ACTING THROUGH REGISTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) use generative visual representations to condition robot actions, yet how these representations organize action-relevant information remains poorly understood. We identify persistent high-norm spatial outliers in WAM representations and find that their action relevance persists even after normalization and K/V projection attenuate their extreme magnitude. Motivated by this observation, we introduce RegisterWAM, which adds learnable non-spatial registers to the world model and uses them directly for action conditioning. Registers alone do not fully eliminate spatial outliers; pairing them with observation flow-matching (FM) loss supervision suppresses spatial outliers while concentrating high-norm responses in registers. Beyond this reorganization, observation FM loss supervision also improves action learning without registers, motivating an additional noised-observation FM objective that learns current-scene recovery from corrupted inputs. The resulting model achieves 81.64% success on LIBERO-Plus without additional embodied pretraining data or foundation models, reaches 93.0% average success on RoboTwin 2.0, and improves real-world manipulation performance. Code and models will be made public.

1 INTRODUCTION

World Action Models (WAMs) adapt generative video models to jointly model visual dynamics and robot actions (Cheang et al., 2024; Liang et al., 2025; Zhu et al., 2025; Kim et al., 2026). The video backbone encodes the current scene and its possible evolution, and the resulting representations are used to condition an action model (Hu et al., 2024; Liao et al., 2025; Pai et al., 2025). This formulation allows robot policies to benefit from large-scale video pretraining and predictive representations of physical dynamics (Yuan et al., 2026b; Zhang et al., 2026b; Ma et al., 2026a; Zhao et al., 2026). However, how these video representations behave inside WAMs, and how such behavior relates to action generation, remain poorly understood.

To investigate how video representations may support action generation, we examine how spatial tokens evolve across the layers of the video backbone. A striking pattern emerges. In the middle-to-late layers, a small number of tokens develop exceptionally large hidden-state norms, approximately an order of magnitude greater than those of the remaining tokens (Fig. 1). High-norm outliers are not entirely absent from pretrained video models, but they are typically weaker and less consistent. In the WAM studied here, they become substantially more pronounced and persist across different observations. Their amplification in a model adapted for action generation points to a potentially meaningful role, yet their function remains unclear. Then, *are these high-norm outliers merely numerical artifacts, or do they participate in the generation of robot actions?*

To answer this question, we trace the outliers through the visual K/V cache that connects the Video DiT to the Action DiT. Interestingly, we observed that, despite their extreme hidden-state norms, normalization and projection substantially attenuate the magnitude gap in the cached keys and values. The unusually large norms are therefore not directly propagated to the action model. Their functional relevance, however, persists. Token-level interventions on the K/V cache, implemented by masking visual positions from action attention, measure the resulting change in predicted actions, which we term action sensitivity. The entries associated with these outliers retain non-trivial action sensitivity (Fig. 1). Their action sensitivity is not necessarily dominant, as tokens aligned with task-relevant objects often produce larger action changes. What distinguishes the outlier tokens is instead the consistency of their action relevance across observations, even when their spatial locations contain little task-relevant information. Taken together, these results suggest that high-norm outliers

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.

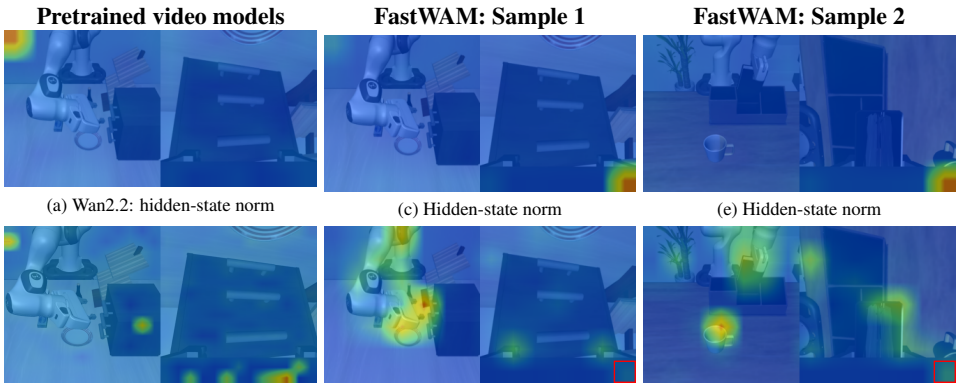


Figure 1: **Spatial outliers and action relevance.** Hidden-state norms and masked-token action sensitivity across video models and FastWAM.

are not simply numerical artifacts. Rather, they provide a visible signature of spatial tokens whose representational roles have become partially decoupled from their nominal spatial locations.

The partial decoupling between token function and spatial location suggests a possible architectural tension. Although visual tokens are organized around local image regions, the world model may also need to represent information that is useful for action generation but not naturally associated with any single location. In the absence of dedicated non-spatial capacity, some spatial tokens may be recruited to serve such roles, with their high norms providing an observable signature of this functional shift. This hypothesis motivates a natural architectural question: *can these representation roles be better accommodated by tokens that have no spatial identity?*

To investigate this possibility, we propose *RegisterWAM*, which introduces a small set of dedicated non-spatial tokens, termed *registers*, into the world model. Unlike visual tokens, registers have no explicit correspondence to image regions. They interact with spatial tokens throughout the world model and directly condition the action model, providing representation capacity that is not tied to any particular spatial location. Registers alone do not fully eliminate spatial outliers. Although they suppress intermediate-layer peaks, outliers remain in the final layers (Appendix B). We therefore introduce observation flow-matching (FM) loss supervision on the dense clean-observation tokens, encouraging them to remain predictive of visual observations while leaving registers free to capture information that need not be spatially localized. Under this joint design, high-norm outliers disappear from the spatial tokens and become concentrated in the registers. Thus, the observable outlier behavior can be separated from spatial representations when dedicated non-spatial capacity is paired with explicit observation FM loss supervision. This reorganization is accompanied by improved policy performance. Register-only and raw-plus-register conditioning achieve broadly comparable performance across the settings in Figure 6. We adopt register-only conditioning to reduce the visual context processed during iterative action denoising. Registers therefore do more than absorb high-norm behavior. They form an effective interface between the world model and the action model, concentrating representations that are directly useful for action generation.

Further analysis shows improved action learning with observation FM loss supervision even without registers, suggesting that its benefits extend beyond spatial outlier removal. This motivates treating the current observation not only as a conditioning input, but also as an auxiliary modeling target. We therefore introduce noised-observation FM loss supervision, which turns current-scene modeling into a recovery task from corrupted visual inputs. This objective provides additional supervision for the world model rather than an additional action-conditioning stream. RegisterWAM combines this auxiliary objective with registers and observation FM loss supervision, using registers to condition action generation.

RegisterWAM achieves 81.64% overall success on the full LIBERO-Plus test set without using additional embodied pretraining data or introducing additional foundation models. On RoboTwin 2.0, it reaches 93.0% average success, compared with 91.9% for FastWAM (Yuan et al., 2026b) and 92.6% for Faster-WAM (Zhao et al., 2026). Across six real-world task settings, it improves the mean task score from 0.3483 to 0.5238 over FastWAM.

