How Representational Priors Shape Generalization in World Action Models

Anonymous authors
Paper under double-blind review

Abstract

World action models (WAMs) are shown to substantially diverge under generalization performance based on their generative backbones. We investigate this gap through the representational geometry that WAMs inherit from their backbones and how policy training reorganizes it. Our central hypothesis is that a useful representational prior would structure geometry that separates nuisance information from the in-distribution manifold in advance of training as a policy. We quantify this with *nuisance orthogonality*; orthogonality of the feature under visual variation outside the empirical in-distribution manifold. Across various image and video backbones, we show that a higher orthogonality from the representational prior is associated with stronger robustness to visual variations after WAM training, enabling policy generalization measurable prospectively for world models. In addition, analyzing the resulting robust WAMs reveals *nuisance suppression*; relative suppression of feature displacement caused by such variations and robust action generation of the policy. Together, these findings provide a geometric perspective on how the representational prior of the world model shapes the generalization performance of WAMs.

1 Introduction

World action models (WAMs) have recently become a major paradigm for a robotic foundation model (Kim et al., 2026; Ye et al., 2026; Li et al., 2026a; Pai et al., 2025; NVIDIA et al., 2026). They inherit the representations and rich spatiotemporal knowledge from generative backbones with scalable pretraining, predict the future with world modeling objective, and generate actions from the features from the reasoning process. It has been shown that they improve generalization to out-of-distribution scenes by predicting how the world changes.

However, recent researches question whether the capabilities of WAMs come from explicit test-time prediction or enriched representations from the world modeling, and whether video prediction is a good representational prior or objective for world modeling. FastWAM finds that representational gains from world modeling, rather than explicit future generation at test time, is the main contributor of WAM's performance (Yuan et al., 2026). Furthermore, ImageWAM shows image editors as backbones led to significant improvements to the policy generalization from edit-aware representation which contain task-relevant visual changes and is evidenced by attention maps (Zhang et al., 2026b). We therefore ask:

> ***Which geometric structure makes a backbone's representational prior useful for WAM generalization, and how does policy adaptation transform that structure?***

Understanding this question is necessary to explain why some inherited representations support stronger generalization. Inspired by the baseline WAMs and fundamental questions about their representations and performances, we start by analyzing the representational geometry of in- and out-of-distribution observations.

Throughout the paper, we introduce *nuisance orthogonality*, which characterized the geometric separation between feature displacement from visual variations and the empirical in-distribution manifold. High nuisance orthogonality therefore measures a potential for disentangled representational prior which presents in the generative backbones before policy training. We show that nuisance

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.

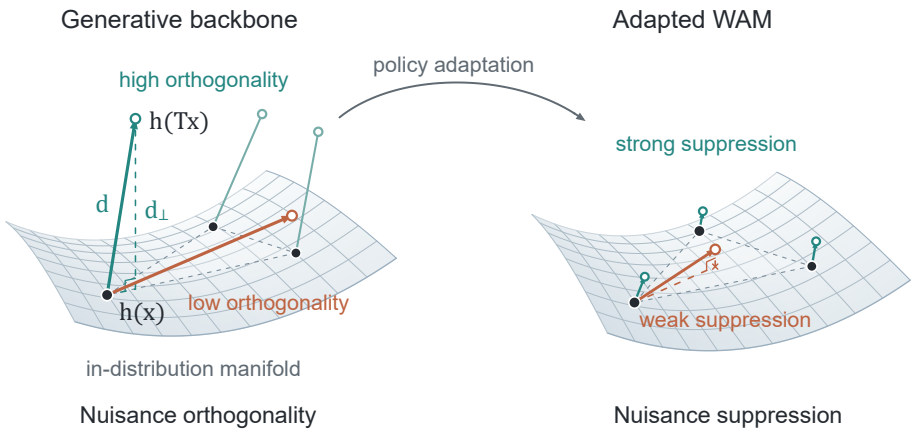


Figure 1: **Nuisance orthogonality and suppression lead to generalization of WAM.** We show that orthogonality of feature displacement with respect to in-distribution data manifold is related to generalization of WAM. This arises from the reorganization of representations during policy adaptation, which we explain through nuisance suppression. Since adaptation reshapes the representation space, the figure provides a conceptual illustration.

orthogonality is highly related to the generalization performance of WAMs, across visual variations and models, allowing prospective diagnostics of world models. Furthermore, we present *nuisance suppression* after policy training, a suppressing behavior of residual displacement so that the policy can remove nuisance information, improve scene retrieval, and enable robust action generation.

Our contributions are as follows:

- **A geometric characterization of representational priors:** Nuisance orthogonality, the fraction of feature displacement orthogonal to the empirical span of in-distribution data, quantifying whether a backbone separates action preserving visual changes from directions that distinguish clean scenes.
- **An evidence from nuisance orthogonality to policy robustness:** Across world model comparisons, higher nuisance orthogonality is associated with stronger downstream visual robustness motivating a diagnostic for assessing world model backbones before WAM policy training.
- **An analysis on how representations reshape after policy adaptation:** We introduce nuisance suppression to measure how representation space change after adaptation, showing that more robust policy exhibits stronger and more consistent nuisance suppression.

