BEYOND FUTURE PREDICTION: DENOISING AS GENERATIVE ADAPTATION FOR ROBOT CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained generative Diffusion Transformers (DiTs) capture rich pixel-level visual and language-conditioned structure through large-scale image and video generation training. A growing line of robot policies builds on this generative prior through generative visual and action co-training, commonly instantiated as future visual prediction, yet recent methods increasingly move this prediction out of the inference and keep it only for co-training. This shift leaves open a more basic question, what a pretrained generative DiT actually contributes to action learning, and how this prior should be adapted for control. We introduce **NowWAM**, a future-target-free co-training formulation that denoises the current observation and predicts robot actions from the same visual stream, directly coupling the native generative objective to the action-facing representation across the denoising trajectory. Under matched controlled settings, past and future visual targets perform comparably, whereas restricting training to the clean endpoint substantially reduces robustness. This suggests that a separate future target is not essential for generative adaptation, but the continuum of denoising states remains an effective interface for control. On LIBERO-Plus, NowWAM reaches 87.7% with FLUX2-Klein, improving over the future-target co-training baseline by 6.1 points while halving training visual tokens (784→392) and reducing step time from 2.85 s to 1.63 s, a 1.8× speedup. With the pure text-to-image Z-Image backbone, NowWAM still reaches 87.8%, confirming that strong control adaptation does not depend on video generation or image-editing backbones.

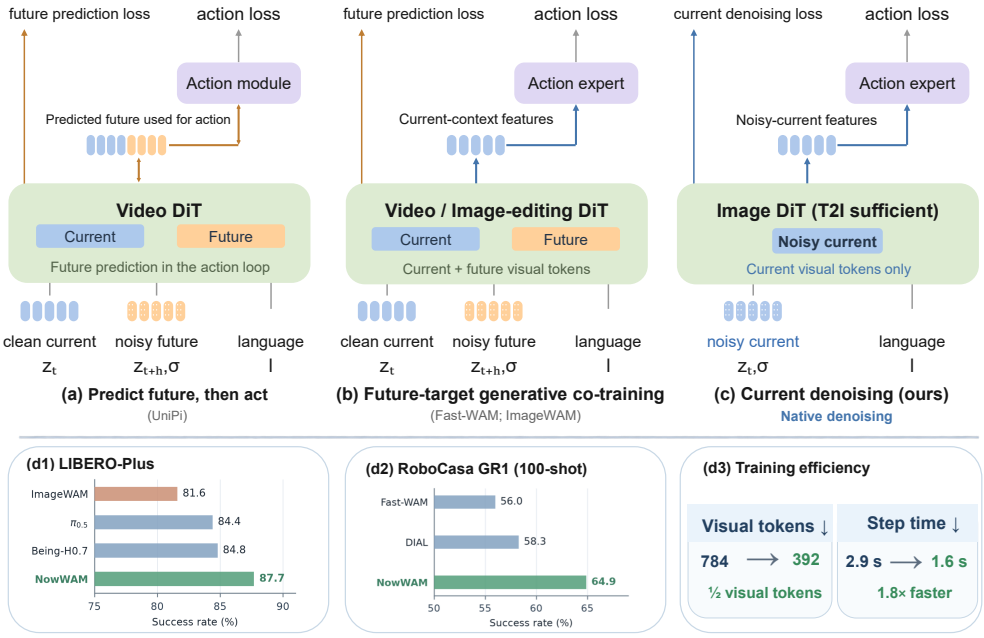


Figure 1: **From future prediction to current denoising.** (a) Future prediction is used for action prediction. (b) Future targets are used for generative co-training. (c) NowWAM denoises the current visual stream directly. (d) NowWAM improves robustness while reducing training cost.

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.

1 INTRODUCTION

Large-scale pretraining has become the foundation of modern robot policies. A representative type of work is vision-language-action models that inherit semantic representations from pretrained vision-language models and adapt them to action prediction (Kim et al., 2024; Black et al., 2024a; Bjorck et al., 2025; Physical Intelligence et al., 2025; Yuan et al., 2026a). Relatively speaking, another parallel line of work instead builds policies on pretrained generative Diffusion Transformers (DiTs), whose pretraining captures rich pixel-level visual and language-conditioned structure for generation. Yet how such generative priors should be transferred to control remains unclear. Existing generative policies commonly rely on future visual prediction, making it a prevailing interface between generative pretraining and action learning (Du et al., 2023; Wu et al., 2024; Ye et al., 2026).

