RoboWM-Bench: A Benchmark for Evaluating World Models in Robotic Manipulation

Abstract

*Recent advances in large-scale video world models have enabled realistic future prediction, suggesting potential for robot learning from imagined videos. However, visual realism does not imply physical plausibility, and behaviors from such videos may fail in real-world execution. Existing benchmarks consider plausibility but mainly focus on perception, lacking an evaluation of whether predicted behaviors can translate into successful actions. We introduce **RoboWM-Bench**, a manipulation-centric benchmark for embodiment-grounded evaluation of video world models. This benchmark converts behaviors from both human-hand and robotic manipulation videos into embodied actions and validates them through execution. It spans diverse manipulation scenarios and establishes a unified and reproducible protocol. Evaluation of state-of-the-art video world models shows that generating physically executable behaviors remains challenging, with failures in spatial reasoning, contact stability, and physical realism. While finetuning on manipulation data yields improvements, physical inconsistencies still persist, suggesting opportunities for more physically grounded video generation for robots.*

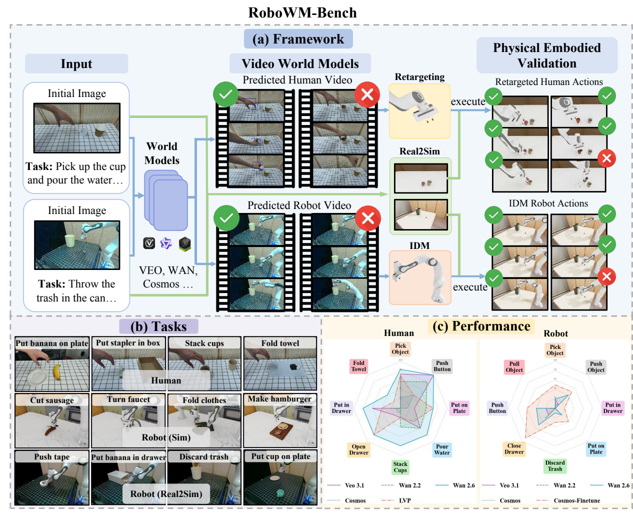

Figure 1. **RoboWM-Bench** is a manipulation-centric benchmark for evaluating video world models under embodied execution. (a) Given the initial scene observations and task descriptions, world models generate manipulation videos with human hands or robot arms. The predicted behaviors are converted into embodied actions and validated in simulation through real-to-sim reconstruction. (b) RoboWM-Bench spans a diverse suite of manipulation tasks. (c) Performance of world models on RoboWM-Bench.

1. Introduction

Recent advances in video world models have enabled realistic and temporally coherent future predictions [4, 10, 11, 27], opening new opportunities for robot learning from generated videos. However, despite their impressive visual fidelity, such models may generate behaviors that violate physical consistency when grounded in real-world dynamics. Motivated by this challenge, a growing body of work has begun to tailor world models to robotic manipulation scenarios, seeking to better capture active embodiment-aware interactions [1, 2, 5, 7, 12, 26]. As both visual realism and physical consistency improve, imagined manipulation videos are increasingly viewed as a scalable data source for robot learning [5, 14, 25]. In this context, reliable evaluation is crucial for assessing world models and ensuring physically grounded policies.

Existing benchmarks for world models primarily emphasize visual fidelity, semantic consistency, and temporal coherence [9, 13, 15, 20, 21, 31]. More recent efforts incorporate physical plausibility, revealing that SOTA models often fail to maintain coherent physical dynamics despite strong perceptual quality [24, 32, 33]. While these evaluations represent important progress, they remain largely perception- or diagnostic-oriented. To reliably assess predicted videos, evaluation should extend beyond perceptual plausibility to consider whether the predicted behaviors can be faithfully executed by embodied agents and successfully accomplish the intended tasks. A recent study [8] moves toward embodiment-grounded evaluation by validating executability on real robots. However, comprehensive evaluation requires broader coverage and greater task diversity, and its reliance on real-world testing makes large-scale, reproducible benchmarking challenging.

In this paper, we introduce **RoboWM-Bench**, a systematic and reproducible manipulation-centric benchmark for embodiment-grounded evaluation of video world models. RoboWM-Bench operationalizes *physical executability* as a principled and measurable evaluation criterion, grounding assessment in verifiable control execution rather than perceptual judgment alone. Specifically, the benchmark evaluates whether interactions predicted in generated videos can be translated into executable action sequences and successfully performed in a physically grounded environment.

