# CSWAM: Better Causal Semantic Representations for Out-of-Distribution Generalization in World Action Models

**Tianbin Liu**[1], **Jian Zhu**[1,*,†], **Taiyi Su**[1], **Jianjun Zhang**[1,2], **Chong Ma**[1,2], **Zitai Huang**[1,2], **Weiyi Lu**[1], **Yi Xu**[1,†]

[1]AIRC, Midea Group [2]Tongji University

liutb27@midea.com, jianzhu823@gmail.com, xuyi42@midea.com

arXiv:2609.18462v2 [cs.CV] 17 Sep 2026

## 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

[*]Project lead.

[†]Corresponding author.

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.

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. 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).

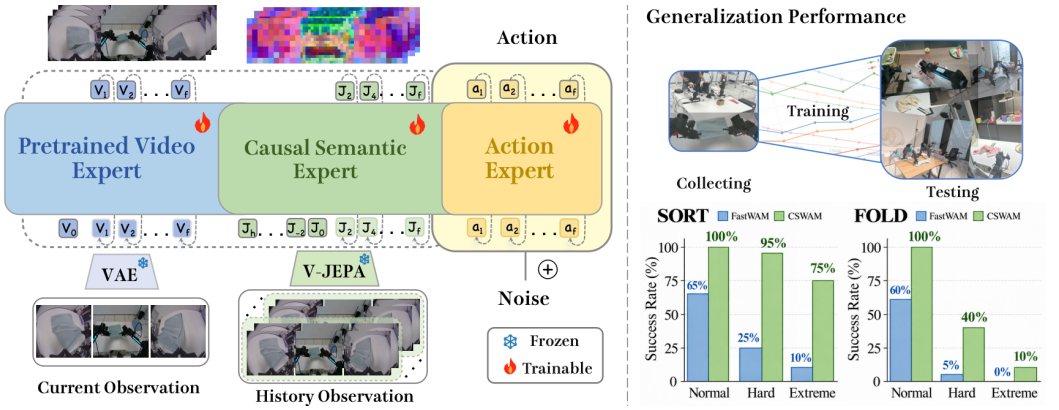


Figure 1: Overview of the proposed CSWAM. Left: CSWAM augments a video–action WAM with a causal semantic expert that encodes observation history through frozen V-JEPA and shares the resulting semantic and motion context with the video and action experts. Right: after training in the collection environment, CSWAM retains substantially higher success than FastWAM as visual shifts increase from Normal to Hard and Extreme on the Sort and Fold tasks. Flames and snowflakes denote trainable and frozen modules, respectively.

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), 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 $\ell$ 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 $J_h = P_J(z_J^h)$.

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 $\tau_J$,

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

where $\alpha_{\tau_J}$ and $\sigma_{\tau_J}$ define the noise path. Its projection, $J_f = P_J(z_{J,\tau_J}^f)$, forms the future semantic 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. We denote the current-video, future-video, and action token groups by $V_c$, $V_f$, and $A$, respectively.

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 $J_h$. It thus learns actions from the appearance and generative dynamics represented by the video expert together with the semantic state and motion represented by the causal semantic expert. All three experts are additionally conditioned on the language instruction.

The resulting information flow is summarized in Fig. 2. Joint video–semantic prediction lets the video stream reuse V-JEPA history, while the action expert learns from current visual dynamics together with history-derived semantic and motion context. The temporal mask keeps future prediction targets separate from the observed context used for action generation.

### 3.4 Joint training and action-only inference

Joint optimization allows the history-derived semantics to be shaped through their interaction with visual dynamics and action prediction, while each stream retains its own learning objective. The video expert predicts future visual dynamics, the semantic expert predicts future V-JEPA features, and the action expert denoises action chunks. Let $m \in \{V, J, A\}$ index these three streams, and let $u_m$ and $\hat{u}_m$ denote the target and predicted flow for stream $m$. Their flow-matching errors are combined as

