analysis of this observation and have not investigated it in depth. Second, our real-world evaluation covers only two pick-and-place tasks on a single UR5 platform, primarily under lighting and background texture shifts. Broader real-world evaluation is needed in future work to establish the scope of the observed benefits.

B ARCHITECTURE AND IMPLEMENTATION DETAILS

B.1 UNI-LAM ARCHITECTURE AND TRAINING DETAILS

Data and input processing. We train Uni-LAM on a mixture of RM2.0-Franka, RM1.0-Franka, RM2.0-UR, and Agibot-World.(All other latent action models compared in this paper are trained on the same datasets for a fair comparison. These datasets are used exclusively to train the latent action models; the World-Action Model itself undergoes no additional embodied pretraining.) Each sample contains two endpoint observations $(o_t, o_{t+\Delta})$ from a single non-wrist camera, typically the primary third-person view. The endpoints are separated by 1.6 seconds, giving a frame-index offset of $\mathrm{round}(1.6 \times \mathrm{FPS})$. RGB frames are center-cropped if their aspect ratio exceeds $4{:}3$ and bilinearly resized to 224×224. During training, we independently augment the same two-frame clip twice: once for inverse-dynamics encoding and once for decoder conditioning and reconstruction targets. Both branches use random resized cropping to 224×224 with area scale $[0.8, 1.0]$ and aspect ratio $[1.0, 1.3]$, followed by color jitter with brightness, contrast, saturation, and hue parameters of 0.2, 0.2, 0.25, and 0.04. Pixels are scaled to $[0, 1]$ and normalized using ImageNet statistics. Validation and latent-target extraction use only deterministic preprocessing and normalization. Proprioceptive inputs include end-effector position, orientation, and gripper state, with min–max or binary normalization as appropriate, and are represented as 32-dimensional vectors.

Dual-modality encoding and latent bottleneck. Frozen DINOv3 ViT-B/16 and DA3-BASE backbones extract semantic and geometric features from the same RGB observations. We use layer-normalized features from their penultimate transformer layers, obtaining 196 patch tokens per frame and modality, with dimensions 768 and 1536, respectively. DA3 processes frames independently to prevent temporal information leakage. Separate linear projections map both modalities to width 1024. Fixed sinusoidal 3D positional encodings over time, height, and width are added per modality before token-wise concatenation, yielding $784 = 2 \times 2 \times 196$ context tokens (two timesteps, two modalities, and 196 patches). A single learnable query first attends to all context tokens through a 16-head cross-attention block. The context and query then enter a 24-layer QFormer-style transformer with width 1024, 16 heads, feed-forward width 4096, pre-normalization, GELU activations, and zero dropout. A modality-blocked mask restricts semantic and geometric context tokens to their respective streams across time, while the query attends to both. The encoder uses neither proprioceptive tokens nor camera-view identifiers. The final query is projected to 32 dimensions and passed through a continuous variational bottleneck with layer normalization and learned mean and log-standard-deviation parameters. Training uses reparameterized samples $\mathbf{z}$; downstream policy alignment uses the posterior mean.

Decoding, objectives, and optimization. Two parameter-independent, 12-layer AdaLN transformer decoders reconstruct future semantic and geometric features, conditioned on their respective current-frame features and the shared 32-dimensional code $\mathbf{z}$. Conditioning features and targets come from the independently augmented decoder branch. Both decoders use width 1024, 16 heads, feed-forward width 4096, GELU activations, and fixed 2D spatial positional encodings. A linear projection of $\mathbf{z}$ modulates each block's self-attention and feed-forward sublayers through AdaLN with zero-initialized gates. Input/output projections connect the hidden width to each modality's feature dimension, with a final layer normalization. An embodiment-specific auxiliary MLP predicts the 32-dimensional proprioceptive change from $\mathbf{z}$ and the initial state. Training combines semantic and geometric Smooth L1 reconstruction losses with unit weights, an auxiliary L2 loss with $\lambda_{\mathrm{aux}} = 1.0$, and a KL penalty to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with fixed $\lambda_{\mathrm{KL}} = 5 \times 10^{-5}$ and no annealing. The semantic Smooth L1 transition parameter is 0.1; diversity regularization is disabled. We train for two epochs using

AdamW with base learning rate 3×10^{-4}, coefficients $(0.9, 0.999)$, and weight decay 0.01, excluding biases and layer-normalization parameters. The learning rate warms up linearly for $10{,}000$ optimizer steps and remains constant thereafter. Training uses 16 GPUs with a per-GPU batch size of 16, yielding an effective global batch size of 256, BF16 mixed precision, and global gradient L2-norm clipping at 1.0.