As illustrated in Figure 1, given initial observations and task descriptions, video world models generate future manipulation videos involving either human hands or robot arms. The predicted behaviors are then converted into embodied action sequences through inverse dynamics modeling [3, 14] for robotic videos, or pose tracking and retargeting [17, 18, 23] for human demonstrations. To enable fair and accessible evaluation across real-world scenarios, we adopt a real-to-simulation (real-to-sim) framework in which scenes are reconstructed in simulation to match their real-world counterparts [6, 29]. The extracted actions are then executed within the reconstructed environments using visually and physically high-fidelity simulation, enabling standardized and reproducible validation of physical executability. RoboWM-Bench spans a broad spectrum of manipulation scenarios, including diverse object dynamics, short- and long-horizon tasks, and both single-arm and bimanual interactions. Through this unified protocol, RoboWM-Bench enables consistent and reproducible comparison of video world models under embodiment constraints. It also provides hierarchical evaluation, incorporating both step-level executability metrics and final task-level success rates, which together enable fine-grained diagnostic analysis as well as holistic performance assessment.

We evaluate video world models on RoboWM-Bench under embodied execution. The results suggest a noticeable gap between visual realism and physical executability, with success rates decreasing as task complexity increases. Qualitative analysis further reveals common inconsistencies in generated videos, including unrealistic object deformation and inaccurate contact prediction, which may lead to dynamically infeasible actions during execution. While fine-tuning on manipulation data improves executability, certain physical inconsistencies remain. These observations suggest that ensuring physically consistent behavior remains a challenge for current video world models, and highlight opportunities for advancing more physically grounded and embodiment-aware world modeling.

2. RoboWM-Bench

2.1. Benchmark Overview

RoboWM-Bench evaluates video world models via physical executability across human and robotic manipulation. Given an observation and task, world-model-predicted videos are converted into actions (Sec. 2.2) and executed in real-to-sim environments for reproducible evaluation (Sec. 2.3). The benchmark spans a diverse suite of tasks (Sec. 2.4), with executability determined by whether the intended task objective is achieved (Sec. 2.5).

2.2. Embodied Video-to-Action Execution

RoboWM-Bench evaluates both human-hand and robotic manipulation videos. While human-centric predictions often exhibit higher visual quality, robotic videos align more directly with downstream control.

2.2.1. Human-Centric Retargeting.

Following [17, 18], we estimate 3D human hand poses via HaMeR [23] and retarget them to robot end-effector (EE) poses. The EE position is defined as the thumb-index midpoint. To improve orientation stability over prior formulations [18], we project the thumb and index fingertips onto a fitted plane; the x-axis connects these projections, while the z-axis is the plane normal, better preserving interaction geometry. For gripper aperture, we use the minimum distance between the thumb and all other fingertips to capture diverse contact points. Finally, we apply trajectory smoothing and temporal denoising to stabilize the retargeted motion [18].

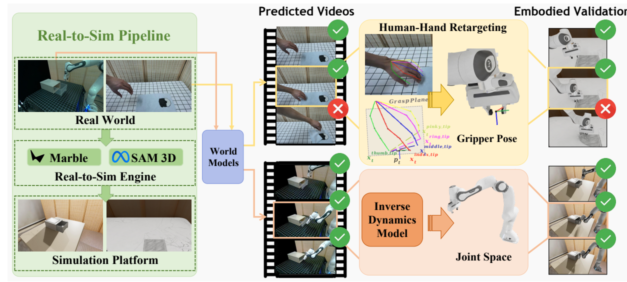

Figure 2. **Pipeline of RoboWM-Bench.** We reconstruct real-world scenes into simulation via a real-to-sim pipeline for reproducible testing. Video predictions are mapped to actions through either human-centric retargeting or robot-centric inverse dynamics. The resulting actions are executed in simulation to evaluate embodied executability.

2.2.2. Robot-Centric Execution.

For robotic videos, we recover joint-space actions via an inverse dynamics model (IDM) [3, 14]. The IDM takes consecutive frames as input to predict intermediate action chunks [14]. To optimize training, we employ a two-stage strategy: (1) **Simulation Pretraining**: We collect large-scale, high-frequency (80Hz) robot trajectories in simulation for smooth motion supervision. To bridge the sim-to-real gap, we adopt a background-masking strategy [26], isolating the robot arm to minimize domain discrepancies. (2) **Real-world Finetuning**: The IDM is finetuned on a small real-world dataset without masking. This pipeline enhances data efficiency and sim-to-real generalization, ensuring reliable action extraction for robotic benchmarks.

