arXiv:2607.28993v1 [cs.RO] 31 Jul 2026

ST-WAM: Semantic-Temporal World Action Model for Robust Manipulation under Visual Distribution Shifts

Mingxin Wang[1,2], Bin Hu[1], Bin Qian[1], Kaitao Jiang[2], Haoning Wu[3], Feng Yan[4], Bowen Jing[5], Ruiyang Hao[6], Enyi Wang[1], Kangning Niu[2], Yandan Yang[2], Mu Xu[2], Yan Wang[1], Houde Liu[1]*, Tianlun Li[2]*

[1]Tsinghua University, [2]AMAP-CV-LAB, Alibaba Group
[3]Shanghai Jiao Tong University, [4]Xi'an Jiaotong University
[5]The University of Manchester, [6]King's College London

Abstract

World Action Models (WAMs) have emerged as a promising paradigm by jointly modeling robot actions and future visual dynamics. However, their reliance on pixel-generative future supervision can entangle action-relevant state transitions with task-irrelevant visual content, limiting robustness under visual distribution shifts. We identify **Training-Distribution Hallucination**, a recurring phenomenon in which futures conditioned on visually shifted observations hallucinate training-domain content rather than remain faithful to the current scene. A controlled frame-triplet diagnosis further shows that DINOv3 features remain more stable across visual shifts while better preserving task-state distinctions than Wan-VAE latents. Rather than correcting the predicted futures, we propose **Semantic-Temporal WAM (ST-WAM)** to improve action robustness by using DINOv3 as a shared semantic representation for future prediction and history retrieval while retaining fine-grained VAE dynamics. Its Dual-Space Future Experts (DSFE) jointly predict future VAE latents and DINO features, while Current-Anchored Intent Retrieval (CAIR) retrieves task-relevant evidence from recent DINO history under the current visual-language context. ST-WAM is trained end-to-end without additional embodied pretraining or task-specific annotations, and requires no explicit future generation at inference. It achieves 98.7% on LIBERO and 92.8% on RoboTwin 2.0; more importantly, compared with Fast-WAM, it improves zero-shot LIBERO-Plus performance by 21.3 percentage points and more than doubles real-world success under visual shifts from 25.8% to 61.5%. These results demonstrate that semantic-temporal modeling effectively complements pixel-generative dynamics for robust manipulation. The project page is available at https://thu-wangmx.github.io/st-wam/.

Introduction

Leveraging world-dynamics priors acquired through large-scale video pretraining, World Action Models (WAMs) (Bi et al. 2026; Kim et al. 2026) jointly model future visual states—typically in VAE latent spaces—and robot actions, providing a promising alternative to VLAs that directly map current observations to actions (Black et al. 2024; Kim et al. 2024; Physical Intelligence et al. 2025). Despite their strong performance on standard manipulation benchmarks, the robustness of video-generative WAMs under visual distribution shifts remains unclear.

*Corresponding author.

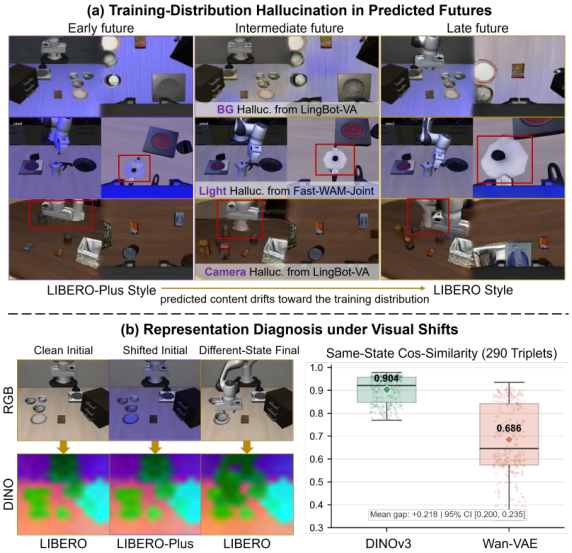

Figure 1: (a) Three representative Training-Distribution Hallucination cases from video-generative WAMs under LIBERO-Plus background-texture, illumination, and camera-viewpoint shifts. (b) A representative diagnostic triplet and a comparison of DINOv3 and Wan-VAE cosine similarities across same-state frames.

To investigate this question, we analyze two representative video-generative WAMs. As illustrated in Fig. 1(a), when LingBot-VA (Li et al. 2026b) and Fast-WAM-Joint (Yuan et al. 2026), both trained only on LIBERO, are evaluated zero-shot on LIBERO-Plus, their predicted videos progressively drift toward LIBERO-style content under perturbations to background textures, illumination, and other scene characteristics. We refer to this phenomenon as **Training-Distribution Hallucination**: when the current observation deviates from the training distribution, the predicted future hallucinates training-domain content rather than remaining faithful to the current scene. To assess its prevalence, we manually audit the predicted futures of both models on 30 randomly sampled cases under each of three visual shifts—background, illumination, and camera viewpoint—

totaling 180 predictions; 70.6% distinctly exhibit Training-Distribution Hallucination. Moreover, Fast-WAM and Fast-WAM-Joint drop from success rates of 97.6% and 98.5% on LIBERO to 51.5% and 59.0% on LIBERO-Plus, respectively (Zhang et al. 2026d). Taken together, the recurring hallucination pattern across two representative models and the substantial zero-shot performance drops reveal a clear robustness limitation of video-generative WAMs under visual distribution shifts.

These observations shift the central question from whether WAMs can predict the future to how the future should be represented for robust control. Many recent video-generative WAMs model future states in VAE latent spaces (Yuan et al. 2026; Li et al. 2026b; Ye et al. 2026b). Optimized primarily for visual reconstruction, these representations can entangle action-relevant transitions with task-irrelevant or hallucinated visual content. Recent studies have explored alternative future representations, including latent action, semantic masks, and spatial value maps (Chen et al. 2026a; Lou et al. 2026; Yu et al. 2026; Fan et al. 2026). However, existing approaches often rely on large-scale embodied pretraining, multi-stage training, or specialized pipelines for auxiliary supervision. We instead seek a semantic representation that can be extracted directly from raw observations while remaining stable under visual shifts and discriminative across task states. Large-scale self-supervised visual encoders, such as DINOv3 (Siméoni et al. 2025), provide a promising basis through their semantically structured features.

