CSWAM: BETTER CAUSAL SEMANTIC REPRESENTATIONS FOR OUT-OF-DISTRIBUTION GENERALIZATION IN WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

FastWAM-style world action models enable efficient action-only inference, but generalize poorly under visual distribution shifts. Their reconstruction-oriented representations emphasize appearance-specific details, limiting generalization to unseen scenes and objects. Without observation history, the model also lacks temporal evidence for robustly identifying task-relevant state changes and motion in unfamiliar visual conditions. To address these limitations, we present the *Causal Semantic World Action Model* (CSWAM), which augments FastWAM with a causal semantic expert built on V-JEPA 2.1. V-JEPA provides temporally grounded representations of semantic state changes and motion with less dependence on appearance-specific details. The expert learns their future evolution from a sparse history of current and past observations and shares the history-derived context with both the video and action streams through causal attention. At inference, CSWAM conditions action denoising on the current video state and observed semantic history, retaining efficient action-only inference. We conduct simulation and real-robot experiments to evaluate generalization under distribution shifts. With embodied pretraining, CSWAM raises Randomized success on RoboTwin 2.0 Clean-to-Randomized transfer from 10.16% to 45.18%, a gain of 35.02 percentage points over FastWAM. Across two real-robot tasks and three OOD difficulty levels, CSWAM improves average success over FastWAM by 42.5 percentage points, from 27.5% to 70.0%.

1 INTRODUCTION

By jointly modeling future visual states and robot actions, world action models (WAMs) Cheang et al. (2024); Li et al. (2026); Ye et al. (2026); Yuan et al. (2026b); Zhang et al. (2026a); Zhu et al. (2026); Wang et al. (2026c) have emerged as a promising alternative to vision-language-action (VLA) models Zitkovich et al. (2023); Kim et al. (2025); Black et al. (2024); Su et al. (2026) for robot manipulation. However, WAMs that first predict future video and then generate actions incur substantial latency at inference. Consequently, a growing line of work adopts a FastWAM-style paradigm that jointly trains video and action prediction but generates only actions at deployment Li et al. (2025); Yuan et al. (2026b); Lin et al. (2026). This design retains the benefits of video co-training while avoiding iterative future-video generation at deployment, making the paradigm attractive for robot control Yuan et al. (2026b).

FastWAM-style models perform well when the visual distribution at deployment resembles that of the training data. However, we find that FastWAM's success rate falls sharply under visual distribution shifts, such as changes in backgrounds and object appearances. Two limitations may contribute to this gap: the visual representation used for action prediction and the absence of observation history. First, the video stream operates on reconstruction-oriented VAE latents, which preserve appearance details but are not explicitly aligned with task-relevant semantics. Changes in background or object appearance can therefore alter the visual context even when the underlying manipulation state is unchanged. Second, the action expert receives only the current observation and cannot use history that reveals motion, state transitions, and task progress. Without stable semantics and temporal evidence, the policy can rely on correlations specific to the training scenes that fail under visual shifts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Generalization to out-of-distribution (OOD) visual conditions is essential for real-world deployment, since collecting demonstrations across every background and object appearance is infeasible. Recent work has introduced pretrained semantic encoders such as DINO Siméoni et al. (2025) to make robot policies and WAMs less sensitive to appearance changes Burns et al. (2025); Wang et al. (2026a). These studies highlight the value of semantic features, but leave open how temporally predictive representations and observation history should be incorporated into efficient action-only WAMs. DINO provides strong spatial semantics, but when applied as a framewise encoder, it does not explicitly represent how a scene evolves. We therefore use V-JEPA 2.1 Bardes et al. (2024); Assran et al. (2025); Mur-Labadia et al. (2026), whose video-native joint-embedding predictive representations more directly capture semantic state changes and motion while reducing reliance on appearance-specific details. Paired with causal observation history, these representations provide both task-relevant semantic cues and richer temporal evidence for action prediction.