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

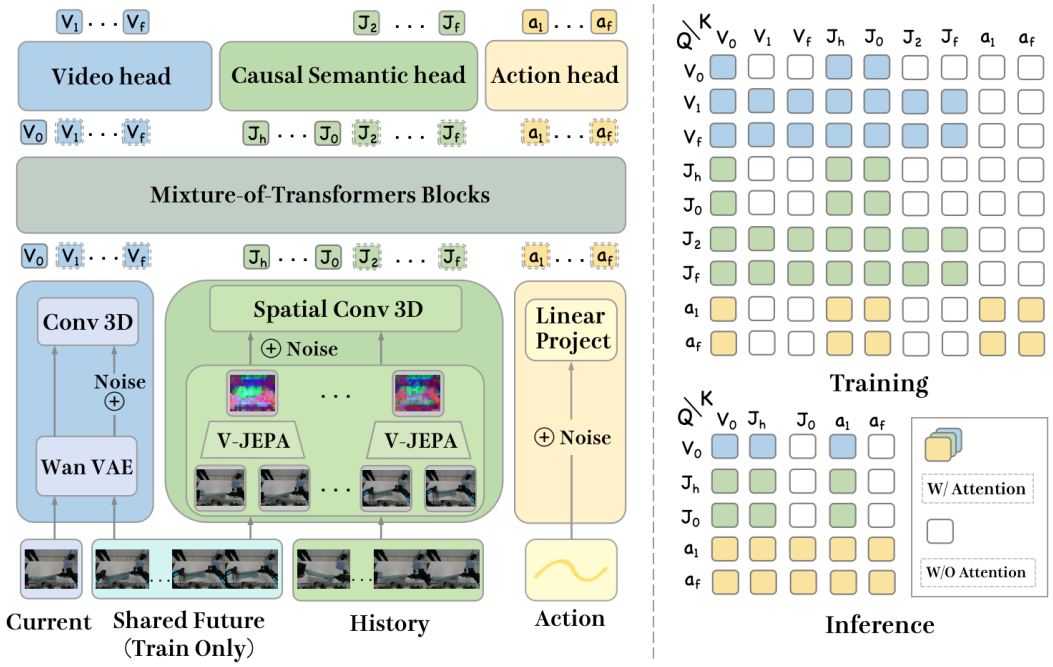


Figure 2: Architecture and attention pattern of CSWAM. Left: The video and causal-semantic streams use the same future clip as training supervision, encoded by the Wan VAE and V-JEPA 2.1 in their respective latent spaces. We use V-JEPA to jointly encode each pair of consecutive sampled frames into one latent time step. The video, semantic, and action experts then exchange information through MoT blocks and produce their modality-specific predictions. Right: The training mask allows future video and semantic tokens to interact for auxiliary dynamics learning, while exposing the semantic history to both the video and action streams. At inference time, only the observed context and action tokens are retained.

where $\lambda_m$ balances the three streams, $\tau_m$ 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 $\tau_A$ independently. Through the shared MoT attention, the three losses jointly shape the interacting representations; the V-JEPA encoder remains frozen.

At inference, the training-only groups $V_f$ and $J_f$ are discarded. The observed states $V_c$ and $J_h$ are computed once and reused across action denoising steps, so only the action expert is evaluated iteratively.

## 4 Experiments

We evaluate CSWAM in RoboTwin 2.0 and on a real dual-arm robot, covering both standard manipulation performance and transfer to unseen visual conditions. RoboTwin provides broad task coverage under mixed Clean and Randomized training as well as a stricter Clean-to-Randomized protocol. The real-robot evaluation complements this controlled simulation setting with two long-horizon tasks under progressively compounded visual shifts. We further use controlled variants to examine the semantic representation, observation history, and generative video pathway.

### 4.1 Evaluation protocols and setup

**Simulation protocols.** We evaluate on 50 bimanual manipulation tasks from RoboTwin 2.0 Chen et al. (2025) using two complementary 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.

Table 1: Real-robot generalization settings. In-distribution (ID) and out-of-distribution (OOD) are defined relative to the task-specific fine-tuning distribution. All settings use an OOD deployment environment; Hard additionally changes lighting and the tabletop, while Extreme also changes task objects and containers.

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

(2026); Yuan et al. (2026b); Wang et al. (2026a), a single policy is trained on 2,500 demonstrations collected in clean scenes and 25,000 demonstrations collected with heavy scene randomization. The same jointly trained checkpoint is evaluated in the Clean and Randomized settings. Because both scene distributions are represented during training, this protocol measures standard mixed-training performance rather than transfer to an unseen distribution.

In **RoboTwin2.0-Clean2Random**, policies are trained using only the 2,500 Clean demonstrations. Clean evaluation is therefore in-distribution (ID), whereas Randomized evaluation is OOD because the policy has not observed the randomized scene distribution during fine-tuning. We evaluate without any adaptation and repeat the comparison with and without embodied pretraining. For *Embodied PT*, FastWAM is first pretrained on a 5,000-hour embodied robotics corpus. CSWAM inherits the resulting video- and action-expert weights. Its causal semantic transformer is warm-started from the shape-compatible ActionDiT parameters, while the semantic input and output layers are initialized separately. The setting without Embodied PT omits this robotics-data pretraining stage.