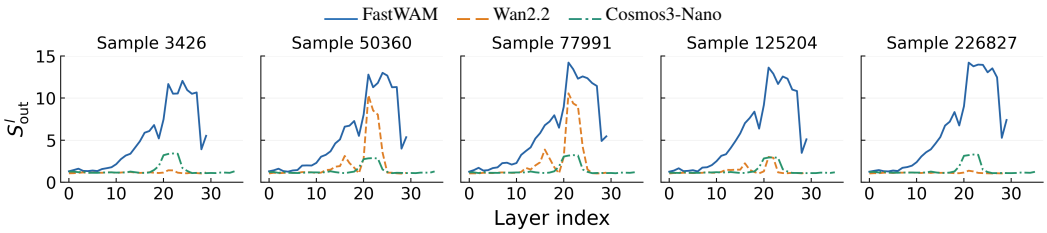


Figure 2: **Layer-wise spatial outlier strength.** High-norm outliers are more pronounced and consistent in the examined FastWAM checkpoint.

Our contributions are threefold: **(i)** We uncover high-norm spatial tokens in WAMs whose action-relevant roles are partially decoupled from their nominal spatial locations. **(ii)** We propose *Register-WAM*, which combines non-spatial registers with observation FM loss supervision to separate these roles from spatial tokens and provide an effective interface for action generation. **(iii)** We introduce auxiliary noised-observation FM loss supervision for current-scene recovery and demonstrate the effectiveness of the resulting framework across robot manipulation benchmarks.

2 From Spatial Outliers to Explicit Registers

2.1 Spatial Outliers in World Action Models

We consider the FastWAM setting (Yuan et al., 2026b), in which a Video DiT produces spatial representations that condition an Action DiT. At layer l, we denote the visual hidden states by $H^l = \{h^l_1, \dots, h^l_N\}$, where $h^l_j \in \mathbb{R}^d$ corresponds to a location in the latent grid. Our input resolution yields a 7×14 grid and thus $N = 98$ spatial tokens. After normalization and projection, these states form a cached K/V context reused throughout iterative action denoising. The same tokens therefore serve as both spatial world representations and conditions for action generation.

We measure the largest spatial outlier at layer l by $S^l_{\text{out}} = \max_j \|h^l_j\|_2 / \operatorname{median}_j \|h^l_j\|_2$, i.e., the ratio between the largest and median spatial-token norms. In FastWAM, this ratio rises sharply in middle-to-late Video DiT layers and the high-norm locations recur across observations (Figure 2). Similar but weaker and less consistent outliers appear in pretrained Wan2.2 (Team Wan et al., 2025) and Cosmos3-Nano (NVIDIA, 2026). We next test whether the pronounced FastWAM outliers have a functional consequence for action generation. Visualization settings and additional examples across models and observations are provided in Appendix B.

2.2 Outliers and Action Generation

The extreme magnitude does not propagate to the Action DiT. A direct explanation for the observed behavior is that the unusually large hidden norm produces an unusually large action-conditioning signal. The visual K/V interface largely removes this scale difference:

$$h_j \longrightarrow \mathrm{Norm}(h_j) \longrightarrow (K_j, V_j). \tag{1}$$

Although $\|h_{\text{out}}\|_2 \gg \|h_{\text{normal}}\|_2$, the corresponding key and value norms remain within the range of ordinary spatial tokens. The extreme hidden-state norm therefore does not directly propagate as an extreme action-facing K/V magnitude. This rules out the simplest explanation that the outlier affects action generation merely because its activation scale is unusually large. Figure 3 compares hidden-state, key, and value norms with masked-token action sensitivity for the same observation.



(a) Hidden-state norm **(b)** Key norm **(c)** Value norm **(d)** Action sensitivity

Figure 3: **From hidden states to action.** The magnitude anomaly vanishes after K/V projection, while its action relevance persists.

Action relevance persists. We next directly measure whether each visual token affects action generation. Let $\mathcal{M}$ denote the default Action DiT attention mask. For visual token j, we construct $\mathcal{M}^{(-j)}$ by masking position j for every Action DiT query, and define the action sensitivity as the resulting change in the predicted action,

$$I_j = \left\| A(K, V; \mathcal{M}) - A(K, V; \mathcal{M}^{(-j)}) \right\|_2 . \tag{2}$$

This intervention removes token j from the action-attention computation, including its contribution to the attention normalization, while leaving the remaining visual representations unchanged. The action-sensitivity maps show that the outlier does not dominate action generation: tokens covering task-relevant objects can cause substantially larger action changes. Nevertheless, the outlier exhibits stable, non-trivial action sensitivity across observations. Its rank is concentrated around 15 among the 98 spatial tokens, rather than behaving like an arbitrary background token (Figure 4). Its extreme hidden-state magnitude is therefore removed at the interface, yet its downstream action relevance persists.

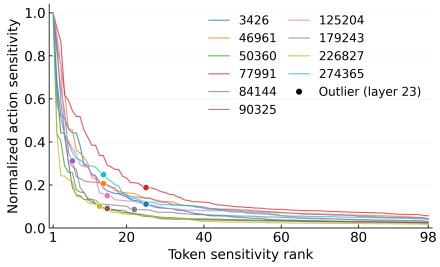


Figure 4: **Action-sensitivity ranks.** Dots mark the outlier token.

Conservative interpretation. The high-norm outlier should not be viewed as an action token, an attention sink (Xiao et al., 2024), or a unique carrier of action-relevant information. Many ordinary spatial tokens also contribute substantially to action generation. Instead, the extreme norm makes one aspect of the model's internal representation behavior directly observable. The persistence of action relevance after the magnitude anomaly is removed suggests that spatial token space can support representation behavior beyond its nominal role of local spatial encoding. The outlier is thus an observable manifestation of this behavior, rather than the behavior itself.

2.3 Registers with Observation FM Loss Supervision

The preceding analysis suggests that spatial tokens in a WAM can support representational roles beyond local spatial encoding, with high-norm outliers providing an observable signature of this behavior. This does not imply that an outlier uniquely carries action information or identify its semantic content. Instead, it motivates a controlled architectural probe: *what changes when the world model is given representation capacity without spatial identity?*

Register tokens. Given the spatial video tokens H, we append M learnable register tokens $R = \{r_1, \ldots, r_M\}$ to form the Video DiT input $Z = [H; R]$. Following the terminology of Darcet et al. (2024), we call these learnable tokens *registers*. The spatial tokens and registers are jointly updated by the same self-attention layers, while registers have no explicit correspondence to image locations. We use the spatial outlier as a probe of whether this additional non-spatial capacity changes the observable organization of the representation.

Registers alone are insufficient. Registers alone do not fully eliminate spatial outliers: outliers remain at the output layer, even though high-norm responses also emerge among the registers (Figure 5a). Layer-wise curves in Appendix B show that intermediate-layer peaks are nevertheless suppressed. The limited output-layer change is accompanied by a modest policy gain in the naive setting. FastWAM conditions its Action DiT on the dense observation states (*Raw*, referring only to observation tokens in this section), $A \leftarrow H$, and achieves 53.64% success. Appending 8 registers and conditioning on both token types, $A \leftarrow [H; R]$, without observation FM loss supervision reaches 54.18%. Thus, simply adding explicit slots neither resolves the spatial anomaly nor substantially improves the policy.