B.2 Iso-WAM Architecture and Training Details

Inputs and action representation. Iso-WAM uses both the primary and wrist-camera observations. Each view is resized to 224×224, normalized to $[-1, 1]$, and concatenated horizontally with the other view to form a 224×448 input to the video VAE. Separate per-camera images are retained for latent-action teacher supervision. Each training window contains 33 observation timesteps and 32 control actions. Sampling video frames at indices $0, 4, \ldots, 32$ yields nine frames aligned with the action sequence; these frames include the current observation and future training targets. The initial 8-dimensional proprioceptive state, comprising six end-effector pose components and two gripper components, is linearly projected into an additional context token and appended to the text embeddings. Each action has seven dimensions: six incremental end-effector pose components and one absolute gripper command. We retain the dataset's native coordinate convention without additional frame transformations. Actions and proprioceptive states use global min–max normalization based on the training pipeline's `dataset_stats.json`, without timestep-specific action statistics.

Mixture-of-Transformers and alignment queries. Iso-WAM combines the 30-layer Wan2.2 TI2V-5B Video DiT with a 30-layer Action DiT through mixed self-attention at every layer. The video and action experts have hidden widths of 3072 and 1024, and feed-forward widths of 14336 and 4096, respectively. Both use 24 attention heads with head dimension 128, but maintain independent projection, normalization, feed-forward, and modulation parameters. Each expert projects its hidden states into a shared 3072-dimensional attention space. Their queries, keys, and values are concatenated along the sequence dimension for masked joint self-attention; the outputs are then split and projected back to each expert's hidden width. Each expert subsequently applies its own cross-attention to the text and proprioceptive context, followed by its feed-forward sublayer. Video tokens use frame-causal attention, with bidirectional attention within each frame and no access to later frames or action/query tokens. Action tokens can attend to all video tokens, whereas alignment queries can directly attend only to video tokens from the current observation frame. Action and query tokens interact bidirectionally. The action stream contains 32 action tokens and 16 learnable alignment queries, indexed by eight future transition positions and two camera views. Queries are interleaved by transition and camera. For transition index $j \in \{0, \ldots, 7\}$, both camera queries share the action rotary-position index $\min(4(j+1), 32) - 1$. A linear alignment head maps query representations from 1024 to 32 dimensions for cosine alignment with the corresponding teacher targets.

Objectives and optimization. The policy is trained on the concatenation of the four LIBERO datasets (Spatial, Object, Goal, and LIBERO-10), using the no-op-filtered LeRobot versions without explicit dataset sampling weights. The objective combines video and action flow-matching losses with latent-action alignment: $\mathcal{L}_{\text{policy}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{action}} + 0.1\,\mathcal{L}_{\text{align}}$. The video and action training schedulers each use 1,000 timesteps and a flow-matching shift parameter of 5.0. We initialize the video expert from Wan2.2 TI2V-5B and the action expert from a Wan2.2-derived, linearly interpolated checkpoint with alpha scaling and hidden width 1024. Both experts, the proprioceptive projection, and the alignment queries and head are trainable. The VAE, text encoder, and latent-action teacher, including its DINO and DA3 backbones, remain frozen; text embeddings are cached. We train for 10 epochs using AdamW with learning rate 10^{-4}, coefficients $(0.9, 0.95)$, and weight decay 0.01. Linear warm-up occupies the first 5% of optimizer steps, followed by cosine decay to 10^{-6}. Training uses eight GPUs with four samples per GPU and two gradient-accumulation steps, yielding an effective global batch size of 64. We use BF16 mixed precision, Accelerate with DeepSpeed ZeRO stage 1, and global gradient-norm clipping at 1.0. Checkpoints are saved every 2,000 optimizer steps.

C Benchmark Descriptions and Evaluation Settings

LIBERO. The standard LIBERO benchmark (Liu et al., 2023) comprises four task suites: Spatial, Object, Goal, and Long. Each suite contains 10 tasks and 500 expert demonstrations, yielding 40 tasks and 2,000 demonstrations in total. We train our policies on the standard LIBERO demonstration data. Evaluation consists of 50 rollouts per task, corresponding to 500 rollouts per suite and 2,000 rollouts per random seed. We report task success rates for each suite and the overall average across the four suites. This benchmark assesses in-distribution task performance and control capabilities.

LIBERO-Plus. LIBERO-Plus (Fei et al., 2025) evaluates manipulation skills under seven categories of distribution shifts: camera viewpoint, robot initial-state, language instruction, lighting, background, image noise, and object layout. Our policies are trained on standard LIBERO data and directly evaluated on LIBERO-Plus without additional training on its perturbed data. Evaluation consists of 10,030 episodes per random seed, and reported success rates are averaged over two seeds. We report results for each perturbation category and the overall average to characterize zero-shot out-of-distribution robustness.