However, the role of future prediction has progressively shifted in generative robot policies. As illustrated in Fig. 1, earlier approaches place predicted futures directly inside the action loop, while more recent methods remove future generation from deployment and retain it only for generative co-training (Du et al., 2023; Wu et al., 2024; Ye et al., 2026; Yuan et al., 2026b; Zhang et al., 2026b). This progression raises a basic question: if the future is no longer used for action inference, does predicting the future itself explain why generative co-training helps? Our controlled comparison offers a simple clue that under matched settings, past and future visual targets yield comparable robustness, so the benefit cannot be uniquely attributed to forward temporal semantics. This in turn motivates us to consider *what a pretrained generative DiT actually contributes to action learning and how this prior should be adapted for control*.

To answer this, we go back to the native denoising process of a generative DiT itself. We find that applying this denoising process directly to the visual stream used for control already provides a useful training signal for action learning, with no auxiliary target required. Motivated by this observation, we propose **NowWAM** (Fig. 2), which removes the separate temporal target and applies denoising directly to the current visual stream used for control. This makes the same visual stream serve both generative adaptation and action learning, rather than maintaining a separate auxiliary target. The action route itself is otherwise unchanged, and the key intervention lies only in where the generative objective is applied. Importantly, this is not simply a current-frame replacement for future-target co-training. The action-facing visual stream itself is sampled along the DiT's denoising trajectory and jointly optimized for generative velocity prediction and action learning. The action expert is therefore trained with visual-language context spanning a continuum of denoising states rather than only the clean endpoint, directly coupling the native generative objective to the representation used for control. This single-stream formulation also removes the additional target tokens used by future-target co-training. At inference, the same policy simply evaluates the clean endpoint in a single forward pass, with no auxiliary visual stream or visual denoising rollout. By reducing generative adaptation to this native interface, NowWAM can further extend to a pure text-to-image DiT, showing that temporal generative pretraining is also not required.

This simpler formulation is also substantially more robust under distribution shift. On LIBERO-Plus, NowWAM achieves 87.7% success with FLUX2-Klein and 87.8% with Z-Image, outperforming state-of-the-art VLA and generative-policy baselines in our comparison while maintaining near-saturated performance on standard LIBERO (Fei et al., 2025; Black Forest Labs, 2026; Cai et al., 2025; Liu et al., 2023). Its gains are concentrated on challenging visual perturbations, consistent with improved robustness of the current-state representation. The strong performance of the pure text-to-image Z-Image backbone further shows that the NowWAM formulation transfers beyond image-editing pretraining to a pure text-to-image DiT. Removing the target stream also halves the number of visual tokens during training (784→392), providing a simpler and more efficient interface between generative pretraining and action learning.

Our contributions are threefold:

- **We dissect how pretrained generative vision priors transfer to robot control.** Controlled analyses show that past and future targets perform comparably, while restricting adaptation to the clean endpoint of denoising process substantially reduces robustness.
- **We introduce NowWAM, a future-target-free generative co-training policy.** NowWAM applies current-frame denoising to the action-facing representation, coupling generative adaptation with action learning in a single stream.

- **We show that this simpler interface generalizes across benchmarks and architectures.** NowWAM improves performance on LIBERO-Plus and RoboCasa, extends from image-editing backbones to a pure text-to-image DiT, and halves the visual tokens used during joint training.

2 Related Work

Pretrained VLM backbones for robot control. Generalist robot policies increasingly adapt large-scale pretrained visual and language representations to action prediction, with modern VLA and diffusion-based foundation models scaling this paradigm across heterogeneous datasets, embodiments, tasks, and action spaces (O'Neill et al., 2024; Team et al., 2024; Brohan et al., 2023; Kim et al., 2024; Black et al., 2024a; Physical Intelligence et al., 2025; Bjorck et al., 2025; Liu et al., 2025; Liang et al., 2026; Wen et al., 2025; Kim et al., 2025; Zheng et al., 2026; Yang et al., 2026; Yuan et al., 2026a; Wang et al., 2026a). These backbones are primarily understanding-oriented, and their pretrained representations are reused for control without altering the underlying visual-language interface.

Generative backbones for robot control. Large generative models, by contrast, also learn transferable visual structure well beyond their native synthesis interface. Representations extracted from pretrained diffusion models support semantic correspondence, open-vocabulary recognition, dense matching, and geometry-aware understanding across architectures and denoising states, and have further been adapted to dense prediction tasks such as depth, geometry, and segmentation (Tang et al., 2023; Luo et al., 2023; Stracke et al., 2025; Xu et al., 2023; Hedlin et al., 2023; Ke et al., 2024; Fu et al., 2024; Kondapaneni et al., 2024; Peebles & Xie, 2023; He et al., 2025; Xu et al., 2025; Wang et al., 2026b; Jiang et al., 2026). These results suggest that the transferable value of pretrained representations is not tied to preserving their original output interface, motivating us to ask how generative representations should be reorganized when jointly adapted to action learning.