Motivated by this analysis, we introduce CSWAM, a Causal Semantic World Action Model that extends the FastWAM-style video–action architecture with a history-conditioned semantic expert (Fig. 1). A frozen V-JEPA 2.1 encoder first extracts semantic features from a sparse history of current and past observations. The semantic expert predicts future semantic features to model how task-relevant states evolve over time. A structured attention mask allows both the video and action streams to use the observed semantic history while preventing the action stream from accessing future targets. At deployment, the future target streams are omitted: the current video state and semantic history are computed once, and only the action stream is denoised. By sharing the same history-derived semantic and motion context with both streams, CSWAM supports future-video learning and action prediction while preserving FastWAM's efficient action-only inference. Experiments in simulation and on real robots demonstrate consistent gains under visual distribution shifts. Under the standard RoboTwin 2.0 mixed-training protocol, CSWAM achieves 94.7% success on Clean and 94.3% on Randomized, respectively. Under the stricter Clean-to-Randomized transfer setting with embodied pretraining, CSWAM raises OOD Randomized success from 10.16% to 45.18%, a gain of 35.02 percentage points over FastWAM. On LIBERO-Plus, CSWAM achieves 77.1% overall success, exceeding ST-WAM Wang et al. (2026a) by 4.3 percentage points. Across the two real-robot tasks and three OOD difficulty levels, CSWAM raises average success from 27.5% to 70.0%, a gain of 42.5 percentage points over FastWAM.

Our contributions are threefold:

- We investigate how temporally predictive representations and observation history support OOD generalization in efficient action-only WAMs, and design controlled visual-shift evaluations in simulation and on real robots.
- We propose CSWAM, whose V-JEPA 2.1-based causal semantic expert predicts future semantic features from observation history and shares the history-derived context with both the video and action streams, while retaining action-only inference.
- Extensive simulation and real-robot experiments show that CSWAM improves both standard manipulation performance and OOD generalization over FastWAM.

2 Related work

2.1 World action models

World action models (WAMs) learn robot policies jointly with predictive visual dynamics, using future-state modeling to provide supervision about scene evolution and action consequences. Representative formulations couple video and action through joint prediction Cheang et al. (2024), shared representations with decoupled prediction heads Li et al. (2025), or causal autoregressive diffusion Li et al. (2026). This paradigm has also been scaled with pretrained video diffusion backbones and heterogeneous robot data Ye et al. (2026). More recent work explores complementary directions, including 4D geometric priors Zhang et al. (2026a), dual-system architectures Zhu et al. (2026), and modular pretraining frameworks Wang et al. (2026c). An important deployment question is whether visual prediction remains active at inference. The Unified Video Action model allows actions to be decoded without generating video Li et al. (2025). FastWAM develops this idea into an action-only WAM: video and action prediction are jointly trained, but future-video denoising is removed at deployment Yuan et al. (2026b). Using a Mixture-of-Transformers architecture Liang et al. (2024),

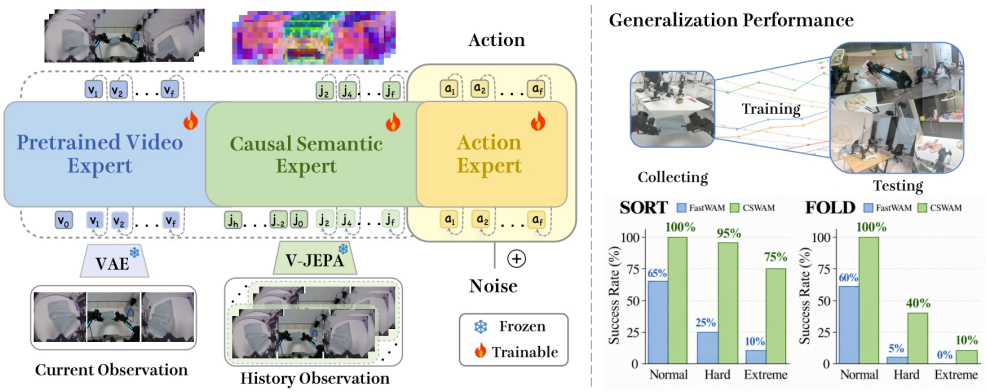


Figure 1: Overview of CSWAM. Left: the causal semantic expert encodes V-JEPA observation history and shares temporal context with the video and action experts. Right: real-robot success under compounded visual shifts.

it conditions action denoising on a fixed context derived from the current observation. Subsequent methods explore alternative trade-offs between future conditioning and inference efficiency. LingBot-VA 2.0 continues to predict future latents during asynchronous closed-loop execution Zhang et al. (2026b), whereas Faster-WAM computes compact future representations once and reuses them during action denoising Zhao et al. (2026). CSWAM instead builds directly on FastWAM's action-only formulation, expanding its observed context with temporally predictive V-JEPA history while retaining inference without future prediction.