2 Related Work

World Action Models. World action models (WAMs) are emerging as a major paradigm for robot foundation models by action generation with predictive modeling of future visual observations (Kim et al., 2026; Ye et al., 2026; Li et al., 2026a; Pai et al., 2025; NVIDIA et al., 2026). Their appeal lies in transferring spatiotemporal priors from generative pretraining to robotics control, while future prediction provides supervision linking actions to their visual consequences. Empirical studies report improved data efficiency and generalization to unseen objects, tasks, and environments, with strong robustness under visual perturbations (Zhang et al., 2026c). Beyond direct policy execution, predictive world modeling can help test-time planning by evaluating the consequences of candidate actions (Kim et al., 2026; Zhou et al., 2026).

Recent work further shows that video prediction co-training is necessary to improve world representation, but explicit test-time generation contributes only to trivial improvements of downstream tasks (Yuan et al., 2026). In addition, some offer an alternative to video generation backbones show-

ing that pretrained image editing representations are helpful for visual generalization (Zhang et al., 2026b; Labs, 2025; Wu et al., 2025; Wang et al., 2025). These researches motivate examining the representational priors of generative backbones and their underlying generalization performance. Our study investigates which geometric properties inherited from visual pretraining are related with robustness and how they reshape after policy adaptation.

Nuisance separation and representation geometry. Nuisance invariance and disentanglement provide a foundation for separating task information from irrelevant variation (Achille & Soatto, 2018), although unsupervised disentanglement is not identifiable without inductive biases or supervision (Locatello et al., 2019). Related objectives remove nuisance factors adversarially (Ganin et al., 2016), separate shared from domain-specific features, or encourage orthogonality between task and nuisance representations (Bousmalis et al., 2016; Sarhan et al., 2020), with applications in domain adaptation and fair representation learning.

In robot manipulation, prior work has decomposed generalization failures into factors of variation and studied which properties of pretrained visual representations predict robustness to distribution shift (Xie et al., 2023; Burns et al., 2023). In visual reinforcement learning, similar ideas enable transfer across visual domains and robustness to task-irrelevant distractors (Higgins et al., 2018; Zhang et al., 2021). More recent work disentangles nuisance factors through explicit or implicit separation in latent action pretraining (Bu et al., 2025; Li et al., 2025; Jeong et al., 2026; Dai et al., 2026; Zhang et al., 2026a). Our diagnostic instead measures nuisance separation in pretrained representations and relates it empirically to policy robustness.

3 Experimental Setup

3.1 Motivation

3.1.1 Why image editing could be beneficial for generalization

Prior work have shown that WAMs from image editing models tend to generalize better than video prediction models (Zhang et al., 2026b). From scalable pretraining, image editors learn to follow instructions so that they can zero-shot inference image editing even in out-of-domain images as shown in Figure 2.

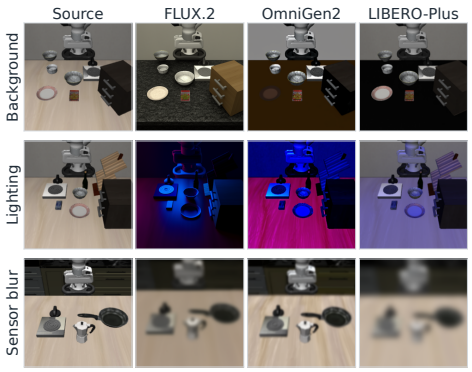


Figure 2: **Zero-shot editing on LIBERO.** Image editing models are already capable of zero-shot editing LIBERO data with visual variations.

During the denoising process, image editors are conditioned on a reference scene and a noisy target image that contains editing modifications, which allow the model to understand the relationship between the source image and edited target image from the early transformer blocks. General knowledge and representational relationships of background modification, illumination, texture, or blurring are more efficiently learnable through scalable image editing pretraining. Video generation models, in contrast, are biased more towards temporal continuity, motion, and dynamics rather than counterfactual changes of the world. This motivates the hypothesis that the generative backbones for robust policy would have more rich representations under domain shift even before policy adaptation.

3.1.2 Representational prior is more than pretraining objective

The generalization performance of ImageWAM has shown significant improvements for using image editing models over video prediction models. Howver, as in Table 1, when training ImageWAM from scratch initialization of FLUX.2 (63.9%), its advantage over Wan2.2 and FastWAM (46.5%) shows that prior knowledge of image editing alone cannot fully explain this comparison. Thus, we assume that representational priors for WAMs are not just what the generative backbone has learned

Table 1: **Nuisance Orthgonality and LIBERO-Plus Success Rate.** Nuisance orthogonality of the generative world models and LIBERO-Plus evaluation suite success rate across visual variations after WAM training. † represents results from the original papers.