Pretrained generative models have been incorporated into robot control through interfaces such as generated visual subgoals, diffusion-derived visuomotor representations, joint visual-action denoising, predictive visual representations, and generative video backbones jointly adapted for control (Black et al., 2024b; Deng et al., 2026; Guo et al., 2024; Hu et al., 2024; Wu et al., 2024; Cheang et al., 2024; Kim et al., 2026b). Among these, future prediction, cast as a world model coupling future visual dynamics with action, is particularly prominent, spanning joint denoising of future observations and actions (Ye et al., 2026; Bi et al., 2026; Zhu et al., 2025; Wu et al., 2024; Cheang et al., 2024; Liang et al., 2025; Kim et al., 2026b; Cai et al., 2026), causal imagine-then-act inference from predicted futures (Du et al., 2023; Zhou et al., 2024; Feng et al., 2025; Li et al., 2026; Bharadhwaj et al., 2024; Zhao et al., 2025; Zhang et al., 2026a), and unified autoregressive world-action prediction (Hu et al., 2024; Cen et al., 2025), together forming the future-in-the-action-loop paradigm in Fig. 1(a), where deployment stays coupled to future prediction. Fast-WAM (Yuan et al., 2026b) instead moves future prediction out of deployment and retains it only as a training-time generative target, with ImageWAM (Zhang et al., 2026b) further reducing that target to a single conditional future image, defining the training-time future-target paradigm in Fig. 1(b). Unlike all these methods, which still rely on some form of future visual target to drive generative adaptation, NowWAM removes the future target altogether and applies the native denoising trajectory directly to the current action-facing representation, showing generative adaptation for control does not require a distinct temporal target.

3 Method

3.1 Generative Co-training with a Separate Visual Target

We study robot policies that jointly adapt a pretrained generative Diffusion Transformer (DiT) and an action predictor on robot demonstrations. Let l denote the language instruction, z_t the latent representation of the current observation, and a the action sequence. A common way to retain the pretrained generative training interface during robot adaptation is to introduce an additional visual target $\mathbf{z}_t^+$, typically corresponding to one or more future observations, and to optimize its generative objective together with action prediction.

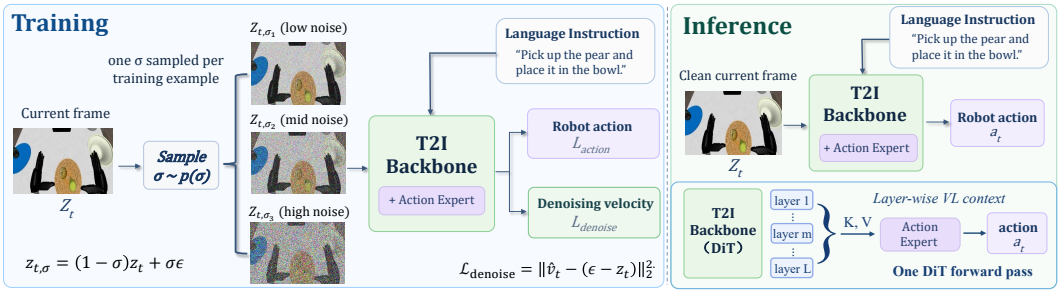


Figure 2: **NowWAM.** During training, the current visual stream is sampled along the pretrained denoising trajectory and jointly supports generative prediction and action learning. At inference, the same policy operates at the clean endpoint ($\sigma = 0$) in a single forward pass.

For a generative noise level σ, the target is sampled along the pretrained denoising trajectory,

$$\mathbf{z}^+_{t,\sigma} = (1-\sigma)\mathbf{z}^+_t + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I). \tag{1}$$

Conceptually, joint training operates on

$$[\, l \mid z_t \mid \mathbf{z}^+_{t,\sigma} \mid a \,]. \tag{2}$$

Here and below, the bracket notation denotes the conceptual training streams rather than literal concatenation into a single transformer. We adopt a Mixture-of-Transformers (MoT) architecture; the DiT-MoT interaction is described in Sec. 3.3. The current observation provides the visual context for action prediction, while the additional target stream provides the generative adaptation path. In the formulation we study, the action stream does not directly consume the target tokens; the target affects action learning through the shared generative backbone and its adaptation during co-training.