2.2 Visual representations for robot manipulation

Visual representations determine what information world modeling contributes to action prediction. Generative WAMs commonly operate in VAE latent spaces optimized for visual reconstruction and prediction Ye et al. (2026); Yuan et al. (2026b). These latents preserve the appearance details required for video generation, but may also retain nuisance variations unrelated to manipulation. Studies of robust visuomotor learning show that pretrained semantic representations can reduce this sensitivity Burns et al. (2025). DINOv3 Siméoni et al. (2025) learns strong dense visual features through image self-supervision and has been adopted as semantic supervision in WAMs Wang et al. (2026a). When applied as a framewise encoder, however, it does not explicitly model temporal relations, leaving state changes and motion to be inferred by the downstream model. Joint-embedding predictive architectures (JEPAs) learn representations by predicting target features from visible context rather than reconstructing pixels. I-JEPA Assran et al. (2023) establishes this principle for images, and V-JEPA Bardes et al. (2024) extended it to video. V-JEPA 2 Assran et al. (2025) scales action-free video pretraining and demonstrates strong motion understandings, while V-JEPA 2.1 Mur-Labadia et al. (2026) improves spatial grounding through dense context loss and deep self-supervision. Concurrent JEPA-WAM Lin et al. (2026) predicts structured V-JEPA targets using a predictor shared with action generation. CSWAM instead uses a dedicated semantic expert to predict future V-JEPA features from sparse observation history and shares the history-derived context with both the video and action streams.

2.3 OOD generalization in robot manipulation

Visual distribution shifts remain a persistent challenge for robot policies trained from demonstrations. Studies of visuomotor behavior cloning show that even pretrained visual encoders can be sensitive to subtle changes in lighting, scene texture, and distractor objects, and that in-distribution accuracy alone does not predict robustness Burns et al. (2025). Scaling to VLAs does not remove this failure mode; similar failures have been reported under controlled visual distribution shifts Fei et al. (2025). Cross-paradigm evaluations further suggest that WAMs can benefit from spatiotemporal video priors, but their robustness depends on how those priors are incorporated into action prediction Zhang et al. (2026c). Within WAMs, recent methods improve robustness through different forms of future

or historical conditioning. ST-WAM Wang et al. (2026a) combines DINOv3 future supervision with current-anchored retrieval from semantic history. Faster-WAM Zhao et al. (2026) retains compact future representations during action denoising to balance robustness and inference efficiency. CSWAM instead preserves FastWAM's action-only inference while using video-native V-JEPA history to provide temporally predictive semantic context to both the video and action streams. Our evaluation separates mixed-training performance from OOD transfer through Clean-to-Randomized simulation and compounded real-robot visual shifts.

3 Causal Semantic World Action Model

3.1 Overview

Our goal is to strengthen the observation representation available to a FastWAM-style policy without giving up its action-only deployment paradigm Yuan et al. (2026b). CSWAM retains the video and action streams and adds a semantic stream derived from observation history. The video stream preserves generative visual dynamics, while the semantic stream provides temporally grounded state and motion context that is less tied to the appearance of the current frame. Separate transformer experts process the three streams and exchange information through a Mixture-of-Transformers (MoT) architecture Liang et al. (2024), as illustrated in Fig. 1. During training, future visual and semantic prediction shape these representations alongside action generation; at deployment, the action expert uses only information computed from current and past observations.

We formalize the resulting action policy as follows. Let o_t be the current multi-view observation, $\mathcal{H}_t$ an observation history ending at o_t, and ℓ a language instruction. The video encoder represents the current observation as $z_V^c = E_V(o_t)$, while the frozen semantic encoder represents the history as $z_J^h = E_J(\mathcal{H}_t)$. Given these observed features, the deployed policy predicts an action chunk $a_{t:t+T_a-1} \in \mathbb{R}^{T_a \times d_a}$ according to

$$\hat{a}_{t:t+T_a-1} \sim p_\theta\big(\cdot \mid z_V^c, z_J^h, \ell\big), \tag{1}$$

where T_a is the action horizon and d_a is the action dimension.

3.2 Causal semantic expert

The generative video stream must retain fine appearance details to predict future frames, yet these details can change substantially across environments without altering the underlying task state. We therefore introduce the causal semantic expert as a complementary pathway that supplies action prediction with temporally predictive visual context. Its design combines a frozen V-JEPA feature space, which emphasizes semantic state changes and motion, with a sparse causal history that reveals how these states evolve up to the current observation. The expert learns to extrapolate future semantic features from this observed history.

