# SEMA-WAM: HIERARCHICAL SEMANTIC DISTILLATION FOR ROBUST ACTION GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Leveraging spatiotemporal priors from pretrained video generators, World Action Models (WAMs) jointly forecast future observations and robot actions so that anticipated scene evolution can inform policy learning. Yet pixel-centric supervision typically overlooks high-level semantics needed for robust decision-making, limiting the semantic grounding of learned actions. Prior work either predicts action-relevant features from pretrained visual encoders as auxiliary targets or injects them as conditioning, leaving their effect on the core action representation largely indirect. Going beyond such indirect formulations, we propose SEMA-WAM, a hierarchical semantics-aware Mixture-of-Transformers that internalizes multi-source semantic knowledge in the action pathway. SEMA-WAM comprises (1) a multi-source semantic bank that fuses SigLIP2, semantic masks, and DINOv3 for action-relevant perception; (2) a hierarchical injection that couples dense mask structure with compressed foundation-model tokens, aligning fine spatial detail with high-level context; and (3) an action-centric distillation that transfers this hierarchy into a compact action pathway, discarding the mask stream and bank at inference. SEMA-WAM achieves competitive performance, reaching $98.3\%$ on standard LIBERO and $77.3\%$ on LIBERO-Plus, outperforming FAST-WAM-JOINT by $8.9$ points on LIBERO-Plus.

## 1 INTRODUCTION

World Action Models (WAMs) jointly forecast future observations and robot actions, allowing anticipated scene evolution to inform policy learning while inheriting spatiotemporal priors from pretrained video generators (Yuan et al., 2026b; Wan et al., 2025). Coupled video–action designs such as Motus and Fast-WAM train video and action experts together under Mixture-of-Transformers architectures (Bi et al., 2026; Yuan et al., 2026b). Seer and Video Prediction Policy likewise link visual foresight to control (Tian et al., 2025; Hu et al., 2024), while DreamZero and LingBot-VA further scale video–action learning on large pretrained generators (Ye et al., 2026; Li et al., 2026a). Unified world models and Cosmos-style fine-tuning similarly couple video dynamics with action (Zhu et al., 2025; Kim et al., 2026). These WAMs complement observation-conditioned VLAs that map language and vision to actions without an explicit future model, including RT-2 and $\pi_0$ (Brohan et al., 2023; Black et al., 2024), as well as $\pi_{0.5}$ and Octo (Intelligence et al., 2025; Team et al., 2024), and open generalist checkpoints such as OpenVLA (Kim et al., 2024). Despite this coupling, however, the futures that most WAMs supervise remain largely pixel-centric, so high-level semantics that are crucial for robust decision-making are only weakly grounded in the representations that drive action generation.

To address this problem, compact representation spaces transfer upstream world knowledge more effectively for action learning (Wang et al., 2026c;a), and the choice of predicted "world" itself strongly shapes control (Chen et al., 2026a). Against this backdrop, recent methods introduce structured futures beyond RGB, yet typically as an auxiliary stream or a conditioning signal rather than as competence owned by the action pathway (Yan et al., 2026a). Privileged foresight distillation further shows that richer training-time generative supervision can be compressed into a leaner inference-time action pathway (Fang et al., 2026). Among these efforts, ST-WAM and Mask World Model place semantic structure at the center of world–action learning, via stable semantic features or mask-based foresight (Wang et al., 2026b; Lou et al., 2026). Together they show that semantics matter for



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

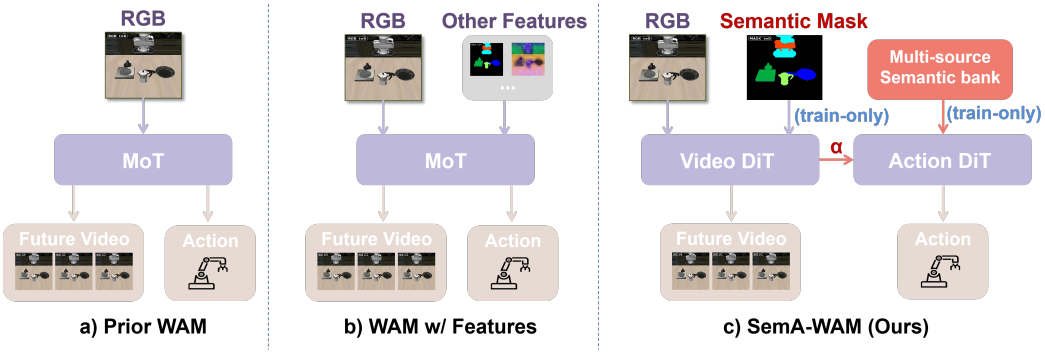



Figure 1: Overview of WAM designs. (a) Prior WAMs decode future video and actions from RGB with a single MoT. (b) Feature-augmented priors concatenate additional cues into the same MoT. (c) SEMA-WAM feeds RGB and a semantic mask into Video DiT, injects a multi-source semantic bank into Action DiT, and couples the two experts with a learnable gate $\alpha$. The semantic mask and the semantic bank are used only during training and are discarded at inference.

control, yet still treat them mainly as an auxiliary target rather than competence internalized by the action expert itself.

SEMA-WAM goes beyond such indirect formulations with a hierarchical, train-time semantic pathway on a Mixture-of-Transformers factorization (Liang et al., 2024; Yuan et al., 2026b), internalizing multi-source semantic knowledge in the action expert. Keeping the Video DiT / Action DiT split of FAST-WAM (Yuan et al., 2026b) and the frozen Wan VAE (Wan et al., 2025) for RGB and colorized masks, we tokenize each modality separately and concatenate them along the sequence. Video DiT co-predicts photometric and semantic mask futures with a shared DiT backbone but independent heads (Peebles & Xie, 2023). Mask tokens serve as keys and enter Action DiT only via a residual branch scaled by a zero-initialized per-layer gate $\alpha$, while asymmetric routing in Video DiT blocks RGB queries from attending to mask keys.