At deployment, the auxiliary target stream is absent. The policy therefore reduces to

$$[\, l \mid z_t \mid a \,], \tag{3}$$

where action prediction depends only on the language instruction and the current observation. This asymmetry raises a simple question: *if the additional visual stream serves primarily to adapt the pretrained generative representation during training, must the adaptation be organized around a distinct future target?*

3.2 NOWWAM

We introduce **NowWAM**, which applies the pretrained generative interface directly to the current visual stream instead of a separate visual target. We sample $u \sim \mathcal{U}(0, 1)$ and set

$$\sigma = \frac{su}{1 + (s-1)u}, \tag{4}$$

where $s = 1$ gives uniform sampling and $s < 1$ shifts samples toward the low-noise end of the denoising trajectory. The current latent is then sampled as

$$z_{t,\sigma} = (1-\sigma)z_t + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

and the corresponding velocity target is

$$v^*_t = \epsilon - z_t. \tag{6}$$

Training now contains only one visual stream,

$$[\, l \mid z_{t,\sigma} \mid a \,]. \tag{7}$$

The same trajectory-conditioned current representation supports both generative prediction and action learning. For each training example, we sample a single noise level σ and use the resulting $z_{t,\sigma}$ for denoising supervision and as visual context for the action expert. **The action expert reads**

language-conditioned context from this generatively adapted representation, rather than a vision-only feature stream. No additional clean visual stream or auxiliary target stream is retained during training. Unlike simply replacing the future target with a separate current-frame target, NowWAM applies the native denoising trajectory directly to the action-facing current representation.

Let $\hat{v}_t$ denote the predicted visual velocity and $\hat{a}$ the predicted robot action. We optimize

$$\mathcal{L} = \lambda_{\text{vis}}\mathcal{L}_{\text{denoise}} + \lambda_{\text{act}}\mathcal{L}_{\text{action}}, \tag{8}$$

with

$$\mathcal{L}_{\text{denoise}} = \|\hat{v}_t - (\epsilon - z_t)\|_2^2 . \tag{9}$$

Here, $\mathcal{L}_{\text{action}}$ is the masked mean-squared loss on the continuous action target, following the standard action supervision used by the policy. In our main configuration, $\lambda_{\text{vis}} = 0.5$ and $\lambda_{\text{act}} = 1.0$. Invalid action dimensions and padded action tokens are excluded from the loss.

3.3 Trajectory-Conditioned Training, Clean-Endpoint Control

The distinction between training and deployment is central to NowWAM. During training, the action expert reads visual-language context derived from $z_{t,\sigma}$, so action learning is coupled to representations spanning a continuum of denoising states. At inference, the same representation trajectory is evaluated at its clean endpoint:

$$\sigma = 0, \qquad z_{t,0} = z_t, \tag{10}$$

and the policy operates on the clean current observation:

$$[\, l \mid z_{t,0} \mid a \,] = [\, l \mid z_t \mid a \,]. \tag{11}$$

Backbone and action expert. Our policy follows a DiT–MoT structure with a pretrained generative DiT backbone and a separate action DiT. The generative backbone processes the language and visual streams and provides layer-wise visual-language context to the action expert, which keeps its own action tokens and parameters and interacts with this context through masked mixed attention. In the pure text-to-image instantiation, the backbone has no conditioning image or temporal input; the current observation is the visual latent being denoised, while language provides the external conditioning. NowWAM leaves the action pathway unchanged and modifies only the visual adaptation interface: during training, the current latent is sampled along the denoising trajectory, while deployment uses its clean endpoint. The action expert is therefore trained on visual-language representations from different generative states, while at inference it uses only the clean current observation in a single forward pass.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate NowWAM in three complementary settings. Standard LIBERO (Liu et al., 2023) serves as an in-distribution reference, where strong pretrained policies are already close to saturation, testing whether removing the separate future target preserves nominal manipulation capability. RoboCasa GR1 Tabletop (Nasiriany et al., 2024; Bjorck et al., 2025) evaluates few-shot transfer to a different simulator and task family; we use the 24 public ID tasks with 100 target-task demonstrations per task and 50 evaluation episodes per task, for 1,200 episodes in total. LIBERO-Plus (Fei et al., 2025) is our primary robustness benchmark, covering seven categories of visual and environmental perturbations.

Evaluation protocol. For the main benchmark results, we follow the full evaluation protocol for each benchmark, including all 10,030 LIBERO-Plus episodes, and we use the final EMA checkpoints for NowWAM. For controlled analyses, we evaluate the non-EMA checkpoints on the same fixed 1,923-episode LIBERO-Plus subset, so that the larger ablation suite remains tractable. Within each controlled comparison, all training and evaluation settings are held fixed except for the stated intervention. We therefore use the main results to compare final policy performance and the controlled analyses to isolate individual design choices.

Table 1: **Standard LIBERO and RoboCasa GR1 manipulation.**

(a) Standard LIBERO

Method	Spatial	Object	Goal	Long	Avg.
$\pi_{0.5}$	98.8	98.2	98.0	92.4	96.9
X-VLA	98.2	98.6	97.8	97.6	98.1
LingBot-VA	98.5	99.6	97.2	**98.5**	**98.5**
Motus	96.8	99.8	96.6	97.6	97.7
Fast-WAM	98.2	**100.0**	97.0	95.2	97.6
ImageWAM	97.2	99.2	**98.8**	98.4	98.4
NowWAM	**99.5**	**100.0**	97.5	96.5	98.4

(b) RoboCasa GR1 Tabletop

Method	Demos	SR (%)
GR00T-N1.6	1000	47.6
StarVLA-OFT	1000	48.8
ABot-M0	1000	58.3
RLDX-1	Full	58.7
DIAL	100	58.3
Fast-WAM	100	56.0
NowWAM	100	**64.9**

External methods use their respective training recipes; demonstration counts refer to target RoboCasa GR1 data. Our RoboCasa evaluation follows the public 24-task ID protocol with 50 episodes per task.