To expose temporal changes over a longer interval without increasing K, we sample the history at a stride of s raw video frames,

$$\mathcal{H}_t = \big(o_{t-(K-1)s}, \ldots, o_{t-s}, o_t\big). \tag{2}$$

We pass the entire history to a frozen V-JEPA 2.1 encoder Mur-Labadia et al. (2026). Its input embedding partitions the sampled observations into non-overlapping temporal tubelets of size q, after which all spatiotemporal tokens are processed jointly. The encoded history is

$$z_J^h = E_J(\mathcal{H}_t) \in \mathbb{R}^{B \times d_J \times T_h \times H_J \times W_J}, \tag{3}$$

where B is the batch size, d_J is the encoder feature dimension, and $T_h \times H_J \times W_J$ is its spatiotemporal feature grid, with $T_h = K/q$. Each temporal slice therefore aggregates a short local transition, while video attention integrates evidence across the full observed history rather than treating its frames as independent descriptors.

The V-JEPA feature grid is spatially dense. Before it enters the semantic expert, a learnable spatiotemporal projection P_J reduces its spatial token count, preserves its temporal order, and maps it to the expert's hidden space. This produces the observed semantic tokens $J_h = P_J(z_J^h)$.

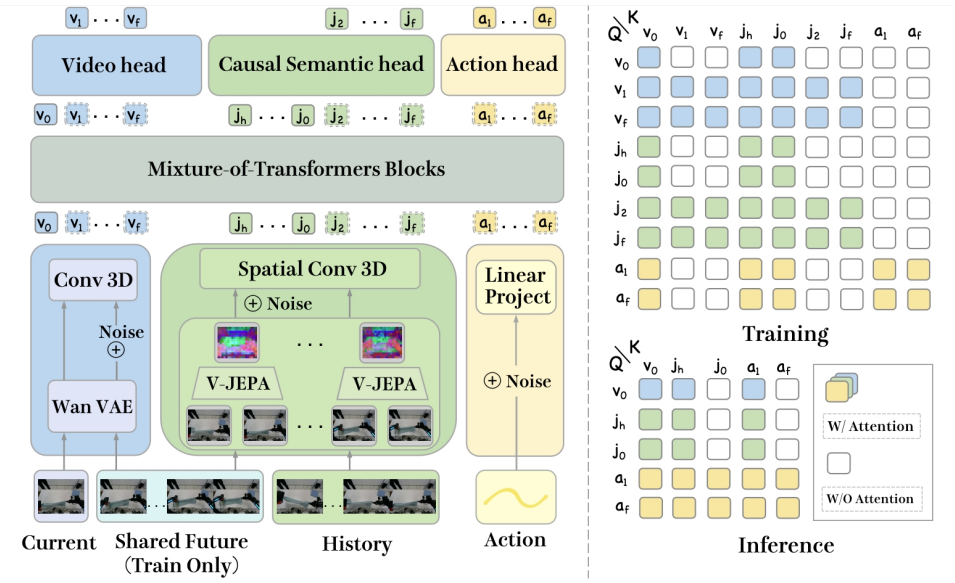


Figure 2: Architecture and attention pattern of CSWAM. During training, future video and semantic tokens provide auxiliary prediction targets, while semantic history conditions both the video and action streams. At inference, the model retains observed context and denoises only actions.

Encoding history makes temporal evidence available, but does not by itself require the trainable semantic pathway to capture predictive state evolution. We therefore train it to extrapolate future V-JEPA features. The same frozen encoder maps a future clip $\mathcal{F}_t$ to the prediction target $z_J^f = E_J(\mathcal{F}_t)$, which we perturb at a sampled semantic noise level τ_J,

$$z_{J,\tau_J}^f = \alpha_{\tau_J} z_J^f + \sigma_{\tau_J} \epsilon_J, \qquad \epsilon_J \sim \mathcal{N}(0, I), \tag{4}$$

where α_{τ_J} and σ_{τ_J} define the noise path, and $J_f = P_J(z_{J,\tau_J}^f)$ denotes the projected future tokens.

We formulate this prediction in the same flow-matching framework as the video and action streams. A diffusion transformer Peebles & Xie (2023) processes $[J_h; J_f]$ and predicts the flow of the future semantic features conditioned on the clean history. An output head maps its prediction to the target feature space, and the loss is applied only to the future tokens. This objective encourages the expert to model semantic state evolution rather than use V-JEPA features as static conditioning.