Observation FM loss supervision reorganizes the representation. The remaining output-layer outliers suggest that extra non-spatial capacity does not by itself constrain how the original spatial

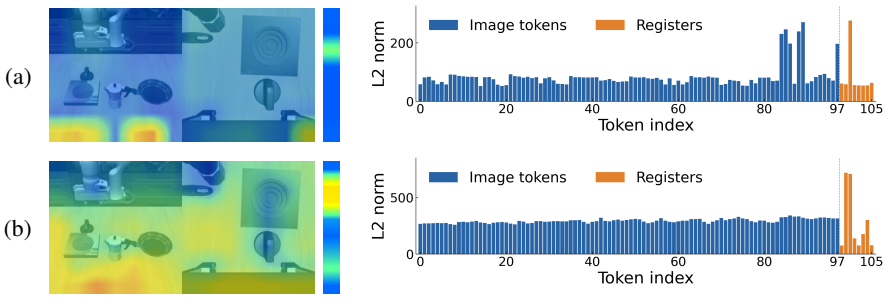


Figure 5: **Output-layer representations with registers.** (a) Without observation FM loss supervision, spatial outliers remain alongside high-norm registers. (b) With observation FM loss supervision, spatial-token norms are regularized while high-norm responses remain concentrated in registers.

tokens are used. We therefore introduce observation FM loss supervision on the dense clean observation tokens. Specifically, we apply the standard flow-matching loss at zero noise level ($\sigma = 0$) to the dense observation-token outputs, excluding registers. Under this supervision, the output-layer spatial-token norms become substantially more uniform: the pronounced spatial outlier is reduced, while high-norm responses remain concentrated in the registers (Figure 5b). We do not interpret this redistribution as evidence that any individual register has a fixed semantic role; rather, it shows that the observable specialization no longer needs to occupy a spatial location. Figure 12 in Appendix B further shows their complementary effects across layers: registers suppress intermediate-layer outliers, observation FM loss supervision primarily suppresses final-layer outliers, and their combination suppresses both.

Action conditioning after outlier reorganization. After observation FM loss supervision suppresses the output-layer spatial outlier, we ask how the reorganized world states should condition action generation. We compare two interfaces while varying the number of registers. The first exposes only the register states to the Action DiT, $A \leftarrow R$, directly testing whether the registers contain sufficient action-relevant information. With 1, 4, 8, 16, and 64 registers, this interface obtains 59.56%, 61.11%, 60.46%, 58.86%, and 58.32%, respectively. Every displayed count remains competitive with the 58.07% raw-token reference trained with observation FM loss supervision, and 4 registers give the strongest register-only result among the displayed settings. Increasing the number of registers does not consistently improve performance.

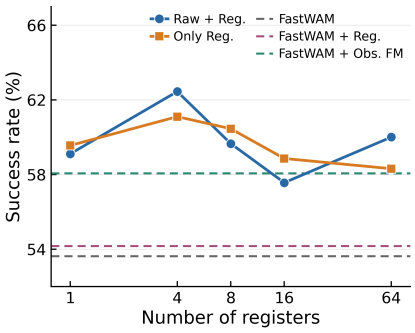


Figure 6: **Register interfaces.** Curves compare Only Reg. and Raw + Reg. conditioning with observation FM loss supervision. Raw denotes dense observation tokens only. Dashed lines show FastWAM, raw plus 8 registers without observation FM loss supervision, and raw tokens with observation FM loss supervision but no registers.

The second interface exposes both the raw observation states and registers, $A \leftarrow [H; R]$. Under the same observation FM loss supervision, the corresponding results are 59.11%, 62.45%, 59.66%, 57.57%, and 60.01%. Raw tokens improve upon the register-only interface at 4 and 64 registers, but not at 1, 8, or 16. They can therefore complement registers in some configurations, although the combined interface is not uniformly superior.

Together, the representation and policy results form a consistent sequence. Registers alone provide a possible destination for non-spatial specialization, but do not fully eliminate spatial outliers or substantially improve the policy in the naive setting. Observation FM loss supervision reorganizes the token norms; after this reorganization, the resulting register states become directly useful as action conditions. A small register set can carry action-relevant information on its own, while raw spatial tokens remain complementary in some configurations. We therefore treat registers as explicit representation slots, not as a universal replacement for dense visual conditioning. Given their broadly comparable performance in Figure 6, we adopt register-only conditioning to reduce the visual context processed during action denoising.

2.4 Beyond Outlier Reorganization

The preceding results resolve the spatial-outlier observation: registers provide non-spatial slots, and observation FM loss supervision causes the visible specialization to move away from spatial locations. A separate control reveals an additional effect of the same supervision. Without introducing any registers, the setting trained with observation FM loss supervision achieves 58.07% (the green dashed reference in Figure 6), compared with 53.64% for the original FastWAM. This improvement does not require redistribution into registers.

This result motivates treating the current observation not only as a conditioning input, but also as an auxiliary modeling target. We extend this idea by perturbing the observation with noise and introducing noised-observation FM loss supervision. The resulting task requires predicting a recovery direction from incomplete visual information, making current-scene modeling an explicit denoising task. Its role is to provide additional supervision for the world model, rather than additional representations for the action model to read. The next section combines noised-observation FM loss supervision with registers and observation FM loss supervision to construct RegisterWAM.

3 RegisterWAM

Building on Section 2, RegisterWAM combines learnable registers with observation FM loss supervision. Registers provide the visual conditioning for the Action DiT. Noised-observation FM loss supervision supplies an auxiliary denoising task. Figures 7 and 8 show the architecture and attention mask.

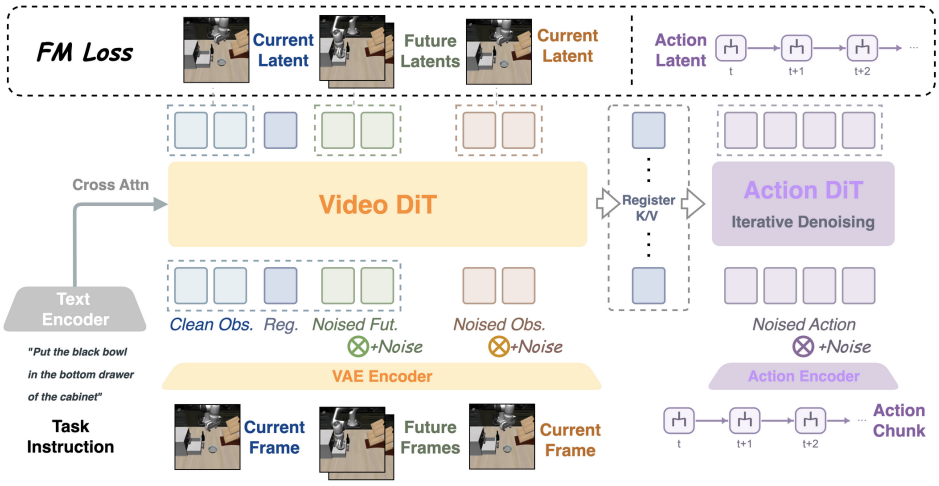


Figure 7: **RegisterWAM architecture.** Registers condition the Action DiT; visual outputs receive FM loss supervision. Encoded robot-state tokens are appended to the text embeddings (omitted for clarity).

