WORLDARENA 2.0: EXTENDING EMBODIED WORLD MODEL BENCHMARKING ON MODALITY, FUNCTIONALITY AND PLATFORM

Anonymous authors
Paper under double-blind review

ABSTRACT

World models have emerged as a central paradigm for embodied intelligence, enabling agents to predict action-conditioned future and reason about environmental dynamics. However, existing embodied world model benchmarks are still largely confined to vision-only prediction, offline embodied applications, and simulator-based evaluation, making them insufficient for assessing increasingly comprehensive world models. In this work, we introduce WorldArena 2.0, an expanded benchmark that systematically broadens embodied world model evaluation along three dimensions: *modality*, *functionality*, and *platform*. Along the modality dimension, WorldArena 2.0 extends evaluation from vision-only to visuotactile modalities, enabling assessment of multimodal perception and prediction. Along the functionality dimension, it extends beyond policy evaluation and planning to assess world models as interactive RL environments for policy optimization. Along the platform dimension, it moves beyond simulator-only evaluation to a diverse suite of simulated and real-world robotic settings across multiple embodiments. Under a standardized protocol, WorldArena 2.0 comprehensively evaluates perceptual quality, interactive utility, and cross-platform performance, providing a comprehensive testbed for tracking progress toward embodied world models. The project page including codes and demos is available at: https://worldarena2.github.io/.

1 INTRODUCTION

World models (Ding et al., 2025; Shang et al., 2026b; Long et al., 2025) are increasingly adopted as the core component for embodied intelligence, as they allow agents to predict future states and reason about the outcomes of their actions. In embodied applications, however, the practical value of a world model depends on more than visual realism. Beyond generating visually accurate future observations, these models need to capture physically grounded dynamics to support downstream reasoning and decision-making, such as action planning and policy learning. This transition from visual predictors to interactive environments naturally requires a corresponding shift in evaluation methods. As a result, modern benchmarks are gradually expanding their focus from visual quality to functional utility, aiming to assess these models under more realistic and demanding conditions.

Existing benchmarks for world models generally fall into two categories. The first one primarily focuses on the quality of video generation, evaluating models based on spatio-temporal consistency and visual fidelity (Li et al., 2025a; Duan et al., 2025; Yue et al., 2025). The second category advances toward embodied functionality, assessing whether world models can capture action-conditioned dynamics to facilitate specific downstream tasks. Notable frameworks in this line include WorldSimBench (Qin et al., 2024), WorldEval (Li et al., 2025c), World-in-World (Zhang et al., 2026), WoW-World-Eval (Fan et al., 2026) and WorldArena (Shang et al., 2026a). Among these, WorldArena (Shang et al., 2026a) represents an important step toward unified embodied world model evaluation. It provides a systematic benchmark that jointly evaluates perceptual quality and downstream functional utility, moving beyond purely visual prediction to examine whether world models can support embodied decision-making. In particular, WorldArena introduces a standardized evaluation framework in which world models are assessed not only by the fidelity of generated videos, but also by their usefulness for embodied tasks, including data engine, action planning and policy evaluation. This design establishes an explicit bridge between generative prediction and end-to-end

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.

system utility. Rather than treating world models as isolated video predictors, our benchmark assesses their practical effectiveness when integrated with downstream control interfaces.

While these benchmarks provide a foundation for evaluating visual quality and basic task functionality, they do not fully capture the interactive and physical requirements of actual embodied deployment. Specifically, there remain three major limitations. First, prevailing benchmarks are predominantly confined to vision-only settings. This overlooks the multimodal nature of embodied interactions, where tactile feedback is essential for resolving contact-rich dynamics and physical friction. Second, downstream evaluation is largely restricted to open-loop planning (Zhang et al., 2026) or static policy evaluation (Li et al., 2025c), rarely investigating the capacity of a world model to serve as an interactive reinforcement learning environment that supports iterative policy improvement through imagined rollouts. Finally, benchmark results are currently obtained almost exclusively in simulation, leaving it entirely unclear whether strong performance in controlled virtual environments translates to real-world deployment.