For both protocols, we run 100 evaluation rollouts per task and report success rate averaged across tasks. We fine-tune for five epochs and select the checkpoint with the lowest held-out validation loss.

**Real robot.** We conduct experiments on a dual-arm platform equipped with two ARX Robotics manipulators. The evaluation covers two long-horizon bimanual tasks: sorting toys into containers (Sort) and folding a towel (Fold). The FastWAM baseline is pretrained on the 5,000-hour corpus described above. CSWAM loads its pretrained video- and action-expert weights and warm-starts its causal semantic transformer from the ActionDiT parameters; the semantic input and output layers are initialized separately. Both models are then fine-tuned on the same task-specific demonstrations—12 hours for Sort and 4 hours for Fold. Each task is evaluated under the three controlled distribution shifts defined in Table 1.

A rollout is successful if it satisfies the task-specific completion criterion within the time limit. To avoid evaluating a single fixed layout, we randomize object initialization at every difficulty level. For Sort, toys are scattered across the tabletop and containers are placed within their respective side regions; for Fold, the towel is initialized at a random position inside the basket. These layout variations accompany the controlled visual shifts in Table 1. Detailed task definitions, success criteria, and time limits are provided in the supplementary material. We conduct 20 trials for each task, difficulty, and model combination, yielding 240 real-robot rollouts, and report success rate together with mean completion time over successful trials.

**Implementation.** For all experiments, the three camera views are composed into a single mosaic and resized to $384 \times 320$ before visual encoding. The video expert is initialized from pretrained Wan2.2-TI2V-5B weights, and the ActionDiT transformer from a dimension-adapted version of the same backbone. Depending on the experimental setting, these experts are either fine-tuned directly on RoboTwin or first pretrained on the 5,000-hour embodied corpus. CSWAM uses the corresponding video- and action-expert weights. Its causal semantic transformer is warm-started from shape-compatible ActionDiT parameters, while the modality-specific input and output layers are initialized separately. Optimization is performed in bfloat16 precision with a cosine learning-rate schedule, a peak learning rate of $4 \times 10^{-4}$, weight decay of $10^{-2}$, and a maximum gradient norm of 1.0.

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 3: OOD generalization on RoboTwin2.0-Clean2Random. All models are fine-tuned only on Clean demonstrations. Embodied PT denotes additional embodied-data pretraining before this fine-tuning stage.

| 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** |

### 4.2 Simulation: standard performance and OOD transfer

**Standard mixed-training performance.** We first evaluate whether the proposed representation improves robustness under broad training coverage without compromising standard task performance.

On RoboTwin2.0-Full, CSWAM achieves the best performance in both settings, with 94.7% on Clean, 94.3% on Randomized, and 94.5% on average (Table 2). It surpasses the strongest listed VLA baseline, Qwen-RobotManip, and the strongest listed WAM baseline, ABot-M0.5. Relative to FastWAM, CSWAM improves the average success rate by 2.6 percentage points; relative to the closely related ST-WAM, it improves Clean, Randomized, and average success by 1.6, 1.8, and 1.7 points, respectively. These results establish that the semantic pathway and causal history do not trade away performance when broad scene variation is already covered during training. They do not by themselves establish OOD transfer, however, because randomized scenes are part of the RoboTwin2.0-Full training distribution.

**Clean-to-Randomized OOD transfer.** Strong performance under mixed training does not establish generalization to unseen visual distributions. We therefore turn to the stricter RoboTwin2.0-Clean2Random protocol, in which both methods are trained only on Clean demonstrations and tested on Randomized scenes without adaptation. Table 3 reports results with and without embodied pretraining.

Without embodied pretraining, CSWAM improves FastWAM from 63.68% to 78.48% on Clean and from 1.82% to 9.36% on Randomized, increasing the average by 11.17 percentage points.

Embodied pretraining benefits both architectures, but its effect on OOD transfer differs markedly. For FastWAM, it raises Clean, Randomized, and average success by 13.00, 8.34, and 10.67 points, respectively. For CSWAM, the corresponding gains are 5.98, 35.82, and 20.90 points. Within this

Table 4: Real-robot success rate (SR, %) and mean completion time over successful trials (seconds). N/A indicates that FastWAM had no successful rollout in that setting.

| | | 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 |

comparison, the stronger CSWAM architecture therefore extracts substantially more OOD benefit from the same pretrained video and action initialization. This suggests that embodied pretraining and causal semantic modeling are complementary: broad priors in the pretrained video and action experts become more transferable when combined with history-conditioned semantic dynamics.

With embodied pretraining, CSWAM consequently improves over FastWAM by 7.78 points on Clean and 35.02 points on Randomized, raising average success from 43.42% to 64.82%. The much larger gain on the unseen Randomized distribution than on Clean provides direct evidence that the improvement is not merely better fitting of the fine-tuning distribution. It instead supports our hypothesis that causal semantic modeling improves transfer under visual shift.

### 4.3 Real-world OOD generalization

RoboTwin2.0-Clean2Random isolates visual transfer at scale, but simulation alone cannot capture the compounded changes encountered in deployment. We therefore evaluate whether the same advantage persists on a physical robot. All three difficulty levels use an unseen deployment environment. Normal changes the environment alone, Hard additionally changes lighting and tabletop appearance, and Extreme further replaces task objects and containers.

Under the controlled shifts in Table 4, CSWAM improves success in every task–difficulty combination. On Sort, the gains over FastWAM are 35, 70, and 65 percentage points from Normal to Extreme, yielding an average SR of 90.0% versus 33.3%. On Fold, the corresponding gains are 40, 35, and 10 points, yielding 50.0% versus 21.7%. Across all 120 rollouts per model, CSWAM reaches 70.0% success compared with 27.5% for FastWAM, an absolute gain of 42.5 points. Restricting the comparison to the four Hard and Extreme settings, where multiple shifts are compounded, CSWAM raises average success from 10% to 55%. The improvement therefore extends beyond synthetic randomization to two physically distinct manipulation skills. Extreme Fold, with simultaneous environment, appearance, object, and container changes, remains challenging despite the overall gain.

Among successful rollouts, CSWAM averages 44.2 seconds across the three Sort settings, compared with 64.8 seconds for FastWAM. For Fold, the corresponding averages over Normal and Hard, where both methods record successful rollouts, are 145.2 and 177.3 seconds. The 39.8-second result for CSWAM on Extreme Sort reflects the success-conditioned nature of this metric. Under milder shifts, a rollout may recover through multiple attempts and still contribute a long completion time; under Extreme, similarly difficult rollouts are more likely to fail and are excluded, leaving relatively direct successes in the reported average.

### 4.4 Ablation studies

We compare CSWAM with two ablated baselines under the RoboTwin2.0-Full protocol. *Without video expert* removes the generative video pathway while retaining the semantic and action experts. *Without history observations* removes all history frames while retaining the remaining architecture. All variants use the same training data and optimization budget.

Table 5: 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** |

**History observations.** Removing the history input reduces Clean and Randomized success by 1.7 and 1.8 points, respectively. The similar degradation in both settings indicates that the benefit is not tied to a particular visual distribution. Past observations provide direct evidence of object displacement, gripper–object interaction, and task progress that cannot be fully recovered from the current observation alone.

**Video expert.** Removing the video expert produces a substantially larger drop of 6.1 points on Clean and 8.1 points on Randomized. Predicting visual dynamics supplies fine-grained supervision about geometry, contact, and local motion, whereas the semantic pathway emphasizes higher-level state evolution. The larger degradation on Randomized further suggests that this detailed dynamics signal remains important under appearance variation. The full model performs best by combining the two pathways rather than replacing video prediction with semantic prediction.

### 4.5 Component analysis

We further analyze two design choices within the semantic pathway while keeping the rest of CSWAM unchanged. The encoder comparison isolates the semantic representation under a matched downstream token budget, and the history study fixes V-JEPA 2.1 while varying the amount of temporal context.

Table 6: 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** |

**Semantic encoder.** Table 6 compares V-JEPA 2.1 with DINOv3. 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, supporting the value of jointly encoded temporal features. The consistent gains in both Clean and Randomized settings also indicate that the improvement is not specific to one appearance regime. Unlike independently encoded image descriptors, V-JEPA tubelets integrate changes across adjacent observations before they enter the semantic expert, making state-transition and motion cues more directly available to the policy.

Table 7: 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** |

**Number of history frames.** As shown in Table 7, 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. Overall, eight history frames improve Clean and Randomized success over zero history by 1.7 and 1.8 points. The larger incremental gain from 4 to 8 frames

shows that the useful temporal context is not saturated at four frames: the wider observation window exposes additional state changes and motion evidence. Together with the encoder comparison, these results show that both the temporal representation and the amount of observed context contribute to performance.

## 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.

## 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. $\pi_0$: 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.