Real-World Robot Deployment. We evaluate Iso-WAM and Fast-WAM-Joint on a 6-DoF UR5 manipulator using two pick-and-place tasks: placing a banana onto a green box and placing an orange into a white box. For each task, we collect 80 kinesthetic demonstration trajectories under standard laboratory conditions, with normal white lighting and a plain blue tabletop. Evaluation includes normal conditions and two types of unseen visual shifts. For *lighting shifts*, the standard ceiling lights are switched off and replaced with intense blue or purple LED illumination. For *texture shifts*, the plain tabletop is replaced with either a red-and-white striped cloth or a black-and-white striped cloth. The policies are evaluated under these shifted conditions without additional fine-tuning. We conduct 20 rollouts per task under each condition, and report success rates averaged across the two tasks. This setting assesses whether the policies preserve manipulation performance under real-world visual distribution shifts.

RoboTwin 2.0. RoboTwin 2.0 (Chen et al., 2025) is a simulation benchmark for bimanual robotic manipulation with diverse tasks and domain randomization. We adopt a clean-to-random evaluation protocol: policies are trained on demonstrations collected in clean environments without domain randomization and directly evaluated in randomized environments without additional fine-tuning. The randomized setting introduces scene clutter, lighting variations, background textures, and tabletop-height changes. This protocol assesses zero-shot out-of-distribution generalization under combined visual and environmental shifts.

D Detailed Description of Baselines

We organize our baseline comparisons around four experimental settings: in-distribution manipulation, out-of-distribution generalization, real-world deployment, and representation alignment ablations. The comparisons include external robotic policy baselines and controlled alignment variants built upon Fast-WAM-Joint. Fast-WAM and Fast-WAM-Joint serve as the direct baselines for our action-only and joint video–action inference settings, respectively. The baselines used in each setting are described below.

- **In-Distribution Manipulation on LIBERO.** On the four standard LIBERO suites (Spatial, Object, Goal, and Long), we compare against OpenVLA (Kim et al., 2024), π_0 (Black et al., 2024), $\pi_{0.5}$ (Physical Intelligence et al., 2025), π_0-FAST (Pertsch et al., 2025), UniVLA (Bu et al., 2025), WorldVLA (Cen et al., 2025), and Motus (Bi et al., 2026). These methods provide reference performance across vision-language-action and world-model-based policy approaches. We additionally compare with Fast-WAM and Fast-WAM-Joint from the Fast-WAM framework (Yuan et al., 2026b). Fast-WAM generates actions without explicitly generating future video at inference time, whereas Fast-WAM-Joint jointly denoises future visual latents and actions. Their comparisons with Iso-WAM-uncond and Iso-WAM, respectively, assess the effect of dynamic representation alignment under the two inference paradigms.

- **Out-of-Distribution Generalization on LIBERO-Plus.** For zero-shot robustness evaluation, we compare against UniVLA, OpenVLA-OFT (Kim et al., 2025), π_0, π_0-FAST, WorldVLA, Fast-WAM, and Fast-WAM-Joint. The evaluation covers the seven LIBERO-Plus perturbation categories: camera viewpoint, robot initial-state, language instruction, lighting, background, image noise, and object layout. Our LIBERO-trained policies are evaluated on LIBERO-Plus without additional training on its perturbed data. The two Fast-WAM variants provide direct references for assessing robustness improvements under action-only and joint video–action inference.
- **Real-World Robotic Manipulation.** We use Fast-WAM-Joint as the direct baseline for deployment on the UR5 manipulator. The comparison covers two tasks: placing a banana onto a green box and placing an orange into a white box. Both methods are trained using demonstrations collected under standard laboratory conditions and evaluated under normal conditions, blue and purple lighting, and two unseen patterned tabletop textures. This comparison examines whether dynamic representation alignment improves physical manipulation robustness under visual distribution shifts.
- **Representation Alignment Ablations.** To examine the choice of alignment target, we construct the following variants under the Fast-WAM-Joint paradigm. *No Alignment* retains the base policy objective without auxiliary representation alignment. *DINO Feature* and *DA3 Feature* use static future-frame features from DINOv3 (Siméoni et al., 2025) and Depth Anything 3 (Lin et al., 2025), respectively. *DINO Latent Action* and *DA3 Latent Action* instead use transition representations derived from the corresponding single-modality latent action models. *Optical Flow* uses an optical-flow-based alignment target to examine an alternative motion representation. *Dual Latent Action* simultaneously supervises the policy with separate semantic and geometric latent-action targets. In contrast, our *Uni-Latent Action* variant aligns the policy with a single target produced by Uni-LAM, which combines both modalities before policy alignment. All alignment variants are evaluated on both standard LIBERO and LIBERO-Plus to characterize their in-distribution performance and out-of-distribution robustness.

  RoboTwin 2.0 Clean-to-Random Generalization. We compare against StarVLA (Ye et al., 2026a), GR00T-N1.7 (Bjorck et al., 2025), LDA-1B (Lyu et al., 2026), FastWAM (Yuan et al., 2026b), and LingBot-VA (Li et al., 2026a). We additionally evaluate Iso-WAM on both FastWAM and LingBot-VA to assess its applicability across different WAM backbones.

E Detailed Results on RoboTwin 2.0

As shown in Table 5, Iso-WAM consistently improves performance on both FastWAM and LingBot-VA. For FastWAM, success rates increase from 77.8% to 79.8% under the clean setting and from 1.9% to 3.6% under the random setting. For LingBot-VA, Iso-WAM improves clean success from 81.2% to 82.0% and random success from 29.8% to 37.4%, achieving a 7.6-percentage-point gain under domain randomization. These improvements demonstrate the applicability of dynamic representation alignment across different WAM backbones. By extending evaluation to bimanual manipulation under combined environmental perturbations, these experiments broaden the scope of our robustness evaluation and complement the LIBERO and LIBERO-Plus results, providing further evidence that dynamic representation alignment improves out-of-distribution generalization.

F Detailed Ablation Results on LIBERO

In this section, we provide the comprehensive evaluation results on the in-distribution (ID) standard LIBERO benchmark for the ablation study discussed in the main text, serving as a supplementary explanation. As shown in Table 6, we detail the success rates across four distinct task suites: Spatial, Object, Goal, and Long (LIBERO-10). The results demonstrate that our proposed model still guarantees the highest success rates on the ID tasks.

G Details of Real-World Experiments

Table 5: **Comparison on RoboTwin clean→random** (Chen et al., 2025).

Method	clean	random
StarVLA (Ye et al., 2026a)	58.1	10.6
GR00T-N1.7 (Bjorck et al., 2025)	43.6	20.7
LDA-1B (Lyu et al., 2026)	53.0	12.3
FastWAM (Yuan et al., 2026b)	77.8	1.9
+ Iso-WAM	79.8	3.6
LingBot-VA (Li et al., 2026a)	81.2	29.8
+ Iso-WAM	**82.0**	**37.4**

Table 6: Detailed ablation results on the in-distribution standard LIBERO benchmark. Success rates (%) are evaluated across four task suites: Spatial, Object, Goal, and Long.

Alignment Target	**ID (Standard LIBERO)**				
	Spatial	**Object**	**Goal**	**Long**	**Avg.**
Baseline (No Alignment)	99.4	98.2	98.8	95.6	98.0
Static Feature Alignment					
DINO Feature	99.2	98.4	98.8	97.2	98.4
DA3 Feature	99.8	99.0	99.0	95.8	98.4
Dynamic Latent Action Alignment					
DINO Latent Action	99.0	99.2	98.6	98.0	98.7
DA3 Latent Action	99.4	99.2	99.0	96.4	98.5
Optical Flow	99.2	98.8	99.2	96.6	98.5
Dual Latent Action	98.2	97.8	98.6	97.4	98.0
Uni-Latent Action (Ours)	98.8	99.6	97.8	98.6	98.7

Due to constraints in experimental time and hardware resources, our real-world evaluations are exclusively conducted on a single-arm 6-DoF UR5 manipulator. The visual observation system consists of dual RGB cameras: one static third-person camera positioned in front of the robotic arm and one egocentric camera mounted on the robot's wrist. During the data collection phase, all demonstrations were recorded in a standard laboratory setting featuring normal white room lighting and a plain blue, unpatterned tabletop. We collected exactly 80 expert trajectories for each manipulation task. During the inference phase, we rigorously conducted 20 evaluation rollouts per task under each environmental condition. To further increase the task difficulty and evaluate policy robustness, various unseen distractor objects—such as red blocks, grapes, and pears—were randomly scattered across the workspace alongside the target objects.

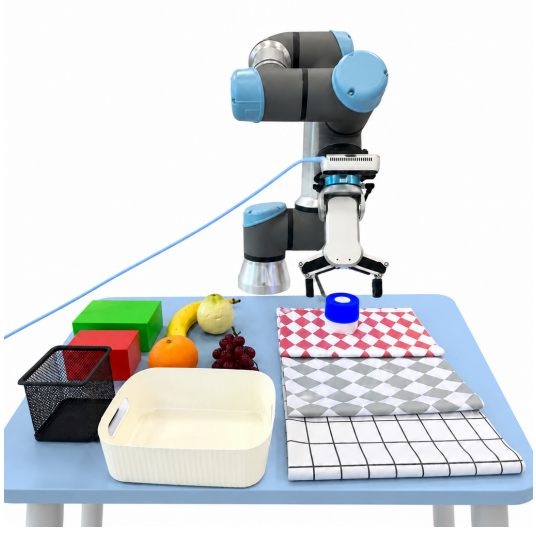


Figure 6: Real-world robot setup.

Lighting Shifts. To construct the lighting shift scenarios, we introduced drastic photometric perturbations by completely turning off the standard laboratory ceiling lights and replacing them with intense colored LED lighting. Specifically, we evaluated the policies under two extreme out-of-distribution (OOD) lighting conditions: a deep blue illumination and a stark purple illumination. These dramatic lighting shifts significantly alter the overall color distribution and cast complex, high-contrast shadows.

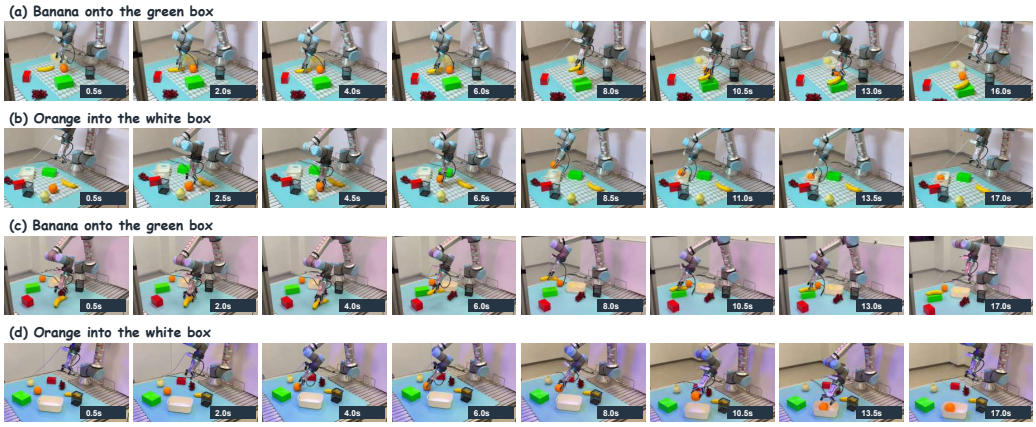


Figure 7: Further Visualizations of Real-World Experiments.

Texture Shifts. For the texture shift scenarios, we assessed the models' robustness against background texture variations by drastically altering the visual appearance of the workspace. We replaced the standard plain blue tabletop with visually noisy tablecloths. Specifically, we utilized two distinct high-frequency patterned cloths: a red-and-white striped tablecloth and a black-and-white striped tablecloth. These strong geometric distractors severely disrupt standard appearance-based representations.

Task Descriptions. We designed two representative pick-and-place manipulation tasks for this real-world deployment. The first task requires the robot to pick up a banana and accurately place it on top of a green box. The second task involves grasping an orange and securely placing it inside a white box. Both tasks demand precise 3D spatial localization and stable end-effector control.

Failure Analysis. Finally, we conducted a qualitative failure analysis of the baseline policies. For Fast-WAM-Joint, the primary cause of failure in both tasks under OOD conditions is inaccurate spatial perception near the target object, which is often exacerbated by the surrounding distractor objects. This visual entanglement frequently causes the baseline to mistakenly grasp the wrong distractor or execute an empty grasp. For instance, when attempting to grasp the banana in the OOD environments, Fast-WAM-Joint struggles to correctly estimate the distance to the banana, leading to severely misaligned gripper closures. In contrast, thanks to its superior dynamic transition isolation mechanism, Iso-WAM exhibits remarkably strong generalizability and robustness, achieving a significant performance improvement over Fast-WAM-Joint.

H Analysis of Dual-Target Alignment.

To further examine the optimization difficulty discussed in the main-text ablation study, we compare dual-target alignment with its single-target counterparts. In the dual-target variant, a shared set of queries is projected through two separate MLP heads to predict the DINO and DA3 latent-action targets, respectively. We observe that the mean query-to-target distances are larger for both targets than in their corresponding single-target variants. These comparisons are made within each target space, rather than across the DINO and DA3 spaces. This observation is consistent with the hypothesis that simultaneously fitting two distinct transition representations makes policy alignment more difficult, potentially limiting the integration of complementary information. It provides some supporting evidence for the performance degradation observed with dual-target alignment. Uni-LAM addresses this difficulty by integrating both modalities into a shared latent space before policy alignment, providing a single unified supervision target.

I ADDITIONAL REPRESENTATION DIAGNOSTICS

We test Uni-LAM's motion information under appearance changes and object–gripper spatial readout (Figures 8 and 9). For these motion and spatial readout experiments, all teachers are frozen; we use their deterministic 32-dimensional posterior means. Main-camera RGB images (224×224, range $[0, 1]$) are resized to the teacher's native input size: 256 for DINO-LAM and 224 for DA3-LAM and Uni-LAM. We additionally visualize the 32-dimensional Uni-LAM variant's sensitivity to local temporal changes in Figure 10.

Shared protocol. Appearance transforms are applied identically to both endpoints: blue cast scales RGB by $(0.40, 0.65, 1.35)$ with clipping to $[0, 1]$; dimming scales all channels by 0.35. Ridge probes standardize inputs and targets using clean training data and select regularization from $\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ by clean-validation 3D error. Perturbed inputs use the same probe without adaptation. These diagnostics compare existing teachers, not matched-training ablations of fusion; 95% bootstrap intervals are exploratory, without multiple-comparison correction.

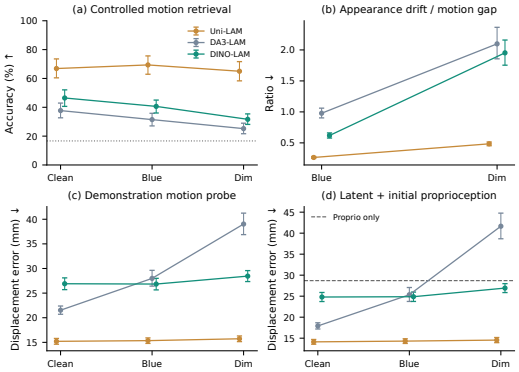


Figure 8: **Motion robustness.** (a,b) Controlled direction retrieval and appearance drift; (c,d) held-out displacement probes without/with proprioception. Dotted/dashed lines: chance/proprioception only. Bars: 95% task-bootstrap intervals.

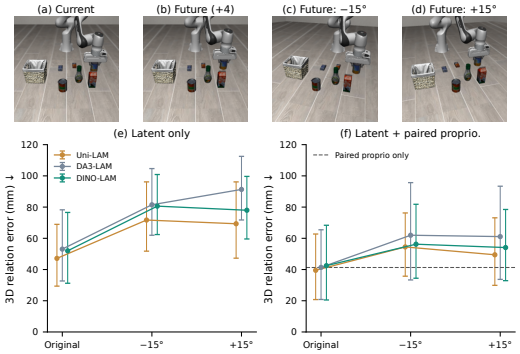


Figure 9: **Spatial readout.** (a–d) A held-out pair and perturbed views; (e,f) spatial probes without/with paired proprioception. Dashed line: proprioception only. Bars: 95% task/layout-bootstrap intervals.

I.1 MOTION INFORMATION UNDER APPEARANCE CHANGES

Controlled motion. We evaluate 480 fixed simulator transitions: 40 tasks, two initial states, and six Cartesian translation directions, each executed for four control steps. Direction retrieval selects the nearest clean latent (Euclidean distance) among six directions from the other initial state of the same task; chance is $1/6$. Appearance variants reuse these physical transitions. We normalize clean-to-perturbed latent distance by the median distance to the other five directions at the same initial state. This raw-latent drift-to-motion ratio differs from the centered, unit-normalized t-SNE metric.

Table 7: Controlled direction retrieval (%). Brackets: 95% task-bootstrap intervals.

Teacher	Clean	Blue	Dim
DINO-LAM	46.5 [40.6, 52.1]	40.6 [36.0, 45.0]	31.7 [28.1, 35.4]
DA3-LAM	37.7 [32.7, 42.9]	31.5 [27.1, 35.8]	25.2 [21.7, 29.0]
Uni-LAM	66.9 [60.4, 73.5]	69.4 [62.9, 75.6]	65.0 [58.3, 71.7]

Uni-LAM achieves the highest retrieval accuracy in every condition (66.9/69.4/65.0% for clean/blue/dim). Its blue/dim drift-to-motion ratios are 0.265/0.485, below DINO-LAM's 0.619/1.953 and DA3-LAM's 0.978/2.098 (Figure 8a,b): motion remains distinguishable despite appearance changes. Clean effective ranks are 5.20/3.74/4.78 for DINO/DA3/Uni-LAM, computed as $\exp(-\sum_j p_j \log p_j)$ from normalized squared singular values of centered features; these non-constant codes should be interpreted in light of the cohort's six commanded directions.

Displacement probes. We sample 20 demonstration episodes per task across 40 LIBERO tasks, with eight four-step pairs per episode: 6,400 pairs in total. Per-task episode splits are 12/4/4 for training/validation/test (3,840/1,280/1,280 pairs). Earlier appearance-visualization episodes are excluded; no pair crosses splits. This is a probe holdout, not a teacher pretraining holdout. One ridge probe per representation predicts $\Delta\mathbf{p} = \mathbf{p}_{t+4} - \mathbf{p}_t$ across tasks, with initial-state, and within-task shuffled-label controls. Euclidean errors are averaged within tasks and then equally across tasks; probe intervals use 10,000 task-bootstrap draws.

Table 8: Held-out demonstration displacement error (mm; lower is better). Brackets: 95% task-bootstrap intervals.

Probe input	Clean	Blue	Dim
Initial proprio.	28.7 [27.4, 30.0]	28.7 [27.4, 30.0]	28.7 [27.4, 30.0]
DINO-LAM	26.9 [25.7, 28.1]	26.8 [25.7, 28.0]	28.4 [27.3, 29.6]
DA3-LAM	21.5 [20.7, 22.4]	28.0 [26.4, 29.6]	39.0 [36.9, 41.2]
Uni-LAM	15.2 [14.6, 15.8]	15.3 [14.8, 15.9]	15.7 [15.1, 16.3]
DINO, shuffled labels	30.6 [28.9, 32.1]	30.7 [29.0, 32.3]	30.7 [29.1, 32.3]
DINO + proprio.	24.8 [23.7, 25.9]	24.9 [23.7, 26.0]	26.9 [25.9, 28.0]
DA3 + proprio.	17.9 [17.2, 18.7]	25.3 [23.7, 27.1]	41.7 [38.7, 44.8]
Uni + proprio.	14.1 [13.6, 14.7]	14.3 [13.7, 14.9]	14.5 [13.9, 15.2]

Results. Uni-LAM has the lowest displacement error in all conditions, with or without proprioception (Table 8). Its latent-only error changes from 15.2 mm to just 15.3/15.7 mm under blue/dim inputs. With proprioception, clean error falls from 28.7 to 14.1 mm, supporting motion information beyond the initial robot state. DINO-LAM also improves over its shuffled-label control (26.90 vs. 30.57 mm); its paired improvement over proprioception alone is -3.89 $[-4.89, -2.86]$ mm. Together, these diagnostics support Uni-LAM's more accurate, appearance-robust motion readout in the tested settings.

I.2 Sensitivity to local temporal changes

The motion probes above measure what can be read out from a latent action. We complement them by examining which spatial regions' temporal information affects the frozen 32-dimensional Uni-LAM teacher (Figure 10). For each region r of a 7×7 grid, we replace the contextual features at both endpoints by their temporal mean, simultaneously in the semantic and geometric streams. The response is the resulting latent change, $s_r = \|z(F) - z(T_r(F))\|_2$, where T_r denotes this replacement. This visualization uses the 32-dimensional variant; the motion and spatial readout experiments in the adjacent subsections use 32-dimensional teachers. Reference and intervened forward passes use matching batch sizes.

Visualization protocol. We show ten transitions, four recorded steps apart, selected from 40 fixed task examples for scene and motion-location diversity, without ranking sensitivity–flow agreement. We normalize each response map by its spatial sum, $\widetilde{s}_r = s_r / \sum_q s_q$, and use a common color range $[0, 0.309]$. A square-root color mapping ($\gamma = 0.5$) and overlay opacity 0.78 improve visibility; colorbar ticks retain the original normalized values. Bilinear interpolation is only for display. RAFT-large (Teed & Deng, 2020) flow magnitude is scaled per example and serves only as a visual reference; it does not enter the sensitivity computation.

Interpretation. Responses appear near moving robot regions in several examples, but also at some image edges and background regions. Across all 40 cases, the mean spatial Spearman correlation with estimated flow is 0.525 (task-bootstrap 95% interval: 0.483–0.570). These maps support sensitivity to temporal information in the teacher's contextual features. They are not raw attention weights or pixel-level motion masks, and do not establish that the latent ignores background information.

I.3 Object-related three-dimensional spatial readability

Protocol. We collect 20 scripted pick-and-place attempts for each of eight LIBERO-Object tasks, including failures. Each uses a fresh simulator/controller, benchmark initial states 2–21, and inde-

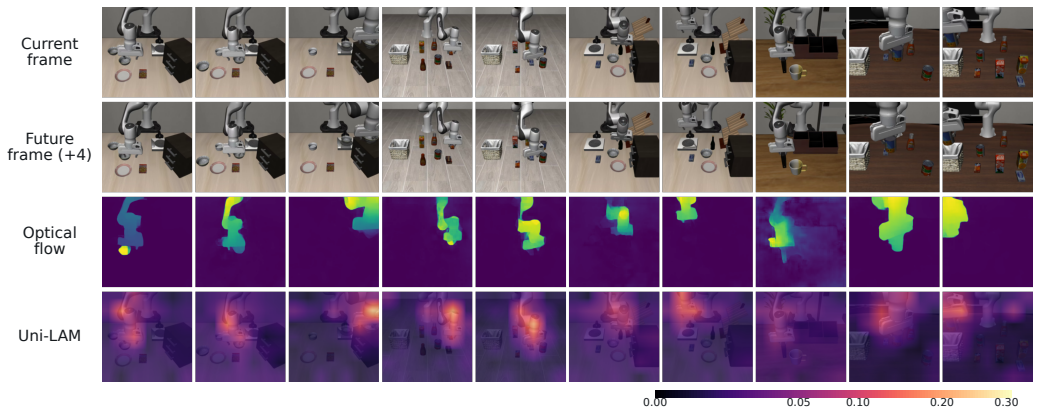


Figure 10: **Uni-LAM temporal sensitivity.** Rows show current frames, future frames (+4 steps), estimated optical flow, and sensitivity maps for ten examples. Brighter regions indicate larger latent changes after removing local temporal feature differences; display details are given in the text.

pendent uniform ± 35 mm target-object x/y offsets before settling. Target and container identities are fixed by the task. The script executes approach, descent, closure, lift, transport, lowering, and release; commanded stages are not success labels. Fourteen four-step windows per trajectory yield 2,240 pairs, with 12/4/4 train/validation/test layouts per task. All windows and camera variants of a trajectory stay in the same split.

A separate ridge probe per task predicts the future object reference point relative to the gripper in robot-base coordinates:

$$\mathbf{r}_{t+4} = \mathbf{R}^{\mathsf{T}}_{\text{world}\leftarrow\text{base}}(\mathbf{p}^{\text{object}}_{t+4} - \mathbf{p}^{\text{gripper}}_{t+4}). \tag{6}$$

Labels use the simulator object root body and grip site, not surface clearance or estimated depth; restoration checks reproduce images pixel-exactly and labels within floating-point precision. We compare latent-only inputs and latents with both endpoint proprioceptive states against paired proprioception alone. The agent-view camera orbits $\pm 15^\circ$ about world z around $(0, 0, 0.10)$ m, leaving states and labels fixed. Probes use clean views only; errors weight tasks equally, with 10,000 hierarchical task/layout-bootstrap draws.

Table 9: Held-out-layout 3D relative-position error (mm; lower is better). Brackets: 95% task/layout-bootstrap intervals.

Probe input	Original view	-15°	$+15^\circ$
Paired proprio.	41.3 [21.4, 65.5]	41.3 [21.4, 65.5]	41.3 [21.4, 65.5]
DINO-LAM	52.0 [31.2, 76.6]	80.6 [62.4, 100.9]	78.0 [59.6, 99.7]
DA3-LAM	53.1 [32.7, 78.2]	81.5 [62.0, 104.7]	91.3 [71.7, 112.5]
Uni-LAM	47.1 [29.4, 68.9]	71.7 [51.8, 96.2]	69.3 [47.3, 96.1]
DINO + paired prop.	42.5 [20.5, 68.4]	56.2 [34.4, 81.8]	54.1 [32.9, 78.5]
DA3 + paired prop.	41.4 [20.8, 65.5]	62.0 [33.3, 95.7]	61.1 [33.7, 93.4]
Uni + paired prop.	39.5 [20.8, 62.7]	54.5 [35.7, 76.2]	49.5 [29.9, 73.1]

Results. Uni-LAM has the lowest mean error among the teachers in every view, both with and without proprioception. Clean latent-only error is 47.14 mm, versus 51.99/53.08 mm for DINO/DA3-LAM; with paired proprioception it is 39.5 mm versus 41.3 mm for proprioception alone. The clean Uni-minus-DINO difference is -4.85 $[-10.12, 0.92]$ mm, so the lower mean does not establish a statistically significant advantage. For DINO plus proprioception, the difference from proprioception alone is 1.17 $[-1.86, 4.97]$ mm. These results suggest useful spatial readout, while all teachers remain view-sensitive. With four test layouts per task and task-specific probes, the result concerns within-task readout, not zero-shot object selection. Linear latent probes also omit the current-frame features available to Uni-LAM's reconstruction decoders.