Model	Camera		Background		Light		Sensor Noise		Average	
	Orth.	SR (%)	Orth.	SR (%)	Orth.	SR (%)	Orth.	SR (%)	Orth.	SR (%)
Baselines										
Wan2.2†	0.673	16.4	0.420	53.7	0.631	78.2	0.653	37.7	0.594	46.5
FLUX.2	0.803	77.6	0.774	84.9	0.845	98.5	0.773	94.9	0.799	89.0
Scratch FLUX.2	0.754	65.3	0.584	56.8	0.778	71.4	0.778	62.0	0.724	63.9
OmniGen2†	0.802	80.0	0.700	69.4	0.712	82.6	0.768	77.1	0.746	77.3
Ovis-U1†	0.791	63.3	0.718	66.7	0.772	86.3	0.806	75.2	0.772	72.9
Additional Models										
FlashMotion	0.696	52.3	0.515	50.9	0.682	90.9	0.685	61.7	0.645	64.0
Kiwi-Edit	0.790	62.3	0.741	68.0	0.811	94.4	0.802	83.8	0.786	77.1
SANA-Video	0.715	60.8	0.696	62.3	0.803	94.4	0.753	84.3	0.742	75.4
MAGI-1	0.814	67.9	0.662	66.4	0.683	87.4	0.841	83.2	0.750	76.2

through the pretraining objective, but is affected by collection of all design choices including vision and language tokenizers, architecture and initialization, language injection pipeline, pretraining datasets. To analyze this, we start by analyzing the geometry of internal representations of generative backbones when various observations are encoded as a conditioned frame.

3.2 Experimental Setting

Diagnostic dataset and visual shifts. We construct a diagnostic dataset from LIBERO-Plus using paired observations that share the same underlying simulator state but differ in visual appearance (Liu et al., 2023; Fei et al., 2025). The dataset covers four categories of visual shift: lighting, background, sensor noise, and camera viewpoint, each evaluated at multiple strengths. We use 40 clean reference observations to estimate the empirical in-distribution feature subspace introduced in Section 4. For these reference states, we collect 1,200 corresponding out-of-distribution observations and use the resulting pairs to measure variation-induced feature displacement. Generalization requiring changes in language understanding, task composition, or action semantics is outside the scope of this study.

Generative backbones and WAMs. The experiments use Wan2.2, FLUX.2, OmniGen2 and Ovis-U1 from the prior works and scratch FLUX.2 as generative backbones (Wan et al., 2025; Yuan et al., 2026; Labs, 2025; Wu et al., 2025; Wang et al., 2025; Zhang et al., 2026b). Generalization performances are presented with their policy success rates in LIBERO-Plus trained with LIBERO dataset under the configurations from FastWAM and ImageWAM, with total of $5,418$ evaluation trajectories.

In addition, to show that our work is generalizable to other generative backbones, we extend FlashMotion, Kiwi-Edit, SANA-Video 2.0 5B, and MAGI-1 4.5B with a comparable model capacity scale and extend them to a generative backbone for WAMs (Li et al., 2026b; Lin et al., 2026; Chen et al., 2026; ai et al., 2025). They are extended to WAM with MoT architecture with action expert, trained with future prediction loss respecting their original design and flow matching action prediction loss (Black et al., 2026). Specific model configurations and training settings are in the Appendix A.1.

4 Nuisance Orthogonality

With the above motivation, we analyze the displacement of world model representations caused by the visual variations from the clean image. We hypothesize that displacement direction and their orthogonality from the in-distribution manifold would measure disentanglement of information from the generative backbones, which would be geometrically easier to remove nuisance information and less suffer from shortcut learning during policy adaptation.

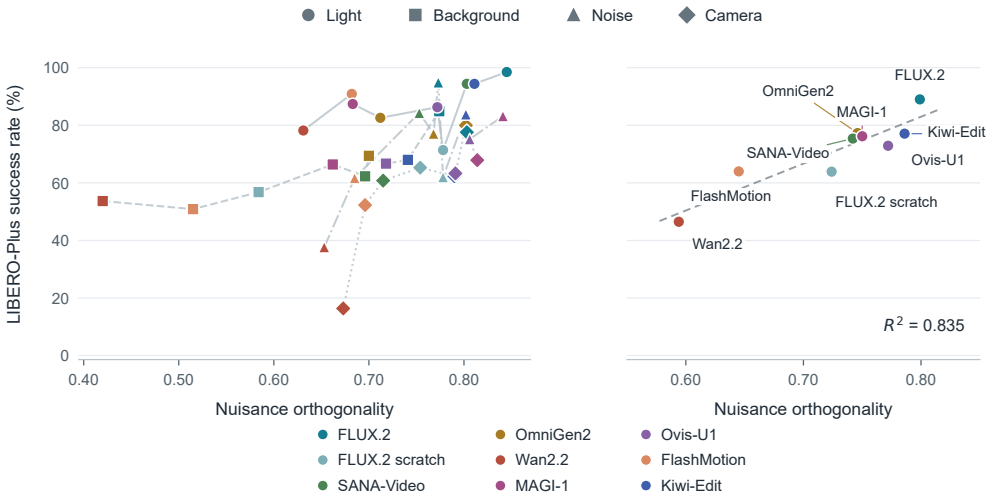


Figure 3: **Visualization of Nuisance Orthgonality and LIBERO-Plus Success Rate.** Overview and visualization of nuisance orthogonality and LIBERO-Plus generalization performance in Table 1. The left figure shows tendency across variation types and the right shows the mean.

4.1 Definition

Let h_ℓ be intermediate representations from transformer block at layer ℓ of generative backbone before WAM training. We extract $h_\ell(x)$ spatially averaged with the clean observation x. With mean of the reference features $\mu_\ell = \frac{1}{n}\Sigma_{i=1}^{n} h_\ell(x_i)$, define

$$H_\ell = \begin{bmatrix} (h_\ell(x_1) - \mu_\ell)^\top \\ \cdots \\ (h_\ell(x_n) - \mu_\ell)^\top \end{bmatrix}, \qquad P_\ell = V_r V_r^\top \tag{1}$$

where V_r contains the retained right singular vectors of $H_\ell = U\Sigma V^\top$ with retained rank r, P_ℓ becoming a projection matrix to the empirical representation subspace at layer ℓ.

Then, we extract $h_\ell(Tx)$ which are features of corresponding sample and action-preserving transformation T as a condition frame. Computing the orthogonality of the displacement vector

$$d_\ell(x,T) = h_\ell(Tx) - h_\ell(x), \qquad q_\ell(x,T) = 1 - \frac{\|Pd_\ell(x,T)\|^2}{\|d_\ell(x,T)\|^2} \tag{2}$$

which can be aggregated over model layers and our diagnosis dataset, resulting the nuisance orthogonality of generative model under transformation T.

$$\mathcal{O}(T) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_x[q_\ell(x,T)]. \tag{3}$$

A high orthogonality score means that the displacement is mostly orthogonal to sampled in-distribution subspace.

4.2 Analysis

4.2.1 Nuisance Orthogonality Aligns with Generalization

If nuisance orthogonality is a useful inherited prior, generative backbones with higher orthogonality should tend to become more robust policies. We test this on the baseline models with an average success rate of LIBERO-Plus tasks as a generalization performance after WAM training. We retain full rank $r = 39$ for 40 reference scenes preserving all nonzero empirical singular directions.

Table 1 presents the results with nuisance orthogonality compared to the evaluation success rate. Figure 3 further shows that there is a strong tendency for the orthogonality to the downstream WAM generalization performance in general, while the sensitivity differs between visual variation types and difficulties. We further ablate this by the retained rank in Figure 4, showing that the nuisance orthogonality is consistent with the changes in the retained rank. We note the value for the orthogonality increases as retained rank decreases because fewer directions contribute to the projected displacement.

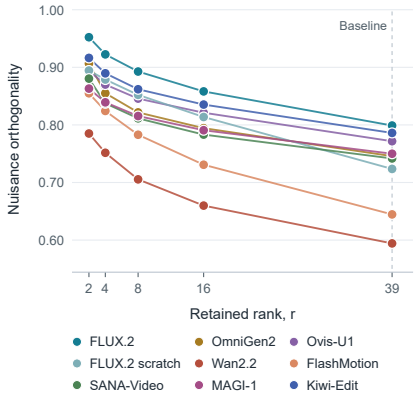


Figure 4: **Rank ablation.** Nuisnace orthogonality is consistent to retained rank changes.

4.2.2 Implications for model selection

To show that this tendency is generalized to other generative backbones, we conduct identical analysis to world models including FlashMotion, Kiwi-Edit, SANA-Video 2.0 5B and MAGI-1 4.5B (Li et al., 2026b; Lin et al., 2026; Chen et al., 2026; ai et al., 2025). Each model is trained as a WAM policy respecting their architecture and action expert designs in FastWAM and ImageWAM, with video prediction loss from the original backbones and flow matching loss for action prediction. Specific designs are in the Appendix A.1.

The results in Table 1 and Figure 3 show that we could predict the performance in advance of WAM training by measuring the nuisance orthogonality. Moreover, video models being capable of high generalization performance supports the hypothesis that any model with rich representation regardless of their pretraining objective could have a useful representational prior.

5 Nuisance Suppression

A useful representational prior for generalization should support policy adaptation that preserves task-relevant scene distinctions while reducing sensitivity to action-preserving appearance changes. We therefore measure how displacements within and outside the empirical in-distribution manifold change after WAM training.

5.1 Definition

WAM adaptation can completely change representational geometry and scale of both displacements and features. We therefore normalize them using the empirical covariance of in-distribution features in the corresponding model. The same in-distribution observations define the reference set for both checkpoints, but we fit the basis and variances at each checkpoint representations separately.

Using the centered clean matrix H from Equation 1, compute $H = U\Sigma V^\top$, retained rank r, projection matrix $P = V_r V_r^\top$, and feature displacement $d = h(Tx) - h(x)$. Each $\lambda_k = \sigma_k^2/n$ is the variance of $v_k^\top h(x_i)$ across the n in-distribution observations. All of them are checkpoint-specific.

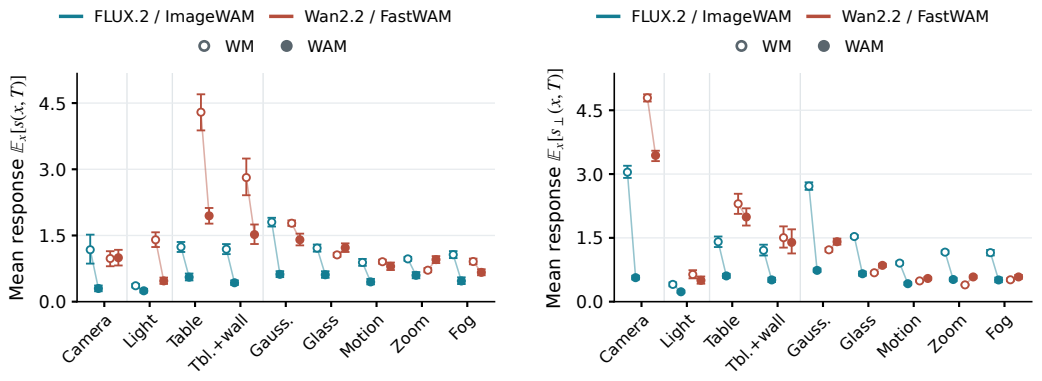
For transformation T, define the standardized response within the in-distribution manifold and its WAM/WM ratio as

$$s(x,T) = \frac{1}{r}\sum_{k=1}^{r}\frac{\langle v_k, d\rangle^2}{\lambda_k}. \qquad \rho_T = \frac{\mathbb{E}_x[s_{\mathrm{WAM}}(x,T)]}{\mathbb{E}_x[s_{\mathrm{WM}}(x,T)]}. \tag{4}$$

And to measure the orthogonal component, define

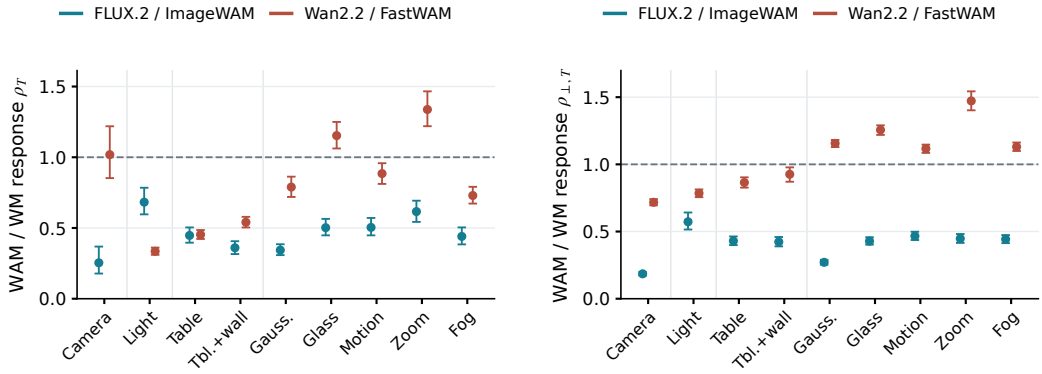
$$s_\perp(x,T) = \frac{\|(I-P)d\|^2}{\sum_{k=1}^{r}\lambda_k}, \qquad \rho_{\perp,T} = \frac{\mathbb{E}_x[s_{\perp,\mathrm{WAM}}(x,T)]}{\mathbb{E}_x[s_{\perp,\mathrm{WM}}(x,T)]}. \tag{5}$$

Where the denominator is the total variance of the clean reference features for normalization.



(a) Absolute in-distribution response.

(b) Absolute orthogonal response.



(c) Relative in-distribution response.

(d) Relative orthogonal response.

Figure 5: **Nuisance responses before and after WAM adaptation.** (a) and (b) show the absolute mean responses $\mathbb{E}_x[s(x,T)]$ and $\mathbb{E}_x[s_\perp(x,T)]$ for each WM and WAM. (c) and (d) show the corresponding WAM/WM ratios ρ_T and $\rho_{\perp,T}$. Suppression takes place when the values are under the dashed line marking 1.0.

A value $\rho_T < 1$ or $\rho_{\perp,T} < 1$ indicates relative *nuisance suppression* for the corresponding component. Within-span suppression means smaller displacements in coordinates whitened by the covariance of clean reference features, measuring reduced response along directions that distinguish the sampled clean observations. Orthogonal suppression means less residual outside that span per unit of total reference variance, indicating less sensitivity.

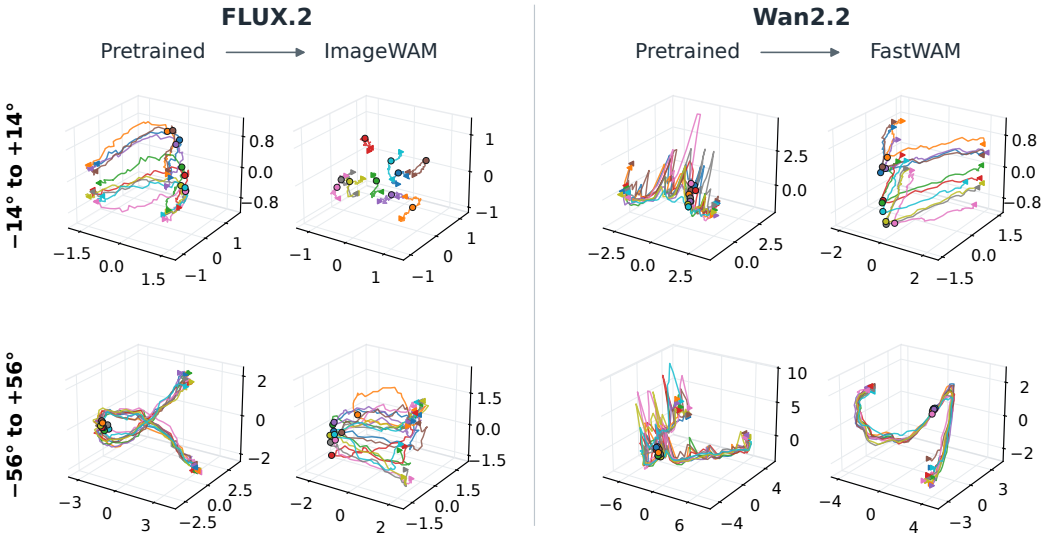
5.2 Analysis

Nuisance Suppression of Robust Policy. If nuisance suppression is what happens after robust policy training, we can expect suppression on both components. We test this property on FLUX.2 to ImageWAM and Wan2.2 to FastWAM on our diagnosis dataset. For simplicity and readability, we present metrics from the features extracted from *single_07* for FLUX.2 and ImageWAM, and *block_14* for Wan2.2 and FastWAM.

Figure 5 shows how the responses change after policy adaptation. The absolute scores in the top figures decrease consistently from FLUX.2 to ImageWAM across all tested variations, measuring consistent nuisance suppression. The corresponding ratios in the bottom figures confirm this pattern by showing ρ_T and $\rho_{\perp,T}$ being below 1 for both along in-distribution manifold directions and orthogonal direction.



(a) **Camera sweep examples.** Examples of nine views from -56° to $+56^\circ$.



(b) **3D PCA Visualization.** Top is features from $\pm 14^\circ$ range and bottom from $\pm 56^\circ$ range. Colors identify the ten scenes, circles mark 0° view and triangles mark the camera range endpoints.

Figure 6: **Representations of Camera variation.** We present example for camera sweep experiments in (a) and 3D PCA Visualizations in (b). There are fitted and scaled independently in each panel for visibility.

In contrast, FastWAM shows a less consistent pattern, with weaker suppression compared to ImageWAM. Where suppression does not even take place for some of the visual variation types. Also, the absolute score for camera perturbation is not suppressed enough explaining FastWAM's weak performance under camere perturbation (16.4%). The comparison therefore presents the stronger robustness of ImageWAM with more consistent and stronger nuisance suppression.

Controlled Experiment and Visualization. Camera perturbation is the case of LIBERO-Plus with lowest success rate for both FLUX.2 and Wan2.2. We compare FLUX.2 to ImageWAM and Wan2.2 to FastWAM with representations of 64 intermediate camera views ranging over weak ($[-14^\circ, +14^\circ]$) and strong ($[-56^\circ, +56^\circ]$) perturbations on 10 scenes from identical task as in Figure 6a. Effects of adaptation on the representations are visualized with 3D PCA in Figure 6b.

In weak perturbations, ImageWAM representations are suppressed enough so that 0° angle scenes are identifiable from the features of the perturbed scene contracting into compact groups near distinct original views. Their representation approaches camera invariance over this weak perturbation range without even seeing them. On the other hand, FastWAM representations tries to disentangle the camera trajectory from scene identity rather than suppressing it.

Continued in strong perturbations, ImageWAM representation paths vary along the scenes and camera viewpoints but each scenes remain distinguishable from each other. However, FastWAM representations are dominated by the camera variations. This means that models learn to ignore the nuisance information without any assumptions if the representational prior is properly shaped and geometrically easy to do so. Supported by our experiments, we measure this representational prior via orthogonality and ignorance of nuisance via suppression.

Reshaped Geometry of Nuisance Suppression. We analyze that WAM adaptation reduces relative influence of nuisance variation compared to in-distribution variation by strengthening distinctions among the in-distribution samples in order to precisely predict the action. In our controlled experiments, adapting FLUX.2 to ImageWAM increases the total in-distribution variance $\sum_{k=1}^{r} \lambda_k$ by a factor of 13.4, while the visual variation cause the displacement of feature with in-distribution variance and orthogonal direction increased by 2.73 and 2.48. However, adapting Wan2.2 to FastWAM increases total in-distribution variance by a factor of 4.42, while each displacement component increases by 5.63 and 3.18, with much weaker suppression compared to ImageWAM.

6 Conclusion

We studied how the representational geometry of generative backbones relates to visual generalization in world action models. Nuisance orthogonality measures whether action-preserving visual variations displace features outside the empirical subspace of in-distribution observations, providing a diagnostic before computation intensive policy training. Across the evaluations, higher orthogonality is generally related with stronger policy robustness in LIBERO-Plus, with further evaluations of additional world models support examining representational geometry beyond the distinction between image and video pretraining. After adaptation, ImageWAM capable of robust action generation exhibits consistent nuisance suppression of displacement, whereas FastWAM shows less consistent and weak suppression. Together, these findings suggest that useful representational priors separate visual nuisance from in-distribution samples and that policy adaptation reduces its relative influence.

Limitations Our evidence is restricted to simulated LIBERO manipulation and the visual shifts evaluated in LIBERO-Plus. We exclude generalization benchmarks for language instructions or action compositionalities to study with limited scope on action-preserving visual generalization. Also the experimental results are observational, our results do not isolate every causal effect of orthogonality or of pretraining objective, but study representational geometry and its relationship with various compounding effects and generalization performance.

AI use statement

In this work, we used generative AI tools to aid or polish writing, to enhance the readability of the figures and tables, to draft sections of the paper and to generate prompts for the zero-shot image editing experiment in Figure 2. The details are described in the Appendix A.2. We have not used generative AI tools for research ideation or execution, generating synthetic datasets, proving mathematical claims. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work studies the robustness of simulated robot manipulation using existing benchmarks and pretrained models. Our work does not establish safety in physical deployment, robustness to extreme conditions, or reliability in interactions with people.

Reproducibility statement

We strongly encourage the reproducibility of our work. We provide the maximum level of detail regarding our experimental setup in the Section 3.2 and Appendix A.1. This includes a description of the datasets used, the model architectures, the training procedures and hyperparameter settings.

Acknowledgments

Acknowledgements have been removed for the purpose of blind review and will be included in the final version.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL http://jmlr.org/papers/v19/17-646.html.

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, 2026. Released July 24, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 343–351, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation?, 2023. URL https://arxiv.org/abs/2312.12444.

Junsong Chen, Jincheng Yu, Yitong Li, Shuchen Xue, Haozhe Liu, Jingyu Xin, Yuyang Zhao, Tian Ye, Zhangjie Wu, Zian Wang, Daquan Zhou, Ping Luo, Song Han, and Enze Xie. Sana-video 2.0: Hybrid linear attention with attention residuals for efficient video generation, 2026. URL https://arxiv.org/abs/2607.21553.

Weisheng Dai, Kai Lan, Jianyi Zhou, Bo Zhao, Xiu Su, Junwen Tong, Weili Guan, and Shuo Yang. Conla: Contrastive latent action learning from human videos for robotic manipulation, 2026. URL https://arxiv.org/abs/2602.00557.

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.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning, 2018. URL https://arxiv.org/abs/1707.08475.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Youngjoon Jeong, Junha Chun, and Taesup Kim. Learning to act robustly with view-invariant latent actions, 2026. URL https://arxiv.org/abs/2601.02994.

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, 2026. URL https://arxiv.org/abs/2601.16163.

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

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. *arXiv preprint arXiv:2601.21998*, 2026a.

Quanhao Li, Zhen Xing, Rui Wang, Haidong Cao, Qi Dai, Daoguo Dong, and Zuxuan Wu. Flashmotion: Few-step controllable video generation with trajectory guidance, 2026b. URL https://arxiv.org/abs/2603.12146.

Zuolei Li, Xingyu Gao, Xiaofan Wang, and Jianlong Fu. Latbot: Distilling universal latent actions for vision-language-action models, 2025. URL https://arxiv.org/abs/2511.23034.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance, 2026. URL https://arxiv.org/abs/2603.02175.

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.

Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. URL http://proceedings.mlr.press/v97/locatello19a.html. Best Paper Award.

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*, 2026.

NVIDIA, :, Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei

Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sieniawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiaxiang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang, Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://arxiv.org/abs/2606.02800.

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

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations, 2020. URL https://arxiv.org/abs/2003.05707.

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.

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.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation, 2023.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

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*, 2026. URL https://arxiv.org/abs/2603.16666.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction, 2021. URL https://arxiv.org/abs/2006.10742.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Tianqiu Zhang, Muyang Lyu, Yufan Zhang, Fang Fang, and Si Wu. Dila: Disentangled latent action world models, 2026a. URL https://arxiv.org/abs/2605.15725.

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.

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

Pengfei Zhou, Shengcong Chen, Di Chen, Jiaxu Wang, Rongjun Jin, Bingwen Zhu, Yike Pan, Songen Gu, Kuanning Wang, Shufeng Nan, Xingyu Qiu, Chenhao Qiu, Pu Yang, Yunuo Cai, Jianxiong Gao, Yifan Li, Yanwei Fu, Xiangyu Yue, Zhi Chen, and Jianlan Luo. τ_0-wm: A unified video-action world model for robotic manipulation, 2026. URL https://arxiv.org/abs/2606.01027.

A Appendix

A.1 Model and Training Configurations

Tables 2 summarize model specifications and training configurations. We present summary of the baseline and additional WAMs that follows the setting of FastWAM and ImageWAM configurations.

Table 2: **Model specifications.**

Model	Type	Backbone & Size		Action Expert Size
FastWAM (Yuan et al., 2026)	Video	Wan2.2-TI2V (Wan et al., 2025)	5.00B	1.02B
ImageWAM (FLUX.2) (Zhang et al., 2026b)	Image	FLUX.2 (Labs, 2025)	3.88B	0.64B
ImageWAM (Scratch FLUX.2)	Image	Scratch FLUX.2	3.88B	0.64B
ImageWAM (OmniGen2)	Image	OmniGen2 (Wu et al., 2025)	3.97B	0.76B
ImageWAM (Ovis-U1)	Image	Ovis-U1 (Wang et al., 2025)	1.13B	1.04B
Additinal WAMs	Video	FlashMotion Li et al. (2026b)	5.00B	1.21B
	Video	Kiwi-Edit Lin et al. (2026)	5.00B	1.02B
	Video	SANA-Video 2.0 Chen et al. (2026)	4.47B	0.75B
	Video	MAGI-1 ai et al. (2025)	4.46B	0.64B

Table 3: **Additional Backbone Model Specifications and Training configurations.**

Backbone	FastWAM design		ImageWAM design	
	FlashMotion	Kiwi-Edit	SANA-Video 2.0	MAGI-1
Action Expert Configurations				
Layers	30	30	30	30
Hidden Dimension	1024	1024	1024	1024
FFN Dimension	4096	4096	4096	4096
Activation	GELU	GELU	SwiGLU	SwiGLU
QKV Heads	24/24	24/24	20/20∗, 10/10	24/8
Action chunk size	16	16	16	16
Action token injection	ZeroConv	Joint Attention	- (IDM)	- (IDM)
Initialization source	ControlNet	Backbone	Backbone	Backbone
Training Configurations				
Batch size	128	128	80	80
Epochs	10	10	10	10
Optimizer	AdamW	AdamW	AdamW	AdamW
Betas	(0.9, 0.95)	(0.9, 0.95)	(0.9, 0.95)	(0.9, 0.95)
Learning rate	1e-4	1e-4	1e-4	1e-4
Weight decay	1e-2	1e-2	1e-2	1e-1†, 1e-2
Scheduler	Cosine	Cosine	Cosine	Cosine
Warmup steps	5%	5%	5%	5%
Final learning rate	1e-6	1e-6	1e-6	1e-6

∗ For linear attention layers. † For video backbone.

Table 3 shows WAM Extension with additional backbones and training configurations. We implement them with some model-specific modifications based on FastWAM and ImageWAM action expert designs. We use action expert initialization preprocessing for training stability and predict video frames interpolated to temporal compression rates of VAE in each model. (Yuan et al., 2026; Zhang et al., 2026b; Li et al., 2026a).

For FlashMotion and Kiwi-Edit, which uses adapted Wan2.2-TI2V-5B, we use FastWAM action expert design to directly extend it to WAM. To respect FlashMotion's trajectory conditioning, we initialize action expert with preprocessed ControlNet branch weight parameters and inject action expert features to video prediction branch using ZeroConv adaptor layers instead of joint attention (Zhang et al., 2023). This is why its action expert has more parameters than FastWAM. Moreover, Kiwi-Edit supports our motivation in Section 3.1 where editing has property that helps generalization, but is not limited to image editing models.

For the additional backbones, SANA-Video 2.0 and MAGI-1, we use ImageWAM (FLUX.2) action expert design but limit the number of action expert layers to 30, less than their backbone, which attends to middle layers to limit their capacity and improve stability inspired by prior works (Team et al., 2026; Ma et al., 2026). We match attention strategy to full IDM setting of ImageWAM instead of the original model's which include autoregressive denoising or block causal generation of multiple timeframes.

Computation resource. All training runs used 4 Nvidia H100s with gradient accumulation of 2 to match identical effective batch size of each training configuration settings, taking up to 38 hours for each WAM training runs.

A.2 Zero-shot Editing

We used commercial LLM (Claude Opus 5 xhigh) to generate prompts from original LIBERO and corresponding LIBERO-Plus sample with our editing context, then use them as a zero-shot prompt for image editing models to generate edited samples in Figure 2 (Anthropic, 2026). For image editing, we used CFG of 3.0 and inference step of 50 (Ho & Salimans, 2022). Example prompts for each row are as below.

- Replace the entire visible tabletop edge to edge, including the surface beneath every object, with large irregular cloudy gray cobblestone texture. This is a surface material, not a rug or mat: leave no border and no original wood visible. Replace only the background wall with muted dusty light-blue plaster. Preserve the exact camera, lighting, robot, cabinet, stove, rack, bottle, bowl, plate, cheese box, silhouettes, geometry, and positions
- Change only the scene illumination: moderately dimmer illumination, a cool blue cast and a subtle magenta cast, substantially softer contrast and shadows; mean-light ratio approximately 0.70x.
- Apply moderate Gaussian defocus blur.