3.1 Registers as the World–Action Interface

World representation. We encode the current observation into VAE latents and append M learnable register tokens to the visual sequence. The Video DiT processes these tokens together with future tokens, whose inputs are pure Gaussian noise during both training and inference. During training, encoded future frames provide the targets for the future FM loss. Language instructions are encoded into text embeddings, followed by encoded robot-state tokens; the resulting sequence supplies cross-attention conditioning.

Action conditioning. Let R^{ℓ} denote the register states at Video DiT layer ℓ. We construct the corresponding action-conditioning keys and values from these register states, following the layer-wise K/V conditioning of the base WAM. This register K/V cache supplies the visual conditioning for the Action DiT. Thus, the world model retains dense visual representations while the visual interface to the action model consists of M registers.

3.2 Observation-Side FM Loss Supervision

Observation FM loss supervision. Motivated by the outlier analysis, we apply the observation FM loss $\mathcal{L}_{\text{obs}}$ to the outputs corresponding to the dense clean observation tokens. Together with registers, this supervision reduces spatial outliers and concentrates high-norm responses in registers (Section 2). Registers themselves are not supervised by visual FM losses. The improvement with observation FM loss supervision even without registers further motivates extending observation FM loss supervision to perturbed inputs.

Noised-observation FM loss supervision. Given the current observation latent x, we construct

$$x_\sigma = (1-\sigma)x + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where σ controls the noise level. The corresponding tokens are processed by the same Video DiT, with the noised-observation FM loss $\mathcal{L}_{\text{noised-obs}}$ on their outputs. The attention mask in Figure 8 isolates this branch, which provides an auxiliary denoising task whose gradients update the Video DiT parameters, not an additional visual conditioning stream.

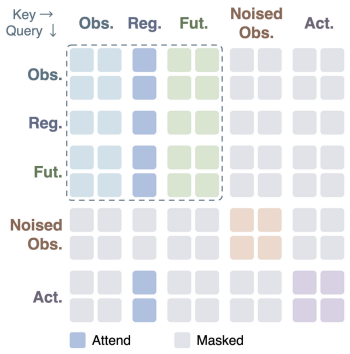


Figure 8: **Training attention mask.** Action tokens attend to registers and themselves. Gray entries are masked.

3.3 Training and Inference

We jointly train the Video DiT and Action DiT with

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{action}} + \lambda_v \big(\mathcal{L}_{\text{obs}} + \mathcal{L}_{\text{future}}\big) + \lambda_n \mathcal{L}_{\text{noised-obs}}, \tag{4}$$

where $\mathcal{L}_{\text{action}}$ supervises action-chunk denoising and the other three terms supervise the corresponding dense visual outputs. All visual losses exclude register tokens. Future inputs remain at the maximum noise level during training to match their inference-time inputs, while noised-observation FM loss supervision uses a separate noise level σ. At inference, the Video DiT runs once to construct the register K/V cache, which the Action DiT reuses throughout its iterative action denoising. The action model therefore accesses a fixed-size register interface without requiring iterative video generation.

4 Experiments

Section 2 establishes the FastWAM-base representation findings. Here we evaluate the resulting model, compare action-conditioning choices, ablate observation-side FM loss supervision, and test transfer to another generative backbone.

4.1 Experimental Setup

We evaluate on LIBERO-Plus (Fei et al., 2025), RoboTwin 2.0 (Chen et al., 2025), and 6 real-world task settings. The world model builds on Wan2.2, while the policy uses an Action DiT with 10 denoising steps. Unless noted otherwise, RegisterWAM uses 8 registers interacting with observation and future tokens, and noise level $\sigma = 0.5$ for the noised-observation FM loss supervision. Comparisons include published FastWAM and other WAMs, representative vision–language–action policies, and our future-conditioned baseline. For our LIBERO-Plus experiments, all models are trained exclusively on LIBERO (Liu et al., 2023) data, without using LIBERO-Plus data for training. The ablation studies are evaluated on the same 20% subset of the LIBERO-Plus test data. All models with registers use an action-loss weight $\lambda_a = 0.2$. For the FastWAM analysis in Section 2, the 58.07% no-register reference with observation FM loss supervision also uses $\lambda_a = 0.2$, whereas the original 53.64% FastWAM uses $\lambda_a = 1.0$. Changing only the action-loss weight from 1.0 to 0.2 leaves the original FastWAM's performance essentially unchanged. All settings in the final-model ablations use $\lambda_a = 0.2$.

4.2 Main Results

Table 1 reports performance across the seven LIBERO-Plus distribution shifts. RegisterWAM achieves 81.64% overall success on the full LIBERO-Plus test set of 10,030 episodes, exceed-

Table 1: LIBERO-Plus success rates (%). Emb. PT.: embodied-policy pretraining; Extra Prior: external pretrained models, modalities, or geometric/semantic/motion supervision.

Method	Emb. PT.	Extra Prior	Camera	Robot	Lang.	Light	Backg.	Noise	Layout	Avg.
$\pi_{0.5}$ Black et al. (2025)	✓	–	75.4	77.5	85.6	96.9	94.6	89.7	85.7	85.7
FastWAM Yuan et al. (2026b)	–	–	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
Faster-WAM Ma et al. (2026a)	–	–	67.9	49.0	92.1	94.3	57.0	82.3	82.7	75.0
Faster-WAM Zhao et al. (2026)	–	–	53.8	71.6	94.7	96.3	61.3	63.6	79.1	73.6
4D-WAM Yang et al. (2026)	–	✓	45.2	64.3	90.6	94.3	57.7	69.1	79.2	71.0
ST-WAM Wang et al. (2026)	–	✓	55.4	60.1	79.3	93.0	74.2	79.5	74.3	72.8
DreamWAM Yuan et al. (2026a)	–	✓	53.8	63.6	94.8	96.7	71.6	67.2	80.7	75.5
GaussianWAM Zhang et al. (2026c)	–	✓	54.7	61.6	75.4	90.8	58.2	83.0	76.7	71.3
Robust-WAM Yan et al. (2026b)	–	✓	28.7	43.4	73.6	90.2	62.6	56.8	67.8	58.9
PILOT Ma et al. (2026c)	–	✓	76.0	70.0	84.0	92.0	85.0	78.0	82.0	81.0
RegisterWAM	–	–	**59.2**	**71.9**	**93.7**	**95.7**	**87.9**	**85.1**	**84.3**	**81.6**

Table 2: RoboTwin 2.0 success rates (%).