3.3 Cross-stream context integration

The semantic pathway should inform the original WAM computation rather than serve only as an auxiliary prediction task. We therefore integrate it with the video and action pathways through masked MoT attention. The video expert takes the current VAE latent z_V^c as a clean prefix and learns a flow field for noised future video latents, retaining the appearance detail and generative dynamics of the pretrained video model. ActionDiT operates on a noised action chunk and predicts the corresponding action flow. Let V_c, V_f, and A denote the current-video, future-video, and action tokens.

Together with the causal semantic expert, these pathways form a three-expert MoT. Each expert retains its modality-specific projections and feedforward layers, while masked attention allows information to be exchanged across streams without collapsing their modality-specific representations. The mask is designed both to share observed context and to isolate future prediction targets. In particular, the video tokens attend to the observed V-JEPA history J_h. The semantic history therefore not only conditions the semantic expert, but also enriches the video representation used to learn future visual dynamics. During training, the future video and semantic streams can further exchange information, allowing their complementary prediction objectives to be learned jointly.

The action expert attends to both the current-video tokens V_c and the semantic-history tokens J_h, combining visual dynamics with history-derived semantic state and motion. All three experts are also conditioned on language; Fig. 2 summarizes the resulting information flow.

3.4 JOINT TRAINING AND ACTION-ONLY INFERENCE

Joint training couples the three representations while retaining a separate objective for each stream. The video and semantic experts predict future latent flows, while the action expert predicts the action flow. For stream $m \in \{V, J, A\}$, let u_m and $\hat{u}_m$ denote its target and predicted flow. The joint objective is

$$\mathcal{L} = \sum_{m \in \{V,J,A\}} \lambda_m \, \mathbb{E}\big[w_m(\tau_m) \|\hat{u}_m - u_m\|_2^2\big] \,, \tag{5}$$

where λ_m balances the three streams, τ_m is the sampled noise level, and $w_m(\tau_m)$ is its noise-dependent weight. Each squared error is averaged over valid elements after applying the padding mask of its modality. The video and semantic streams share a sampled noise level, $\tau_V = \tau_J$, aligning the corruption levels of their future targets, whereas the action stream samples τ_A independently. Through the shared MoT attention, the three losses jointly shape the interacting representations; the V-JEPA encoder remains frozen.

At inference, we discard V_f and J_f, compute V_c and J_h once, and reuse them while the action expert performs iterative denoising.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOLS AND SETUP

RoboTwin 2.0. We evaluate 50 bimanual tasks from RoboTwin 2.0 Chen et al. (2025) under two protocols. In **RoboTwin2.0-Full**, following the standard multi-task setting used by Motus, LingBot-VA, FastWAM, and ST-WAM Bi et al. (2026); Li et al. (2026); Yuan et al. (2026b); Wang et al. (2026a), a single policy is trained on 2,500 Clean and 25,000 Randomized demonstrations and evaluated in both settings; this measures mixed-training performance rather than transfer. In **RoboTwin2.0-Clean2Random**, policies are trained only on the 2,500 Clean demonstrations and tested on Randomized scenes without adaptation. We fine-tune for five epochs, select the checkpoint with the lowest validation loss, and report success over 100 rollouts per task.

Clean2Random is evaluated with and without embodied pretraining (PT). For Embodied PT, FastWAM is first pretrained on a 5,000-hour robotics corpus, and CSWAM inherits its video- and action-expert weights. The causal semantic transformer is warm-started from shape-compatible ActionDiT parameters, while its input and output layers are initialized separately.

LIBERO-Plus. We additionally use LIBERO-Plus Fei et al. (2025), which evaluates seven controlled shifts in camera viewpoint, robot initial state, language instruction, lighting, background texture, sensor noise, and object layout. Following the benchmark protocol, we report category-wise and overall success rates. CSWAM uses the embodied-pretrained initialization described above.

Real robot. We use a dual-arm platform with two ARX Robotics manipulators for two long-horizon tasks: sorting toys into containers (Sort) and folding a towel (Fold). Both models use the embodied-pretrained initialization above and are fine-tuned on the same task data—12 hours for Sort and 4 hours for Fold. We evaluate the shifts in Table 1; ID and OOD are defined relative to each task's fine-tuning distribution, and object initialization is randomized at every level. Each task, level, and model is tested for 20 trials (240 total); we report success rate and mean completion time over successful trials. Figure 3 shows the corresponding scenes.