Complementing this dense stream, a multi-source semantic bank supplies Action DiT alone with a compact readout. Frozen SigLIP2 (Tschannen et al., 2025) encodes the current dual-view observation and task sentence into a memory compressed by learned slots (Jaegle et al., 2021); mid-depth Video DiT mask tokens serve as keys, and frozen DINOv3 features (Siméoni et al., 2025), aggregated in the spirit of Representation Autoencoders (Singh et al., 2026), as values. Cross-attention fuses these cues into a short bank appended after proprioception. Figure 1 contrasts this hierarchy with prior MoT WAMs. Both the mask stream and the semantic bank are training-only and dropped at inference. We refer to this as *action-centric privilege distillation*: privileged pathways supervise Action DiT during training and are removed at test time, so semantic competence is retained by the action pathway through joint optimization, in the spirit of PFD (Fang et al., 2026) and MECo-WAM (Zhang et al., 2026).

We evaluate this distilled setting on standard LIBERO and on LIBERO-Plus (Fei et al., 2025), and further on a Franka robot. With no mask stream and no semantic bank at test time, SEMA-WAM reaches 77.3% success on the official 10,030-episode LIBERO-Plus split, exceeding FAST-WAM-JOINT (68.4%) by 8.90 percentage points, while retaining 98.3% on standard LIBERO. Our contributions are:

- A hierarchical semantics-aware Mixture-of-Transformers that organizes dense mask futures and foundation-model cues so multi-source semantic evidence is internalized by the action pathway rather than kept as an online auxiliary channel.
- We propose an *action-centric privilege distillation* protocol in which the mask stream and semantic bank act as training-only privileged pathways that are removed at inference, transferring their benefit into Action DiT by co-training rather than by classical feature knowledge distillation.
- Empirical evidence on LIBERO, LIBERO-Plus, and real-robot evaluation that this pathway improves robustness and yields more task-focused attention and trajectories, with each semantic module contributing measurably under ablation.

## 2 Related Work

**Vision-language-action models.** Mapping observations and language to actions is a central route to generalist manipulation. RT-2 transfers vision–language priors into policies (Brohan et al., 2023), and $\pi_0/\pi_{0.5}$ scale flow-based action generation (Black et al., 2024; Intelligence et al., 2025). Open generalist recipes such as Octo and OpenVLA further release trainable checkpoints (Team et al., 2024; Kim et al., 2024). SemanticVLA improves efficiency by semantic-aligned sparsification and enhancement of visual tokens for manipulation (Li et al., 2026b). These VLA systems show that multimodal priors transfer well to control, yet they typically act from the current observation without modeling how the visual world will evolve, which motivates world action models that couple future prediction with action generation.

**Early world action models.** World action models jointly predict future visual states and robot actions. Seer and Video Prediction Policy link visual foresight to control (Tian et al., 2025; Hu et al., 2024), while Motus and Fast-WAM train video and action experts together under Mixture-of-Transformers designs (Bi et al., 2026; Yuan et al., 2026b). DreamZero and LingBot-VA further scale video–action learning on pretrained generators (Ye et al., 2026; Li et al., 2026a). These early WAMs mainly define the future in pixel or VAE-reconstruction space, where task-relevant state is entangled with appearance, motivating a shift toward richer representations for action learning.

**Representation-aware world action models.** Recent WAMs therefore explore structured alternatives to RGB futures. FlowWAM uses optical flow (Chen et al., 2026b); Flex-$\pi$ and DreamWAM add geometry and semantic streams alongside RGB (Yan et al., 2026a; Yuan et al., 2026a); Mask World Model and MaskWAM study mask futures and prompting (Lou et al., 2026; Yu et al., 2026); and GeoSem-WAM and Robust-WAM inject geometric or semantic objectives into world–action learning (Ma et al., 2026; Yan et al., 2026b). SG-WAM further supplies text-grounded, spatially aware semantic guidance for world–action learning (He et al., 2026). Closest to our motivation, ST-WAM shows that semantic features remain more stable under visual shifts than reconstruction latents (Wang et al., 2026b). These studies of representation mechanisms inspire our hierarchical semantic pathway for action generation.

## 3 Method

SEMA-WAM jointly predicts future RGB and semantic masks with a Video DiT, and predicts actions with an Action DiT (Figure 2). It retains the Mixture-of-Transformers factorization (Liang et al., 2024) of FAST-WAM (Yuan et al., 2026b), so the action expert reads visual tokens together with language and proprioception. Mask evidence enters Action DiT only through an asymmetric residual scaled by a learnable per-layer gate $\alpha$. A multi-source semantic bank aggregates vision and language cues with mid-depth mask geometry, and this bank is provided to Action DiT alone.

### 3.1 Preliminaries

A training sample comprises an RGB video $x^{\mathrm{rgb}}$, a colorized semantic-mask video $x^{\mathrm{mask}}$, a language instruction, proprioception, and an action chunk $a$. Index $t{=}0$ denotes the current observation; $t{>}0$ are futures to be predicted. The frozen Wan VAE (Wan et al., 2025) maps each video to a 48-channel latent. A shared 3-D patch embedding with patch size $1{\times}2{\times}2$ tokenizes each latent into $N$ tokens of width $d{=}3072$ for the 30-layer Video DiT (Peebles & Xie, 2023), whose token positions use 3-D rotary embeddings. Action DiT instead linearly maps the chunk to width $d_a{=}1024$ and applies 1-D rotary position encodings along the action sequence, also with 30 layers. Flow matching corrupts a clean latent $z$ with noise $\epsilon$ as $z_\sigma = (1-\sigma)z + \sigma\epsilon$ and regresses a velocity field toward $\epsilon - z$ (Yuan et al., 2026b). We keep this objective and the pretrained Wan / Action DiT initialization; the changes below concern only how semantic evidence is represented and routed.