Backbones and training. We use FLUX.2-Klein-4B (Black Forest Labs, 2026) as our primary generative backbone and additionally evaluate NowWAM with Z-Image-6B (Cai et al., 2025), a pure text-to-image DiT, to further test whether the formulation depends on image-editing or video-generation pretraining. Unless otherwise stated, the generative backbone and action expert are jointly optimized.

Baselines. We compare against representative generalist VLA and diffusion-based policies, including π_0 and $\pi_{0.5}$ (Black et al., 2024a; Physical Intelligence et al., 2025), OpenVLA-OFT (Kim et al., 2025), X-VLA (Zheng et al., 2026), ABot-M0 (Yang et al., 2026), and GR00T (Bjorck et al., 2025), as well as recent generative and world-action policies including LingBot-VA (Li et al., 2026), Motus (Bi et al., 2026), Cosmos-Policy (Kim et al., 2026b), Fast-WAM (Yuan et al., 2026b), and ImageWAM (Zhang et al., 2026b). We additionally include StarVLA/StarVLA-OFT (StarVLA Community, 2026), Being-H0.7 (Luo et al., 2026), RLDX-1 (Kim et al., 2026a), and DIAL (Chen et al., 2026) where reported for the corresponding benchmark.

4.2 In-Distribution and Few-Shot Manipulation

Before studying robustness under distribution shift, we first evaluate whether NowWAM preserves nominal manipulation performance and transfers beyond LIBERO with limited target-task data. Table 1 reports standard LIBERO and RoboCasa GR1 Tabletop results.

In-distribution manipulation. Standard LIBERO is close to saturation for strong pretrained policies. NowWAM reaches 98.4% average success, matching the future-target generative baseline and remaining competitive with the strongest reported policies. Thus, removing the separate future target does not compromise nominal manipulation performance.

Few-shot cross-benchmark transfer. RoboCasa provides a more challenging transfer setting under a different simulator and task family. Using only 100 target-task demonstrations per task, NowWAM reaches 64.9% success over 1,200 evaluation episodes. Under the same 100-shot target-data budget, this improves over DIAL (58.3%) and Fast-WAM (56.0%), providing evidence that the NowWAM formulation transfers beyond LIBERO. Methods trained with larger target-task budgets are included in Table 1 for reference.

4.3 Robustness under Distribution Shift

We next evaluate on LIBERO-Plus, our primary robustness benchmark, which stresses policies with seven categories of visual and environmental perturbations over 10,030 episodes and is substantially more discriminative than the near-saturated standard LIBERO setting. NowWAM reaches 87.7% success with FLUX2-Klein, improving the future-target generative baseline from 81.6% by 6.1 points. The gains are concentrated on challenging visual shifts, including camera, robot, and background perturbations, while language performance remains comparable, indicating that NowWAM primarily improves the robustness of the current visual representation. With the pure T2I Z-Image

Table 2: **Success rates (%) on LIBERO-Plus across seven perturbation categories.**

Method	Perturbation category							**Overall**
	Camera	Robot	Lang.	Light	Bkg.	Noise	Layout	SR (%)
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$	78.4	**73.6**	80.8	96.2	94.1	89.0	84.5	84.4
StarVLA	52.5	49.8	88.5	95.7	**95.7**	73.0	76.9	74.1
OpenVLA-OFT	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
ABot-M0	60.4	67.9	86.4	96.2	91.6	86.4	82.6	80.5
Cosmos-Policy	75.8	63.3	81.7	96.5	88.9	92.7	82.2	82.2
Being-H0.7	82.0	59.0	82.8	97.8	90.0	93.5	**88.5**	84.8
Fast-WAM	15.5	43.4	67.1	79.3	52.4	39.3	60.0	49.5
ImageWAM (FLUX2-Klein-4B)	77.7	48.3	**89.9**	97.5	86.0	95.2	81.8	81.6
NowWAM (FLUX2-Klein-4B)	**88.6**	70.5	87.5	**98.3**	93.0	97.4	82.4	87.7
NowWAM (Z-Image-6B)	84.9	72.5	87.6	96.7	94.0	**97.5**	85.3	**87.8**

backbone, NowWAM further reaches 87.8%, demonstrating that the formulation also transfers to a pure text-to-image backbone.