As illustrated in Fig. 1(b), we conduct a controlled representation diagnosis using 290 frame triplets from LIBERO and LIBERO-Plus. Each triplet contains two initial frames from the same task with identical robot and object states but different visual conditions, together with a final frame from the same LIBERO demonstration as a different-state reference. DINOv3 achieves an average cosine similarity of 0.904 between the same-state initial frames, compared with 0.686 for Wan-VAE latents. Moreover, when comparing the shifted initial frame with the other two frames, DINOv3 yields higher similarity to the state-matched clean frame than to the different-state final frame in 95.2% of triplets, versus 60.0% for Wan-VAE. DINOv3 therefore exhibits both stronger same-state stability under visual shifts and better different-state discriminability than Wan-VAE latents. Additional details are provided in the supplementary material.

These properties make DINOv3 suitable for two complementary temporal roles: as a future prediction target, it supervises task-relevant state evolution; as a history representation, it provides evidence of recent task progress when visual shifts make current-frame cues unreliable. Rather than correcting hallucinated future videos, we use these complementary semantic cues to improve action robustness.

In this paper, we propose **Semantic-Temporal WAM (ST-WAM)**, an end-to-end WAM that uses DINOv3 as a shared semantic representation across two complementary temporal directions. Prospectively, **Dual-Space Future Experts (DSFE)** jointly model future VAE latents and DINO features in a three-branch Mixture-of-Transformers with the action expert, coupling fine-grained visual dynamics with visually stable semantic transitions. Retrospectively, **Current-Anchored Intent Retrieval (CAIR)** uses the current visual-language context to retrieve task-relevant evidence from recent DINO history for action generation. ST-WAM requires no additional embodied pretraining, multi-stage training, or task-specific semantic annotations.

Extensive experiments across LIBERO, LIBERO-Plus, RoboTwin 2.0, and five real-world tasks demonstrate the effectiveness of ST-WAM while retaining efficient inference. Without embodied pretraining, ST-WAM achieves 98.7% on LIBERO and outperforms Fast-WAM by 21.3 percentage points under zero-shot transfer to LIBERO-Plus. It further achieves leading performance on RoboTwin 2.0 and more than doubles real-world success under visual distribution shifts from 25.8% with Fast-WAM to 61.5%.

Our main contributions are summarized as follows:

- We identify **Training-Distribution Hallucination** in video-generative WAMs, where predictions conditioned on visually shifted observations drift toward training-domain content, and provide a controlled frame-triplet diagnosis showing that, compared with VAE latents, DINOv3 offers greater same-state stability under visual shifts while better distinguishing different task states.
- We propose **ST-WAM**, which unifies semantic modeling across complementary prospective and retrospective directions: Dual-Space Future Experts (DSFE) jointly model future dynamics in VAE and DINO spaces, while Current-Anchored Intent Retrieval (CAIR) extracts task-relevant evidence from recent DINO history for action generation.
- Extensive evaluations in simulation and the real world demonstrate that, without additional embodied pretraining, ST-WAM maintains strong in-distribution performance while substantially improving robustness under visual distribution shifts.

Related Work

Vision-Language-Action Models

VLA models transfer semantic knowledge from pretrained vision-language models to directly map observations and instructions into robot actions (Kim et al. 2024; Black et al. 2024; Physical Intelligence et al. 2025, 2026; Yan et al. 2025; Du et al. 2026). Recent methods incorporate temporal context or future prediction: IntentVLA models short-horizon intent from recent history, while DreamVLA, VLA-JEPA, and DeFI introduce predictive objectives into VLA learning (Lian et al. 2026; Zhang et al. 2026b; Sun et al. 2026; Zhang et al. 2026c; Song et al. 2026). In contrast, we study how future and historical information should be represented within WAMs.

Video-Generative World Action Models

World Action Models jointly model future visual states and robot actions. Representative methods include DreamZero (Ye et al. 2026b), LingBot-VA (Li et al. 2026b), and Motus (Bi et al. 2026), which couple video prediction with action generation. Fast-WAM (Yuan et al. 2026) and GigaWorld-Policy (Ye et al. 2026a) omit explicit future

video generation during deployment to improve inference efficiency. Despite their different inference designs, their future supervision remains rooted in pixel-generative objectives, potentially entangling action-relevant dynamics with task-irrelevant visual factors.

WAMs with Alternative Future Representations

Beyond pixel-generative objectives, recent works explore semantic or spatially structured future representations. Some methods predict semantic masks (Yu et al. 2026; Lou et al. 2026), while others model geometric-semantic cues, spatial value maps, or compact latent conditions (Fan et al. 2026; Ma et al. 2026; Su et al. 2026; Luo et al. 2026; Li et al. 2026a; Liu et al. 2026; Zhang et al. 2026a). LDA-1B (Lyu et al. 2026) and LaWAM (Chen et al. 2026a) model future states directly in DINO feature space, but rely on large-scale embodied pretraining and multi-stage training. In contrast, ST-WAM retains fine-grained VAE dynamics while incorporating both DINOv3 future supervision and DINO-based history retrieval within an end-to-end WAM, achieving competitive performance without embodied pretraining.

Methodology

Problem Formulation

Given a current multi-view observation $\mathbf{o}_t$, proprioceptive state $\mathbf{s}_t$, and language instruction ℓ, the policy predicts an action chunk $\mathbf{a}_{t:t+H-1}$. We also use a short history $\mathcal{H}_t$ of M preceding observations and a future observation sequence $\mathbf{o}_{t+1:t+K}$. Let E_{VAE} and E_{DINO} denote the frozen Wan2.2 VAE (Wan et al. 2025) and DINOv3 encoders, respectively. We encode the observations into two complementary spaces:

$$\mathbf{z}^v = E_{\mathrm{VAE}}(\mathbf{o}_{t:t+K}), \quad \mathbf{z}^s = E_{\mathrm{DINO}}(\mathbf{o}_{t:t+K}). \tag{1}$$