### 3.2 Semantic mask prediction

Simply concatenating mask and RGB latents along the channel axis would expand the Wan stem from 48 to 96 channels and invalidate the pretrained convolutional weights. SEMA-WAM therefore keeps the native 48-channel VAE interface for both modalities and couples them only after

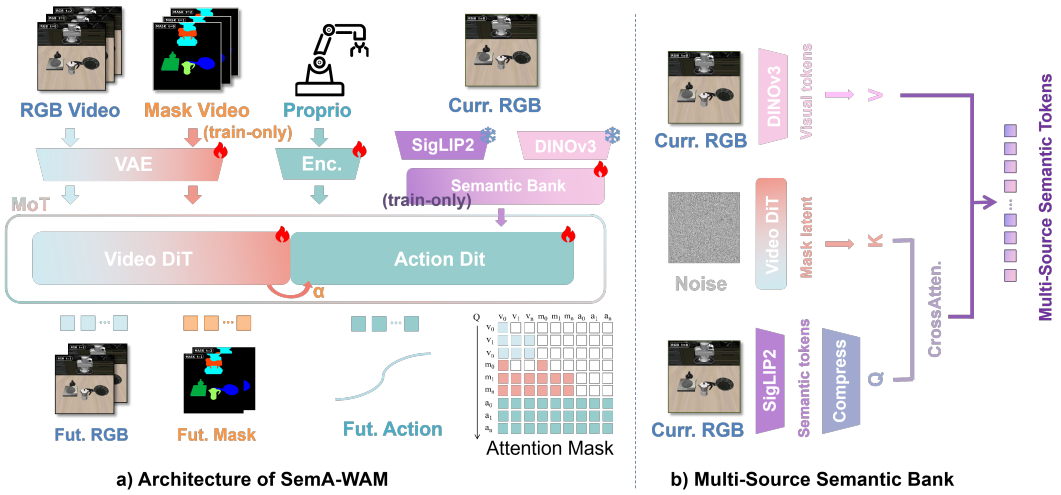



Figure 2: Architecture of SEMA-WAM. (a) A Mixture-of-Transformers jointly forecasts future RGB frames, future semantic masks, and future actions. Video DiT and Action DiT are coupled by a learnable residual gate $\alpha$, and a multi-source semantic bank conditions Action DiT alone; the attention map illustrates the resulting cross-modal token routing. The mask stream and the semantic bank are training-only and discarded at inference. (b) The semantic bank compresses SigLIP2 tokens into queries, uses mid-depth Video DiT mask latents as keys and DINOv3 visual tokens as values, and aggregates them via cross-attention into multi-source semantic tokens that are appended after proprioception in the Action DiT context.

tokenization. Let

$$z^{\mathrm{rgb}} = \mathrm{VAE}(x^{\mathrm{rgb}}), \qquad z^{\mathrm{mask}} = \mathrm{VAE}(x^{\mathrm{mask}}) \tag{1}$$

denote the frozen Wan latents. A shared 3-D patch embedding $E_{\mathrm{patch}}$ maps each latent to $N$ tokens of width $d$, which are then concatenated along the sequence axis,

$$h^{\mathrm{rgb}} = E_{\mathrm{patch}}\big(z^{\mathrm{rgb}}\big), \qquad h^{\mathrm{mask}} = E_{\mathrm{patch}}\big(z^{\mathrm{mask}}\big), \tag{2}$$

$$h^{\mathrm{dual}} = \big[h^{\mathrm{rgb}};\, h^{\mathrm{mask}}\big] \in \mathbb{R}^{2N\times d}. \tag{3}$$

Both streams share $E_{\mathrm{patch}}$ and the Video DiT backbone, and 3-D rotary positions assigned locally within each stream distinguish mask tokens from RGB tokens without a stream-type embedding. After the shared backbone, SEMA-WAM reuses Wan's pretrained prediction head for RGB and a duplicated copy with separate parameters for the mask stream, each trained independently to predict $v_{\mathrm{rgb}}$ and $v_{\mathrm{mask}}$. The model thus learns a semantic-centric future in parallel with the photometric future.

Semantic mask video is available only at training time (Figure 2). The first mask frame $x_0^{\mathrm{mask}}$, i.e., $z_0^{\mathrm{mask}}$, is a condition rather than a reconstruction target; only $t{>}0$ mask frames are supervised. During training, SEMA-WAM drops the first-frame mask condition with probability $p{=}0.5$ by replacing it with zeros, while still supervising future mask prediction.

$$z_0^{\mathrm{mask}} \xleftarrow{p=0.5} 0. \tag{4}$$

### 3.3 GATED MASK–ACTION COUPLING

The dual-stream design does not by itself regulate how strongly mask evidence influences action generation. A single shared softmax over RGB and mask keys allows an imperfect mask stream to dominate the action representation. SEMA-WAM therefore factors each Action DiT layer $\ell$ into an RGB base branch and a mask residual branch with separate softmaxes,

$$y_\ell^{\mathrm{base}} = \mathrm{Attn}\big(Q_\ell,\, K_\ell^{\mathrm{rgb}},\, V_\ell^{\mathrm{rgb}};\, a\big) \tag{5}$$

$$y_\ell^{\mathrm{mask}} = \mathrm{Attn}\big(Q_\ell,\, K_\ell^{\mathrm{mask}},\, V_\ell^{\mathrm{mask}}\big) \tag{6}$$

$$y_\ell = y_\ell^{\text{base}} + \alpha_\ell \, y_\ell^{\text{mask}} \tag{7}$$

In both branches, $Q_\ell$ is formed from action tokens. The ; $a$ in equation 5 denotes concatenation of RGB and action tokens as the base key/value set. equation 6 keeps the same action query but restricts keys and values to mask tokens only. The scalar $\alpha_\ell$ is a learnable per-layer gate initialized at 0, so Action DiT starts from the RGB-only pathway while the mask stream can still be trained by $\mathcal{L}_{\text{mask}}$. Inside Video DiT, asymmetric routing further prevents RGB queries from attending to mask keys, which preserves the pretrained photometric pathway. During training, Action DiT attends to the full dual-stream token sequence, including predicted future RGB and mask tokens.