4.2 SIMULATION: STANDARD PERFORMANCE AND OOD TRANSFER

Standard mixed-training performance. On RoboTwin2.0-Full, CSWAM achieves 94.7% on Clean, 94.3% on Randomized, and 94.5% on average, the best results among the listed VLA and WAM baselines (Table 2). It improves over FastWAM by 2.6 points on average and over ST-WAM by 1.6, 1.8, and 1.7 points on Clean, Randomized, and average, respectively. Because both settings appear in training, these results measure mixed-training robustness rather than OOD transfer.

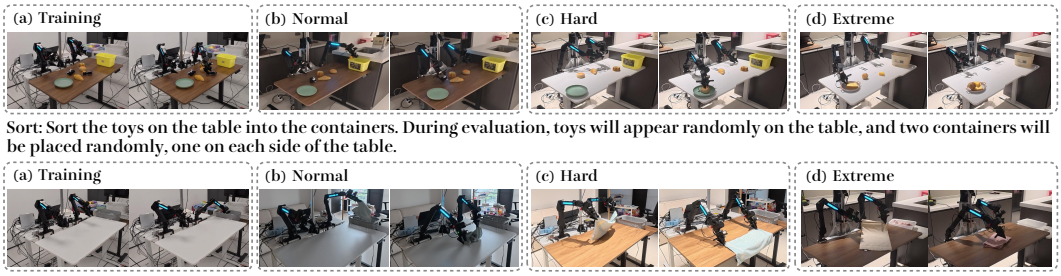


Figure 3: Real-robot evaluation scenes for Sort (top) and Fold (bottom), shown for training and the Normal, Hard, and Extreme OOD settings.

Table 1: Real-robot evaluation settings.

Level	Env.	Lighting	Tabletop	Object	Container
Normal	OOD	ID	ID	ID	ID
Hard	OOD	OOD	OOD	ID	ID
Extreme	OOD	OOD	OOD	OOD	OOD

Table 3: OOD success (%) on RoboTwin2.0-Clean2Random. Embodied PT precedes task fine-tuning.

Method	Embodied PT	Clean	Randomized	Average
FastWAM	No	63.68	1.82	32.75
FastWAM	Yes	76.68	10.16	43.42
CSWAM	No	78.48	9.36	43.92
CSWAM	Yes	**84.46**	**45.18**	**64.82**

Clean-to-Randomized OOD transfer. Mixed-training performance does not establish transfer to unseen visual distributions, so we evaluate the stricter RoboTwin2.0-Clean2Random protocol: both methods are trained only on Clean demonstrations and tested on Randomized scenes without adaptation (Table 3). Without embodied PT, CSWAM achieves 43.92% average success versus 32.75% for FastWAM, an 11.17-point gain independent of embodied-data pretraining. Embodied PT improves both models, but its benefit is substantially larger for CSWAM: average success increases by 10.67 points for FastWAM and by 20.90 points for CSWAM, including a 35.82-point gain on Randomized. With embodied PT, CSWAM ultimately outperforms FastWAM by 7.78 points on Clean and 35.02 points on Randomized, raising the average from 43.42% to 64.82%. The much larger improvement on unseen Randomized scenes indicates that embodied priors and causal semantic modeling are complementary, rather than merely improving fit to the fine-tuning distribution.

LIBERO-Plus robustness. Table 4 compares CSWAM with VLA and WAM baselines under the seven LIBERO-Plus shifts. CSWAM achieves the highest overall success rate of 77.1%, outperforming the strongest baseline, ST-WAM, by 4.3 percentage points. It performs particularly well under camera-viewpoint and lighting shifts, reaching 85.9% and 93.4%, respectively, while its 52.1% success under robot-initial-state shifts identifies variation in the initial robot state as the most challenging category.

4.3 Real-world OOD generalization

We compare CSWAM with FastWAM on Sort and Fold under three increasingly challenging OOD settings (Table 5): Normal changes the environment, Hard additionally changes lighting and tabletop appearance, and Extreme also replaces task objects and containers. CSWAM improves success in every task–difficulty combination.

Table 5: Real-robot success rate (SR, %) and mean completion time for successful trials (seconds).

		SR (%)		Time (s)	
Task	Level	FastWAM	CSWAM	FastWAM	CSWAM
Sort	Normal	65	100	41.1	42.6
	Hard	25	95	73.8	50.3
	Extreme	10	75	79.5	39.8
Fold	Normal	60	100	114.5	136.3
	Hard	5	40	240.0	154.0
	Extreme	0	10	N/A	176.5

Table 2: Success rate (%) on RoboTwin2.0-Full under standard mixed Clean and Randomized training. Baselines are grouped into VLA and WAM methods.