Figure 3 visualizes representative execution failures under two challenging perturbations. Under RoboInit, the future-target baselines misidentify the grasp target and fail to transfer the instructed black bowl, whereas NowWAM successfully grounds, grasps, and places the target bowl. Under the camera-view perturbation, the baselines fail to push open the top drawer and therefore cannot complete the required interaction, while NowWAM successfully opens the drawer and completes the full placement sequence.

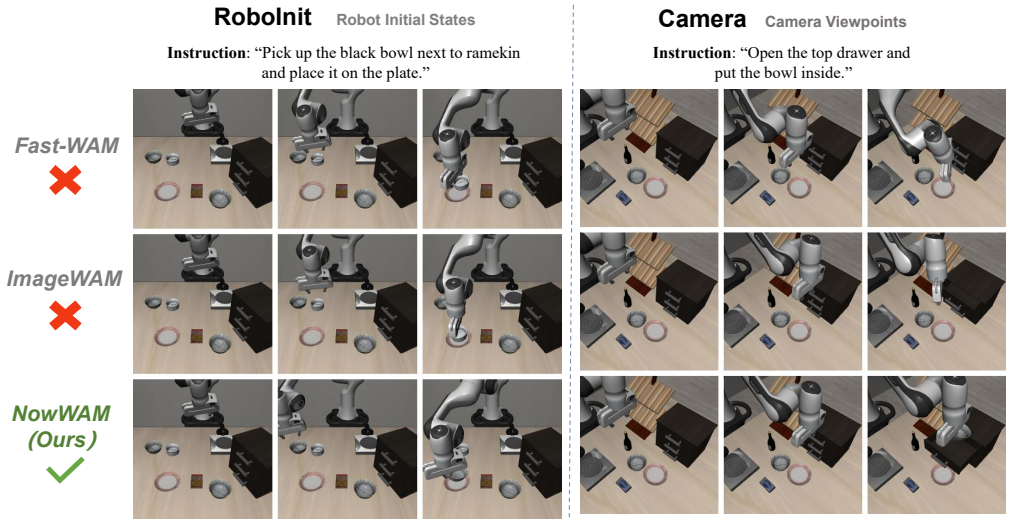


Figure 3: **Qualitative rollouts under distribution shift.** We show representative trajectories under RobotInit and camera-view perturbations in LIBERO-Plus. Fast-WAM and ImageWAM exhibit distinct grounding and execution failures, whereas NowWAM completes the corresponding tasks. These examples qualitatively illustrate the robustness gains observed in Table 2.

4.4 Ablations and Analysis

We conduct controlled experiments on LIBERO-Plus to distinguish three design choices in generative adaptation for control: generative initialization, the temporal semantics of the visual target, and

Table 3: **All rows are strictly matched; only the stated intervention changes.** Results use non-EMA checkpoints on the same fixed 1,923-episode subset.

(a) Generative initialization

Init.	Visual loss	SR (%)
Random	Edit	58.14
Pretrained	Edit	83.05
Random	None	50.55
Pretrained	None	75.81

(b) Auxiliary visual target

Target	SR (%)
Current target (t)	79.10
Future ($t + 16$)	82.89
Future ($t + 32$)	83.70
Past ($t - 16$)	83.80

(c) Denoising adaptation

Training condition	SR (%)
Uniform	83.46
Low-noise shifted	84.97
No visual loss	78.73
Clean only	77.48

training along the denoising trajectory. Table 3 summarizes the results. All rows follow the identical experimental contract described in Sec. 4.1, with only the stated intervention changed. All analyses use non-EMA checkpoints.

Generative initialization. Pretrained initialization substantially improves robustness both with and without a generative objective (Table 3a). With the Edit objective, pretraining raises success from 58.14% to 83.05%; even without the visual objective, it improves performance from 50.55% to 75.81%. This large gain without visual supervision shows that the pretrained generative backbone already provides a strong robustness prior, while the additional generative objective further improves adaptation on top of this initialization. These results establish generative pretraining as a major source of robustness while leaving open how that prior should be adapted to control.

Auxiliary visual targets. Changing only the temporal target shows that past and future targets perform comparably (Table 3b). Future targets at $t + 16$ and $t + 32$ reach 82.89% and 83.70%, while the past target at $t - 16$ reaches 83.80%. Thus, forward temporal prediction is not uniquely responsible for the benefit of the auxiliary generative target. The current-target control retains the same two-stream formulation with a separate auxiliary target, and is therefore distinct from NowWAM's single-stream current denoising.