### 3.4 Multi-source semantic bank

Mask futures capture object-level structure, but the current observation still carries complementary semantic cues. SEMA-WAM aggregates three sources into a fixed-length bank $S$ that is consumed only by Action DiT (Figure 2b). Video DiT continues to read language and proprioception and remains unaffected by the semantic bank.

**Construction.** The current RGB frame $I_0$ enters SEMA-WAM as a single image, formed by horizontally concatenating the two camera views. The semantic bank then splits $I_0$ along its width into the two views and encodes each view independently. Frozen SigLIP2 (Tschannen et al., 2025) yields non-pooled vision patch tokens $P_v$. For language, the semantic bank first removes the long Wan instruction wrapper and keeps only the task sentence. SigLIP2 then encodes that sentence as token-level last hidden states $H_t$, rather than a single pooled contrastive embedding. The semantic bank projects both features into a shared width $d_s$=1024 and concatenates them into one source that jointly carries visual and semantic information,

$$M = \big[\text{proj}_v(P_v); \text{proj}_t(H_t)\big]. \tag{8}$$

A learned view embedding is added to the vision tokens before this concatenation, so that the two cameras remain distinguishable after their tokens are merged into $M$.

To obtain a compact readout, the semantic bank compresses $M$ once into sixteen tokens with learned queries, each carrying a fixed sinusoidal slot identity (Jaegle et al., 2021),

$$Q = \text{LN}\big(\text{CrossAttn}(q,\, M,\, M)\big) \in \mathbb{R}^{16\times d_s}. \tag{9}$$

Here $q$ denotes the sixteen learned queries. Cross-attention lets these queries select task-relevant information from the full vision and text sequence, instead of forwarding every token and its redundant content. The same restriction applies to the keys. The semantic bank uses only the current mask-condition tokens, the $t$=0 mask frame after condition dropout, reads them through the first 15 Video DiT blocks, detaches the activations, and projects

$$K = \text{proj}_m\big(\text{LN}(h_{15}^{\text{mask}})\big). \tag{10}$$

Limiting the keys to this condition frame supplies the observed mask geometry without reintroducing the remaining mask tokens, which Action DiT already receives through the residual branch.

Values are appearance features from frozen DINOv3 (Siméoni et al., 2025). Inspired by the last-$K$ aggregation in Representation Autoencoders (Singh et al., 2026), the semantic bank averages the last four transformer layers of $I_0$ and projects them to the same width,

$$V = \text{proj}_d\big(\text{LN}(\text{DINO}_{\text{last-4}}(I_0))\big). \tag{11}$$

Before routing, $V$ is linearly resampled to the length of $K$. Cross-attention then yields

$$S = \text{CrossAttn}(Q,\, K,\, V) \in \mathbb{R}^{16\times d_s}. \tag{12}$$

This assignment routes language-grounded queries through mask geometry toward DINO appearance (Figure 2b). The resulting $S$ is thus a fixed set of task-conditioned tokens that fuse language, mask geometry, and DINO appearance from these sources.

**Injection.** SEMA-WAM linearly maps $S$ into the Action DiT text-context dimension and appends it after the proprioceptive token,

$$c^{\text{action}} = \big[c^{\text{T5}};\, c^{\text{prop}};\, \text{proj}_{S\to c}(S)\big]. \tag{13}$$

SigLIP2 and DINOv3 remain frozen. The trained bank parameters are the projections, the view embedding, the query compressor, the router, and the map into the action context.

Table 1: Success rates (%) on the standard LIBERO benchmark.

| Method | Spatial | Object | Goal | LIBERO-10 | Average |
|---|---|---|---|---|---|
| OpenVLA | 84.7 | 88.4 | 79.2 | 53.7 | 76.5 |
| $\pi_0$ | 96.8 | 98.8 | 95.8 | 85.2 | 94.1 |
| SemanticVLA | 98.6 | <u>99.6</u> | 97.6 | 94.8 | 97.7 |
| Fast-WAM | 96.8 | 99.4 | 96.4 | 94.4 | 96.8 |
| Fast-WAM-Joint | 98.6 | 98.6 | 94.4 | 96.6 | 97.0 |
| MWM | <u>98.8</u> | **100.0** | <u>98.2</u> | 96.0 | 98.3 |
| MaskWAM | <u>98.8</u> | **100.0** | <u>98.2</u> | 96.4 | <u>98.4</u> |
| ST-WAM | **99.0** | **100.0** | **99.0** | <u>96.8</u> | **98.7** |
| **Sema-WAM (ours)** | 98.0 | 99.2 | 97.2 | **98.4** | 98.3 |