For each $r \in \{v, s\}$, we partition $\mathbf{z}^r$ into current conditioning tokens $\mathbf{z}^r_{\mathrm{cur}}$ and future prediction targets $\mathbf{z}^r_{\mathrm{fut}}$, i.e., $\mathbf{z}^r = [\mathbf{z}^r_{\mathrm{cur}}; \mathbf{z}^r_{\mathrm{fut}}]$. During training, ST-WAM models the following conditional joint distribution:

$$p_\theta(\mathbf{a}_{t:t+H-1}, \mathbf{z}^v_{\mathrm{fut}}, \mathbf{z}^s_{\mathrm{fut}} \mid \mathbf{o}_t, \mathbf{s}_t, \ell, \mathcal{H}_t)\,. \tag{2}$$

At inference, ST-WAM reduces to an action-only policy for efficient deployment:

$$\mathbf{a}_{t:t+H-1} \sim \pi_\theta(\cdot \mid \mathbf{o}_t, \mathbf{s}_t, \ell, \mathcal{H}_t)\,. \tag{3}$$

Dual-Space Future Experts

Unified Visual–Semantic Future Modeling. DSFE models future states in complementary VAE-latent and DINO-feature spaces. The frozen VAE of Wan2.2-TI2V-5B (Wan et al. 2025) encodes the observation sequence into visual latents $\mathbf{z}^v$, while a frozen DINOv3 encoder (Siméoni et al. 2025) extracts dense, frame-wise semantic features $\mathbf{z}^s$. After modality-specific embedding, the pretrained Wan2.2 Video DiT models $\mathbf{z}^v_{\mathrm{fut}}$ to preserve the fine-grained visual dynamics inherited from video pretraining, whereas a semantic future DiT models $\mathbf{z}^s_{\mathrm{fut}}$ to learn semantic state transitions.

Three-Branch Mixture-of-Transformers. As shown in Fig. 2(a), the visual and semantic future DiTs, together with the action DiT, form a three-branch Mixture-of-Transformers (Liang et al. 2024). Each branch retains its own parameters and prediction head, while layer-wise mixed attention enables mutual refinement between the two future spaces and allows the action branch to integrate current evidence from both. During flow-matching training, the three experts jointly denoise future VAE latents, future DINO features, and action tokens, with branch-specific heads estimating flow velocities in their respective spaces.

Structured Cross-Branch Attention Mask. As illustrated in Fig. 3, we apply an asymmetric mask at each mixed-attention layer. Clean current VAE and DINO tokens interact within and across their spaces but cannot read future or action tokens, forming leakage-free anchors. The two noisy future streams attend to both current anchors and each other, enabling mutual refinement across the two future spaces while remaining isolated from action tokens. Action tokens attend to both current streams and themselves but cannot access either future stream. This routing prevents future-target leakage into action generation and action-token interference with future modeling, while allowing the future streams to be omitted at inference for efficient deployment.

Current-Anchored Intent Retrieval

Current-Anchored Semantic Queries. As shown in Fig. 2(b), given the current observation $\mathbf{o}_t$ and language instruction ℓ, a frozen Qwen3-VL model extracts its final-layer multimodal hidden states:

$$\mathbf{H}^q_t = E_{\mathrm{Qwen}}(\mathbf{o}_t, \ell) \in \mathbb{R}^{B \times L_q \times d_q}, \tag{4}$$

where L_q is the sequence length and d_q is the hidden dimension of Qwen3-VL. A bank of N_I learnable queries $\mathbf{Q}$, each with dimension d_r, attends to the Qwen3-VL features to produce a compact set of current semantic tokens:

$$\mathbf{U}^0_t = \mathrm{MHA}\left(\mathbf{Q}, P_q\mathbf{H}^q_t, P_q\mathbf{H}^q_t\right). \tag{5}$$

Here, $\mathbf{Q}$ serves as the query, while the projected Qwen3-VL features serve as keys and values. P_q is a learnable linear projection from the Qwen3-VL hidden dimension d_q to the hidden dimension d_r, and $\mathbf{U}^0_t \in \mathbb{R}^{B \times N_I \times d_r}$ denotes the current semantic tokens before history fusion. These tokens jointly encode the current scene and instruction, serving as semantic anchors that guide subsequent intent retrieval.

Semantic History Retrieval. For each observation in the short history $\mathcal{H}_t$, the frozen DINOv3 encoder extracts dense patch features, providing a visually stable semantic representation from which task-relevant historical evidence can be retrieved. The features from all M observations are linearly projected, augmented with learnable temporal embeddings, and concatenated into a history-token sequence $\mathbf{R}_t$.

Starting from $\mathbf{U}^0_t$, CAIR applies L cross-attention blocks, using the current semantic tokens as queries and $\mathbf{R}_t$ as keys and values. This current-conditioned interaction selectively retrieves historical evidence relevant to the present task state while reducing sensitivity to task-irrelevant visual variations.

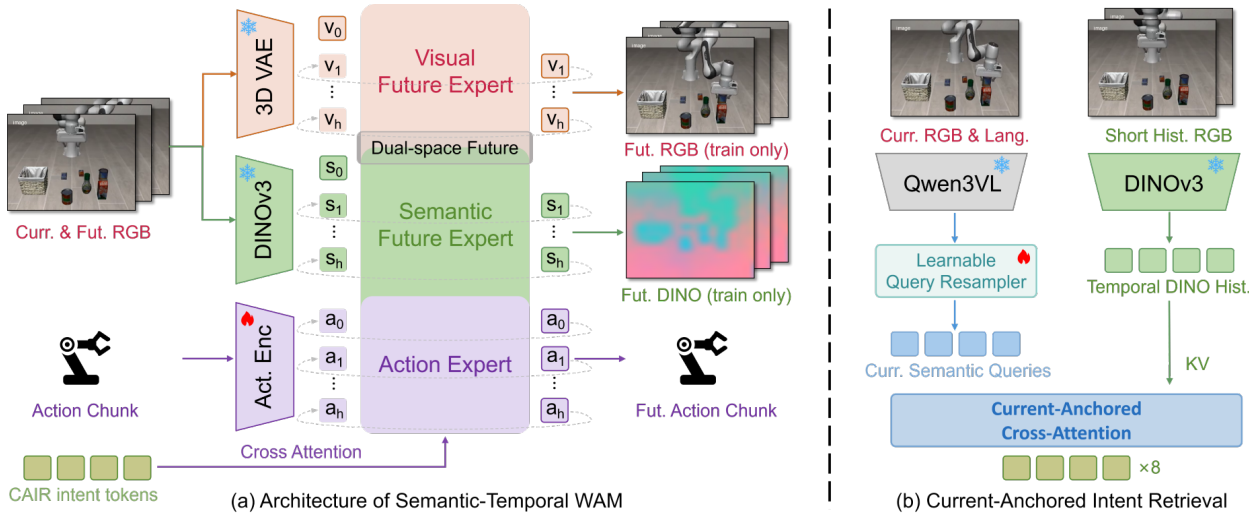

Figure 2: Overview of ST-WAM. (a): Dual-Space Future Experts (DSFE) jointly model future dynamics in the VAE visual-latent space and the DINOv3 semantic space. The visual and semantic future experts interact with the action expert through a three-branch Mixture-of-Transformers. (b): Current-Anchored Intent Retrieval (CAIR).

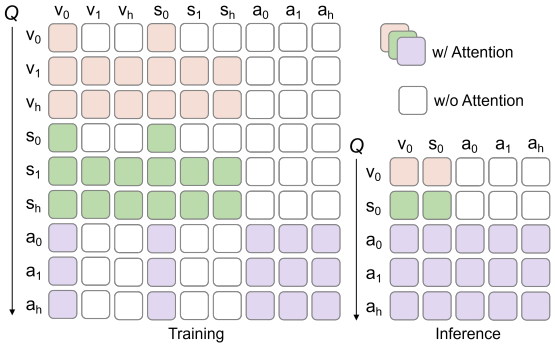

Figure 3: Structured cross-branch attention masks during training and inference.

After L blocks, the refined tokens $\mathbf{U}_t^L$ are projected into the action-context space:

$$\mathbf{I}_t = P_o\left(\mathbf{U}_t^L\right) \in \mathbb{R}^{B\times N_I\times d_c}, \tag{6}$$

where P_o is the output projection and d_c is the context dimension of the action expert. We refer to $\mathbf{I}_t$ as short-horizon intent tokens: a latent, label-free summary of recent task progress relevant to current action decision, rather than an explicitly supervised variable.

For expert conditioning, the frozen T5 encoder maps the language instruction into the shared language context $\mathbf{C}_\ell = E_{\mathrm{T5}}(\ell)$. After appending the projected proprioceptive token $P_p(\mathbf{s}_t)$, where P_p is a learnable projection, the contexts of the three experts are

$$\mathbf{C}_v = \mathbf{C}_s = [\mathbf{C}_\ell; P_p(\mathbf{s}_t)], \qquad \mathbf{C}_a = [\mathbf{C}_\ell; P_p(\mathbf{s}_t); \mathbf{I}_t]. \tag{7}$$

Each expert receives its corresponding context through cross-attention at every DiT block. Thus, the short-horizon intent tokens are injected only into the action expert and optimized end-to-end through action flow matching, allowing semantic evidence from history to guide action generation without altering the conditioning contexts of the DSFE branches.

Together, DSFE and CAIR use DINOv3 in complementary temporal directions: DSFE prospectively supervises future semantic dynamics, whereas CAIR retrospectively retrieves task-relevant evidence from recent semantic history for robust action generation under visual distribution shifts.

Joint Flow-Matching Objective

We jointly train the visual future, semantic future, and action experts using flow matching (Lipman et al. 2022). Let the clean branch targets be $\mathbf{x}^v = \mathbf{z}^v_{\mathrm{fut}}$, $\mathbf{x}^s = \mathbf{z}^s_{\mathrm{fut}}$, and $\mathbf{x}^a = \mathbf{a}_{t:t+H-1}$. For each branch $r \in \{v, s, a\}$, we sample a timestep $\tau_r \in [0, 1]$ and Gaussian noise $\boldsymbol{\epsilon}^r \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noisy input is constructed through linear interpolation:

$$\mathbf{x}^r_{\tau_r} = (1-\tau_r)\mathbf{x}^r + \tau_r\boldsymbol{\epsilon}^r. \tag{8}$$

Here, $\tau_r = 0$ corresponds to the clean target and $\tau_r = 1$ to pure noise. The corresponding target velocity is

$$\mathbf{u}^r = \boldsymbol{\epsilon}^r - \mathbf{x}^r. \tag{9}$$

Because the two future experts exchange information through mixed attention, we set $\tau_v = \tau_s = \tau_f$ to synchronize their denoising stages, while sampling τ_a independently from the same timestep distribution. Gaussian noises remain independent across all three branches, and action denoising is therefore decoupled from the two training-only future flows.

The clean current tokens, noisy future tokens, and noisy action tokens are processed in a single MoT forward pass under the structured attention mask, producing branch-specific velocity estimates $\hat{\mathbf{u}}^r_\theta$. For each branch, we optimize

$$\mathcal{L}_r = \mathbb{E}\left[w(\tau_r)\,\|\hat{\mathbf{u}}^r_\theta - \mathbf{u}^r\|_2^2\right], \qquad r \in \{v, s, a\}, \tag{10}$$

where $w(\tau_r)$ denotes the timestep-dependent weighting of the flow scheduler. The overall training objective is

$$\mathcal{L} = \lambda_v\mathcal{L}_v + \lambda_s\mathcal{L}_s + \lambda_a\mathcal{L}_a, \tag{11}$$

where $\mathcal{L}_v$, $\mathcal{L}_s$, and $\mathcal{L}_a$ denote the visual, semantic, and action expert losses, respectively, and λ_v, λ_s, and λ_a are their corresponding loss weights.

Experiments

Experimental Setup

Benchmarks and Protocols. We evaluate in-distribution manipulation performance on the four LIBERO suites (Liu et al. 2023): Spatial, Object, Goal, and Long, covering 40 tasks with 50 evaluation rollouts per task. For out-of-distribution evaluation, we directly evaluate the LIBERO-trained policy on LIBERO-Plus (Fei et al. 2025) without fine-tuning, which comprises 10,030 test cases spanning seven perturbation dimensions. We additionally evaluate bimanual manipulation on RoboTwin 2.0 (Chen et al. 2025) and ST-WAM is trained on a mixture of 2,500 clean and 25,000 heavily randomized demonstrations and each task is evaluated over 100 trials in both clean and randomized settings.

Baselines. We compare ST-WAM against a broad range of representative methods. These include VLAs such as π_0 (Black et al. 2024), $\pi_{0.5}$ (Physical Intelligence et al. 2025); video-generative WAMs such as Fast-WAM (Yuan et al. 2026), Motus (Bi et al. 2026), and LingBot-VA (Li et al. 2026b); and methods exploring alternative future representations, such as LaWAM (Chen et al. 2026a), Mask World Model (Lou et al. 2026), MaskWAM (Yu et al. 2026).

Real-World Evaluation. We evaluate ST-WAM on an Agilex Piper 6-DoF single-arm robot using 50 demonstrations per task under fixed nominal visual conditions. We consider five tasks with diverse temporal and geometric requirements: (1) *Arrange Flowers*, inserting three bouquets into a vase; (2) *Drawer Organization*, opening a drawer, placing a pen inside, and closing it; (3) *Bean Scooping*, transferring beans from a plate to a bowl and returning the spoon; (4) *Arrange Fruits*, placing five fruits into a basket; and (5) *Hang Mug*, hanging a mug on a rack. We compare ST-WAM with π_0 (Black et al. 2024) and Fast-WAM (Yuan et al. 2026), with all methods post-trained separately for each task using the same demonstrations. We test them under nominal and four visual-shift conditions without further fine-tuning: *Background*, which replaces the tabletop texture with unseen patterns; *Lighting*, which changes the illumination intensity; *Object Appearance*, which changes object colors or instances while preserving their geometry and task function; and *Compound*, which applies all three shifts simultaneously. Each method is evaluated over 30 trials per task and condition using predefined object initializations to ensure fair comparison.

Implementation Details. The visual future expert uses the 5B Video DiT of Wan2.2-TI2V-5B (Wan et al. 2025), while the semantic and action experts use 1B DiTs initialized from the pretrained Wan2.2 weights. The Wan2.2 VAE and T5 encoder, DINOv3 ViT-S/16 visual encoder (Siméoni et al. 2025), and Qwen3-VL-4B-Instruct (Bai et al. 2025) remain frozen throughout training.

We set the action horizon to $H = 32$ and the future horizon to $K = 8$, with future observations sampled every four control steps. For CAIR, we set $M = 4$ and construct $\mathcal{H}_t$ from observations at frame indices $\{t-24, t-16, t-8, t-1\}$. The retrieval uses $L = 2$ cross-attention blocks and $N_I = 8$ learnable queries, producing eight intent tokens.

We train on LIBERO and RoboTwin 2.0 for 10 and 5 epochs with global batch sizes of 128 and 1,024, respectively. We use AdamW ($\mathrm{lr} = 1 \times 10^{-4}$, weight decay 0.01), cosine decay, BF16 precision, gradient clipping at 1.0, and a shifted flow-matching schedule with shift 5.0; the loss weights are $(\lambda_v, \lambda_s, \lambda_a) = (1.0, 0.02, 1.0)$. At inference, we use 10 flow-integration steps and execute 10 actions before replanning; additional details are provided in the supplementary material.

Method	Emb. PT.	Spat.	Obj.	Goal	Long	Avg.
Vision-Language-Action Models						
π_0 (Black et al. 2024)	Yes	98.0	96.8	94.4	88.4	94.4
$\pi_{0.5}$ (Physical Intelligence et al. 2025)	Yes	98.8	98.2	98.0	92.4	96.9
IntentVLA (Lian et al. 2026)	No	99.3	99.7	98.1	97.4	98.6
Video-Generative World Action Models						
Fast-WAM (Yuan et al. 2026)	No	98.2	**100.0**	97.0	95.2	97.6
Motus (Bi et al. 2026)	Yes	96.8	99.8	96.6	97.6	97.7
LingBot-VA (Li et al. 2026b)	Yes	98.5	99.6	97.2	**98.5**	98.5
WAMs with Alternative Future Representations						
Mask World Model (Lou et al. 2026)	No	98.8	**100.0**	98.2	96.0	98.3
MaskWAM (Yu et al. 2026)	No	98.8	**100.0**	98.2	96.4	98.4
GeoSem-WAM (Ma et al. 2026)	No	99.0	**100.0**	98.2	97.0	98.6
LaWAM (Chen et al. 2026a)	Yes	**99.4**	99.6	98.4	97.0	98.6
ST-WAM (Ours)	No	99.0	**100.0**	**99.0**	96.8	**98.7**

Table 1: Success rates (%) on the four LIBERO suites. **Emb. PT.** indicates large-scale pretraining on robot trajectories or embodied videos before LIBERO adaptation.

Method	Emb. PT.	Clean	Random	Avg.
π_0 (Black et al. 2024)	Yes	65.92	58.40	62.16
$\pi_{0.5}$ (Physical Intelligence et al. 2025)	Yes	82.74	76.76	79.75
GigaWorld-Policy (Ye et al. 2026a)	Yes	87.00	85.00	86.00
Motus (Bi et al. 2026)	Yes	88.66	87.02	87.84
LaWAM (Chen et al. 2026a)	Yes	92.64	89.80	91.22
Fast-WAM (Yuan et al. 2026)	No	91.88	91.78	91.83
LingBot-VA (Li et al. 2026b)	Yes	92.90	91.50	92.20
GeoSem-WAM (Ma et al. 2026)	No	92.94	92.14	92.54
ST-WAM (Ours)	No	**93.06**	**92.48**	**92.77**

Table 2: Success rates (%) on RoboTwin 2.0 under the standard mixed clean-and-randomized training setting.

Main Results

Performance on LIBERO. As shown in Table 1, ST-WAM achieves an average success rate of 98.7%, the highest among the compared methods. Without embodied pretraining, it outperforms recent strong WAMs such as Motus (97.7%) (Bi et al. 2026) and LingBot-VA (98.5%) (Li et al. 2026b). Thus, ST-WAM further improves in-distribution performance.

Performance on RoboTwin 2.0. As shown in Table 2, ST-WAM achieves success rates of 93.06% and 92.48% in the clean and randomized settings, respectively, yielding the highest average success rate of 92.77% among the compared methods. Without embodied pretraining, ST-WAM outperforms both the embodied-pretrained LingBot-VA (Li et al.

Method	Emb. PT.	Camera	Robot	Lang.	Light	BG	Noise	Layout	Overall
UniVLA (Bu et al. 2025)	Yes	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0 (Black et al. 2024)	Yes	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-FAST (Pertsch et al. 2025)	Yes	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA (OFT) (Tan et al. 2025)	Yes	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT (Kim, Finn, and Liang 2025)	Yes	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
X-VLA (Zheng et al. 2025)	Yes	23.4	**89.7**	75.7	88.2	**96.0**	62.7	71.8	71.4
Fast-WAM (Yuan et al. 2026)	No	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
Fast-WAM-Joint (Yuan et al. 2026)	No	34.0	55.1	**88.9**	90.0	44.9	33.3	73.4	59.0
ST-WAM (Ours)	No	55.4	60.1	79.3	**93.0**	74.2	**79.5**	**74.3**	**72.8**

Table 3: Zero-shot success rates (%) on LIBERO-Plus. Baseline results from (Zhang et al. 2026d; Chen et al. 2026b).

2026b) (92.20%) and the closely related Fast-WAM (Yuan et al. 2026) (91.83%). These consistent results demonstrate the effectiveness of ST-WAM for bimanual manipulation across clean and randomized environments.

Zero-Shot Generalization on LIBERO-Plus. As shown in Table 3, ST-WAM achieves an overall success rate of 72.8%. Without embodied pretraining, it surpasses several embodied-pretrained VLAs, including OpenVLA-OFT (69.6%), RIPT-VLA (68.4%), and X-VLA (71.4%). More importantly, ST-WAM improves the closely matched Fast-WAM baseline (Yuan et al. 2026) from 51.5% to 72.8%, a gain of 21.3 percentage points. This improvement is consistent across all seven perturbation categories, with particularly large gains of 39.0 and 41.8 percentage points under camera and sensor-noise perturbations, respectively. It also outperforms Fast-WAM-Joint on all six non-language perturbations. Together, these results suggest that complementing fine-grained VAE dynamics with DINO-based semantic supervision reduces reliance on low-level visual cues and substantially improves out-of-distribution robustness.

Inference Efficiency. On RoboTwin 2.0, we benchmark the complete action-chunk inference call, including all model components active at deployment, on a single NVIDIA A100-80GB GPU using BF16 precision and 10 flow-integration steps. Averaged over 20 synchronized runs, ST-WAM generates a 32-step action chunk in 756.17 ms, compared with 609.30 ms for Fast-WAM (Yuan et al. 2026). This 1.24× latency represents a moderate overhead for improved robustness while retaining sub-second inference.

Real-World Generalization. As shown in Fig. 4 and Table 4, ST-WAM achieves 79.3% average success under the nominal condition, outperforming Fast-WAM and π_0 by 14.6 and 32.0 percentage points, respectively. Under visual distribution shifts, ST-WAM achieves 61.5%, surpassing Fast-WAM and π_0 by 35.7 and 28.7 points, respectively, and retaining a substantial advantage under the compound shift (48.0% vs. 15.3%). Notably, Fast-WAM drops by 38.9 points from the nominal to shifted conditions, compared with only 17.8 points for ST-WAM, suggesting that pixel-generative future representations are particularly sensitive to visual distribution shifts. Removing the semantic future expert or CAIR reduces shifted-condition performance to 41.0% and 43.7%, respectively, confirming their complementary contributions to real-world robustness. Task-wise results under visual shifts are provided in the supplementary material.

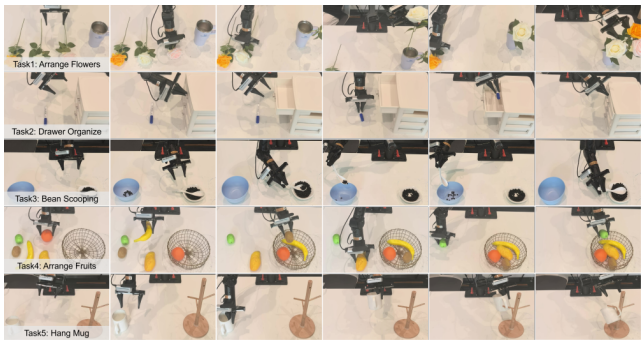

Figure 4: Real-world evaluation of ST-WAM on five tasks with diverse temporal and geometric requirements.

Ablation and Qualitative Analysis

Beyond the component-level real-world ablations, we conduct finer-grained controlled studies on LIBERO and LIBERO-Plus, as shown in Table 5, to disentangle the design choices underlying DSFE and CAIR. Specifically, we address three critical questions:

Q1: Are VAE and DINO future representations complementary or interchangeable? We first compare different future representation spaces without intent conditioning. The *DINO Future Only* variant achieves 39.7% on LIBERO-Plus, below the 51.5% of the VAE-based Fast-WAM. In contrast, jointly modeling future states in the VAE and DINO spaces improves the success rate to 66.4% even without CAIR. These results indicate that DINO semantics cannot directly replace VAE dynamics: VAE latents preserve fine-grained visual and motion information, whereas DINO features provide complementary object- and state-level semantics.

Q2: Is explicit semantic future prediction necessary? Keeping CAIR fixed, *w/o Semantic Future Expert* removes the semantic stream and achieves 63.5% on LIBERO-Plus. The parameter-matched *Semantic Expert w/o Future Obj.* retains the semantic DiT, current-DINO conditioning, and mixed-attention interactions, but removes its future target and loss, yielding 62.9%. In contrast, the full model reaches

Method	Nominal Environment						Visual Distribution Shifts				
	Flower	Drawer	Scoop	Fruit	Hang	Avg.	BG	Light	Obj. App.	Comp.	Avg.
π_0 (Black et al. 2024)	56.7	46.7	30.0	46.7	56.7	47.3	33.3	40.7	35.3	22.0	32.8
Fast-WAM (Yuan et al. 2026)	70.0	66.7	46.7	66.7	73.3	64.7	27.3	35.3	25.3	15.3	25.8
w/o Semantic Future Expert	76.7	73.3	56.7	73.3	80.0	72.0	43.3	50.0	41.3	29.3	41.0
w/o CAIR	80.0	76.7	60.0	76.7	83.3	75.3	46.0	52.7	44.0	32.0	43.7
ST-WAM (Ours)	**86.7**	**80.0**	**66.7**	**76.7**	**86.7**	**79.3**	**66.0**	**70.0**	**62.0**	**48.0**	**61.5**

Table 4: Real-world success rates (%) under the nominal condition and visual distribution shifts. Results for each visual shift are averaged across all five tasks. "Comp." combines background, lighting, and object-appearance shifts.

Variant	Future Prediction	Intent Conditioning	LIBERO	LIBERO-Plus
Fast-WAM (Yuan et al. 2026)	VAE	None	97.6	51.5
DINO Future Only	DINO	None	96.3	39.7
Dual-Space w/o CAIR	VAE + DINO	None	97.8	66.4
w/o Semantic Future Expert	VAE	CAIR (DINO history)	97.3	63.5
Semantic Expert w/o Future Obj.	VAE	CAIR (DINO history)	95.8	62.9
Naive History Retrieval	VAE + DINO	Unanchored DINO history	96.3	56.5
Qwen Current Only	VAE + DINO	Qwen current	96.5	62.3
CAIR with VAE History	VAE + DINO	CAIR (VAE history)	96.3	64.7
ST-WAM (Ours)	VAE + DINO	**CAIR (DINO history)**	**98.7**	**72.8**

Table 5: Ablation results on LIBERO and LIBERO-Plus. "Intent Conditioning" denotes the alternative action-expert contexts used to ablate the design choices of CAIR.

72.8%, showing that the gain cannot be explained by current DINO features or additional model capacity alone, but requires explicit future-semantic prediction.

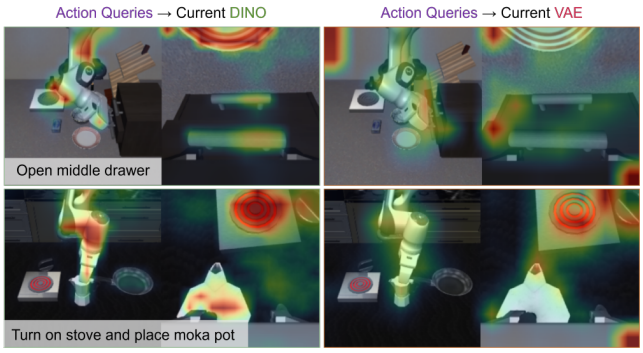

Figure 5: Attention heatmaps on two representative tasks. Warmer colors indicate stronger relative attention within each map.

Q3: How should short-horizon history be represented and retrieved? The *Naive History Retrieval* variant compresses DINO history with unanchored learnable queries and achieves only 56.5% on LIBERO-Plus, versus 72.8% for the full model, demonstrating the necessity of a current semantic anchor. Removing history entirely in *Qwen Current Only* yields 62.3%, ruling out VLM-derived current-frame semantics alone as the source of improvement. Replacing DINO history with Wan-VAE latents in *CAIR with VAE History*, while retaining the same Qwen anchor and retrieval architecture, obtains 64.7%. All three alternative conditioning schemes underperform the *Dual-Space w/o CAIR* baseline (66.4%), indicating that improperly represented or retrieved context can be detrimental; only current-anchored retrieval from DINO history surpasses this baseline, reaching 72.8%. Together, these results show that effective intent modeling requires both a current visual-language anchor and a visually stable DINO representation of recent history. Detailed results for each subset are provided in the supplementary material.

Cross-Branch Attention Visualization. Fig. 5 visualizes the attention from action queries to the current VAE and DINO tokens in the MoT mixed self-attention. Across both tasks, action-to-DINO attention aligns more closely with the manipulated objects and interaction regions, whereas action-to-VAE attention is distributed over broader scene areas. This qualitative pattern suggests that DINO provides task-focused semantic cues complementary to the fine-grained visual dynamics represented by VAE latents.

Conclusion

Motivated by the robustness limitations exposed by **Training-Distribution Hallucination**, we introduced **ST-WAM** to improve video-generative World Action Models under visual distribution shifts. ST-WAM uses DINOv3 in complementary temporal directions: DSFE complements fine-grained VAE dynamics with visually stable future semantics, while CAIR retrieves task-relevant intent from recent semantic history under the current visual-language context. Ex-

tensive simulation and real-world experiments demonstrate substantially improved robustness while preserving strong in-distribution performance, efficient action-only inference, and freedom from large-scale embodied pretraining. Overall, our results highlight semantic-temporal modeling as an effective direction for robust world-action learning beyond pixel-centric futures. Future work will extend ST-WAM beyond visual distribution shifts to changes in physical dynamics and embodiments.

References

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Bi, H.; Tan, H.; Xie, S.; Wang, Z.; Huang, S.; Liu, H.; Zhao, R.; Feng, Y.; Xiang, C.; Rong, Y.; et al. 2026. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 35101–35113.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*.

Chen, J.; Wang, K.; Chen, K.; Chen, S.; Gao, F.; Tang, W.; Li, Z.; Liu, W.; Yao, Z.; Li, B.; et al. 2026a. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*.

Chen, R.; Yang, Y.; Tang, Z.; Huo, D.; Lin, T.; Wu, H.; Liu, H.; Chen, Y.; Zheng, L.; Yuan, B.; et al. 2026b. Abot-m0. 5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*.

Chen, T.; Chen, Z.; Chen, B.; Cai, Z.; Liu, Y.; Li, Z.; Liang, Q.; Lin, X.; Ge, Y.; Gu, Z.; et al. 2025. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*.

Du, F.; Yan, F.; Wu, J.; Xu, X.; Zhang, W.; Wang, W.; Guo, Y.; Qian, B.; He, Z.; Wang, F.; and Yang, H. 2026. CF-VLA: Efficient Coarse-to-Fine Action Generation for Vision-Language-Action Policies. arXiv:2604.24622.

Fan, L.; Xu, Z.; Cao, C.; Zhang, W.; Yuan, M.; and Chen, J. 2026. Aim: Intent-aware unified world action modeling with spatial value maps. *arXiv preprint arXiv:2604.11135*.

Fei, S.; Wang, S.; Shi, J.; Dai, Z.; Cai, J.; Qian, P.; Ji, L.; He, X.; Zhang, S.; Fei, Z.; et al. 2025. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*.

Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

Kim, M. J.; Gao, Y.; Lin, T.-Y.; Lin, Y.-C.; Ge, Y.; Lam, G.; Liang, P.; Song, S.; Liu, M.-Y.; Finn, C.; et al. 2026. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Li, B.; Yin, X.; Lin, M.; Zhang, Y.; and Xu, D. 2026a. EgoWAM: World Action Models Beyond Pixels with In-the-Wild Egocentric Human Data. In *Robot World Models*.

Li, L.; Zhang, Q.; Luo, Y.; Yang, S.; Wang, R.; Han, F.; Yu, M.; Gao, Z.; Xue, N.; Zhu, X.; et al. 2026b. Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998*.

Lian, S.; Yu, B.; Lin, X.; Shen, Z.; Yang, L. T.; Jin, Y.; Liu, H.; Wu, C.; Yuan, H.; Huang, C.; et al. 2026. IntentVLA: Short-Horizon Intent Modeling for Aliased Robot Manipulation. *arXiv preprint arXiv:2605.14712*.

Liang, W.; Yu, L.; Luo, L.; Iyer, S.; Dong, N.; Zhou, C.; Ghosh, G.; Lewis, M.; Yih, W.-t.; Zettlemoyer, L.; et al. 2024. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. In *The eleventh international conference on learning representations*.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791.

Liu, Y.; Sun, P.; Li, S.; Xie, Y.; Zhang, L.; Chao, X.; Dong, S.; Chen, F.; Zhang, X.-P.; and Ding, W. 2026. Oa-wam: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*.

Lou, Y.; Chi, X.; Zhang, X.; Qian, Z.; Li, C.; Zhang, R.; Lyu, Y.; Song, G.; Fu, C.; Xu, H.; et al. 2026. Mask World Model: Predicting What Matters for Robust Robot Policy Learning. *arXiv preprint arXiv:2604.19683*.

Luo, H.; Zhang, W.; Feng, Y.; Zheng, S.; Xu, H.; Xu, C.; Xi, Z.; Fu, Y.; and Lu, Z. 2026. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*.

Lyu, J.; Liu, K.; Zhang, X.; Liao, H.; Feng, Y.; Zhu, W.; Shen, T.; Chen, J.; Zhang, J.; Dong, Y.; et al. 2026. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*.

Ma, F.; Peng, D.; Yue, W.; Cao, J.; Wang, B.; Zhang, Q.; and Ma, J. 2026. GeoSem-WAM: Geometry-and Semantic-Aware World Action Models. *arXiv preprint arXiv:2606.03188*.

Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; and Levine, S. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Physical Intelligence; Ai, B.; Amin, A.; Aniceto, R.; Balakrishna, A.; Balke, G.; Black, K.; Bokinsky, G.; Cao, S.; Charbonnier, T.; et al. 2026. $\pi_{0.7}$: A Steerable Generalist Robotic Foundation Model with Emergent Capabilities. *arXiv preprint arXiv:2604.15483*.

Physical Intelligence; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. 2025. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

Siméoni, O.; Vo, H. V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Song, W.; Zhou, Z.; Zhao, H.; Chen, J.; Ding, P.; Yan, H.; Huang, Y.; Tang, F.; Wang, D.; and Li, H. 2026. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 18549–18557.

Su, Y.; Chen, S.; Shi, H.; Liu, M.; Zhang, Z.; Huang, N.; Zhong, W.; Zhu, Z.; Liu, Y.; and Liu, X. 2026. World guidance: World modeling in condition space for action generation. *arXiv preprint arXiv:2602.22010*.

Sun, J.; Zhang, W.; Qi, Z.; Ren, S.; Liu, Z.; Zhu, H.; Sun, G.; Jin, X.; and Chen, Z. 2026. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*.

Tan, S.; Dou, K.; Zhao, Y.; and Krähenbühl, P. 2025. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Yan, F.; Liu, F.; Huang, Y.; Guan, Z.; Zheng, L.; Zhong, Y.; Feng, C.; and Ma, L. 2025. RoboTron-Mani: All-in-One Multimodal Large Model for Robotic Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13707–13718.

Ye, A.; Wang, B.; Ni, C.; Huang, G.; Zhao, G.; Li, H.; Li, H.; Li, J.; Lv, J.; Liu, J.; et al. 2026a. GigaWorld-Policy: An Efficient Action-Centered World–Action Model. *arXiv preprint arXiv:2603.17240*.

Ye, S.; Ge, Y.; Zheng, K.; Gao, S.; Yu, S.; Kurian, G.; Indupuru, S.; Tan, Y. L.; Zhu, C.; Xiang, J.; et al. 2026b. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*.

Yu, H.; Lin, H.; Zhang, J.; Zhang, W.; Gu, C.; Li, H.; and Tan, P. 2026. Maskwam: Unifying mask prompting and prediction for world-action models. *arXiv preprint arXiv:2606.13515*.

Yuan, T.; Dong, Z.; Liu, Y.; and Zhao, H. 2026. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*.

Zhang, J.; Zhu, J.; Su, T.; Ma, C.; Huang, Z.; Xu, Y.; and Wang, H. 2026a. Learning 4D Geometric Priors for Inference-Efficient World Action Models. *arXiv preprint arXiv:2607.05468*.

Zhang, W.; Liu, H.; Qi, Z.; Wang, Y.; Yu, X.; Zhang, J.; Dong, R.; He, J.; Wang, H.; Zhang, Z.; et al. 2026b. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228.

Zhang, W.; Zhang, B.; Qi, Z.; Zeng, W.; Jin, X.; and Zhang, L. 2026c. Disentangled robot learning via separate forward and inverse dynamics pretraining. *arXiv preprint arXiv:2604.16391*.

Zhang, Z.; Li, Z.; Rahmati, B.; Yang, R. H.; Ma, Y.; Rasouli, A.; Pakdamansavoji, S.; Wu, Y.; Zhang, L.; Cao, T.; et al. 2026d. Do world action models generalize better than vlas? a robustness study. *arXiv preprint arXiv:2603.22078*.

Zheng, J.; Li, J.; Wang, Z.; Liu, D.; Kang, X.; Feng, Y.; Zheng, Y.; Zou, J.; Chen, Y.; Zeng, J.; et al. 2025. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*.