Denoising trajectory. Although NowWAM consumes clean observations at inference, training only at the clean endpoint reaches 77.48%, substantially below training along the denoising trajectory (Table 3c). Here, *Clean only* fixes $\sigma = 0$ while retaining the visual objective, whereas *No visual loss* retains noisy-current training but removes visual supervision. Uniform sampling reaches 83.46%, while shifting the distribution toward lower noise further improves success to 84.97%. Removing the visual denoising objective reduces success to 78.73%. Thus, neither clean-endpoint training nor noisy-current exposure alone explains the gain. Effective adaptation therefore requires coupling action learning to the pretrained generative objective across the denoising trajectory, rather than only at the clean endpoint.

Taken together, these analyses separate three factors in generative adaptation for control. Pretrained generative initialization provides a strong robustness prior, while past and future auxiliary targets perform similarly, indicating that forward temporal semantics are not uniquely privileged. At the same time, both clean-only training and noisy-current training without visual supervision perform substantially worse than trajectory-conditioned denoising. These results suggest that the key ingredient is not future prediction itself, but retaining generative supervision on the action-facing representation across the native denoising trajectory.

Table 4: **Training efficiency.** Same 2×H200 setup with global batch size 64. Both methods use FLUX2-Klein-4B as the backbone.

Method	Training proxy	Vis. tokens ↓	Step time (s) ↓	Peak mem. (GiB) ↓
ImageWAM	Future image	784	2.85	78.9
NowWAM	**Current denoising**	**392 (50% fewer)**	**1.63 (1.8× faster)**	**74.3 (5.8% lower)**

4.5 Training Efficiency

We finally compare the training cost of the two formulations under the same 2×H200 setup with global batch size 64. Both measurements use the same FLUX2-Klein-4B backbone and action expert. Text embeddings and VAE latents are precomputed and cached, so the reported step time

and peak memory measure joint DiT–action training only and exclude text/VAE encoding. The difference therefore comes from the visual training interface: future-target co-training processes both the current reference and a separate target stream (784 visual tokens), whereas NowWAM keeps only the current stream (392 tokens).

As shown in Table 4, removing the second visual stream cuts step time from 2.85s to 1.63s ($1.8\times$) and peak memory from 78.9 to 74.3GiB. These gains come directly from the NowWAM formulation, without additional acceleration.

4.6 Attention Visualization

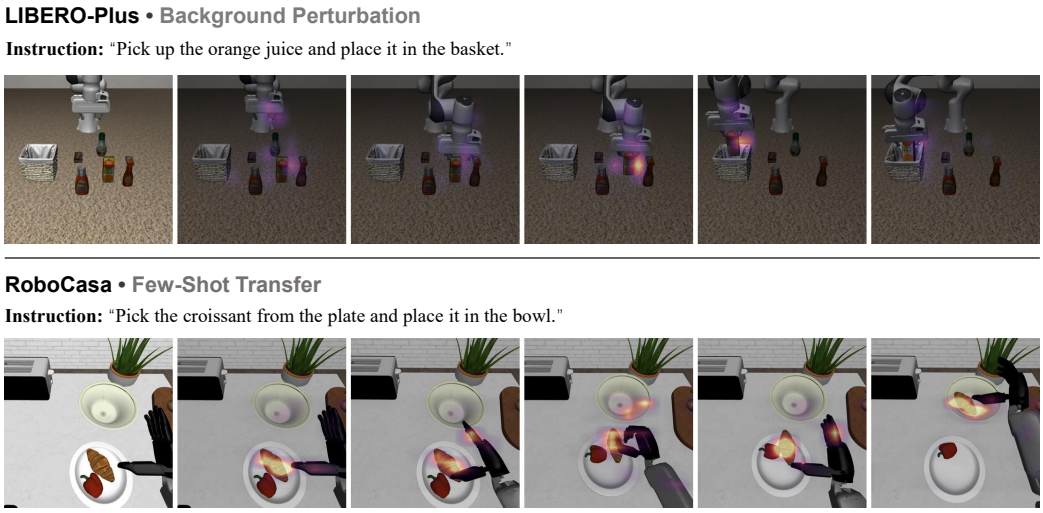


Figure 4: **Task-conditioned attention across robustness and transfer settings.** Each row starts with the RGB observation, followed by attention maps over the rollout.

Figure 4 visualizes task-conditioned attention over successful rollouts on LIBERO-Plus and RoboCasa. Across execution stages, attention focuses on task-relevant objects and interaction regions, qualitatively illustrating the representations learned by NowWAM.