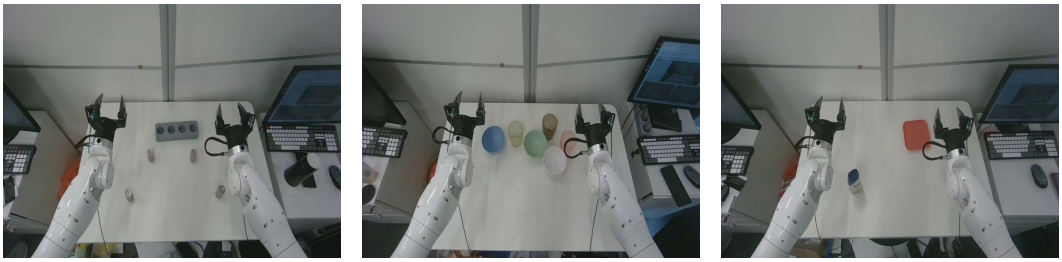
Method	Emb. PT.	Clean	Rand.	Avg.
LingBot-VA Li et al. (2026b)	✓	92.9	91.5	92.2
FastWAM Yuan et al. (2026b)	–	91.9	91.8	91.9
Faster-WAM Zhao et al. (2026)	–	92.8	92.3	92.6
RegisterWAM	–	**92.9**	**93.0**	**93.0**

Table 3: ImageWAM transfer (%).

Method	Avg.
ImageWAM (FLUX.2 4B)	83.10
ImageWAM + Ours (FLUX.2 4B)	**85.42**

ing the listed FastWAM and Faster-WAM results while using the same Wan2.2 backbone, without embodied-policy pretraining or additional geometric or semantic teacher models. We separately mark methods that use additional pretrained models, modalities, or externally supplied geometric, semantic, or motion supervision, excluding FM losses on the original visual data, since their reported scores reflect a different resource setting. Its highest category scores are under language, lighting, and background changes. The same model achieves 82.52% on the 20% test subset used for the ablation studies.

On RoboTwin 2.0 (Table 2), RegisterWAM achieves 92.9% in the clean setting and 93.0% under randomization, for a 93.0% average. This slightly exceeds the compared FastWAM and Faster-WAM variants without using embodied-policy pretraining.



Battery insertion | Bowl-and-cup stacking | Bottle handover

Figure 9: Scenes for the three real-world ID tasks. The three OOD variants modify the bottle-handover setting.

Real-world evaluation. We compare RegisterWAM with FastWAM on a three-camera Tianji dual-arm platform on 3 in-distribution (ID) tasks: battery insertion, bowl-and-cup stacking, and bottle handover and placement (Figure 9), using 300, 60, and 300 demonstrations, respectively. For battery insertion, the left arm inserts the two left batteries into the two left slots, followed by the right arm inserting the two right batteries into the two right slots. The stacking task forms a stack of three bowls and then a stack of three cups. For bottle handover, the left arm grasps and lifts a white shampoo bottle, passes it to the right arm, and the right arm places it into the orange plate. All 3 out-of-distribution (OOD) settings modify this last task: (i) changing the table height while keeping the arm height fixed; (ii) covering the original white table with a blue tablecloth; and (iii) adding a white distractor plate beside the target plate.

We evaluate each of the 6 settings over 20 trials. Using step-wise task scores rather than binary success rates, RegisterWAM improves the mean score from 0.3483 to 0.5238 (Table 4). It improves

Table 4: Real-world task scores (20 trials per setting). OOD variants modify the bottle-handover task; Avg. equally weights all six settings.

	In-distribution			Out-of-distribution			
Method	Batteries	Stacking	Handover	Height	Tablecloth	Distractor	Avg.
FastWAM	0.5000	0.0000	0.6200	0.5500	**0.0000**	0.4200	0.3483
RegisterWAM	**0.6875**	**0.2500**	**0.7550**	**0.7300**	**0.0000**	**0.7200**	**0.5238**

Table 5: Cumulative component ablation.

Configuration	Interface	Suc. (%)
Baseline	Raw	76.29
+ Noised Obs. FM	Raw	78.59
+ Reg. and Obs. FM	Reg.	**82.52**

Table 6: Register count.

Registers	Suc. (%)
4	78.49
8	**82.52**
16	81.03

Table 7: Noise level.

Nominal σ	Suc. (%)
0.25	76.14
0.50	**82.52**
0.75	80.98

five settings, while both methods score zero under the blue-tablecloth shift. Additional evaluation and scoring details, along with the full task instructions, are provided in Appendix C.1.

4.3 Ablation Studies

Establishing a strong baseline. Following future conditioning in Faster-WAM (Zhao et al., 2026) and RIFT (Zhang et al., 2026a), our future-conditioned baseline conditions actions on dense observation and future tokens (*Raw*). We use pure-noise future inputs during both training and inference throughout the ablations. This future-conditioned baseline achieves 76.29% success on the LIBERO-Plus subset.

FM loss supervision and the register interface. Adding isolated noised-observation FM loss supervision raises success from 76.29% to 78.59% while keeping the action conditioning unchanged (Table 5). Replacing the raw interface with 8 registers and adding observation FM loss supervision further improves success to 82.52%.

Key parameters. We vary the register count and noised-observation noise level separately in the full model. Among the tested register counts, 8 registers perform best; increasing the count to 16 does not improve performance (Table 6). For the noise-level sweep, the scheduler's timestep-dependent loss weighting means that varying σ changes both input corruption and the effective auxiliary FM loss weight. Performance peaks at $\sigma = 0.5$ and drops markedly at the lower noise level of $\sigma = 0.25$ (Table 7). We therefore use 8 registers and $\sigma = 0.5$ by default.

Transfer to ImageWAM. We test transfer to ImageWAM (Zhang et al., 2026b) because it uses the FLUX.2 image-editing backbone (Black Forest Labs, 2025) rather than the video-generation backbone used here. Applying the same design improves average success on LIBERO-Plus from 83.10% to 85.42%, a gain of 2.32 percentage points over the original model (Table 3).

Additional register-visualization analysis is provided in Appendix D.1. Related work is discussed in Appendix A.

5 Conclusion

We identified persistent high-norm spatial outliers in the examined WAM and found that their action relevance survives the removal of the magnitude anomaly at the visual K/V interface. Registers alone do not fully eliminate spatial outliers. Observation FM loss supervision of dense observation tokens suppresses output-layer outliers while high-norm responses concentrate in registers. The resulting register states support effective action conditioning. Improved action performance is also observed with observation FM loss supervision without registers, motivating additional noised-observation FM loss supervision. These findings connect explicit non-spatial capacity and observation-side FM loss supervision as two components of WAM design.

AI USE STATEMENT

AI tools provided assistance with manuscript drafting and language polishing, LaTeX formatting and figure preparation, and discussions of presentation. These outputs served as suggestions, not as independent scientific evidence. The authors retained responsibility for research decisions, methodological implementation, and experimental validation, and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model, 2025. URL https://arxiv.org/abs/2512.13030.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, 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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2309.16588. URL https://arxiv.org/abs/2309.16588.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, 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.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. URL https://arxiv.org/abs/2507.12898.

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. URL https://arxiv.org/abs/2412.14803.

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

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, Anh-Quan Cao, Nermin Samet, Tuan-Hung Vu, and Matthieu Cord. Driving on registers, 2026. URL https://arxiv.org/abs/2601.05083.