### 3.5 Training objective

Sema-WAM trains the RGB stream, the mask stream, and the action chunk with the flow-matching velocity loss. A clean latent $z$ is corrupted as $z_\sigma = (1-\sigma)z + \sigma\epsilon$, and the corresponding head regresses the target velocity $\epsilon - z$. Because the first frame is a condition, both visual losses omit $t$=0,

$$\mathcal{L}_{\text{rgb}} = \left\| v^{\text{rgb}} - (\epsilon^{\text{rgb}} - z^{\text{rgb}}) \right\|_2^2, \qquad \mathcal{L}_{\text{mask}} = \left\| v^{\text{mask}} - (\epsilon^{\text{mask}} - z^{\text{mask}}) \right\|_2^2, \tag{14}$$

where $v$ and the latent are the future frames only. The action loss uses the same velocity target on the action chunk,

$$\mathcal{L}_{\text{action}} = \| v_a - (\epsilon_a - a) \|_2^2. \tag{15}$$

The total objective uses unit weights,

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{action}}. \tag{16}$$

$\mathcal{L}_{\text{rgb}}$ and $\mathcal{L}_{\text{mask}}$ update the shared Video DiT and its two prediction heads. $\mathcal{L}_{\text{action}}$ updates Action DiT and the per-layer gates $\alpha_\ell$. Because action queries attend to video keys, this action loss also updates Video DiT. The semantic bank receives gradients only from $\mathcal{L}_{\text{action}}$, since $S$ enters through the action context. SigLIP2 and DINOv3 stay frozen, and the mask features used as keys are detached, so the bank does not update Video DiT along that path.

## 4 Experiments

### 4.1 Setup

**Benchmarks.** For training, simulation uses the official Fast-WAM LIBERO no-noop split with 1,712 demonstrations across Spatial, Object, Goal, and LIBERO-10, and colorized masks from the simulator renderer; the real-robot set has 487 Franka FR3 demonstrations over 12 tasks, with masks from OWLv2 (Minderer et al., 2023) and SAM2 (Ravi et al., 2025). For evaluation, we report success on standard LIBERO and LIBERO-Plus in simulation, and success on a Franka FR3 arm in the real world. Sema-WAM is evaluated without ground-truth or online masks.

**Training.** Sema-WAM is initialized from Wan2.2-TI2V-5B and an interpolated Action DiT, and trained with AdamW, learning rate $10^{-4}$, cosine decay, weight decay $10^{-2}$, shift 5.0, and 1,000 timesteps. Mask-condition dropout is $p_{\text{drop}} = 0.5$, and the global batch size is 128. On LIBERO, Sema-WAM uses 4 GPUs with batch size 2 and accumulation 16 for 10 epochs, about 21,750 steps. On the Franka set, the same objective uses 8 GPUs with batch size 4 and accumulation 4 for 20,000 steps.

### 4.2 Main results

Tables 1 and 3 report success on standard LIBERO and zero-shot success on the official 10,030-episode LIBERO-Plus split. All simulation methods are trained only on standard LIBERO demonstrations. Table 1 covers the four standard suites, Spatial, Object, Goal, and LIBERO-10. Table 3

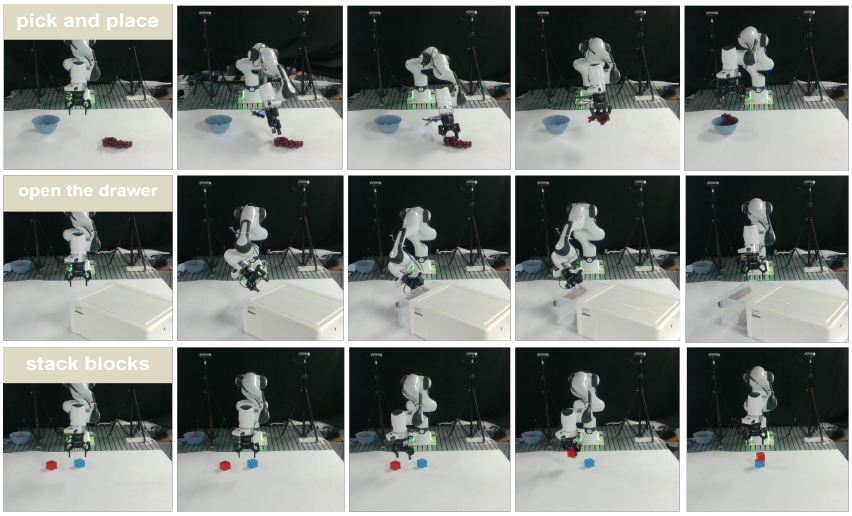



Figure 3: Overview of tasks on Franka.

Table 2: Success counts on Franka.

| Method | Stack blocks | Open/Close drawer | Pick and Place |
|---|---|---|---|
| FAST-WAM | 6/50 | 28/50 | 21/50 |
| FAST-WAM-JOINT | 13/50 | 24/50 | 24/50 |
| **SEMA-WAM (ours)** | **17/50** | **35/50** | **32/50** |

reports the results over the seven perturbation axes. On standard LIBERO, SEMA-WAM reaches 98.3%, within the 96.8%–98.7% range of prior WAMs. On LIBERO-Plus, SEMA-WAM reaches 77.3%, above FAST-WAM (49.9%), FAST-WAM-JOINT (68.4%), and ST-WAM (72.8%). Table 2 further evaluates transfer on a Franka arm under three task families: stack blocks, open/close drawer, and pick and place, with 50 trials per family. Figure 3 provides an overview of real-world Franka rollouts.