Method	Clean	Randomized	Average
VLA			
X-VLA Zheng et al. (2025)	72.8	72.8	72.8
$\pi_{0.5}$ Physical Intelligence et al. (2025)	82.7	76.8	79.8
ABot-M0 Yang et al. (2026)	86.1	85.1	85.6
Qwen-VLA Wang et al. (2026b)	86.1	87.2	86.7
Galaxea G0.5 Liu et al. (2026)	93.7	92.8	93.3
Qwen-RobotManip Yuan et al. (2026a)	93.7	94.0	93.9
WAM			
Motus Bi et al. (2026)	88.7	87.0	87.8
FastWAM Yuan et al. (2026b)	91.9	91.8	91.9
LingBot-VA Li et al. (2026)	92.9	91.6	92.2
ST-WAM Wang et al. (2026a)	93.1	92.5	92.8
LingBot-VA 2.0 Zhang et al. (2026b)	93.8	93.4	93.6
ABot-M0.5 Chen et al. (2026)	94.0	94.2	94.1
CSWAM (ours)	**94.7**	**94.3**	**94.5**

Table 4: Success rate (%) on LIBERO-Plus.

Method	Emb. PT	Camera	Robot	Lang.	Light	BG	Noise	Layout	Overall
UniVLA	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	Yes	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA (OFT)	Yes	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT	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
FastWAM Yuan et al. (2026b)	No	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
ST-WAM Wang et al. (2026a)	No	55.4	60.1	79.3	93.0	74.2	**79.5**	**74.3**	72.8
CSWAM (ours)	No	**85.9**	52.1	**80.2**	**93.4**	91.8	72.6	73.6	**77.1**

On Sort, CSWAM's gains over FastWAM are 35, 70, and 65 percentage points from Normal to Extreme, yielding an average success rate of 90.0% versus 33.3%. On Fold, the corresponding gains are 40, 35, and 10 points, yielding 50.0% versus 21.7%. Averaged across tasks at each difficulty, CSWAM reaches 100.0%, 67.5%, and 42.5% success under Normal, Hard, and Extreme, compared with 62.5%, 15.0%, and 5.0% for FastWAM. Thus, the advantage persists as visual changes are compounded, although Extreme Fold remains challenging for both methods; FastWAM records no successful Extreme Fold rollout, hence the N/A completion time. The largest average margin occurs under Hard shifts, where CSWAM exceeds FastWAM by 52.5 points across the two tasks. Under Extreme shifts, it retains 42.5% average success while FastWAM falls to 5.0%, suggesting that semantic and temporal evidence becomes especially valuable as appearance cues increasingly depart from training. Across the two tasks and three difficulty levels, CSWAM raises average success from 27.5% to 70.0%; over the four Hard and Extreme settings alone, it raises the average from 10% to 55%. Among successful rollouts, CSWAM averages 44.2 seconds on Sort versus 64.8 seconds for FastWAM, and 145.2 versus 177.3 seconds on the Normal and Hard Fold settings where both methods succeed. These completion times indicate that the success-rate improvement does not require consistently longer successful trajectories. The 39.8-second Extreme Sort result is success-conditioned: difficult failures are excluded, whereas recoveries under milder shifts can contribute longer completion times. Together with the Clean2Random results, these physical experiments show that the benefit extends from controlled simulation shifts to compounded changes in real environments, objects, and appearance.

4.4 Ablation studies

On RoboTwin2.0-Full, we compare CSWAM with two matched ablations: *Without video expert* retains only the semantic and action experts, while *Without history observations* sets the history length to zero; all variants use the same data and optimization budget. Removing history reduces Clean and Randomized success by 1.7 and 1.8 points, showing the value of temporal evidence about object displacement, interaction, and task progress. Removing the video expert causes larger drops of 6.1 and 8.1 points, indicating that fine-grained visual dynamics complement the semantic pathway's higher-level state modeling.

Table 6: Ablation results on RoboTwin2.0-Full.