Jiajun Li, Tiecheng Guo, Yifan Ye, Rongyu Zhang, Xiaowei Chi, Qianpu Sun, Ying Li, Yunfan Lou, Yan Huang, Zhihe Lu, Meng Guo, and Shanghang Zhang. Efficient-wam: A 1b-parameter world-action model with low-cost future imagination. *arXiv preprint arXiv:2606.10040*, 2026a. URL https://arxiv.org/abs/2606.10040.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control, 2026b. URL https://arxiv.org/abs/2601.21998.

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

Ziang Li, Dongzhou Cheng, Yibin Wang, Shiyue Wang, Xiaoyang Xu, Lingxuan Weng, Juan Wang, and Jiaqi Wang. Light-wam: Efficient world action models with state-fusion action decoding. *arXiv preprint arXiv:2606.08242*, 2026c.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Jianman Lin, Shailesh Shailesh, Zhongyi Luo, and Jiafei Duan. Breaking the vision–action shortcut: Latent interface training for generalizable robotics foundation models, 2026. URL https://arxiv.org/abs/2609.12641.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

Liheng Ma, Rui Heng Yang, Zhanguang Zhang, Mateo Clemente, Ziwen Hu, Tongtong Cao, and Yingxue Zhang. Faster-WAM: Do world action models need deep action modules?, 2026a. URL https://arxiv.org/abs/2608.02365.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026b.

Xiangkai Ma, Yue Ma, Junjie Wang, Sheng Xu, Mingyang Li, Han Zhang, Yuzheng Zhuang, Wenzhong Li, and Zhihao Yuan. Decoupling intention from trajectory: A representational deduction framework for world action models, 2026c. URL https://arxiv.org/abs/2608.06994.

Motubrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, Louis Liu, Mengchen Cai, Rongxu Cui, Ruowen Zhao, Runqing Wang, Shuhe Huang, Yao Feng, Yinze Rong, Zeyuan Wang, and Jun Zhu. Motubrain: An advanced world action model for robot control, 2026. URL https://arxiv.org/abs/2604.27792.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025. URL https://arxiv.org/abs/2512.15692.

Nikita Starodubcev, Ilia Sudakov, Ilya Drobyshevskiy, Artem Babenko, and Dmitry Baranchuk. Registers matter for pixel-space diffusion transformers, 2026. URL https://arxiv.org/abs/2605.16147.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Mingxin Wang, Bin Hu, Bin Qian, Kaitao Jiang, Haoning Wu, Feng Yan, Bowen Jing, Ruiyang Hao, Enyi Wang, Kangning Niu, Yandan Yang, Mu Xu, Yan Wang, Houde Liu, and Tianlun Li. ST-WAM: Semantic-temporal world action model for robust manipulation under visual distribution shifts, 2026. URL https://arxiv.org/abs/2607.28993.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Ge Yan, Jinghao Liu, Yuzhi Fan, Lei Cai, Minwen Liao, Jesse Zhang, and Dieter Fox. Flex-π: A multi-stream world-action model with compute flexibility, 2026a. URL https://arxiv.org/abs/2608.10860.

Haodong Yan, Junfeng Li, Junjie He, Zhide Zhong, MingMing Yu, Wenxuan Song, Jiaguan Zhu, Yangyang Zheng, Yuqiao Du, Jiadi You, Yingjie Cai, Xu Yan, Guanyi Zhao, Bingbing Liu, and Haoang Li. Robust-WAM: Bridging generative pretraining and semantic foresight in world-action models, 2026b. URL https://arxiv.org/abs/2608.05903.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Pengwei Wang, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing spatiotemporal awareness into world action models through trajectory fields, 2026. URL https://arxiv.org/abs/2608.08023.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, Min Cao, Peng Li, Qiuping Deng, Wenjun Mei, Xiaofeng Wang, Xinze Chen, Xinyu Zhou, Yang Wang, Yifan Chang, Yifan Li, Yukun Zhou, Yun Ye, Zhichao Liu, and Zheng Zhu. Gigaworld-policy: An efficient action-centered world-action model. *arXiv preprint arXiv:2603.17240*, 2026a.

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*, 2026b. URL https://arxiv.org/abs/2602.15922.

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, 2026a. URL https://arxiv.org/abs/2608.04996.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination?, 2026b. URL https://arxiv.org/abs/2603.16666.

Chushan Zhang, Jinguang Tong, Xuesong Li, Yikai Wang, and Hongdong Li. Keep the future, drop the rollout: RIFT for world action models, 2026a. URL https://arxiv.org/abs/2608.11521v1.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *CoRR*, abs/2507.04447, 2025. doi: 10.48550/ARXIV.2507.04447. URL https://doi.org/10.48550/arXiv.2507.04447.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. ImageWAM: Do world action models really need video generation, or just image editing?, 2026b. URL https://arxiv.org/abs/2606.19531.

Zijian Zhang, Yuqing Jiang, Weitao Zhou, Minglei Li, Jinhao Zhang, Yao Mu, Xiaofan Li, Hao Zhao, and Haibao Yu. GaussianWAM: Distilling geometry and semantics from 3D gaussian fields into world-action models, 2026c. URL https://arxiv.org/abs/2608.24714.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models, 2025. URL https://arxiv.org/abs/2503.22020.

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, 2026. URL https://arxiv.org/abs/2608.04404.

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

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets, 2025. URL https://arxiv.org/abs/2504.02792.

A Related Work

World Action Models. WAMs connect visual prediction to robot control through generated visual plans (Du et al., 2023; Bharadhwaj et al., 2024; Zhou et al., 2024; Feng et al., 2025), predictive features (Hu et al., 2024; Liao et al., 2025; Pai et al., 2025; Ma et al., 2026b), or joint video–action learning (Cheang et al., 2024; Liang et al., 2025; Li et al., 2025; Zhu et al., 2025). Other video–action architectures extend this direction (Kim et al., 2026; Bi et al., 2025; Li et al., 2026b; Ye et al., 2026b; Motubrain Team et al., 2026). Related VLA approaches predict visual goals or structured world knowledge before action generation (Zhao et al., 2025; Zhang et al., 2025). Recent designs reduce inference cost through cached representations and lighter action modules (Yuan et al., 2026b; Ma et al., 2026a; Zhao et al., 2026; Zhang et al., 2026a), action-centered causal modeling (Ye et al., 2026a), compact backbones (Li et al., 2026c;a), or image editing instead of video prediction (Zhang et al., 2026b). Others add geometric, semantic, motion, or multi-stream supervision (Yang et al., 2026; Wang et al., 2026; Yuan et al., 2026a; Zhang et al., 2026c; Yan et al., 2026b;a; Ma et al., 2026c). Their common premise is that visual modeling supplies useful representations or learning signals for action prediction. We complement these efforts by studying token-level representation organization and the effects of registers and observation FM loss supervision on spatial outliers and action learning.