5 Conclusion

Pretrained generative DiTs provide a rich source of visual and language-conditioned structure, yet how their native generative interface should be adapted to robot control remains unclear. Existing approaches commonly use future visual prediction for this adaptation, motivating us to ask whether forward temporal semantics are essential, or whether the generative trajectory itself can serve as a more direct interface to action learning. Our controlled studies show that past and future visual targets perform comparably, while restricting adaptation to the clean endpoint substantially reduces robustness, suggesting that future prediction is not uniquely privileged whereas the denoising trajectory remains important. Building on this finding, we introduced **NowWAM**, which applies native denoising directly to the current visual stream and removes the separate future-target branch. Across LIBERO-Plus and RoboCasa, NowWAM improves robustness and few-shot transfer, extends from image-editing to pure text-to-image backbones, and reduces training cost while retaining competitive in-distribution performance. An important limitation is that our study focuses on generative adaptation in manipulation, rather than settings where explicit future prediction is part of the control mechanism. Future modeling may remain useful for dynamics reasoning, planning, or long-horizon decision making, and how current-centric denoising interacts with these settings remains open. In particular, combining the simpler adaptation interface studied here with explicit predictive modules may be valuable when temporal reasoning is required. In the manipulation settings studied here, however, generative pretraining can be effectively adapted through a simpler, current-centric interface without requiring a distinct future target.

AI USE STATEMENT.

Generative AI tools assisted with language refinement and manuscript organization. The authors reviewed all AI-assisted material and take full responsibility for the final manuscript.

ETHICS STATEMENT.

This work does not involve human subjects, personally identifiable information, or other research activities requiring specific ethical approval.

REPRODUCIBILITY STATEMENT.

We provide detailed descriptions of the method and evaluation protocols in the main paper, with implementation details and hyperparameter settings in the appendix. We plan to release the code and configurations upon publication.

REFERENCES

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. 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. 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, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024a.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 33431–33452, 2024b.

Black Forest Labs. FLUX.2 [klein]: Towards Interactive Visual Intelligence. `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jisong Cai, Long Ling, Shiwei Chu, Zhongshan Liu, Jiayue Kang, Zhixuan Liang, Wenjie Xu, Yinan Mao, Weinan Zhang, Xiaokang Yang, et al. Aha-wam: Asynchronous horizon-adaptive world-action modeling with observation-guided context routing. *arXiv preprint arXiv:2606.09811*, 2026.

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.

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.

Yi Chen, Yuying Ge, Hui Zhou, Mingyu Ding, Yixiao Ge, and Xihui Liu. Dial: Decoupling intent and action via latent world modeling for end-to-end vla. *arXiv preprint arXiv:2603.29844*, 2026.

Yu Deng, Yufeng Jin, Xiaogang Jia, Jiahong Xue, Gerhard Neumann, and Georgia Chalvatzaki. Robot-dift: Distilling diffusion features for geometrically consistent visuomotor control. *arXiv e-prints*, pp. arXiv–2602, 2026.

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.

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.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

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.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89454–89467, 2025.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36:8266–8279, 2023.

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. *arXiv preprint arXiv:2412.14803*, 2024.

Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, Jingdong Wang, et al. Deforming videos to masks: Flow matching for referring video segmentation. In *International Conference on Learning Representations*, volume 2026, pp. 65586–65601, 2026.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502. IEEE, 2024.

Dongyoung Kim, Huiwon Jang, Myungkyu Koo, Suhyeok Jang, Taeyoung Kim, Beomjun Kim, Byungjun Yoon, Changsung Jang, Daewon Choi, Dongsu Han, et al. Rldx-1 technical report. *arXiv preprint arXiv:2605.03269*, 2026a.

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.

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*, 2026b.

Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13883–13893. IEEE, 2024.

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.

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*, 2025.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liuao Pei, Tian Nian, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies. In *Forty-third International Conference on Machine Learning*, 2026.

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.

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. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026.

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. *arXiv preprint arXiv:2406.02523*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 117–127. IEEE, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026a.

Zanyi Wang, Xin Lin, Haodong Li, Dengyang Jiang, and Yijiang Li. From rgb generation to dense field readout: Pixel-space dense prediction with text-to-image models. *arXiv preprint arXiv:2607.06553*, 2026b.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Guangkai Xu, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, Chunhua Shen, et al. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations*, volume 2025, pp. 6786–6799, 2025.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2955–2966. IEEE, 2023.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

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*, 2026.

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 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.

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, 2026a.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. Imagewam: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713. IEEE, 2025.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

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*, 2025.

A Appendix

Training details. Unless otherwise specified, all experiments use the shared optimization and training configuration summarized in Table 5.

Table 5: Default training hyperparameters used across experiments unless otherwise specified.

Hyperparameter	Value
Optimizer	AdamW
Adam coefficients (β_1, β_2)	$(0.9,\ 0.95)$
Peak learning rate	1×10^{-4}
Learning-rate schedule	5% warmup, then constant
Weight decay	0
Global batch size	64
Training epochs	9
Gradient clipping	1.0
EMA decay	0.9999
Visual loss weight λ_{vis}	0.5
Action loss weight λ_{act}	1.0
Noise-distribution shift s	0.25
Random seed	42

The noise sampling formulation and clean-endpoint inference setting are described in Sec. 3.2 and Sec. 3.3.