2.3. High-Fidelity Real-to-Sim Framework

RoboWM-Bench conducts evaluation in open-source simulation to ensure accessibility and reproducibility. Built on the LeHome framework [19], our pipeline supports high-fidelity rendering and realistic physics. For real-to-sim reconstruction, we adopt a modular pipeline: (1) **Scene Reconstruction.**: Backgrounds are represented via 4D Gaussians [29] for visual and spatial consistency. (2) **Object Assets**: Rigid geometries are acquired through 3D segmentation [6], while articulated and deformable models follow [19]. (3) **State Initialization**: Initial poses are estimated via FoundationPose [28] or MegaPose [16], with camera calibration performed using FEEPE [30]. This

Table 1. Task-level and step-level embodied execution success rates (%) on RoboWM-Bench across human-hand and robotic tasks.

Method	Human (Task Level)								Human (Step Level)				
									Put in Drawer				
	Pick Object	Push Button	Put on Plate	Pour Water	Stack Cups	Open Drawer	Put in Drawer	Fold Towel	contact	lift	above drawer	in drawer	close drawer
Cosmos	23%	40%	15%	0%	10%	10%	10%	0%	80%	20%	20%	20%	10%
Wan 2.2	57%	80%	55%	60%	40%	0%	20%	0%	**100%**	60%	60%	40%	20%
Wan 2.6	**83%**	**100%**	**70%**	**80%**	**80%**	**80%**	**80%**	**40%**	**100%**	**80%**	**80%**	**80%**	**80%**
Veo 3.1	73%	**100%**	30%	60%	20%	20%	60%	0%	**100%**	70%	70%	60%	60%
LVP	70%	40%	**70%**	40%	20%	**80%**	40%	20%	**100%**	70%	60%	50%	40%

Method	Robot (Task Level)								Robot (Step Level)				
									Put in Drawer				
	Close Drawer	Pick Object	Push Object	Push Button	Put on Plate	Discard Trash	Pull Object	Put in Drawer	contact	lift	above drawer	in drawer	close drawer
Cosmos	0%	10%	10%	10%	10%	0%	0%	0%	10%	0%	0%	0%	0%
Wan 2.2	30%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Wan 2.6	50%	20%	40%	40%	20%	10%	0%	0%	30%	0%	0%	0%	0%
Veo 3.1	20%	20%	10%	20%	10%	0%	0%	0%	30%	0%	0%	0%	0%
Cosmos-FT	**90%**	**50%**	**50%**	**60%**	**40%**	**30%**	**40%**	**20%**	**60%**	**20%**	**20%**	**20%**	**20%**

framework ensures simulation faithfully preserves the physical structure of real scenes, enabling scalable and reproducible evaluation without costly physical hardware.

2.4. Manipulation Task Suite

RoboWM-Bench comprises a suite of manipulation tasks with varying levels of complexity, designed to systematically evaluate the embodied reasoning and physical consistency of video world models. Built on the LeHome simulation engine [19], the benchmark enables reproducible and physically realistic interactions. RoboWM-Bench spans diverse object types and interaction regimes, including rigid-object tasks, articulated interactions, and deformable-object manipulation. It further includes long-horizon tasks requiring multi-stage planning, as well as bimanual tasks that introduce coordination constraints. Together, these tasks enable systematic evaluation.

2.5. Evaluation of Embodied Executability

We define *embodied executability* as the translation of predicted behaviors into dynamically feasible, task-completing actions. Our protocol employs a hierarchical assessment: (1) Step-level checks of key interactions (e.g., contact, lifting). (2) Task-level success requires passing all step-level checks and fulfilling the final objective. This design enables fine-grained failure analysis and clear task-level evaluation.

3. Experiments

We conduct extensive experiments with video world models across diverse tasks (3.1). Specifically: (1) We quantify the embodied executability of world models and provide analysis (3.2). (2) We compare against existing benchmarks and show that RoboWM-Bench provides a more embodiment-grounded evaluation (3.3). (3) We validate RoboWM-Bench by analyzing the consistency of action extraction and simulation-based execution (3.4).

3.1. Environment Setup

We evaluate video world models on human and robotic manipulation tasks. The environment is built on LeHome [19, 22], supporting high-fidelity rendering and deformable objects. For each task, we report average accuracy over 10 episodes with randomized object initialization. We evaluate several SOTA video world models across two categories. **General-purpose models** include closed-source Veo3.1 [11] and Wan2.6 [27], alongside open-source Wan2.2 [27] and Cosmos-Predict2.5 [2]. **Interaction-oriented models** feature LVP [5], designed for complex human behaviors. Additionally, we evaluate **Cosmos-Finetune**, a variant finetuned on our real-world manipulation dataset (50 trajectories per task), to assess the impact of domain-specific data on embodied grounding.

3.2. Embodied Executability of World Models

Table 1 summarizes execution success rates (simulation results in Appendix). We observe several key trends: **(1) Human-hand vs. Robot Embodiment.** Human-centric videos outperform robotic ones, likely due to human-dominated pretraining biases. Furthermore, human hands exhibit superior geometric stability, whereas robotic manipulators often suffer from structural distortions that trigger execution failure. **(2) Interaction Complexity.** Success rates decline as tasks transition from short-horizon interactions (e.g., *Push Button*) to long-horizon tasks (e.g., *Put in Drawer*) due to cumulative multi-step errors. *Fold Towel* remains the most challenging, indicating that deformable object interactions are particularly difficult for current world models. **(3) Impact of Finetuning.** Finetuning Cosmos with minimal data (50 trajectories/task) improves consistency and reduces artifacts. Yet, persistent bottlenecks in 3D spatial reasoning that cause localization and grasping failures suggest a key direction for future improvement.

As shown in Figure 3, Wan2.6 achieves the strongest

Table 2. Accuracy of action extraction methods.

Method (Human)	Pick Object	Stack Cups	Pour Water	Open Drawer	Fold Towel	Put on Plate	Put in Drawer	Average
Retargeting	**100%**	**90%**	**90%**	**100%**	**100%**	**100%**	**100%**	**97.1%**
Method (Robot)	Pick Object	Pull Object	Push Object	Discard Trash	Close Drawer	Put on Plate	Put in Drawer	Average
IDM_{Real}	70%	70%	80%	70%	90%	70%	50%	71.4%
$\text{IDM}_{\text{Sim+Real}}$	**100%**	**90%**	**100%**	**90%**	**100%**	**100%**	**90%**	**95.7%**

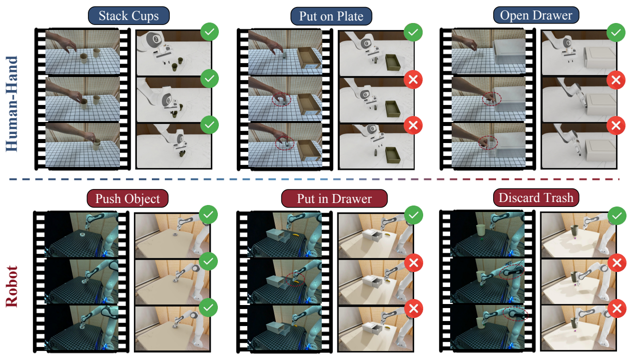

Figure 3. **Qualitative execution results on RoboWM-Bench.** For each task, predicted videos (left) are converted into robot actions and executed in simulation (right).

performance, yet **even strong world models often generate visually plausible but physically inconsistent interactions.** In *Put on Plate*, the model depicts lifting an object despite only touching it without a stable grasp, while in *Open Drawer*, predicted motions resemble closing rather than grasping. These predictions result in execution failures in simulation. For robotic tasks, unrealistic contact behaviors are compounded by geometric distortions in the predicted manipulator, which further reduces reliability.

3.3. Visual Plausibility vs. Embodied Execution

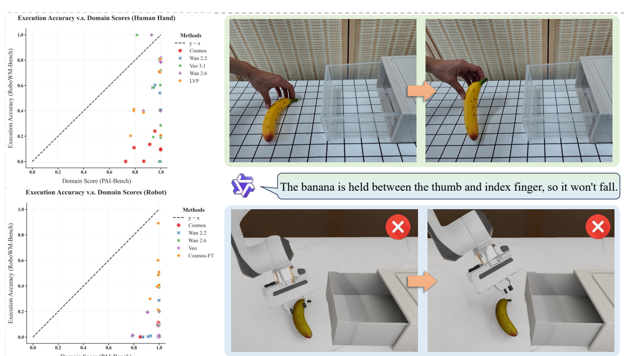

Figure 4. Comparison between PAI-Bench and RoboWM-Bench.

We compare RoboWM-Bench execution accuracy with PAI-Bench scores for perceptual plausibility. As shown in Fig. 4, videos achieve near-saturated scores on PAI-Bench while RoboWM-Bench yields more discriminative outcomes. This discrepancy arises because actions that appear visually plausible may remain physically infeasible, which are captured by embodied execution but often missed by perceptual metrics, demonstrating that RoboWM-Bench provides a more direct measure of physical executability.

3.4. Robustness of RoboWM-Bench

To assess the robustness of RoboWM-Bench, we evaluate two key components of the evaluation pipeline: (1) the accuracy of action extraction from videos, including pose tracking and retargeting for human-hand videos, and the inverse dynamics model (IDM) for robotic manipulation videos; (2) the fidelity of reconstructed simulation environments with respect to their corresponding real-world scenes.

3.4.1. Action Extraction Accuracy.

To verify action extraction accuracy, we execute actions from real-world trajectories in simulation. As shown in Table 2, the human-centric retargeting pipeline achieves high reliability, though minor failures in *Stack Cups* and *Pour Water* arise from precision mismatches between robot grippers and human fingertips. Regarding robotic videos, $\text{IDM}_{\text{Sim+Real}}$ significantly outperforms IDM_{Real}, confirming that simulation pretraining provides essential motion priors for inverse dynamics. Remaining failures are attributed to minor prediction errors that compromise stability in tasks involving precision-dependent contacts.

3.4.2. Simulation Reconstruction Fidelity.

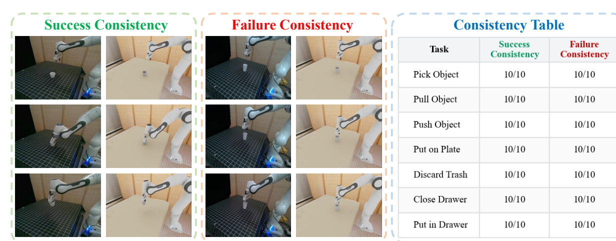

Figure 5. **Real-to-sim consistency.** Identical trajectories are executed in real-world scenes and reconstructed simulation environments, yielding consistent success and failure outcomes.

We evaluate the consistency between real-world trajectories and their reconstructed simulation counterparts. Ideally, identical actions should yield matching outcomes across domains. Since the reconstruction pipeline is shared, we use robotic tasks as a representative benchmark. To quantify consistency, we replay 10 successful and 10 failed real-world trajectories in simulation. As shown in Figure 5, execution outcomes are faithfully reproduced, confirming that our simulation environments preserve the physical structure and interaction dynamics of real scenes. This validation demonstrates that RoboWM-Bench enables reliable evaluation independent of the original physical setups.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1

[2] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 1, 3

[3] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022. 1, 2

[4] Tim Brooks, William Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Clarence Ng, Ricky Wang, Aditya Ramesh, et al. Video generation models as world simulators. `https://openai.com/research/video-generation-models-as-world-simulators`, 2024. OpenAI Technical Report. 1

[5] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025. 1, 3

[6] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 2

[7] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025. 1

[8] Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, et al. Wow, wo, val! a comprehensive embodied world model evaluation turing test. *arXiv preprint arXiv:2601.04137*, 2026. 1

[9] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 1

[10] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 1

[11] Google DeepMind. Veo: a text-to-video generation system (veo-3 technical report). Technical Report Veo-3-Tech-Report, Google DeepMind, 2025. Technical Report. 1, 3

[12] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Y Liao, P Gao, H Li, M Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation (2025). *arXiv preprint arXiv:2501.01895*, 2025. 1

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1

[14] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025. 1, 2

[15] Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5325–5335, 2024. 1

[16] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 2

[17] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025. 2

[18] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *URL https://arxiv. org/abs/2503.00779*, 2, 2025. 2

[19] Zeyi Li, Jade Yang, Jingkai Xu, Shangbin Xie, Tianxing Chen, Yuran Wang, Zhenhao Shen, Yan Shen, Yukun Zheng, Wenjun Li, et al. Lehome: A simulation environment for deformable object manipulation in household scenarios. In *IROS 2025-5th Workshop on RObotic MAnipulation of Deformable Objects: holistic approaches and challenges forward*. 2, 3

[20] Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. Vmbench: A benchmark for perception-aligned video motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13087–13098, 2025. 1

[21] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22139–22149, 2024. 1

[22] NVIDIA Corporation. Nvidia isaac sim: High-fidelity simulation for robotics. `https://developer.nvidia.com/isaac-sim`, 2023. Accessed: 2026-03-03. 3

[23] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2

[24] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026. 1

[25] GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025. 1

[26] GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied ai. *arXiv preprint arXiv:2511.19861*, 2025. 1, 2

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3

[28] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17868–17879, 2024. 2

[29] World Labs. Marble: A Multimodal World Model, 2025. Accessed: 2026-02. 2

[30] Tianshu Wu, Jiyao Zhang, Shiqian Liang, Zhengxiao Han, and Hao Dong. Foundation feature-driven online end-effector pose estimation: A marker-free and learning-free approach. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1921–1928. IEEE, 2025. 2

[31] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025. 1

[32] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 1

[33] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025. 1