Register tokens and latent action interfaces. Registers accommodate non-spatial computation in ViTs (Darcet et al., 2024) and also benefit pixel-space DiTs without naturally occurring patch-token outliers (Starodubcev et al., 2026). High-norm outliers differ from attention sinks, which concern attention mass (Xiao et al., 2024). Driving on Registers (Kirby et al., 2026) and Latent Interface Training (Lin et al., 2026) explore compact latent interfaces for control. Our focus is not latent-token conditioning itself, but spatial outliers and their persistent action relevance in WAMs: we study how observation FM loss supervision reorganizes high-norm responses between dense spatial tokens and registers, and use registers as the world–action interface.

B Additional Outlier Analysis

Visualization settings. For each Video DiT layer, we compute the ℓ_2 norm independently for every spatial token and reshape the resulting values to the latent spatial grid. The heatmaps in the main paper and below are resized to the input-image resolution only for visualization and shown alone or overlaid on the corresponding observation; no spatial smoothing is used for the reported token statistics. The main-paper visualizations use representations from layers 21–23. Hidden-state, key, and value maps show $\|h_j\|_2$, $\|K_j\|_2$, and $\|V_j\|_2$, respectively. Action-sensitivity maps show I_j from Section 2.2, obtained by masking one visual position from all Action DiT queries at a time. Colors should be compared spatially within each panel rather than interpreted as a common absolute scale across images; cross-model magnitude comparisons use the normalized statistic S_{out}. Unless otherwise specified, the main-paper outlier visualizations use future noise level $\sigma_{\text{future}} = 1$, i.e., pure-noise future inputs.

First-frame outliers under different future noise levels. Figure 10 examines the pretrained Wan2.2 backbone while varying the noise level of its future-frame inputs. The horizontal axis indexes Video DiT layers, the vertical axis gives the future noise level, and color shows S_{out} computed over spatial tokens of the first frame. Thus, the varied noise belongs to the future inputs, whereas the measured outliers belong to the first-frame representation. This sweep is distinct from the noised-observation objective in Section 3.

The strongest first-frame outliers appear at high future noise levels, particularly $\sigma_{\text{future}} = 0.9$–$1.0$, with a pronounced peak around layers 21–23 and a weaker response around layers 16–17. The peak exceeds an order of magnitude relative to the median spatial-token norm. At intermediate noise levels these responses weaken, and most lower-noise settings show much smaller norm disparities. The dependence is not strictly monotonic: weak localized responses are also visible at zero noise. These results qualify the observation in Section 2.1: pretrained Wan2.2 can exhibit first-frame outliers, but their strength depends on the future-input noise level and network depth. The default $\sigma_{\text{future}} = 1$ used in the main paper lies in the strong-noise regime where this behavior is most visible. The

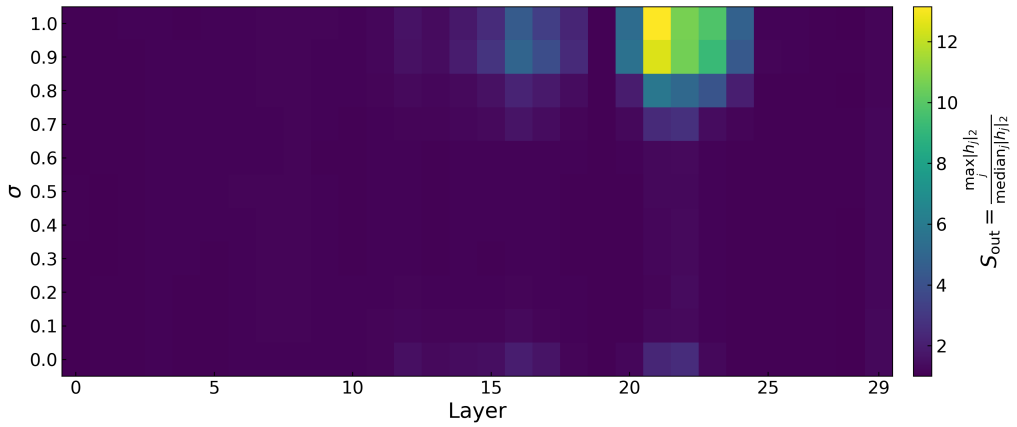


Figure 10: **Wan2.2 first-frame outliers versus future noise.** Color denotes first-frame spatial outlier strength S_{out}. Pronounced peaks occur mainly in middle-to-late layers under strong future noise.

heatmap does not by itself establish how frequently the pattern occurs across observations or what information the outlier tokens encode.

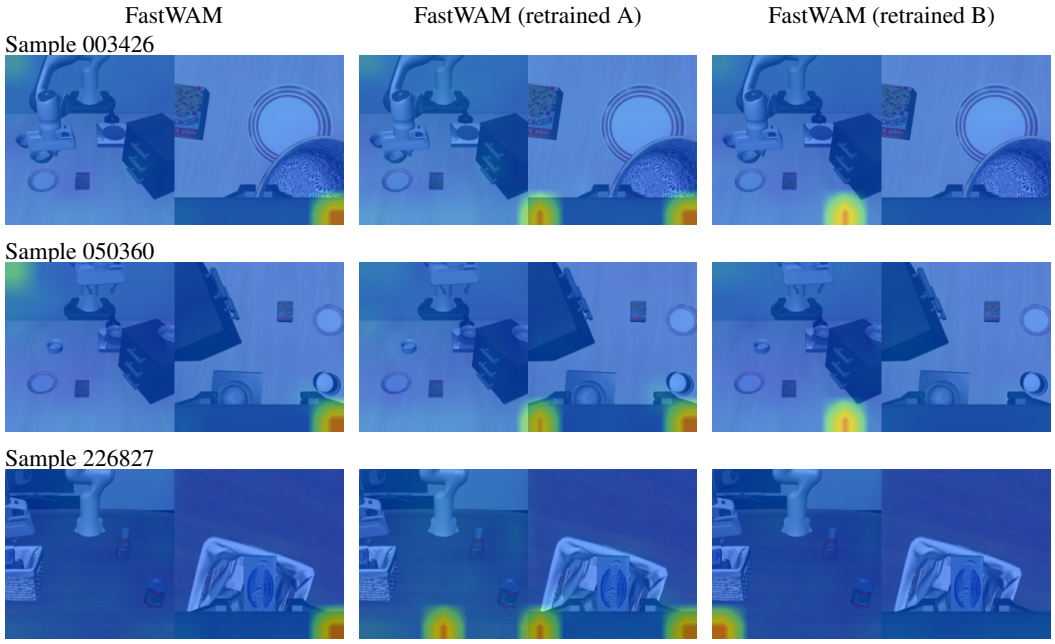


Figure 11: **Additional FastWAM outlier visualizations.** Hidden-state norms for three observations in FastWAM and two retrained variants. All panels show norms overlaid on the corresponding observations. Colors should be compared within each panel.

Additional observations. Figure 11 compares hidden-state norms for three additional observations in FastWAM and two retrained variants. The maps suggest that each model favors a small set of recurring spatial locations for high-norm responses, while the active locations and number of peaks can vary across samples. Thus, these outliers exhibit model-specific spatial preferences rather than an identical pattern for every observation.

Complementary effects across layers. Figure 12 separates the effects of registers and observation FM loss supervision across ten observations. Registers alone substantially suppress intermediate-layer spatial outliers, but elevated norms persist in the final layers. Observation FM loss supervision alone primarily suppresses final-layer outliers, while pronounced intermediate-layer peaks remain. Combining both keeps S_{out} low throughout the network in all ten examples, indicating complementary effects on spatial outliers at different depths.

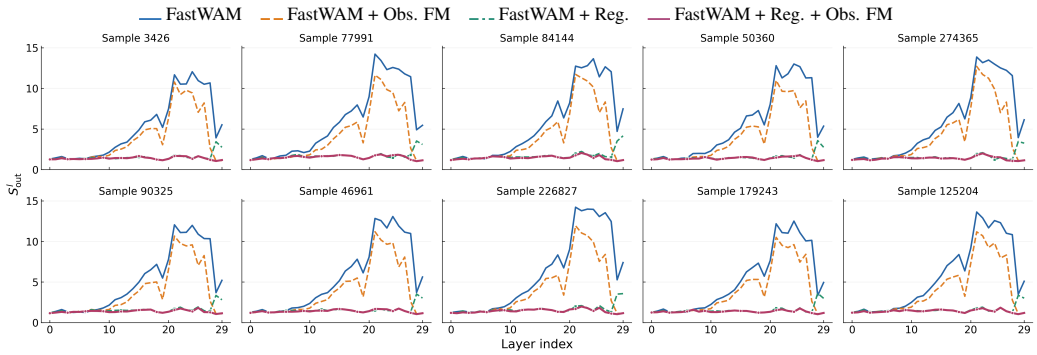


Figure 12: **Complementary effects of registers and observation FM loss supervision.** Layer-wise spatial outlier strength S_{out} across ten observations. Registers suppress intermediate-layer peaks, while observation FM loss supervision suppresses outliers in the final layers; together, they suppress pronounced peaks throughout the network.

Spatial norm maps in RegisterWAM. Figure 13 shows two additional observations from RegisterWAM. Unlike the isolated high-norm peaks in the FastWAM maps, these spatial norm maps show broader variations without conspicuous localized outliers, consistent with the suppression of spatial outliers when registers and observation FM loss supervision are combined.

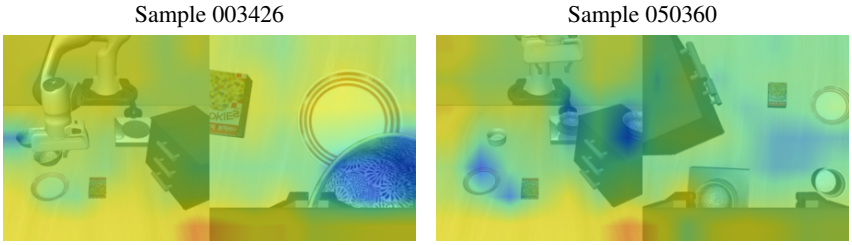


Figure 13: **RegisterWAM spatial norm maps.** Hidden-state norms overlaid on two observations, without conspicuous localized outliers. Colors are interpreted within each panel.

C Benchmark Details

Real-world evaluation is detailed below.

C.1 Real-World Evaluation

Evaluation and scoring. Each setting is evaluated over 20 trials. We report the mean task score, which awards partial credit for completed steps, rather than the fraction of fully successful trials. The overall average in Table 4 is the arithmetic mean of the six setting-level scores, including zero-score settings. RegisterWAM improves all three ID tasks and the height and distractor variants. Both methods obtain zero score with the blue tablecloth, indicating a shared failure under this appearance shift in the tested trials.

Task instructions. The ID tasks use the following instructions, reproduced from the experiment record.

Battery insertion. "First, use the left arm. Pick up the two batteries on the left, one at a time, and insert them into the two left holes. After the left arm finishes, use the right arm. Pick up the two batteries on the right, one at a time, and insert them into the two right holes."

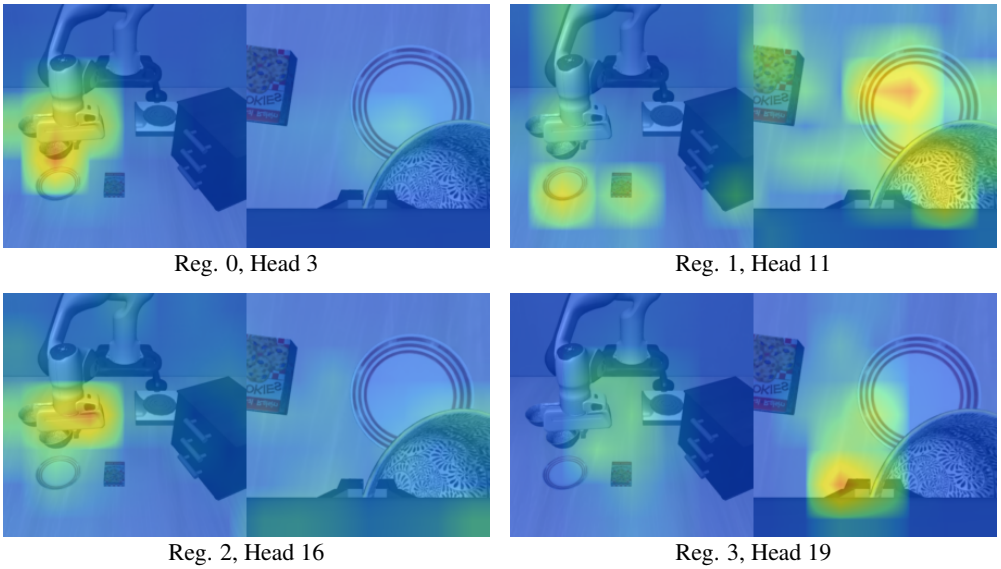
Bowl-and-cup stacking. "The robot first grasps a bowl from the table and places it on the table. It then grasps another bowl and stacks it on top. This stacking action is repeated once more. After that, the robot grasps a cup from the table and places it on the table, then grasps another cup and stacks it. Finally, the robot grasps one more cup and stacks it together with the others."

Bottle handover. "The left arm grasps the white bottle and lifts it into the air, passes it to the right arm, and the right arm places the white bottle into the orange plate on the right side of the table."

D Additional Qualitative Results

D.1 Visualizing Register Attention

Figure 14 visualizes attention from registers to spatial observation tokens for four selected register–head pairs. Reg. 0 (Head 3) and Reg. 2 (Head 16) emphasize the gripper region, with different spatial extents. Reg. 1 (Head 11) distributes attention across several objects, including the plates on the right, whereas Reg. 3 (Head 19) emphasizes a lower-right region beside them. These maps illustrate diverse and partially overlapping spatial attention across register–head pairs; they do not imply fixed semantic roles for individual registers.



Reg. 0, Head 3 | Reg. 1, Head 11

Reg. 2, Head 16 | Reg. 3, Head 19

Figure 14: **Register attention over spatial tokens.** Four selected register–head pairs, overlaid on observation 003426. Register and head indices are shown below each panel.

E Limitations

Our evidence supports functional specialization but does not fully establish what information outlier or register tokens encode. The experiments cover a limited family of WAM architectures, and the observed behavior may depend on model scale, training objective, or optimization history.