Test-Time Scaling for World Action Models via Zero-Shot Geometric Verification

Zesen Zhao[1] Minkyoung Cho[1] Hui Shen[1]
Boyuan Zheng[1] Kunxiao Gao[1] Yulong Cao[2] Z. Morley Mao[1]
[1]University of Michigan [2]NVIDIA
{hymanzzs, minkycho, huishen, boyuann, kunxiao, zmao}@umich.edu
yulongc@nvidia.com

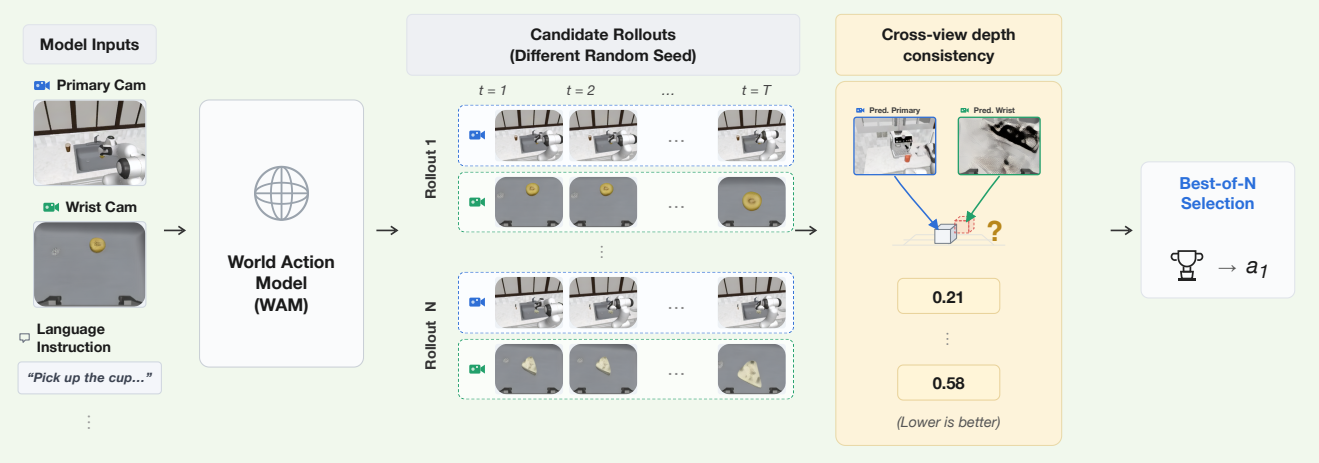

Figure 1. Overview of our approach. WAM rollouts from multiple views are scored by a frozen VGGT model via a cross-view depth reprojection consistency error. The most geometrically consistent candidate is selected for execution.

Abstract

World Action Models (WAMs) jointly predict future video frames and actions from multi-view observations, but their generations can contain geometrically implausible scenes, degrading action quality. For WAMs, on-the-fly rollout selection often relies on model-specific value heads, limiting cross-model reuse. To address this, we propose a training-free, model-agnostic rollout verifier that repurposes VGGT as a frozen geometric evaluator. It scores WAM rollouts using a cross-view depth reprojection consistency error and selects the most geometrically plausible candidate via Best-of-N test-time scaling. Our key insight is that if a world model faithfully captures 3D structure, its multi-view predictions must be geometrically consistent; violations can indicate hallucination. On RoboCasa with Cosmos Policy, our verifier achieves an AUROC of 0.736 for single-step failure detection, substantially outperforming the co-trained value head. Consequently, our test-time scaling improves overall task success rates across various benchmarks and WAM baselines.

1. Intro & Background

World Action Models (WAMs) [4, 11] have recently emerged as a promising paradigm for visuomotor control by jointly predicting future frames and actions from multi-view inputs. However, their predicted futures can contain artifacts that mislead action generation.

Although test-time scaling and verifier-based selection have been explored for robot policies and world models [2, 3, 7], and VGGT features have been used to improve geometric consistency during video world-model training [10], existing methods do not directly leverage WAMs' multi-view predicted futures as a training-free geometric verifier.

We exploit a simple cue: by reprojecting VGGT-predicted 3D points across views, we obtain a zero-shot geometric consistency score for Best-of-N rollout selection without executing the action, enabling on-the-fly test-time scaling for WAMs.

Table 1. Test-time scaling results: success rate (%) across benchmarks and WAMs. The value head is only available for Cosmos Policy.

				VGGT BoN (ours)		
Benchmark	**WAM**	**Baseline**	**Value BoN (N=8)**	N=2	N=4	N=8
RoboCasa [8]	Cosmos Policy	57.1	58.3	62.3	62.8	**66.4**
LIBERO Long [6]	Cosmos Policy	96.2	95.7	98.3	98.6	**99.4**
LIBERO Long	LingBotVA [5]	92.8	N/A	92.7	93.2	**93.3**
RoboTwin 2.0 [1]	LingBotVA	65.3	N/A	56.1	66.7	**67.0**

Table 2. Failure detection AUROC for different signals on RoboCasa with Cosmos Policy.

Signal	**Training-free?**	**AUROC**
Cosmos value head	✗	0.587
Per-view depth confidence	✓	0.584
Cross-view depth consistency (ours)	✓	**0.736**
LPIPS wrist–GT (GT-dependent)	n/a	0.81

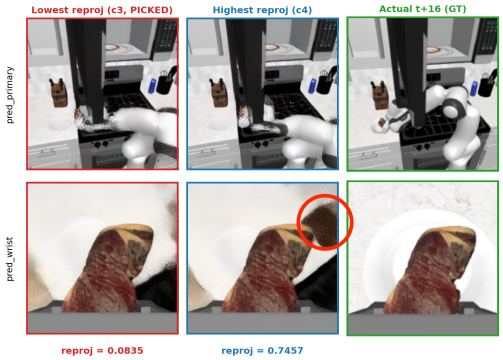

Figure 2. Qualitative candidates; worst rollout shows artifacts.

2. Method

Problem setup. A WAM receives multi-view observations (primary camera I^{pri} and wrist camera I^{wri}) along with a language task description, and jointly generates predicted future frames $(\hat{I}^{\text{pri}}, \hat{I}^{\text{wri}})$ and an action chunk $\mathbf{a}_{1:H}$. Given N candidate rollouts sampled with different random seeds, we seek a selection criterion that identifies the most faithful rollout without action execution.

Cross-view depth reprojection consistency. We feed the predicted frame pair $(\hat{I}^{\text{pri}}, \hat{I}^{\text{wri}})$ into VGGT [9]. Although WAMs predict a sequence of future frames, we score the final predicted frame pair in the rollout. We compute a cross-view depth reprojection consistency error as follows. 3D points from the wrist camera's point map are projected into the primary camera's coordinate frame, yielding projected depth values d^{proj}. These are compared against the VGGT-predicted depth d^{vggt} for the primary view:

$$e_{\text{depth}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left| \log \frac{d^{\text{proj}}(p)}{d^{\text{vggt}}(p)} \right|, \tag{1}$$

where Ω denotes valid projected pixels with VGGT confidence of at least 50%. A low e_{depth} indicates that both predicted views describe a consistent 3D scene. Since no ground-truth depth or executed observation is available at test time, this score is used as a VGGT-induced relative consistency signal rather than an absolute geometric error.

Test-time scaling via Best-of-N. At each action query step, we sample N candidate rollouts with different random seeds, compute e_{depth} for each, and select the candidate with the minimum depth inconsistency.

3. Experiments

Qualitative and correlation analysis. Table 2 evaluates on-the-fly failure detection on RoboCasa with Cosmos Policy. Our cross-view depth consistency score achieves a **0.736 AUROC**, outperforming both the co-trained Cosmos value head and a training-free per-view depth-confidence baseline. This suggests that failure prediction is not explained by VGGT confidence alone; rather, cross-view geometric agreement provides a stronger signal for detecting implausible WAM rollouts. Our score is also close to the GT-dependent LPIPS [12] reference, which requires access to ground-truth future frames and is unavailable at test time. Qualitative examples show that the worst-ranked candidates exhibit visible geometric artifacts (Figure 2).

Test-time scaling and cross-model generalization. Table 1 reports task success rates across benchmarks, WAMs, and candidate budgets. On RoboCasa [8] with Cosmos Policy [4], our geometric Best-of-N selector reaches 66.4% success at $N = 8$, improving over the single-sample baseline by 9.3 percentage points and over the value-based selector by 8.1 percentage points. The gains are largest on RoboCasa, and smaller but positive on high-baseline LIBERO settings. For RoboTwin 2.0, we use a stricter evaluation protocol than the original LingBotVA report. These results suggest that cross-view geometric consistency is a useful proxy for WAM rollout quality, enabling training-free test-time selection without a model-specific value head.

References

[1] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 2

[2] Mingtong Dai, Lingbo Liu, Yongjie Bai, Yang Liu, Zhouxia Wang, Rui Su, Chunjie Chen, Liang Lin, and Xinyu Wu. RoVer: Robot reward model as test-time verifier for vision-language-action model. *arXiv preprint arXiv:2510.10975*, 2025. 1

[3] Suhyeok Jang, Dongyoung Kim, Changyeon Kim, Youngsuk Kim, and Jinwoo Shin. Verifier-Free Test-Time Sampling for Vision-Language-Action Models. In *ICLR*, 2026. 1

[4] 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. 1, 2

[5] 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*, 2026. 2

[6] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *NeurIPS*, 2023. 2

[7] Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. World Action Verifier: Self-improving world models via forward-inverse asymmetry. *arXiv preprint arXiv:2604.01985*, 2026. 1

[8] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024. 2

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 2

[10] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry Forcing: Marrying video diffusion and 3D representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025. 1

[11] 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. *arXiv preprint arXiv:2602.15922*, 2026. 1

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2