**Key Finding 1: Semantic pathways raise robustness, especially under semantic perturbations.** We evaluate SEMA-WAM against recent World Action Models on zero-shot LIBERO-Plus in Table 3, where policies trained only on standard LIBERO face appearance, embodiment, and instruction shifts. While in-distribution LIBERO remains saturated and SEMA-WAM is not the strongest method there, it attains the best overall success on the official LIBERO-Plus split, surpassing the matched joint RGB baseline and contemporary semantic WAMs. This stronger zero-shot transfer indicates improved generalization under distribution shift. The advantage is most pronounced under semantically and appearance-loaded perturbations, where SEMA-WAM leads on the robot-appearance and lighting axes as the hierarchical semantic pathway supplies stable task-relevant structure for Action DiT. Both the mask stream and the semantic bank are discarded at inference. Table 4 further examines how the hierarchical pathway is assembled. Early channel-wise fusion of RGB and mask latents drops sharply on LIBERO-Plus, as stacking modalities before tokenization breaks the pretrained Wan visual prior and prevents the model from reaching the best performance. Sequence-level mask tokens without the residual gate $\alpha$ likewise underperform, indicating that uncontrolled mask injection interferes with RGB-conditioned control. Gated mask coupling already raises LIBERO-Plus substantially, and restoring future RGB further improves zero-shot success. The multi-source semantic bank further strengthens Action DiT robustness under lighting and camera shifts by supplying fused language and appearance cues on top of the mask pathway. The full per-axis breakdown is reported in Table 6 in the appendix. Altogether, the robustness arises from the complete train-time hierarchy that distills multi-source semantics into Action DiT.

Table 3: Zero-shot success rates (%) on the official LIBERO-Plus full set of 10,030 episodes.

| Method | Camera | Robot | Layout | Light | BG | Lang. | Noise | Avg. |
|---|---|---|---|---|---|---|---|---|
| OpenVLA | 0.8 | 3.5 | 28.5 | 8.1 | 34.8 | 23.0 | 15.2 | 15.6 |
| $\pi_0$ | 13.8 | 6.0 | 68.9 | 85.0 | 81.4 | 58.8 | 79.0 | 53.6 |
| Fast-WAM | 15.6 | 43.7 | 60.8 | 79.5 | 52.9 | 68.8 | 38.7 | 49.9 |
| Fast-WAM-Joint | 39.5 | 64.8 | 80.4 | 93.4 | 56.7 | 93.2 | 55.8 | 68.4 |
| DreamWAM | 53.8 | 63.6 | 80.7 | 96.7 | 71.6 | **94.8** | 67.2 | 75.5 |
| SG-WAM | **58.6** | 48.9 | 74.2 | 89.8 | **86.1** | 81.4 | **80.7** | 73.0 |
| ST-WAM | 55.4 | 60.1 | 74.3 | 93.0 | 74.2 | 79.3 | 79.5 | 72.8 |
| **Sema-WAM (ours)** | 55.0 | **81.0** | **81.1** | **97.2** | 73.8 | **94.8** | 63.8 | **77.3** |

**Key Finding 2: Semantic competence is internalized by the action pathway.** Key Finding 1 shows that the hierarchical train-time design improves zero-shot robustness and establishes the effectiveness of the overall architecture. Aggregate success alone, however, does not reveal whether the model itself has acquired semantic competence once privileged cues are removed. We therefore inspect action-to-RGB attention at inference, when neither the mask stream nor the semantic bank is available (Figure 4). On LIBERO-Plus Spatial, Object, and Long suites, Fast-WAM-Joint disperses attention across the scene, whereas Sema-WAM focuses on task-relevant objects and interaction regions. This contrast indicates stronger task-centric perception under visual distribution shift. On Goal, where success depends on following diverse instructions, Sema-WAM favors broader, instruction-adaptive attention rather than a narrowly fixed focus. A similar task-centric pattern appears on the Franka real-robot suite, indicating that the effect transfers beyond simulation. Since inference uses only RGB keys, this attention shift cannot arise from an online semantic channel. It indicates that hierarchical semantic supervision has been internalized by the action pathway.

Table 4: Training-component ablation on LIBERO / LIBERO-Plus. Bold denotes the best result in each column.

| Variant | LIBERO | Plus |
|---|---|---|
| Channel-Concat | 97.3 | 61.0 |
| mask tokens w/o $\alpha$ | 94.7 | 58.3 |
| w/o Fut. RGB | 93.3 | 68.2 |
| w/o Sem. Bank | **98.3** | 76.3 |
| **Sema-WAM (ours)** | **98.3** | **77.3** |

**Key Finding 3: Internalized semantics improve closed-loop execution.** Key Findings 1 and 2 establish that Sema-WAM acquires semantic competence through the hierarchical train-time design and that this competence is expressed in action attention once privileged cues are removed. Success rates and attention maps, however, still leave open whether the same competence reshapes the motion. An attention pattern can become more task-centric without guaranteeing shorter, more direct control; conversely, trajectory regularity would provide behavioral evidence that internalized semantics are used for action generation rather than remaining a perceptual correlate. We therefore compare closed-loop rollouts of Fast-WAM-Joint and Sema-WAM on the full official LIBERO-Plus set of 10,030 episodes. For each episode we record horizon length, end-effector path length, and approach monotonicity toward the final pose (Table 5). Relative to the matched RGB joint baseline, Sema-WAM finishes with shorter horizons and shorter paths, and its approach to the terminal pose is more monotonic. Shorter horizons and paths indicate fewer redundant or corrective detours once the scene departs from the training distribution; higher monotonicity indicates more consistent progress toward the task goal rather than oscillatory or exploratory motion. These trajectory statistics show that the semantic competence identified in Key Findings 1 and 2 is not limited to aggregate robustness or attention focus.