To address these limitations, we introduce **WorldArena 2.0**, which extends WorldArena along three coordinated dimensions: *modality*, *functionality*, and *platform*. First, WorldArena 2.0 expands benchmarking from visual to visuotactile world models. By developing a standardized framework built upon the UniVTAC simulator (Chen et al., 2026a), we enable visual world models to perceive and predict richer, multimodal sensory streams. This extension is critical for embodied manipulation, where tactile signals provide direct evidence of contact, force, slip, and material interaction that cannot be reliably inferred from visual observations alone. By incorporating simulated tactile prediction and a matched real-robot tactile ablation into benchmark evaluation, WorldArena 2.0 evaluates multimodal prediction alongside physical manipulation. Second, WorldArena 2.0 broadens functional evaluation by using world models as interactive RL training environments. Rather than merely evaluating frozen policies, our framework uses learned environments for policy optimization. This setting provides a more demanding test of world model utility: the model must remain stable and action-consistent over iterative interactions, produce reward-relevant state transitions, and support policy optimization without severe compounding errors. With support for diverse action-conditioned world models and modular reward designs, WorldArena 2.0 evaluates whether training in learned dynamics improves policy performance in the ground-truth simulator, rather than merely predicting plausible futures. Finally, along the platform dimension, WorldArena 2.0 moves beyond simulator-centric evaluation to a comprehensive suite of simulated and real-world testbeds. In addition to two multi-embodiment simulation environments, RoboTwin 2.0 (Chen et al., 2025) and LIBERO (Liu et al., 2023), we introduce physical robot evaluation on the AgileX Split-Type ALOHA platform. We benchmark physical manipulation across distinct operational protocols: pouring water and table wiping for cross-platform evaluation, chip grasping and table wiping for physical tactile ablation (Appendix B.4), and a pilot investigation on bottle lifting for policy transfer (Appendix C.10). Together, these extensions establish WorldArena 2.0 as a rigorous diagnostic benchmark capable of assessing world models under grounded perception and physical interaction constraints. Our main contributions are threefold:

- We introduce WorldArena 2.0, an expanded diagnostic benchmark systematically evaluating embodied world models across three dimensions: modality, functionality, and platform.
- We establish standardized evaluation protocols that incorporate visuotactile dynamics and evaluate world models as interactive RL training environments for policy optimization.
- We conduct extensive evaluations across simulated and physical robotic platforms, revealing that while learned models yield tangible policy-optimization gains in simulation, severe reality gaps remain in sustained contact-rich manipulation and physical transfer.

2 Related Work

2.1 Embodied World Models

Embodied world models (EWMs) aim to capture environment dynamics to support decision-making, planning, and counterfactual reasoning in embodied settings (Shang et al., 2026b; Long et al., 2025; Li et al., 2025b). Existing approaches fall into three broad categories. The first learns compact latent dynamics for control and reinforcement learning, exemplified by the Dreamer family (Hafner et al., 2020; 2025), which improves sample efficiency in simulation. The second leverages large-

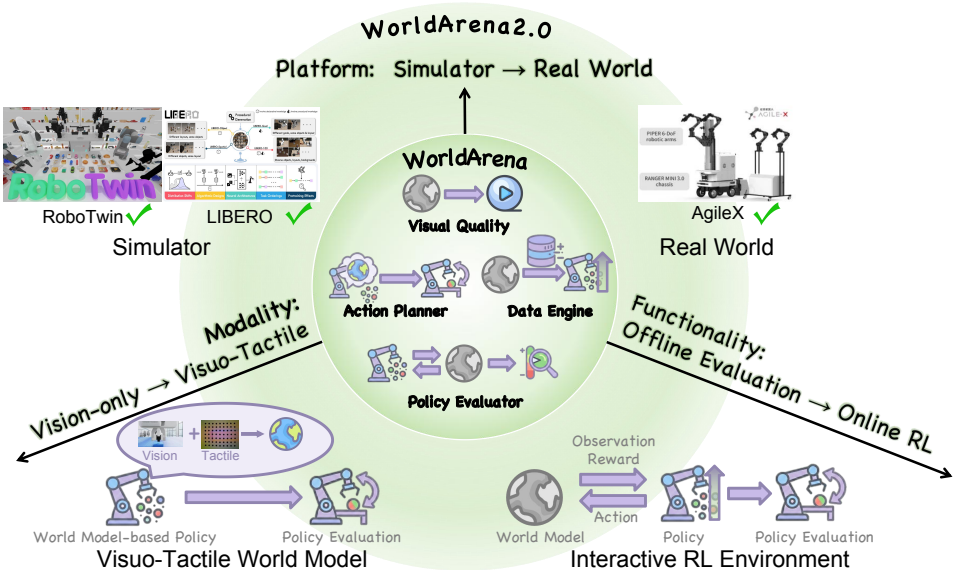


Figure 1: Overview of the extension from WorldArena to WorldArena 2.0 along three dimensions: modality, functionality, and platform.

scale generative models (Wan et al., 2025; Yang et al., 2025; Zhu et al., 2024b) to perform action-conditioned video prediction (Agarwal et al., 2025; Shang et al., 2025; Guo et al., 2025; Chen et al., 2026b), generating future visual observations conditioned on agent actions and language instructions via diffusion (Chi et al., 2025; Feng et al., 2025; Zhu et al., 2024a) or autoregressive architectures (Wu et al., 2024). The third focuses on enhancing physical and interaction fidelity through physics-informed objectives and joint visual-action modeling (Bi et al., 2025; Ye et al., 2026; Li et al., 2026; Yuan et al., 2026b). Together, these directions reflect a shift from low-dimensional dynamics to visually grounded, action-aware prediction.

Despite these advances, visual realism does not guarantee physical validity. Generative rollouts frequently violate basic physical constraints and suffer from error accumulation over extended horizons. This discrepancy demands evaluation paradigms that move beyond visual quality to assess multimodal dynamics grounding and closed-loop interactive decision-making across simulated and physical platforms.

2.2 Benchmarks for Embodied World Models

As generative models are increasingly applied in embodied domains, evaluating their effectiveness as world models remains challenging. Standard video generation metrics (Duan et al., 2025; Lu et al., 2025) emphasize perceptual quality but largely overlook physical realism and action-relevant fidelity. Recent benchmarks can be grouped into three types. The first type focuses on perceptual and spatiotemporal quality, e.g., EWMBench (Yue et al., 2025), which measures scene consistency, motion correctness, and semantic alignment. While informative for visual fidelity, such benchmarks do not assess whether generated dynamics support embodied decision-making. The second type evaluates functional utility, such as WorldSimBench (Qin et al., 2024), which translates generated videos into control signals via inverse dynamics, or WorldEval (Li et al., 2025c) and WoW-World-Eval (Fan et al., 2026), which test policy evaluation or action planning. The third type provides unified frameworks, exemplified by WorldArena (Shang et al., 2026a), which jointly measures perceptual quality and multiple functional aspects, including data engine, policy evaluation, and action planning.

Despite these advances, existing benchmarks have important limitations. They rely exclusively on visual inputs, ignoring tactile feedback that is critical for real-world physical interactions. Their functional evaluations are largely restricted to fixed policies or single-action planning and rarely test whether a world model can support continuous, interactive policy training without error accumulation. Furthermore, most assessments remain simulator-only, leaving the sim-to-real gap largely unexamined. WorldArena 2.0 addresses these gaps by incorporating visuotactile modalities, supporting policy

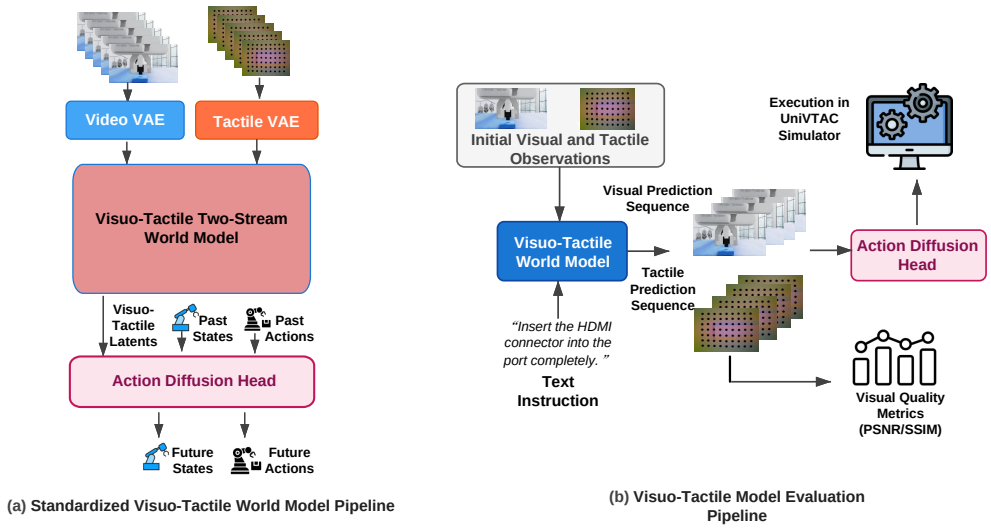


Figure 2: The standardized visuotactile world model architecture design (a) and the visuotactile model evaluation pipeline based on the UniVTAC simulator (b).

optimization in learned environments, and evaluating performance across both simulated and real robotic platforms, bringing benchmark evaluation closer to real-world embodied requirements.

3 WorldArena 2.0 Benchmark

3.1 From WorldArena to WorldArena 2.0

WorldArena (Shang et al., 2026a) pioneered the joint evaluation of perceptual fidelity (16 metrics across 6 dimensions) and offline downstream utility (data engine, policy evaluation, and action planning). However, its scope was restricted to vision-only simulation. WorldArena 2.0 expands this foundation along three coordinated dimensions: (1) **Modality**, incorporating visuotactile dynamics to resolve contact-rich physical interactions; (2) **Functionality**, evaluating world models as interactive RL training environments that support iterative policy optimization; and (3) **Platform**, bridging multi-embodiment simulation (RoboTwin 2.0, LIBERO) and physical execution on an AgileX dual-arm robot. This unified design transforms embodied world model benchmarking from passive video verification into an interactive, physically grounded diagnostic testbed.

3.2 Modality Extension: Evaluating Visuotactile World Models

While contemporary world models (Chi et al., 2025; Feng et al., 2025; Liao et al., 2025; Team et al., 2025) predominantly focus on visual perception, vision-only paradigms remain fundamentally insufficient for contact-rich manipulation tasks, where critical interaction dynamics are only partially observable through visual cues, and diverge substantially from human multimodal sensory experience. Recent efforts have constructed visuotactile world models through joint visuotactile modeling and tactile encoding injection, validating the effectiveness of incorporating tactile information in contact-rich scenarios (Yuan et al., 2026a; Higuera et al., 2026; Zheng et al., 2026). However, these works operate under heterogeneous setups with divergent tactile sensor configurations and inconsistent evaluation protocols. Consequently, the field still lacks a standardized evaluation pipeline for visuotactile world models, and existing benchmarks (Fan et al., 2026; Huang et al., 2024) provide no coverage of tactile modalities. To evaluate the capabilities of existing vision-based embodied world models under multimodal perceptual conditions, WorldArena extends the evaluation scope from vision-only settings to visuotactile world models.

To address the aforementioned challenges, we construct a standardized pipeline that upgrades vision-only world models to visuotactile world models. As illustrated in Figure 2, our pipeline builds

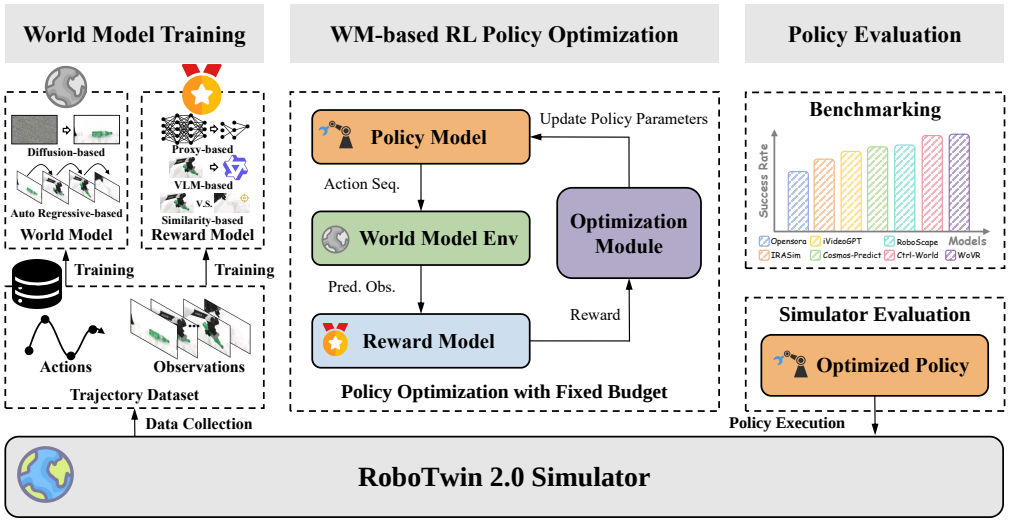


Figure 3: WorldArena 2.0 framework for leveraging world models as RL environments, comprising three stages: world model training, RL policy optimization using the world model, and policy evaluation.

upon an existing video world model, augmenting it with three primary modules: a tactile VAE, a visuotactile two-stream world model, and an action diffusion head. First, the tactile VAE encodes sequences of tactile deformation maps and aligns them into the video latent space of the original world model, enabling plug-in augmentation without architectural surgery. Second, the visuotactile two-stream world model performs synchronized denoising for both video prediction and tactile perception prediction, preserving modality-specific dynamics while enabling cross-modal coordination. Finally, for downstream task execution, the action diffusion head receives past states and actions, together with the predicted visuotactile latents, and performs denoising to directly infer future actions, closing the loop between perceptual prediction and functional manipulation. This standardized pipeline follows a modular design philosophy, augmenting existing video world models (Chi et al., 2025; Feng et al., 2025; Wan et al., 2025; Liao et al., 2025) with plug-in tactile components rather than altering their underlying architectures.

Since the action decoder is trained separately for each world model, downstream task success measures the utility of the complete world-model–decoder system (Appendix B.2). Crucially, our future-information ablation confirms that downstream manipulation is actively driven by predictive dynamics rather than history memorization: conditioning on ground-truth future observations yields 66% task success on Adjust Bottle, whereas removing future inputs (32%) or temporally shuffling future frames (33%) causes an immediate 50% performance drop (Appendix C.9). This verifies that the action decoder genuinely queries future predictive cues, ensuring that control performance reflects world-model dynamics fidelity. This extension is based on the UniVTAC simulator (Chen et al., 2026a). We leverage the official dataset provided by UniVTAC and construct a standardized evaluation pipeline for visuotactile world models within the simulated environment. These extended models are assessed across two complementary dimensions: (1) perceptual quality of tactile video prediction, and (2) functional success rate in downstream manipulation tasks. Through these standardized extensions, we broaden WorldArena's evaluated modality from vision-only perception to the visuotactile domain.

3.3 Functionality Extension: Evaluating World Models as RL Environments

World models have demonstrated strong capability in capturing environment dynamics from real-world data, but their evaluation has mostly been limited to offline prediction. As their fidelity improves, they can serve as interactive proxies for real environments, enabling reinforcement learning (RL) directly within the learned model. Recent studies have begun exploring the use of world models as virtual environments to support RL training (Xiao et al., 2025; Tang et al., 2025; Zhu et al., 2025; Wu et al., 2024). However, existing approaches often rely on system-level designs without

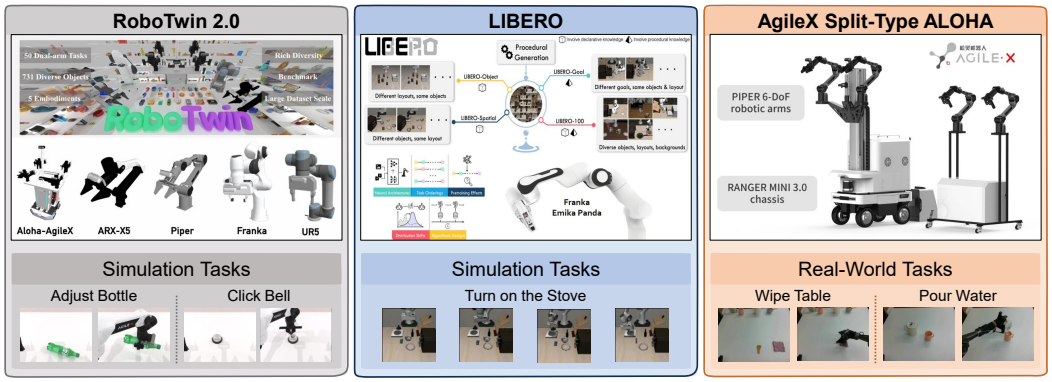


Figure 4: Overview of three test platforms in WorldArena 2.0: RoboTwin 2.0, LIBERO, and a real-world AgileX Split-Type ALOHA robotic platform.

standardized interfaces, making it difficult to isolate the contributions of individual components and compare their impact on final performance.

We propose a standardized framework (Figure 3) with three stages: world model training, policy optimization in the learned environment, and policy evaluation in the ground-truth simulator. The world model $\hat{P}_\phi$ predicts future observations, the policy π_θ selects actions, and the reward model $\hat{R}_\psi$ provides training signals. Let h_t denote the available interaction history, including the task instruction. Their recursive interaction is

$$a_t \sim \pi_\theta(\cdot \mid h_t), \quad \hat{o}_{t+1} \sim \hat{P}_\phi(\cdot \mid h_t, a_t), \quad \hat{r}_t = \hat{R}_\psi(h_t, a_t, \hat{o}_{t+1}). \tag{1}$$

Each component uses the context supported by its interface; this notation does not assume that a single observation is a Markov state. The partial-observability formulation and component-specific inputs are detailed in Appendix C.2.

We first adapt the world and reward models using simulator trajectories. With these models fixed, actor-only GRPO (Shao et al., 2024) updates the policy from group-relative advantages computed over rollouts collected in each learned environment. Finally, optimized policies execute in RoboTwin 2.0 under ground-truth dynamics, and its task-success predicate determines the reported success rates. This measures the utility of a world model as an RL training environment; physical transfer is evaluated separately. Implementation and evaluation protocols are provided in Appendix C.1.

3.4 Platform Extension: Cross-Embodiment Sim-to-Real Evaluation

Most existing benchmarks (Yue et al., 2025; Li et al., 2025c) for embodied world models predominantly evaluate prediction quality on a single simulator with a fixed task set and limited object variation. While this narrow scope simplifies evaluation, it makes results susceptible to overfitting, leading to artificially inflated rankings. Furthermore, conclusions drawn from a single platform are unstable by construction, as they provide no signal on whether a model's learned visual and physical priors can withstand the distribution shifts encountered on different robotic morphologies or real-world sensing conditions. To overcome these limitations, WorldArena 2.0 introduces a cross-embodiment evaluation protocol spanning three distinct platforms, RoboTwin 2.0, LIBERO, and an AgileX split-body ALOHA robot, forming complementary evaluation regimes. Rather than forcing identical, synthetic object configurations across all embodiments, we benchmark on each platform's representative native task distribution (Figure 4). This assesses whether a world model's physical and visual priors maintain consistent functional utility across diverse dynamics regimes rather than overfitting to a single task morphology.

To further assess the functional utility of world models across these platforms, we adopt the two evaluation protocols introduced in WorldArena: (1) **Embodied Data Engine**, where synthetic trajectories generated by a world model are used to train downstream policies, with success rate on the target task measuring the data quality; (2) **Embodied Action Planner**, where an action decoder converts world-model futures into action chunks for repeated planning, and task completion rate reflects its planning reliability. By running these protocols across RoboTwin, LIBERO, and the

Table 1: Visuotactile prediction and task success on UniVTAC. Trial counts and uncertainty are reported in Appendix B.5.

Model	Tactile Prediction Quality		Task Success Rate (%)		
	PSNR ↑	SSIM ↑	Insert HDMI	Lift Bottle	Avg.
ACT (Baseline)	–	–	20	**80**	**50**
Vidar	13.97	0.278	70	0	35
Genie Envisioner	13.36	0.456	0	0	0
Wan2.2	**21.26**	**0.746**	**100**	0	**50**

AgileX ALOHA platform, we systematically measure whether improvements in prediction fidelity translate into real-robot task competence.

4 Experiments

4.1 Visuotactile World Model Evaluation

We conduct a unified evaluation on the UniVTAC simulator across two contact-rich manipulation tasks, *Insert HDMI* and *Lift Bottle*, assessing vision-based embodied world models that have been extended through standardized tactile injection and jointly trained. We select representative vision-based embodied world models, including Vidar (Feng et al., 2025), Wan2.2 (Wan et al., 2025), Genie Envisioner (Liao et al., 2025), and ACT (Zhao et al., 2023) with a tactile tokenizer (Chen et al., 2026a), as our baseline. For tactile prediction quality, we adopt PSNR and SSIM (Wang et al., 2004) as evaluation metrics; for downstream action execution, we measure task success rate. Tactile prediction results and downstream task success rates are summarized in Table 1.

Wan 2.2 (Wan et al., 2025) achieves the highest tactile prediction scores and reaches 100% success on Insert HDMI, outperforming Vidar (70%) and Genie Envisioner (0%). However, all world-model-based policies fail on Lift Bottle, where the ACT baseline attains 80%. Notably, while Genie Envisioner achieves higher tactile SSIM (0.456) than Vidar (0.278), it yields 0% task success compared to Vidar's 70%. This empirical discrepancy demonstrates that standard frame-level reconstruction metrics (PSNR/SSIM) do not monotonically correlate with downstream control. Rather, success hinges on whether the model accurately preserves fine-grained physical dynamics (e.g., contact timing and force transitions), reinforcing the necessity of evaluating closed-loop functional utility alongside perceptual quality.

Real tactile evaluation. Augmenting FastWAM (Yuan et al., 2026b) with tactile sensing increases average physical success from 17.5% to 37.5% across Chip (Grasp chip) and Wiping (Wipe table), with the most pronounced gain on the contact-sensitive Chip task (15% $\rightarrow$ 45%; Table 4). These physical findings complement our simulation diagnosis, providing direct evidence that tactile feedback is indispensable for contact-rich manipulation. Full trial counts, Wilson confidence intervals, and experimental protocols are detailed in Appendices B.4 and B.5.

4.2 World Model Evaluation as RL Environments

We evaluate policies trained in seven world-model environments on two RoboTwin 2.0 tasks under three reward designs. Implementation and evaluation details are provided in Appendix C.1.

Overall performance. Policy optimization inside learned world-model environments consistently improves downstream simulator success over the supervised fine-tuning (SFT) baseline across all tested configurations, although a gap remains relative to policies trained directly in the native simulator (Table 2). To assess the maximal policy-improvement capacity supported by each learned simulator, reported simulator scores represent the optimization ceiling across the budget, evaluated on held-out RoboTwin validation configurations (Appendix C.3). WoVR (Jiang et al., 2026) leads the learned environments on Click Bell, whereas Ctrl-World (Guo et al., 2025) leads on Adjust Bottle, reaching 75.00% and 70.70% under proxy rewards compared to 43.75% and 55.08% for SFT. For the same RL-environment role, DreamX-Phi (DreamX Team et al., 2026) reaches 67.19% (Adjust Bottle task and proxy-based reward) success for a separate policy trained in its learned environment, which

Table 2: Policy success rates (%) on RoboTwin 2.0. **Panel A** reports common baselines: SFT uses no RL reward, while Simulator-based RL is trained with the native RoboTwin task reward. **Panel B** compares policies optimized in learned world-model environments under 3 reward designs. All policies are evaluated using the same RoboTwin task-success criterion.

(A) Common baselines

Method	Training Reward	Click Bell	Adjust Bottle
SFT	–	43.75	55.08
Simulator-based RL	Native reward	87.30	78.90

(B) RL policy optimization in interactive world-model environments

World Model	Proxy-based		VLM-based		Similarity-based	
	Click Bell	Adjust Bottle	Click Bell	Adjust Bottle	Click Bell	Adjust Bottle
OpenSora (Zheng et al., 2024b)	56.25	60.16	55.27	57.03	53.13	58.00
IRASim	53.13	61.33	53.52	58.98	50.78	59.38
iVideoGPT	52.53	56.25	48.44	58.59	52.15	60.93
Cosmos-Predict-2.5(action) (NVIDIA, 2025)	67.38	63.48	54.10	58.40	63.09	61.13
RoboScape	68.75	60.74	55.46	59.38	63.48	59.18
Ctrl-World	69.53	**70.70**	66.80	**65.04**	69.92	**66.02**
WoVR	**75.00**	67.19	**69.38**	64.45	**72.07**	61.35

numerically matches WoVR and lies between Cosmos-Predict-2.5(action) (63.48%) and Ctrl-World (70.70%) in Table 2. Given the substantial computational cost of iterative rollouts (profiling up to 28.8 minutes per iteration on 2×A800 GPUs; Appendix C.7), this comprehensive 42-run study operates under a standardized protocol with fixed iteration budgets. Notably, model rankings remain positively concordant across all three reward formulations (Spearman $\rho \in [0.706, 0.957]$ strictly within the seven learned models, Table 9; and $\rho \geq 0.845$ when pooled across all baselines, Table 8), verifying that relative model utility remains robust across diverse reward specifications.

Reward-model design. We compare a learned visual proxy, a zero-shot VLM evaluator, and goal-image similarity rewards. WoVR and Ctrl-World retain their respective task-leading positions across all three reward families. This agreement supports the leading-model comparison across reward choices, while detailed correlations and remaining reward–generator dependencies are analyzed in Appendix C.5.

Physical transfer. On the real AgileX robot (Lift Bottle), policies optimized inside WoVR and Ctrl-World achieve 25% and 30% multi-stage task progress, substantially outperforming the SFT baseline (10%; Appendix C.10). Extended analyses on rollout horizons, reward stability, and computational budgets are provided in Appendices C.6 and C.7.

4.3 Cross-platform Evaluation and Analysis

Cross-platform video quality evaluation. Following the evaluation protocol established in WorldArena (Shang et al., 2026a), we evaluate video quality across six perceptual dimensions with 16 standardized metrics on RoboTwin, LIBERO, and the real-world AgileX Split-Type ALOHA platform. Detailed results are reported in Tables 11, 12, 13, 14, 15, 16. A consistent pattern emerges across platforms. Commercial generative models such as Veo 3.1 and Wan 2.6 achieve state-of-the-art visual quality, yet they exhibit severe degradation on physical dynamics. On trajectory accuracy (Table 12), specialized embodied world models like CtrlWorld and IRASim outperform Veo 3.1 and Wan 2.6 by more than 3×. This stark divergence highlights that high aesthetic and semantic fidelity does not guarantee physically grounded dynamics, underscoring the critical need for embodied evaluation. Among embodied models, WoW and CtrlWorld consistently rank highly on physics adherence and content consistency.

Table 3: Task success rates (%) of world models as embodied data engines and action planners across three platforms. Real-world Task 1/2 are Pour Water/Wipe Table. Trial counts and uncertainty are reported in Appendix D.5.

Model	RoboTwin 2.0				LIBERO		Real-World			
	Data Engine		Action Planner		Data	Action	Data Engine		Action Planner	
	Task 1	Task 2	Task 1	Task 2	Engine	Planner	Task 1	Task 2	Task 1	Task 2
GigaWorld	2	13	6	19	0	0	0	0	0	0
Genie Envisioner	7	21	10	20	2	6	0	0	0	20
TesserAct (Zhen et al., 2025)	1	35	1	35	34	38	0	0	0	30
Vidar	13	53	2	19	22	14	40	0	30	10
Wan 2.2	15	41	12	20	10	24	10	0	10	0
CogVideoX	3	28	8	16	0	2	10	10	0	50

Cross-platform task success rate. Table 3 evaluates world models as data engines and action planners (Shang et al., 2026a) on RoboTwin, LIBERO, and the real-world AgileX platform. Several architectures achieve near-zero physical success, exposing a substantial gap relative to simulator performance. The extended real-robot suite likewise reveals uneven task competence, with precision-demanding cardboard-box folding remaining more difficult than wiping or garment folding (Appendix D.3). These observations motivate direct evaluation of physical task completion rather than inferring deployment capability from visual quality or simulator performance alone.

Cross-platform performance correlation. Several perceptual dimensions show cross-platform ranking agreement, while functional performance exhibits a less consistent simulator-to-real relationship (Figures 8–11). The observed RoboTwin-to-real task-success association is weak, and low physical success rates highlight the difficulty of transferring functional utility across platforms. These findings motivate direct physical evaluation alongside simulation; correlation estimates and small-sample interpretation are detailed in Appendix D.6.

5 Limitations

Our benchmark is explicitly designed as a broad diagnostic stress test rather than a demonstration of turn-key robotic deployment. To prioritize benchmarking breadth across 12 distinct models, 3 modalities, and multiple platforms under heavy compute constraints (Appendix C.7), our primary RL environment suite operates under standardized fixed-seed protocols per configuration. Furthermore, physical evaluations ($n = 10$ to 20 trials per condition) serve as diagnostic probes to quantify transfer gaps under hardware noise. These experiments establish a rigorous baseline for identifying systemic failure modes, while broad, closed-loop physical deployment remains an open challenge requiring future advances in contact dynamics (Appendix E).

6 Conclusion and Future Work

We introduce WorldArena 2.0, a benchmark built on WorldArena that extends embodied world model evaluation across modality, functionality, and platform. Visuotactile evaluation exposes task-dependent strengths and failures, learned-environment training improves simulator policy performance, and cross-platform comparisons reveal gaps in physical task competence. These complementary evaluations connect predictive quality with practical embodied utility. The current physical studies remain preliminary, motivating broader real-world and closed-loop evaluation in future versions.

AI Use Statement

Generative AI tools were used for language editing, LaTeX format adaptation, and consistency checks. They were not used to generate experimental measurements or to replace author verification of the analyses. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of this work.

REFERENCES

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.

AgileX Robotics. Aloha-agilex: Official code repository. `https://github.com/agilexrobotics/aloha-agilex`, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, et al. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026a.

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.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, et al. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026b.

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.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

DreamX Team, Rui Chen, Xiangxiang Chu, Geng Li, Jifan Li, Qingfeng Shi, Datao Tang, Jing Tang, Jun Wang, and Pengfei Zhang. DreamX-Phi 1.0: Action-conditioned video world model for robotic manipulation. *arXiv preprint arXiv:2608.13489*, 2026. doi: 10.48550/arXiv.2608.13489. URL `https://arxiv.org/abs/2608.13489`.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

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.

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. *arXiv preprint arXiv:2507.12898*, 2025.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026.

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*, pp. 21807–21818, 2024.

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

Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, et al. Wovr: World models as reliable simulators for post-training vla policies with rl. *arXiv preprint arXiv:2602.13977*, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025a.

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

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025b.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025c.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 44776–44791, 2023.

Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yiting Lu, Wei Luo, Peiyan Tu, Haoran Li, Hanxin Zhu, Zihao Yu, Xingrui Wang, Xinyi Chen, Xinge Peng, Xin Li, et al. 4dworldbench: A comprehensive evaluation framework for 3d/4d world generation models. *arXiv preprint arXiv:2511.19836*, 2025.

NVIDIA. Cosmos-predict2.5: Official model and code repository. `https://github.com/nvidia-cosmos/cosmos-predict2.5`, 2025.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

RoboMaster Team. Robomaster dataset. Hugging Face repository, https://huggingface.co/datasets/robomaster2025/RoboMaster, 2025.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

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. *AI Open*, 7:208–226, 2026a.

Yu Shang, Yinzhou Tang, Xin Zhang, Shengyuan Wang, Yuwei Yan, Honglin Zhang, Zhiheng Zheng, Jie Zhao, Jie Feng, Chen Gao, et al. A survey of embodied world models. 2026b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yinzhou Tang, Yu Shang, Yinuo Chen, Bingwen Wei, Xin Zhang, Shu'ang Yu, Liangzhi Shi, Chao Yu, Chen Gao, Wei Wu, et al. Roboscape-r: Unified reward-observation world models for generalizable robotics training via rl. *arXiv preprint arXiv:2512.03556*, 2025.

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.

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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

Xense Robotics. Xense robotics: High-resolution vision-based tactile sensing. https://www.xenserobotics.com/, 2026.

Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-env: Leveraging world model as a virtual environment for vla post-training. *arXiv preprint arXiv:2509.24948*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

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

Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, Quanlu Zhang, Yongji Wu, