P-WAM: PERCEPTION-WORLD-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) have recently shown strong potential for robot manipulation by jointly modeling future visual states and actions, thereby capturing environment dynamics and physical regularities for closed-loop control. However, existing WAMs typically rely on frozen reconstruction-oriented visual VAEs for initial environment perception, which may limit their ability to capture object-level semantics and fine-grained spatial information. In this work, we propose *Perception-World-Action Models (P-WAM)*, a simple and general framework that enhances WAMs with pretrained visual perception encoders while preserving their world dynamics modeling. P-WAM introduces complementary semantic and spatial perceptual features from discriminative visual encoders into the video-action DiT through lightweight connector modules. We instantiate P-WAM with SigLIP and DINOv3 to provide language-aligned semantics and dense spatial representations. Through comprehensive representation probing, experiments on the challenging RoboCasa simulation benchmark, and diverse real-world robot evaluations, we show that P-WAM improves the object-level semantic and fine-grained spatial information encoded in video-action DiT representations and leads to stronger closed-loop policy performance. These results suggest that integrating visual perception with generative world-action modeling is a promising direction for improving WAMs.

1 INTRODUCTION

Compared with standard Vision-Language-Action (VLA) (Zitkovich et al., 2023; Pertsch et al., 2025; Black et al., 2024) models, which primarily learn direct observation-to-action mappings, World Action Models (WAMs) (Kim et al., 2026; Ye et al., 2026b; Li et al., 2026; Ye et al., 2026a) explicitly model world dynamics by predicting how the environment evolves under intervention. By unifying future visual prediction and action modeling within a single framework, WAMs are able to better capture task-relevant spatiotemporal dynamics priors and physical regularities, leading to strong instruction following and generalization capabilities in robot manipulation (Kim et al., 2026; Ye et al., 2026b; Zhang et al., 2026c).

Advanced WAMs (Ye et al., 2026b; Kim et al., 2026) typically fine-tune video generation foundation models (Wan et al., 2025; Ali et al., 2025) on robotic data, jointly predicting future states and actions in a shared representation space. These models are commonly built upon frozen text encoders (Raffel et al., 2020), frozen visual VAEs (Kingma & Welling, 2013; Wan et al., 2025), and a unified video-action diffusion Transformer. Meanwhile, recent studies suggest that visual VAEs trained primarily for reconstruction are limited in semantic representation (Chen et al., 2025a; Yu et al.; Zheng et al., 2025a; Zhang et al., 2026a; Liu et al., 2025c). This limitation has also been recognized in recent studies on multimodal understanding and generation. Representation-enhanced visual generation methods (Wang et al., 2025a; Yu et al.; Zheng et al., 2025a) show that pretrained visual semantic representations can improve generative modeling. In parallel, some advanced unified multimodal understanding and generation models (Deng et al., 2025a; Song et al., 2026; Wu et al., 2025; Li et al., 2025b) choose to rely on distinct visual interfaces, using discriminative visual perception encoders for semantic understanding and visual VAEs for generative modeling. Inspired by these observations, we ask two basic questions on WAMs for robot learning: *(1) Can WAMs that primarily rely on frozen reconstruction-oriented visual VAEs for initial environment perception, learn sufficiently rich perceptual representations for robot manipulation? (2) If not, how can powerful discriminative visual perception encoders be incorporated into WAMs to better capture spatial structure and object semantics?*

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.

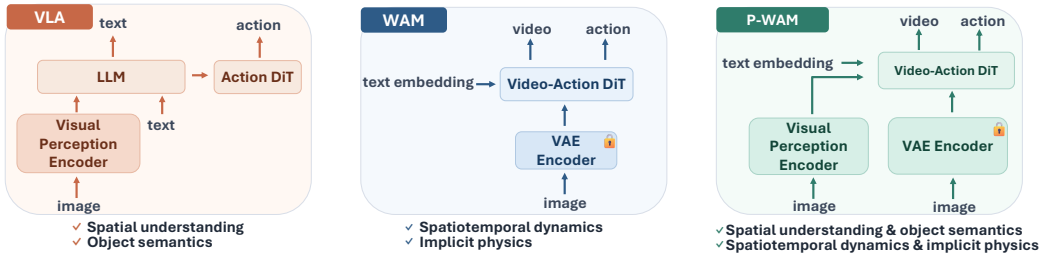


Figure 1: **Comparison of VLA, WAM, and our P-WAM.** VLA models benefit from visual perception encoders for spatial understanding and object semantics, whereas WAMs rely on reconstruction-oriented VAE encoders and video-action modeling to capture spatiotemporal dynamics and implicit physics. P-WAM integrates visual perception encoders with the VAE-based video-action backbone, combining discriminative semantic perception with generative world dynamics modeling.

To systematically investigate these questions, we first adopt Cosmos Policy (Kim et al., 2026), a strong and representative WAM, as our baseline. It is fine-tuned from Cosmos-Predict2-2B (Ali et al., 2025), a video generation foundation model trained on large-scale Physical AI data. Motivated by the observations and analysis above, we further propose Perception-World-Action Models (**P-WAM**), which introduces discriminative visual perception encoders while preserving the ability of WAMs to explicitly model world dynamics. As illustrated in Fig. 1, P-WAM provides a general, simple, and efficient way to integrate semantic and spatial understanding from visual perception encoders with the generative world dynamics modeling capability of WAMs. Specifically, P-WAM instantiates the visual perception branch with two complementary pretrained encoders: DINOv3 (Siméoni et al., 2025), which offers dense visual representations with fine-grained spatial details, and SigLIP (Zhai et al., 2023), which produces visual features well aligned with language instruction. For each view, we extract patch-level visual tokens from both encoders, spatially pool them to a compact grid resolution, and fuse them with a lightweight MLP. The fused tokens from multiple views are then processed by a lightweight transformer connector, which aggregates multi-view visual features and aligns them with the hidden space of the video-action DiT. These aligned tokens are injected into the DiT layers through cross-attention, providing complementary semantic and spatial cues to the shared video-action backbone.

We evaluate our P-WAM and baseline from both representation and policy perspectives to answer the above questions. First, we conduct a frozen-backbone object-level semantic probing experiment on video-action DiT representations, where only a lightweight segmentation decoder is trained to predict task-relevant foreground objects and robot gripper masks. These results show that representations learned by representative WAMs still leave room for stronger object-level semantic and fine-grained spatial understanding, which our P-WAM effectively enhances. Second, we evaluate P-WAM on the challenging RoboCasa (Nasiriany et al.) benchmark and diverse real-world robot manipulation tasks. The results show that the improved perceptual representations translate into better policy performance, suggesting that discriminative visual perception provides a useful complementary signal for generative world action modeling.

Our contributions are summarized as follows:

- We quantitatively identify that representative WAMs relying on reconstruction-oriented visual VAEs for initial environment perception still leave room for improved object-level semantic and fine-grained spatial understanding in their video-action DiT representations.
- We propose P-WAM, a simple, general, and efficient framework that incorporates discriminative visual perception encoders into generative world-action modeling, enhancing semantic and spatial perception while preserving world dynamics modeling.
- We validate P-WAM on the challenging RoboCasa benchmark and diverse real-world robot manipulation tasks, demonstrating improved policy performance.

2 Related Work

Vision-Language-Action Models. Vision-Language-Action (VLA) models bring the semantic understanding and reasoning abilities of large vision-language models (VLMs) (Beyer et al., 2024;

Zhu et al., 2025b; Bai et al., 2025b;a) into embodied intelligence. Existing VLAs can be broadly categorized by their action representation strategy into discrete methods (Zitkovich et al., 2023; Pertsch et al., 2025; Liu et al., 2025b; Wang et al., 2025b; Liang et al., 2025b), which represent actions using specialized action tokenizers or language vocabularies, and continuous methods (Black et al., 2024; Kim et al., 2025; Bjorck et al., 2025; Liu et al., 2025a), which typically employ an additional diffusion transformer as the action expert. Benefiting from strong visual perception encoders inherited from pretrained VLMs (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) and post-training on diverse embodied understanding tasks (Intelligence et al., 2025; Chen et al., 2025b; Liu et al., 2025b; Zhai et al., 2025; Deng et al., 2025b; Zhen et al., 2024; Li et al.) such as spatial grounding, VLA models generally exhibit strong spatial understanding and semantic generalization. Nevertheless, they typically learn reactive observation-to-action mappings, rather than explicitly modeling world evolution under intervention, which limits their ability to capture environment dynamics and physical regularities (Ye et al., 2026b; Wang et al., 2026).

World Action Models. World Action Models (WAMs) integrate world models into the robot action generation pipeline (Wang et al., 2026). Early and intermediate works (Du et al., 2023; Bharadhwaj, Homanga and Dwibedi, Debidatta and Gupta, Abhinav and Tulsiani, Shubham and Doersch, Carl and Xiao, Ted and Shah, Dhruv and Xia, Fei and Sadigh, Dorsa and Kirmani, Sean, 2024; Hu et al., 2025; Pai et al., 2025) largely follow a cascaded paradigm, reformulating action prediction into sequential stages by first forecasting future visual states or goals and then deriving the corresponding action through goal-conditioned policy learning or action decoding. More recent methods (Zhu et al., 2025a; Li et al., 2025a; Kim et al., 2026; Ye et al., 2026b; Li et al., 2026; Ye et al., 2026a; Yuan et al., 2026; Shen et al., 2026; Guo et al., 2024; Cheang et al., 2024) follow a joint paradigm, which produces future states and actions simultaneously, so that future state prediction and action generation are co-optimized within a shared representation space, allowing the model to better inherit spatiotemporal dynamics priors from video generation foundation models and capture the causal interdependencies between environment dynamics and control. Despite their stronger ability to inherit and model environment dynamics and physical priors, WAMs typically rely on frozen visual VAE encoders (Kingma & Welling, 2013), trained mainly for pixel-level reconstruction, to perceive the environment. Although these VAE latents are well suited for generative modeling, recent works (Zheng et al., 2025a; Chen et al., 2025a; Yu et al.) suggest that they remain limited in semantic understanding compared with pretrained VLMs (Bai et al., 2025b) and powerful visual perception encoders (Siméoni et al., 2025; Zhai et al., 2023; He et al., 2022; Radford et al., 2021; Assran et al., 2023; He et al., 2020), especially in capturing object semantics and spatial structure, both of which are crucial for embodied decision-making.

Unified Multimodal Understanding and Generation Models. Unified multimodal models aim to jointly perform multimodal understanding and generation within a single framework (Xie et al., 2026; Cui et al., 2025). To balance semantic understanding and generative fidelity, several recent works (Song et al., 2026; Deng et al., 2025a; Li et al., 2025b; Wu et al., 2025) adopt a decoupled visual interface, using pretrained visual perception encoders from VLMs for understanding while relying on visual VAEs for visual generation, thereby efficiently and effectively unifying the two capabilities. Another line of work (Zhang et al., 2026a; Liu et al., 2025c) seeks to learn unified visual encoders by jointly optimizing reconstruction and semantic alignment objectives, so that a shared visual representation can support both understanding and generation. In the embodied domain, recent works further explore how multimodal understanding and generation can be jointly leveraged for robot action prediction, either through late fusion (Zhang et al., 2026b; Intelligence et al., 2026; Zhao et al., 2025) or earlier integration with a Mix-of-Transformers paradigm (Hu et al., 2026; Cai et al., 2026; Bi et al., 2025). Closest to our work is Cosmos3 (Agarwal et al., 2026), which also couples a reasoning branch with a generative backbone, but its reasoning branch mainly encodes the language instruction for policy rather than the visual observation. Our P-WAM injects visual perception of the observation into a joint WAM with a fully parameter-shared video-action backbone, which preserves spatiotemporal dynamics priors while maintaining a simple and general framework.

3 METHODS

3.1 PRELIMINARIES: WORLD ACTION MODELS

We first review World Action Models (WAMs) through the lens of Cosmos Policy (Kim et al., 2026), a representative WAM for robot manipulation. Let $s_t = \left(p_t, \{I_t^i\}_{i=1}^V\right)$ denote the state, which consists of robot proprioception p_t and multi-view RGB observations $\{I_t^i\}_{i=1}^V$, and let c denote the

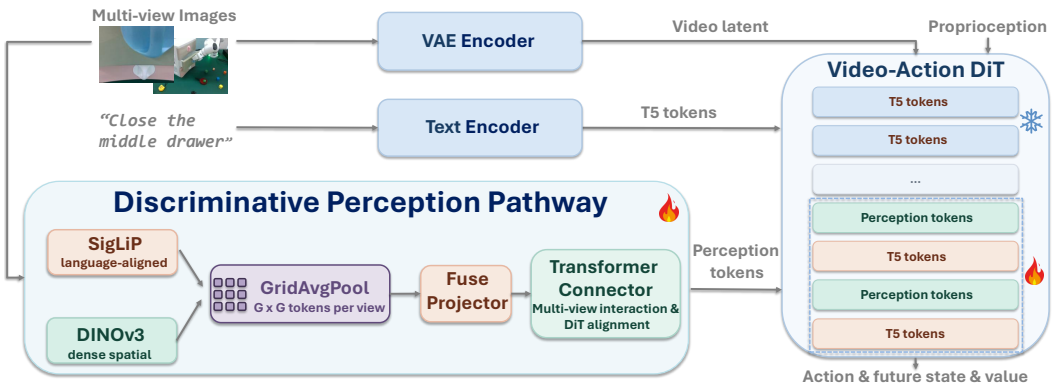


Figure 2: **Overview of P-WAM.** The discriminative perception pathway extracts semantic and spatial perceptual features from pretrained visual perception encoders and converts them into perception tokens. The perception tokens are alternately injected into the last N layers of the video-action DiT.

language instruction. Cosmos Policy adapts a pretrained video generation model (Ali et al., 2025) into a unified video-action DiT. Although the primary goal is unified policy learning and future prediction, the same model also supports action-conditioned dynamics modeling and pure value prediction by changing which latent frames are used as conditioning inputs and which frames are treated as denoising targets:

$$\begin{aligned} &\text{Policy:} && (a_{t:t+H-1}, s_{t+H}, v_{t+H}) = f_\theta(c, s_t)\,, \\ &\text{Dynamics:} && s_{t+H} = f_\theta(c, s_t, a_{t:t+H-1})\,, \\ &\text{Value:} && v_{t+H} = f_\theta(c, s_t, a_{t:t+H-1}, s_{t+H})\,. \end{aligned} \tag{1}$$

Here $a_{t:t+H-1}$ denotes an action chunk, s_{t+H} denotes the future state, and v_{t+H} denotes the value estimate. Despite this unified design, the current visual observations enter the WAM primarily through VAE latents. These latents are well suited for visual reconstruction and generation, but they are not explicitly optimized to encode object-level semantics or dense spatial structure. This motivates P-WAM, which preserves the generative world-action modeling capability of WAMs while introducing complementary discriminative visual perception features into the DiT backbone.

3.2 Discriminative Perception Pathway

As illustrated in Fig. 2, P-WAM augments the original reconstruction-oriented VAE pathway of WAMs with a discriminative perception pathway. Given the current multi-view observations $\{I_t^i\}_{i=1}^V$, this pathway extracts semantic and spatial perception tokens from pretrained visual perception encoders. Importantly, it does not replace the VAE latents used by the original WAM. The VAE latents remain responsible for generative video-action modeling, while the perception tokens provide complementary object-level semantic and fine-grained spatial cues.

In our main instantiation, we use two complementary visual perception encoders: SigLIP (Zhai et al., 2023), which provides language-aligned visual semantics, and DINOv3 (Siméoni et al., 2025), which provides dense spatial representations. For each view I_t^i, we extract patch-level features from both encoders:

$$z_{\text{siglip}}^i = E_{\text{siglip}}(I_t^i), \qquad z_{\text{dino}}^i = E_{\text{dino}}(I_t^i). \tag{2}$$

Instead of reducing each view to a single global token, we spatially pool the patch features into a compact grid of tokens, preserving coarse spatial layout and fine-grained object information:

$$\tilde{z}_{\text{siglip}}^i = \text{GridAvgPool}(z_{\text{siglip}}^i), \qquad \tilde{z}_{\text{dino}}^i = \text{GridAvgPool}(z_{\text{dino}}^i). \tag{3}$$

The pooled SigLIP and DINOv3 tokens are then fused with a lightweight projection module:

$$z_{\text{p}}^i = \phi\big(\tilde{z}_{\text{siglip}}^i, \tilde{z}_{\text{dino}}^i\big)\,, \tag{4}$$

where z_{p}^i denotes the fused perception tokens for view i.

The perception tokens from all views are passed to a lightweight transformer connector:

$$V_{\mathrm{p}} = C_{\psi}\left(z_{\mathrm{p}}^{1}, \ldots, z_{\mathrm{p}}^{V}\right). \tag{5}$$

The connector serves two purposes. First, it aggregates semantic and spatial information across multiple camera views. Second, it aligns the perception tokens to the hidden space of the video-action DiT. The resulting tokens V_{p} are used as perception conditioning tokens for DiT injection.

3.3 Alternating Perception Injection

After obtaining the perception tokens V_{p} from the transformer connector, P-WAM injects them into the video-action DiT through the existing cross-attention interface, as shown in Fig. 2. We keep the early DiT layers unchanged and frozen, so that they preserve the pretrained video-action latent processing and generative dynamics modeling capability inherited from the original WAM. P-WAM only modifies the last N DiT layers, where higher-level video-action representations are more directly involved in action prediction and future-state modeling. In these late layers, P-WAM alternates the cross-attention context between the original T5 (Raffel et al., 2020) language tokens and the perception tokens V_{p}. Following the alternating condition injection strategy used in prior robot diffusion models (Liu et al., 2024; Bjorck et al., 2025), this design avoids injecting both modalities into every layer, which can cause one modality to dominate the other. The modified late DiT layers and the perception pathway are trained, while the frozen early DiT layers retain the original WAM dynamics prior. In this way, P-WAM efficiently integrates discriminative visual perception with generative world-action modeling.

4 Experiments

Our experiments are designed to answer and analyze the following questions:

1. **Do video-action DiT representations still leave room for improved perception?** We first evaluate whether video-action DiT representations in WAMs capture sufficient task-relevant object semantics and fine-grained spatial information. We then examine whether introducing discriminative visual perception through P-WAM improves these representations.
2. **Can improved perceptual representations translate into better policy performance?** We evaluate P-WAM in both simulation and real-world robot deployment to verify whether enhanced visual perception leads to higher closed-loop policy success rates.
3. **How do key design choices of P-WAM affect performance?** We analyze how the choice of visual perception encoder, parameter freezing and fine-tuning strategy, the injection stage, the token types that attend to perception tokens, and the injection strategy of alternating versus concatenating with language tokens affect policy performance.
4. **Can P-WAM benefit from strong pretrained VLMs?** We investigate whether reasoning and semantic priors from a strong pretrained VLM can be transferred to WAMs through the P-WAM framework and improve robot policy performance.

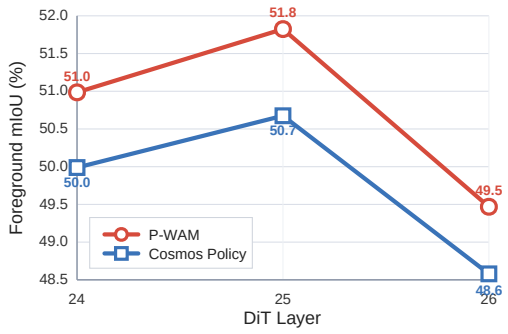


Figure 3: **Representation probing experiments.**

Table 1: **Future visual prediction quality.** Adding the discriminative perception pathway leaves the generative world modeling objective on par or slightly better.

Method	SSIM ↑	FID ↓
Cosmos Policy	0.784	23.7
P-WAM (ours)	**0.784**	**23.5**

Table 2: **RoboCasa simulation results.** We compare P-WAM with representative VLA-based and WAM-based methods. We report the results of the released Cosmos Policy checkpoint and the corresponding seed in our environment. Additional seeds and per-task results are provided in the App. A to show the stability of our improvements and more detailed analysis.

Category	Method	Demos / task	Avg. SR (%)
VLA-based			
	GR00T-N1 (Bjorck et al., 2025)	300	49.6
	GR00T-N1 + DreamGen (Jang et al., 2025)	300	57.6
	DP-VLA (Han et al., 2024)	3000	57.3
	π_0 (Black et al., 2024)	300	62.5
	GR00T-N1.5	300	64.1
	GR00T-N1.5 + HAMLET (Koo et al., 2025)	300	66.4
	GR00T-N1.6	300	66.2
WAM-based			
	UVA (Li et al., 2025a)	50	50.0
	UWM (Zhu et al., 2025a)	1000	60.8
	Video Policy (Liang et al., 2025a)	300	66.0
	FLARE (Zheng et al., 2025b)	300	66.4
	Cosmos Policy (Kim et al., 2026)	50	65.8
	Cosmos Policy + World2Act (Vuong et al., 2026)	50	66.3
	P-WAM (ours)	**50**	**68.3**

We next describe the implementation details of our P-WAM, the RoboCasa simulation evaluation, the real-world evaluation, and the representation probing setups. All comparisons with Cosmos Policy are conducted under identical settings for fairness, with additional details provided in the App. B, App. C, App. D.

P-WAM Implementation Details. For model implementation, we use SigLIP-SO400M with a patch size of 14 and DINOv3-ViT-L/16 as the two visual perception encoders. Patch features from each encoder are spatially pooled to an $N \times N$ grid per view, fused by a two-layer MLP projector, and then processed by a one-layer transformer connector. This perception pathway adds only a modest computational overhead, increasing the single-forward FLOPs by approximately 12% (9.54 vs. 8.53 TFLOPs) compared with Cosmos Policy under the RoboCasa configuration. This accounts for the full multi-view forward pass of both perception encoders. In closed-loop deployment, the per-chunk inference latency increases only marginally, from 1.25 s to 1.30 s (+4%). For training, we initialize P-WAM from pretrained Cosmos Policy (Kim et al., 2026) weights and use the same preprocessed training data, with only 50 human-teleoperated demonstrations per task and no additional generated demonstrations. We retain the original Cosmos Policy learning formulation, including value modeling and future proprioception prediction. Training follows a two-stage procedure: we first freeze the visual perception encoders and train the P-WAM injection modules with the corresponding video-action DiT layers, then unfreeze the visual encoders for joint fine-tuning.

RoboCasa Simulation Experiments Setup. We evaluate our method on the RoboCasa simulation benchmark (Nasiriany et al.). RoboCasa is a challenging benchmark for generalizable manipulation. The benchmark contains 24 static kitchen manipulation tasks with a single Franka Emika Panda robot arm. We follow the same evaluation protocol as Cosmos Policy. For each task, we evaluate success rate over 50 trials across five evaluation scenes with different layouts and visual styles, with 10 trials per scene; additionally, two of the five scenes contain unseen styles that are not encountered in the training data. We report the average success rate across all 24 tasks.

Real-World Experiments Setup. We further evaluate our P-WAM on a set of real-world manipulation tasks covering diverse manipulation skills, as summarized in Tab. 3, with representative task snapshots shown in Fig. 4. The tasks span pick-and-place, long-horizon manipulation, articulated-object interaction, deformable-object manipulation, and contact-rich control. They also evaluate instruction following with respect to object instances, colors, and spatial positions. We collect approximately 1,000 human-teleoperated trajectories in total for real-world training. During evaluation, each method is tested for 25 trials per task under the same robot setup and task-specific success criteria. We compare P-WAM with Cosmos Policy (Kim et al., 2026) and π_0 (Black et al., 2024), and report task success rates.

Table 3: **Real-world manipulation experiment results.** We evaluate P-WAM on diverse real-world tasks, and compare against Cosmos Policy and π_0.

	Task	Cosmos Policy	π_0	P-WAM
Pick-and-place	Lift the cube with the specified color	84.0	68.0	92.0
	Pick the specified object and place into the pot	36.0	28.0	68.0
Long-horizon	Open the pot lid and pour water with the kettle	64.0	72.0	64.0
Articulated object	Open the drawer at the specified level	72.0	60.0	76.0
	Close the drawer at the specified level	76.0	44.0	92.0
Deformable object	Fold the tablecloth	72.0	80.0	72.0
Contact-rich	Wipe the whiteboard with an eraser	44.0	36.0	52.0
Average		64.0	55.4	**73.7**

Table 4: **Ablation studies of P-WAM on RoboCasa.** We ablate the visual encoder choice, visual encoder fine-tuning strategy, the number of DiT layers injected with P-WAM tokens, the token type that attends to the perception tokens, and the perception injection strategy.

(a) Encoder

Encoder	Avg. SR
SigLIP	67.7
DINOv3	67.1
SigLIP + DINOv3	**68.3**
Qwen3-VL-8B	67.4

(b) Fine-tuning

SigLIP	DINOv3	Avg. SR
		66.5
	✓	66.8
✓		67.4
✓	✓	**68.3**

(c) Injection Stage

Last N Blocks	Avg. SR
28	66.8
20	67.4
12	**68.3**
4	66.8

(d) Attending Tokens

Perception attended by	Avg. SR
Action tokens only	67.5
Video tokens only	67.5
Video + Action (joint)	**68.3**

(e) Injection Strategy

Strategy	Avg. SR
Concatenation with T5	67.3
Alternating (ours)	**68.3**

Representation Probing Experiment Setup. To evaluate whether frozen video-action DiT representations encode task-relevant object-level semantic and fine-grained spatial information, we formulate representation probing as a foreground segmentation task. The probe is trained to segment the union of the task-relevant interacted objects and the robot gripper. The task-dependent foreground requires distinguishing manipulation-relevant entities while preserving their spatial extent. The probe reads only the video tokens of the DiT self-attention sequence; perception tokens enter as external cross-attention context and are never read directly, so any gain reflects information written into the shared video-action representation. We exclude the substantially easier wrist-view results and report foreground mIoU averaged across tasks on the two more challenging third-person left and right views, where task-relevant objects are typically smaller and more visually cluttered. This setup provides a stringent test of how strongly frozen video-action representations encode task-relevant object-level semantics and fine-grained spatial information.

Concretely, we use a lightweight segmentation probe consisting of two convolutional layers with a GELU activation followed by PixelShuffle upsampling, with 9.7M parameters in total. Only the probe is trained, while the WAM backbone remains frozen, ensuring that the evaluation reflects information already encoded in the video-action DiT representations rather than end-to-end adaptation. We train and evaluate the probe on our re-rendered RoboCasa segmentation data, whose camera viewpoints and background textures differ from the Cosmos Policy preprocessed data used for policy training. For each task, we use only 10 videos for probe training and reserve the remaining videos for testing. We extract probe features at a fixed relatively high diffusion noise level, following the same noise-scaling rule as the Cosmos Policy training configuration.

Finding 1: Representation Quality. Video-action DiT representations in WAMs encode task-relevant perceptual information, and this information can be further strengthened by our P-WAM. As shown in Fig. 3, we observe that the 25th DiT layer yields the best probing performance and therefore report its nearby middle-to-late layers for comparison. Across these competitive layers, P-WAM consistently improves average foreground mIoU over Cosmos Policy. Since the WAM backbone is

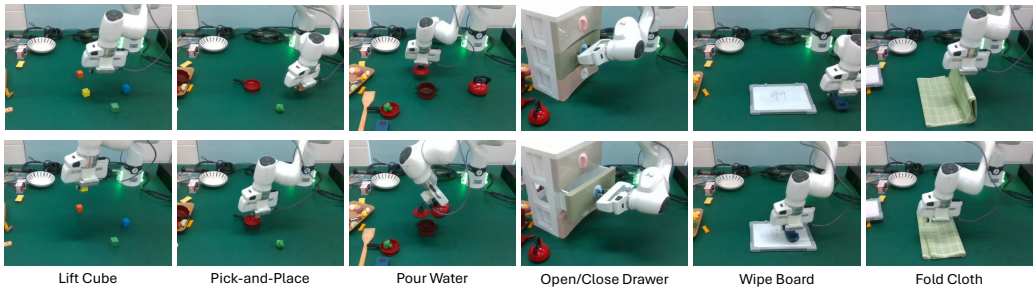


Figure 4: **Task examples used in our real-world robot evaluations.**

frozen and only the lightweight segmentation probe is trained, higher foreground mIoU indicates that task-relevant interacted objects and gripper regions are more accurately decodable from the underlying video-action representations. This suggests that integrating discriminative visual perception into generative world-action modeling efficiently and effectively improves the object-level semantic and fine-grained spatial information encoded in video-action DiT features.

World modeling is preserved. Because P-WAM only adds an external perception conditioning pathway and leaves the VAE latent pathway, denoising targets, and generative objective untouched, we next verify that this added perceptual supervision does not degrade the model's generative world-modeling ability. We evaluate future-frame prediction on the same held-out RoboCasa evaluation set and report the results in Tab. 1: P-WAM attains an SSIM of 0.784 and an FID of 23.5, matching the SSIM of Cosmos Policy (0.784) while slightly improving its FID (23.7). The generative branch is therefore on par with, and in terms of distributional fidelity marginally better than, the baseline it is built upon. This is consistent with observations in representation-enhanced generation (Yu et al.), where stronger semantic perception and generation quality reinforce rather than trade off against each other: the improved object-level and spatial representations are not obtained at the expense of the future-prediction capability, but accompany it.

Finding 2: Policy Performance. Improved perceptual representations translate into stronger policy performance. As shown in Tab. 2, P-WAM achieves an average success rate of 68.3% on RoboCasa using only 50 human-teleoperated demonstrations per task, improving over Cosmos Policy by 2.5 percentage points and Cosmos Policy + World2Act by 2.0 percentage points under the same data setting. This improvement is stable rather than seed-specific: across five random seeds, P-WAM outperforms Cosmos Policy on every seed, raising the average success rate from 65.7% to 67.7% (the App. A), with a paired t-test of $t = 8.16$, $p < 0.01$, and a 95% confidence interval of [1.3, 2.7]. It is also not an artifact of the extra optimization P-WAM receives by continuing from the Cosmos Policy checkpoint: continuing to train Cosmos Policy itself for the same additional steps reaches only 65.6%, below the 65.8% released checkpoint we compare against, and replacing the pretrained perception encoders with randomly initialized encoders of identical architecture under the same two-stage schedule reaches 66.3%, comparable to the baseline and 2.0 points below P-WAM. P-WAM also outperforms representative VLA-based and WAM-based methods, many of which use substantially larger training datasets, suggesting that integrating visual perception into WAMs improves policy learning in a data-efficient manner. We observe particularly clear gains on the most challenging pick-and-place tasks in RoboCasa, where object-level semantics and fine-grained pose detection are critical; detailed per-task results are provided in the App. A. We further evaluate P-WAM on real-world robot manipulation tasks against Cosmos Policy and π_0, with results reported in Tab. 3. All three methods are fine-tuned using the same human-teleoperated dataset and evaluated under the same multi-task protocol. Under this setting, P-WAM improves five out of seven tasks across object pick-place, articulated-object manipulation, and contact-rich control, achieving the highest average success rate of 73.7% compared with Cosmos Policy (64.0%) and π_0 (55.4%).

Finding 3: Design Ablations. Tab. 4 shows that the performance of P-WAM is affected by several key design choices. **(1) Encoder choice.** As shown in Tab. 4(a), combining SigLIP and DINOv3 achieves the best average success rate of 68.3%, outperforming SigLIP alone (67.7%) and DINOv3 alone (67.1%). This indicates that language-aligned semantic features and dense spatial representations provide complementary benefits. Natural to ask is whether the reconstruction-oriented latent could simply be replaced by a semantic one (Zheng et al., 2025a); it reaches only 47.9%, below both P-WAM (68.3%) and the VAE baseline (65.8%), because a semantic latent optimized for discrimi-

native structure carries less of the low-level detail. **(2) Fine-tuning strategy.** As shown in Tab. 4(b), jointly fine-tuning both encoders achieves the best performance (68.3%), while freezing both encoders is clearly less effective (66.5%). Interestingly, partial fine-tuning of the dual-encoder model is also suboptimal: fine-tuning only SigLIP obtains 67.4%, lower than the SigLIP-only model in Tab. 4(a) (67.7%); similarly, fine-tuning only DINOv3 obtains 66.8%, lower than the DINOv3-only model (67.1%). This suggests that when both encoders are used, they should be adapted jointly; otherwise, the frozen branch may provide less well-aligned features that reduce the benefit of the trainable branch. **(3) Injection depth.** As shown in Tab. 4(c), injecting perception tokens into the last 12 of 28 DiT blocks achieves the best result (68.3%), outperforming injection into the last 4 blocks (66.8%) and broader injection into the last 20 or 28 blocks (67.4% and 66.8%). This indicates that intermediate-to-late injection provides a better trade-off between introducing perceptual guidance and preserving the pretrained world-action dynamics knowledge. **(4) Attending tokens.** Because the video-action backbone is a shared DiT that processes both video and action tokens within the same self-attention sequence, perception tokens can in principle be attended to by either token type. As shown in Tab. 4(d), restricting cross-attention to action tokens only or to video tokens only both yield 67.5%, below the joint design (68.3%). This indicates that the shared video-action DiT benefits from allowing both token types to attend to the perception context, rather than routing discriminative perception through a single pathway. **(5) Injection strategy.** Alternating the cross-attention context between T5 and perception tokens outperforms concatenating the two into a single context (68.3% vs. 67.3%), as shown in Tab. 4(e). Concatenation keeps the conditioning format consistent with pretraining, but it also lengthens the context and risks having the newly introduced perception tokens dominated by the well-established T5 language context, weakening the attention allocated to perception. Alternating instead gives each modality dedicated layers, so that the perceptual guidance is not diluted.

Finding 4: VLM Transfer. P-WAM can benefit from strong pretrained VLMs without directly fine-tuning the VLM backbone. As shown in Tab. 4(a), replacing the SigLIP and DINOv3 perception encoders with Qwen3-VL-8B still achieves a competitive average success rate of 67.4% (vs. 65.8% for Cosmos Policy), demonstrating that P-WAM can transfer useful perceptual priors from a pretrained VLM into a WAM. In this variant, we keep the Qwen3-VL-8B (Bai et al., 2025a) backbone frozen and introduce a small set of learnable query tokens to read out task-relevant visual-language representations from the VLM. This result highlights the generality of the P-WAM framework: the perception pathway is not tied to a specific encoder architecture, but can incorporate different discriminative perception sources. Notably, this variant does not yet surpass the SigLIP+DINOv3 instantiation (68.3%), suggesting that a frozen VLM read out by a few query tokens provides less spatially dense supervision than dedicated perception encoders; how to better exploit VLM reasoning within a WAM remains an open direction.

5 Limitations and Discussion

P-WAM demonstrates that introducing discriminative visual perception into WAMs can improve both video-action DiT representations and closed-loop policy performance, while preserving the original generative world-action modeling pathway. This suggests a simple and effective direction for building perception-enhanced WAMs. In this work, we intentionally keep the perception pathway lightweight, using grid pooling, simple feature fusion, and a transformer connector to isolate the effect of adding pretrained visual perception features. We do not explore more complex perception modules, such as multi-scale feature fusion, object-centric feature extraction, or temporal perception memory. These designs may further strengthen the interaction between visual perception and world-action modeling, and we leave them to future work.

6 Conclusion

In this paper, we quantitatively demonstrate that representative WAMs still have limitations in capturing task-relevant object-level semantics and fine-grained spatial information. To address this, we have proposed P-WAM, a simple and general framework that integrates discriminative visual perception with generative world-action modeling. Through representation probing, RoboCasa simulation experiments, and real-world robot evaluations, we show that P-WAM improves video-action DiT representations and leads to stronger closed-loop policy performance. Additionally, our VLM transfer results further suggest that P-WAM offers a promising path toward unifying VLA-style perceptual reasoning with WAM-style generative dynamics modeling.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of parts of the software code and with setting up and configuring the code execution environment. We did not use generative AI tools for research ideation, hypothesis formulation, experimental design, data construction or analysis, interpretation of experimental results, or autonomous execution of experiments, and AI-assisted data collection and AI-assisted peer review are not applicable to this work. Additionally, we used generative AI tools for language polishing and LaTeX formatting.

We reviewed all AI-assisted work. All AI-assisted code and environment configurations were inspected, tested, and verified by the authors, and all AI-assisted text was reviewed and revised before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects and requires no IRB approval. All experiments are conducted on public simulation benchmarks or in controlled laboratory environments under standard safety protocols, and we do not anticipate additional societal, dual-use, or fairness risks beyond those common to robotic manipulation research.

REPRODUCIBILITY STATEMENT

We describe the model architecture, training schedule, hyperparameters, and evaluation protocol in Sec. 3, App. B, App. C and App. D. The simulation benchmark, RoboCasa, is publicly available, and we follow the Cosmos Policy data preprocessing and evaluation protocol without modification. All reported simulation results use the same evaluation protocol and are additionally reported across five random seeds in App. A to allow assessment of stability. Real-world task descriptions, demonstration counts, evaluation splits, and success criteria are provided in App. D. Overall, our training procedure is fully reproducible under the same execution environment; simulation evaluations may exhibit only minor numerical variations due to the underlying physics engine.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Bharadhwaj, Homanga and Dwibedi, Debidatta and Gupta, Abhinav and Tulsiani, Shubham and Doersch, Carl and Xiao, Ted and Shah, Dhruv and Xia, Fei and Sadigh, Dorsa and Kirmani, Sean. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

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, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025a.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025b.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025a.

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. In *Conference on Robot Learning*, pp. 1004–1029. PMLR, 2025b.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.

ByungOk Han, Jaehong Kim, and Jinhyeok Jang. A dual process vla: Efficient robotic manipulation leveraging vlm. *arXiv preprint arXiv:2410.15549*, 2024.

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

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

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning*, pp. 24328–24346. PMLR, 2025.

Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, et al. Bagelvla: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, et al. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. In *Conference on Robot Learning*, pp. 5170–5194. PMLR, 2025.

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

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

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

Myungkyu Koo, Daewon Choi, Taeyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. Hamlet: Switch your vision-language-action model into a history-aware policy. *arXiv preprint arXiv:2510.00695*, 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*, 2026.

Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, et al. Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. In *9th Annual Conference on Robot Learning*.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025a.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025a.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, et al. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025b.

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025a.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Yi Liu, Sukai Wang, Dafeng Wei, Xiaowei Cai, Linqing Zhong, Jiange Yang, Guanghui Ren, Jinyu Zhang, Maoqing Yao, Chuankang Li, et al. Unified embodied vlm reasoning with robotic action via autoregressive discretized pre-training. *arXiv preprint arXiv:2512.24125*, 2025b.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025c.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *RSS 2024 Workshop: Data Generation for Robotics*.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

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. *arXiv preprint arXiv:2501.09747*, 2025.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, 38:95597–95621, 2026.

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.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

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

An Dinh Vuong, Tuan Van Vo, Abdullah Sohail, Haoran Ding, Liang Ma, Xiaodan Liang, Anqing Duan, Ivan Laptev, and Ian Reid. World2act: Latent action post-training via skill-compositional world models. *arXiv preprint arXiv:2603.10422*, 2026.

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.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025a.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, et al. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026.

Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11089–11099, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

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

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 *The Thirteenth International Conference on Learning Representations*.

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

Andy Zhai, Brae Liu, Bruno Fang, Chalse Cai, Ellie Ma, Ethan Yin, Hao Wang, Hugo Zhou, James Wang, Lights Shi, et al. Igniting vlms toward the embodied space. *arXiv preprint arXiv:2509.11766*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Letian Zhang, Sucheng Ren, Yanqing Liu, Xianhang Li, Zeyu Wang, Yuyin Zhou, Huaxiu Yao, Zeyu Zheng, Weili Nie, Guilin Liu, et al. Openvision 3: A family of unified visual encoder for both understanding and generation. *arXiv preprint arXiv:2601.15369*, 2026a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026b.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, et al. Do world action models generalize better than vlas? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026c.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning*, pp. 61229–61245. PMLR, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025a.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, et al. Flare: Robot learning with implicit world modeling. In *Conference on Robot Learning*, pp. 3952–3971. PMLR, 2025b.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

APPENDIX

A ROBOCASA RESULTS ANALYSIS

A.1 IMPROVEMENT STABILITY ANALYSIS

Table 5: RoboCasa results across different random seeds.

Seed	Cosmos Policy	P-WAM (Ours)
191	67.4	**69.6**
192	65.4	**67.3**
193	64.3	**66.6**
194	65.4	**66.5**
195	65.8	**68.3**
Average	65.7	**67.7**

In the main paper, we report the RoboCasa result using seed 195, following the seed used in the Cosmos Policy (Kim et al., 2026). To further examine the stability of the improvement, we additionally evaluate Cosmos Policy and P-WAM under five random seeds from 191 to 195. As shown in Tab. 5, P-WAM consistently outperforms Cosmos Policy across all five seeds, improving the average success rate from 65.7% to 67.7%. These results indicate that the performance gain of our P-WAM is stable across different seeds.

A.2 IN-DEPTH PERFORMANCE ANALYSIS

To better understand where the overall improvement comes from, we further group the RoboCasa tasks into four categories according to their dominant manipulation requirements. Pick-and-place includes the eight PnP tasks, CoffeeSetupMug, and CoffeeServeMug; articulated-object interaction includes opening and closing single or double doors, opening and closing drawers, and operating the sink faucet or spout; twisting knobs includes TurnOnStove and TurnOffStove; pressing buttons includes CoffeePressButton, TurnOnMicrowave, and TurnOffMicrowave. As shown in Tab. 6, P-WAM achieves clear gains on pick-and-place tasks (+4.6 points), twisting-knob tasks (+5.0 points), and button-pressing tasks (+2.7 points). These tasks require accurate object recognition, target localization, and fine-grained spatial alignment, suggesting that the improved object-level semantics and spatial representations introduced by the discriminative perception pathway translate into stronger policy performance. On articulated-object interaction tasks, P-WAM performs comparably to Cosmos Policy, with a small difference of -0.4 points.

Table 6: Success rate (%) comparison grouped by RoboCasa task category. Δ is computed as P-WAM minus Cosmos Policy.

Category	# Tasks	P-WAM (Ours)	Cosmos Policy	Δ
Pick-and-place	10	55.2	50.6	+4.6
Articulated-object interaction	9	82.0	82.4	-0.4
Twisting knobs	2	38.0	33.0	+5.0
Pressing buttons	3	91.3	88.7	+2.6
Average	24	68.3	65.8	+2.5

B MODEL ARCHITECTURE DETAILS

Our P-WAM builds on the Cosmos Policy (Kim et al., 2026) diffusion transformer and enhances it with a discriminative perception pathway. The Wan-VAE (Wan et al., 2025) video tokenizer is utilized to encode spatiotemporal visual latents. The resulting latent sequence is interpreted with a structured slot layout, where slots correspond to a blank placeholder, current proprioception, current

multi-view observations, action chunks, future proprioception, future multi-view observations, and value estimation. These latent tokens are processed by a 28-layer Video-Action DiT operating with a hidden dimension of 2048 and 16 attention heads.

The discriminative perception pathway operates concurrently on the current multi-view observations. For each view, raw images are independently processed by twin foundation backbones: SigLIP-SO400M (Zhai et al., 2023) to capture language-aligned semantic features, and DINOv3-ViT-L/16 (Siméoni et al., 2025) to extract dense, fine-grained spatial representations. To standardize the feature resolution, the patch-level tokens emitted by both encoders are spatially pooled into a uniform $N \times N$ grid per view. These pooled features are subsequently concatenated channel-wise and mapped through a Fuse Projector, yielding exactly N^2 perception tokens for each individual camera view. Subsequently, a single-layer Transformer Connector contextualizes the aggregated these tokens to capture cross-view spatial interactions and projects them to the cross-attention dimension of the downstream Video-Action DiT.

While the original Cosmos Policy DiT baseline uses T5 (Raffel et al., 2020) language tokens as the cross-attention context throughout the video-action DiT, P-WAM introduces perception tokens through the existing cross-attention interface in the intermediate-to-late stages of the backbone. Concretely, in the 28-layer video-action DiT, the first 16 blocks retain the original Cosmos Policy conditioning scheme and attend to T5 tokens. In the last 12 blocks, we alternate the cross-attention context between the original T5 language tokens and the perception tokens produced by the discriminative perception pathway. As a result, six of the last twelve blocks attend to perception tokens, while the remaining six continue to attend to language tokens. This design injects object-level semantic and fine-grained spatial cues without changing the DiT self-attention token sequence, denoising targets, or unified video-action latent layout. The modification is therefore confined to the external conditioning context supplied to selected cross-attention layers. Taking the 3-view RoboCasa (Nasiriany et al.) configuration as a concrete example, the Transformer Connector receives $N^2 \times 3$ perception tokens in total. Tab. 7 summarizes the tensor shapes through each module of P-WAM under this setup, and Tab. 8 lists the key architectural hyperparameters.

Table 7: Tensor shapes through the P-WAM model architecture (under the 3-view RoboCasa setup, where B denotes batch size).

Stage / Module	Shape
Wan-VAE Input	$(B, 3, 41, 224, 224)$
Wan-VAE Output	$(B, 16, 11, 28, 28)$
DiT Visual Tokens	$(B, 11, 14, 14, 2048)$
Visual Encoder Input	$(B, 3, 224, 224, 3)$
Perception Tokens	$(B, 192, 1024)$
T5 Context Tokens	$(B, 512, 1024)$
DiT Output Latents	$(B, 16, 11, 28, 28)$

C Simulation Experimental Details

C.1 RoboCasa Training Details

We train P-WAM on the RoboCasa manipulation benchmark (Nasiriany et al.) following the Cosmos Policy (Kim et al., 2026) RoboCasa setup. The training set contains 24 static kitchen manipulation tasks with 50 human-teleoperated demonstrations per task, and we do not use any additional generated demonstrations. The tasks span pick-and-place, cabinet and drawer manipulation, stove and sink interaction, coffee-machine use, and microwave operation. Each training sample contains synchronized multi-view RGB observations, robot proprioception, a language instruction, and the corresponding robot action trajectory. All RGB observations are resized to 224×224 and augmented with random resized cropping, small in-plane rotation, and color jittering. Proprioceptive states and actions are normalized using empirical statistics computed from the training demonstrations.

We keep the original Cosmos Policy learning formulation unchanged. Besides action chunk prediction, the model is trained with the same future visual latent prediction, future proprioception

Table 8: Key architectural hyperparameters of the P-WAM model.

Component	Hyperparameter
DiT Layers	28
DiT Hidden Dimension	2048
DiT Attention Heads	16
DiT Cross-Attention Dimension	1024
Perception Injection Blocks	Last 12
Alternating Context Schedule	T5 / Perception
Fuse Projector MLP	$(d_{\text{SigLIP}} + d_{\text{DINO}}) \rightarrow 2048 \rightarrow 2048$
Connector Layers	1
Connector Hidden Dimension	2048
Connector Attention Heads	32
Connector Output MLP	$2048 \rightarrow 1024 \rightarrow 1024$

prediction, and value prediction objectives as the baseline. Therefore, P-WAM uses the same data, video-action latent layout, and objectives as Cosmos Policy; the only architectural difference is the added discriminative perception pathway and its alternating cross-attention conditioning in the last 12 DiT blocks.

Training is performed in two stages. In Stage 1, we initialize the video-action backbone from the Cosmos-Policy-RoboCasa-Predict2-2B checkpoint. The SigLIP-SO400M and DINOv3-ViT-L/16 perception encoders are frozen, while the Fuse Projector and Transformer Connector are trained to map pooled multi-view visual features into perception context tokens. On the DiT side, we fine-tune the last 12 DiT blocks together with the output projection layers, while keeping the remaining pre-trained policy parameters frozen. In Stage 2, we resume from the Stage-1 checkpoint and unfreeze the SigLIP and DINOv3 backbones for joint fine-tuning. The trainable DiT components remain the same as in Stage 1. Both stages use AdamW with a weight decay of 0.1. The learning rate is linearly warmed up from 1×10^{-6} to the stage-specific peak learning rate over the first 300 optimization steps, and then linearly decayed to 20% of the peak value. Training is conducted on a single node with 8 NVIDIA A800 GPUs. We use a per-GPU batch size of 16 and gradient accumulation over 6 steps, resulting in an effective global batch size of 768. The full schedule is summarized in Tab. 9.

Table 9: Two-stage RoboCasa fine-tuning schedule.

Setting	Stage 1	Stage 2
Initialization	Cosmos Policy checkpoint	Stage-1 checkpoint
SigLIP / DINOv3 encoders	Frozen	Trainable
Fuse Projector	Trainable	Trainable
Transformer Connector	Trainable	Trainable
Trainable DiT blocks	Last 12	Last 12
DiT output projections	Trainable	Trainable
Other policy parameters	Frozen	Frozen
Training steps	2,500	1,500
Peak learning rate	2×10^{-5}	6×10^{-6}
Final learning rate	4×10^{-6}	1.2×10^{-6}
Optimizer	AdamW, weight decay 0.1	
Warmup	300 steps from 1×10^{-6}	
Hardware	8 NVIDIA A800 GPUs	
Batching	16/GPU, gradient accumulation 6, effective batch 768	

C.2 RoboCasa Evaluation Details

We follow the Cosmos Policy evaluation protocol on the RoboCasa simulation benchmark. The evaluation covers 24 single-stage kitchen manipulation tasks using a single Franka Emika Panda

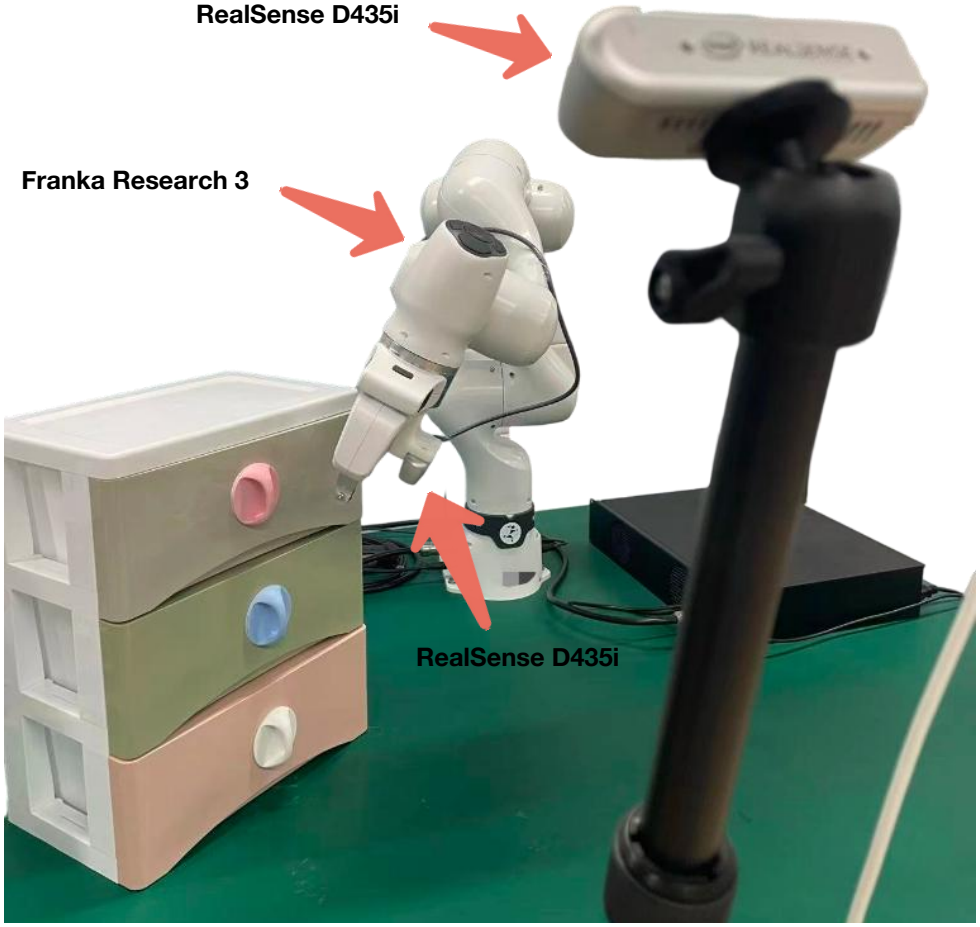


Figure 5: Overview of the real-world experimental workspace. The setup uses a Franka FR3 robot equipped with a Franka Hand two-finger gripper, with RGB observations captured from a wrist-mounted camera and a fixed third-person camera.

robot arm. For each task, we run 50 rollout trials across five evaluation scenes with different layouts and visual styles, assigning 10 trials to each scene. The evaluation uses held-out object instances, and two of the five evaluation scenes include visual styles that are unseen during training. The final metric is the mean task success rate averaged over all 24 tasks.

During evaluation, the policy receives the language instruction, proprioception, and three RGB camera views: left third-person, right third-person, and wrist. Images are rendered at 224×224 resolution. The policy predicts the 7-dimensional arm-and-gripper action used by the Franka robot. At each query, the model predicts a 32-step action chunk, and the first 16 actions are executed open-loop before querying the policy again. Episode success is determined by the official RoboCasa task-success checker. Each task is evaluated independently on a single NVIDIA RTX A6000 GPU.

D Real-World Experimental Details

D.1 Real-World Experimental Setup

We conduct real-world experiments on a 7-DoF Franka FR3 robot equipped with a Franka Hand two-finger gripper. An overview of the setup is shown in Fig. 5. The scene is observed by two Intel

RealSense D435i cameras: one wrist-mounted camera attached to the Franka Hand and one fixed third-person camera. Although the cameras provide RGB-D sensing, we only use RGB images in our experiments. The captured images are stored at a resolution of 480×640 and resized to 224×224 before being fed into the policy.

The policy operates in a Cartesian end-effector action space. Each action is represented as a 7-dimensional vector, consisting of relative end-effector translation, relative end-effector rotation, and a gripper action. The predicted action is executed through an operational-space pose controller at a control frequency of 5 Hz, which provides a practical trade-off between temporal abstraction and motion continuity for our real-world manipulation tasks.

D.2 Real-World Training Details

We train all real-world policies on the same set of human teleoperated demonstrations collected with a 3D SpaceMouse. The dataset contains 1,000 successful trajectories across seven real-world manipulation tasks, covering pick-and-place, long-horizon manipulation, articulated-object interaction, deformable-object manipulation, and contact-rich control. Tab. 10 summarizes the number of demonstrations for each task. Each trajectory contains synchronized two-view RGB observations, robot proprioception, actions, and a corresponding language instruction. The proprioception is represented as an 8-dimensional vector comprising the gripper width and the end-effector pose, where the pose is parameterized by 3D position and 4D quaternion orientation.

For image preprocessing, all RGB frames are resized to 224×224. During model forwarding, the image sequence used by the video VAE is scaled from $[0, 255]$ to $[-1, 1]$ before latent encoding. The additional SigLIP and DINOv3 visual encoders receive the same RGB observations, which are processed by the corresponding pretrained image processors for model-specific normalization. Actions and proprioception are normalized using statistics computed from the real-world training set. Specifically, each dimension is min-max scaled to $[-1, 1]$. During deployment, the current proprioception is normalized with the same statistics, and the predicted actions are unnormalized back to the original robot action scale before being sent to the controller.

For a fair comparison, Cosmos-Policy, π_0 (Black et al., 2024), and P-WAM are trained on the same real-world dataset and use the same action chunk length of 8. π_0 is trained for 10k steps with a batch size of 256, while Cosmos-Policy is trained for 10k steps with a batch size of 480. P-WAM is initialized from the trained Cosmos-Policy checkpoint and further trained with the same batch size of 480 in two stages: the SigLIP and DINOv3 visual encoders are frozen during the first 1.5k steps, and then unfrozen for another 1k steps. Cosmos Policy converged under this data and schedule: its real-world success rate plateaued by approximately 8k steps and did not improve further through 10k. All real-world policies are trained on one 8-GPU A6000 node.

Table 10: Real-world demonstration dataset used for policy training.

Task	# Demonstrations
Lift the cube with the specified color	200
Pick the specified object and place it into the pot	150
Open the pot lid and pour water with the kettle	200
Open the drawer at the specified level	100
Close the drawer at the specified level	100
Fold the tablecloth	100
Wipe the whiteboard with an eraser	150
Total	1000

D.3 Real-World Evaluation Details

We evaluate all policies in the same real-world environment and on the same set of tasks. For a fair comparison, Cosmos-Policy, π_0, and P-WAM are evaluated using the same task set, initial-state distribution, language instructions, observation interfaces, action chunk length, and control

frequency. No human intervention is allowed during a trial. After each success or failure, the scene is reset before the next trial.

We report task success rate as the main evaluation metric. Each task is evaluated over 25 trials with varied initial states. The evaluation includes both in-distribution (ID) and out-of-distribution (OOD) scenes. ID scenes use backgrounds, object positions, and object orientations consistent with those observed in the training demonstrations, whereas OOD scenes use object placements and orientations not present in the training data. Both ID and OOD scenes include diverse distractor objects to evaluate robustness under cluttered real-world conditions. The detailed evaluation splits and success criteria are as follows:

- **Lift the cube with the specified color.** We use 10 ID trials and 15 OOD trials. A trial is considered successful if the robot identifies the cube with the specified color, grasps it, and lifts it clearly above the tabletop.
- **Pick the specified object and place it into the pot.** We use 15 ID trials and 10 OOD trials. A trial is considered successful if the robot identifies the object described by the instruction, grasps it, and places it into the pot.
- **Open the pot lid and pour water with the kettle.** We use 15 ID trials and 10 OOD trials. A trial is considered successful if the robot grasps the pot lid and places it on the table, then grasps the kettle, performs a clear pouring motion toward the pot.
- **Open the drawer at the specified level.** We use 15 ID trials and 10 OOD trials. A trial is considered successful if the robot identifies the drawer level specified by the instruction and clearly pulls out the corresponding drawer without remaining stuck for an extended period.
- **Close the drawer at the specified level.** We use 15 ID trials and 10 OOD trials. A trial is considered successful if the robot identifies the drawer level specified by the instruction and closes the opened drawer without remaining stuck for an extended period.
- **Fold the tablecloth.** We use 15 ID trials and 10 OOD trials. A trial is considered successful if the robot grasps one side of the tablecloth and successfully folds it over.
- **Wipe the whiteboard with an eraser.** We use 10 ID trials and 15 OOD trials. A trial is considered successful if the robot grasps the eraser and wipes the writing on the whiteboard clean.