Table 5: Trajectory stats on 10,030 LIBERO-Plus tasks. Horizon: steps. Path: EE path length (m). Mono.: approach monotonicity. Bold denotes the best result in each column.

| Method | Hor. ↓ | Path ↓ | Mono. ↑ |
|---|---|---|---|
| Fast-WAM-Joint | 245.4 | 1.191 | 0.682 |
| **Sema-WAM** | **239.9** | **1.136** | **0.686** |

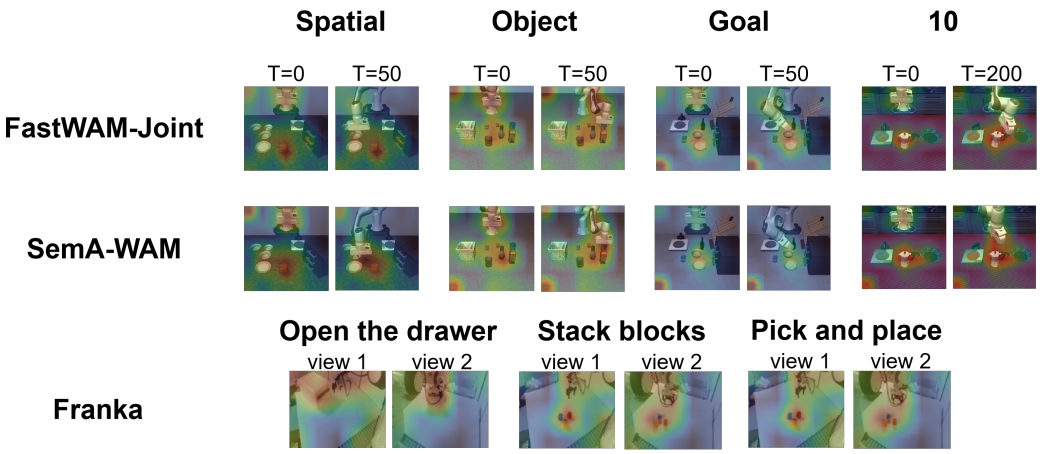



Figure 4: Action-to-RGB attention maps. Top two rows: representative LIBERO-Plus subtasks from the Spatial, Object, Goal, and Long suites, comparing FAST-WAM-JOINT (top) with SEMA-WAM (middle). Bottom: SEMA-WAM on the Franka real-robot suite. Warmer colors indicate stronger attention from Action DiT queries to RGB keys.

Overall, the three findings support a single conclusion. Semantic competence for robust control arises from a train-time hierarchy that is removed at inference, internalized by the action pathway, and expressed jointly in decision attention and executed behavior under distribution shift.

## 5 LIMITATIONS AND BROADER CONSIDERATIONS

SEMA-WAM improves the internalization of semantic evidence in the action pathway and yields measurable gains under distribution shift. Nevertheless, our results also indicate that stronger semantic access is not unconditionally beneficial, since Action DiT must continually trade off privileged semantic cues against the native RGB observation that remains available at deployment. How to allocate, gate, and schedule this balance, so that semantics sharpen task-relevant control without overriding reliable visual evidence, remains an open question. Future work should therefore study adaptive mechanisms for this trade-off, toward more robust semantic use under real-world shift.

## 6 CONCLUSION

In this paper, we present SEMA-WAM, a hierarchical semantics-aware World Action Model that internalizes multi-source semantic knowledge in the action pathway rather than treating semantics as a persistent auxiliary modality. We introduce a train-time semantic hierarchy on a Mixture-of-Transformers: Video DiT co-predicts RGB and mask futures, a multi-source semantic bank fuses foundation-model and mask cues for Action DiT, and a learnable gate residualizes mask evidence into the action expert. Through action-centric privilege distillation, the mask stream and the bank are removed at inference, so their benefit is absorbed into action generation instead of remaining as an online semantic channel. Across multiple evaluations, hierarchical semantic supervision strengthens robustness under distribution shift and encourages more task-centric perception and execution. At the same time, Action DiT must continually balance privileged semantic supervision against the RGB observation that remains available at deployment; how to allocate, gate, and schedule this trade-off is still unresolved. We therefore regard SEMA-WAM as a structured train-time foundation for future work on adaptive semantic–RGB allocation, broader real-world validation, and more reliable semantic use under physical and embodiment shift.

### AI USE STATEMENT

Generative AI tools were used for language editing, outline refinement, and LaTeX formatting assistance. All technical claims, equations, experimental values, and code-facing details were checked against the project materials by the authors, who take responsibility for the final manuscript.

### REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results in the main text and appendix. Section 3 specifies the model architecture, training objectives, and the inference-time privilege-distillation contract. The experimental setup records datasets, evaluation protocols on LIBERO / LIBERO-Plus / Franka, initialization, optimizer settings, and compute schedules. Appendix A details the Channel-Concat baseline, real-robot hardware and closed-loop evaluation settings, and the Franka dataset with mask preprocessing. Training and evaluation scripts, configuration files, and checkpoint-loading instructions will be released publicly upon acceptance.

## REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, et al. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026a.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, et al. Flowwam: Optical flow as a unified action representation for world action models. *arXiv preprint arXiv:2607.13017*, 2026b.

Pengcheng Fang, Hongli Chen, and Xiaohao Cai. Privileged foresight distillation: Zero-cost future correction for world action models. *arXiv preprint arXiv:2604.25859*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Junjie He, Junfeng Li, Zhide Zhong, Haodong Yan, Ruixin Li, Yangyang Zheng, Jiaguan Zhu, Tianran Zhang, Yuqiao Du, Wen Chen, et al. Sg-wam: Text-grounded and spatial-aware semantic guidance for world-action models. *arXiv preprint arXiv:2608.08839*, 2026.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

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

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Wei Li, Renshan Zhang, Rui Shao, Zhijian Fang, Kaiwen Zhou, Zhuotao Tian, and Liqiang Nie. Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18397–18405, 2026b.

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

Yunfan Lou, Xiaowei Chi, Xiaojie Zhang, Zezhong Qian, Chengxuan Li, Rongyu Zhang, Yaoxu Lyu, Guoyu Song, Chuyao Fu, Haoxuan Xu, et al. Mask world model: Predicting what matters for robust robot policy learning. *arXiv preprint arXiv:2604.19683*, 2026.

Fulong Ma, Daojie Peng, Wenjun Yue, Jiahang Cao, Bintao Wang, Qiang Zhang, and Jun Ma. Geosem-wam: Geometry-and semantic-aware world action models. *arXiv preprint arXiv:2606.03188*, 2026.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

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

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 92033–92052, 2025.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junke Wang, Qihang Zhang, Shuai Yang, Yiming Luo, Yujun Shen, Zuxuan Wu, Yu-Gang Jiang, and Yinghao Xu. Repwam: World action modeling with representation visual-action tokenizers. *arXiv preprint arXiv:2606.13674*, 2026a.

Mingxin Wang, Bin Hu, Bin Qian, Kaitao Jiang, Haoning Wu, Feng Yan, 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*, 2026b.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026c.

Ge Yan, Jinghao Liu, Yuzhi Fan, Lei Cai, Minwen Liao, Jesse Zhang, and Dieter Fox. Flex-$\pi$: A multi-stream world-action model with compute flexibility. *arXiv preprint arXiv:2608.10860*, 2026a.

Haodong Yan, Junfeng Li, Junjie He, Zhide Zhong, MingMing Yu, Wenxuan Song, Jiaguan Zhu, Yangyang Zheng, Yuqiao Du, Jiadi You, et al. Robust-wam: Bridging generative pretraining and semantic foresight in world-action models. *arXiv preprint arXiv:2608.05903*, 2026b.

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

Hanyang Yu, Haitao Lin, Jingbo Zhang, Wenyao Zhang, Chenghao Gu, Heng Li, and Ping Tan. Maskwam: Unifying mask prompting and prediction for world-action models. *arXiv preprint arXiv:2606.13515*, 2026.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. Dreamwam: Beyond rgb future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 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*, 2026.

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

## A Additional Experimental Details

**Channel-Concat baseline.** Table 4 reports a Channel-Concat control that fuses RGB and mask before tokenization, rather than after it. Both videos still pass through the frozen Wan VAE, yielding two $48$-channel latents. These latents are stacked along the channel axis into a single $96$-channel tensor and fed to Video DiT at the original sequence length $N$. To accept this input, the pretrained $48$-channel patch embedding is replaced by a $96$-channel convolution whose first half copies the Wan weights and whose second half is zero-initialized; the video prediction head is likewise expanded so that one forward pass emits concatenated RGB and mask velocities. Training otherwise matches the SEMA-WAM protocol, including first-frame mask-condition dropout with probability $0.5$ and a forced zero first-frame mask at evaluation. In contrast, SEMA-WAM keeps the native $48$-channel stem for each modality, tokenizes RGB and mask separately, concatenates the resulting tokens along the sequence axis to length $2N$, and predicts the two futures with independent $48$-channel heads that share only the backbone. Channel-Concat therefore stress-tests early fusion under a broken Wan stem, while SEMA-WAM tests late fusion that preserves the pretrained convolutional interface.

**Detail of ablation.** Table 6 reports the per-axis LIBERO-Plus success of each training-component variant.

Table 6: Detail of ablation on LIBERO-Plus.

| Variant | Camera | Robot | Layout | Light | BG | Lang. | Noise | Avg. |
|---|---|---|---|---|---|---|---|---|
| Channel-Concat | 40.4 | 52.3 | 70.4 | 84.2 | 61.9 | 73.0 | 52.0 | 61.0 |
| mask tokens w/o $\alpha$ | 24.5 | 58.4 | 69.4 | 82.0 | 60.0 | 75.2 | 47.0 | 58.3 |
| w/o Fut. RGB | 40.4 | 75.0 | 74.3 | 89.6 | 63.8 | 80.0 | 59.9 | 68.2 |
| w/o Sem. Bank | <u>48.9</u> | **81.5** | **81.6** | <u>92.6</u> | **77.3** | <u>94.7</u> | <u>63.7</u> | <u>76.3</u> |
| **SEMA-WAM (ours)** | **55.0** | <u>81.0</u> | <u>81.1</u> | **97.2** | <u>73.8</u> | **94.8** | **63.8** | **77.3** |

**Real-robot setup.** We evaluate on a Franka Emika FR3 arm with language-conditioned manipulation. Observations use three RGB views, one wrist and two third-person cameras, each resized to $128\times224$ and composed into the robotwin mosaic of size $384\times320$. The proprioceptive state is 8-D and the action is 7-D end-effector delta control with a binary gripper. Training uses the SEMA-WAM objective with the mask stream and the semantic bank enabled, mask-condition dropout $0.5$, AdamW with learning rate $10^{-4}$, cosine decay, weight decay $10^{-2}$, and bf16. We train on 8 GPUs with per-GPU batch size $4$ and gradient accumulation $4$ (global batch $128$) for $20{,}000$ steps, initializing from Wan2.2-TI2V-5B and the interpolated Action DiT. During closed-loop evaluation, the first-frame mask is forced to zero and the semantic bank is discarded, with $10$ denoising steps and online replanning every $10$ steps.

**Real-robot dataset.** The Franka suite is released as a LeRobot v2.1 dataset with 487 demonstrations over 12 language-conditioned tasks (the *press the switch* task is removed). Teleoperation is recorded at $15$ Hz. Each episode stores three synchronized RGB streams—wrist and two third-person views—at $128\times224$, together with matching colorized mask videos. Proprioception is an 8-D end-effector state $(x, y, z, q_x, q_y, q_z, q_w, \text{gripper})$. Actions are stored as 7-D end-effector deltas $(\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z)$ with an absolute gripper command, in parquet alongside the videos. Semantic masks are not annotated by hand frame by frame: for each RGB video we run open-vocabulary detection with Owlv2 and segment the detections with SAM2, then map the resulting instance masks onto a fixed canonical palette (arm, gripper, objects, and background). The same palette is reused across tasks.