Variant	Clean	Randomized
Without video expert	88.6	86.2
Without history observations	93.0	92.5
CSWAM	**94.7**	**94.3**

4.5 Component analysis

We first compare V-JEPA 2.1 with DINOv3 under a matched downstream token budget (Table 7). Replacing V-JEPA with DINOv3 reduces success by 1.4 points on Clean and 1.5 points on Randomized. We match the downstream token count: V-JEPA jointly encodes eight frames into four tubelets, whereas DINOv3 encodes four sampled frames independently on the same spatial grid. Despite using a DINOv3 ViT-L encoder with approximately $3.5\times$ the parameters of V-JEPA 2.1 ViT-B, the framewise variant performs worse. This supports jointly encoded temporal features, which expose state transitions and motion more directly than independent image descriptors.

Table 7: Comparison of semantic encoders on RoboTwin2.0-Full.

Semantic encoder	Clean	Randomized
DINOv3 Siméoni et al. (2025)	93.3	92.8
V-JEPA 2.1	**94.7**	**94.3**

We next vary only the history length (Table 8). Increasing the history from 0 to 4 frames improves Clean and Randomized success by 0.6 and 0.5 points, and increasing it from 4 to 8 frames provides a further 1.1- and 1.3-point gain. The larger second gain indicates that four frames do not saturate the useful temporal context; the wider window exposes additional state changes and motion.

Table 8: Effect of history length on RoboTwin2.0-Full.

History frames	Clean	Randomized
0	93.0	92.5
4	93.6	93.0
8	**94.7**	**94.3**

5 Conclusion

We presented CSWAM, which augments an action-only WAM with V-JEPA 2.1 representations of causal observation history and predictive semantic modeling. CSWAM achieves 94.5% average success on RoboTwin2.0-Full, raises Clean2Random OOD success from 10.16% to 45.18% over FastWAM, and improves average success from 10% to 55% across the Hard and Extreme real-robot settings. Ablations and component analyses further verify the benefits of V-JEPA features and longer observation history, while showing that semantic and generative video dynamics remain complementary. These results highlight representation and temporal context as key to improving OOD generalization in world action models.

Generative AI Usage Disclosure

Generative AI tools were used solely for language polishing, including grammar correction, wording refinement, and improvements to readability. All scientific ideas, methods, experiments, analyses, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted edits and take full responsibility for the final content of this work.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Mido Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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

Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? In *Conference on Robot Learning*, pp. 4525–4545, 2025.

Chi-Lam Cheang et al. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Ronghan Chen et al. ABot-M0.5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026.

Tianxing Chen et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Ji Li, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, 2025.

Lin Li et al. LingBot-VA: Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

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

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Yihan Lin, Jiawei He, Shifeng Bao, Chen Zhao, Yang Li, Xiaobo Wang, Yan Wang, Cheng Chi, and Jing Zhang. JEPA-WAM: Learning vision-language-action policies with joint-embedding world modeling. *arXiv preprint arXiv:2608.09381*, 2026.

Yicheng Liu et al. G0.5: One autoregressive stream for robot reasoning and action. *arXiv preprint arXiv:2608.11739*, 2026.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Physical Intelligence, Kevin Black, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Taiyi Su, Jian Zhu, Tianjian Wang, Youzhang He, Zitai Huang, Jianjun Zhang, Chong Ma, Hanyang Wang, Tianjiao Zhang, Munan Yin, Weihao Ding, and Yi Xu. DeMaVLA: A vision-language-action foundation model for generalizable deformable manipulation. *arXiv preprint arXiv:2605.31286*, 2026.

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

Qiuyue Wang et al. Qwen-VLA: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026b.

Yuran Wang et al. OpenWAM: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026c.

Yandan Yang et al. ABot-M0: VLA foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

Seonghyeon Ye et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Haoqi Yuan 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.

Jianjun Zhang, Jian Zhu, Taiyi Su, Chong Ma, Zitai Huang, Yi Xu, and Hanli Wang. Learning 4D geometric priors for inference-efficient world action models. *arXiv preprint arXiv:2607.05468*, 2026a.

Qihang Zhang et al. LingBot-VA 2.0: Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026b.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, Feng Wen, Xinyu Wang, Xingyue Quan, and Yingxue Zhang. Do world action models generalize better than VLAs? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026c.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-WAM: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026.

Jinliang Zheng et al. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

Jian Zhu, Jianjun Zhang, Taiyi Su, Tianbin Liu, Zhangyuan Wang, Kai Xie, Zitai Huang, Chong Ma, Youzhang He, Tianjian Wang, Hanyang Wang, Weihao Ding, and Yi Xu. DSWAM: A dual-system world action foundation model for fine-grained robot manipulation. *arXiv preprint arXiv:2607.04927*, 2026.

Brianna Zitkovich et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.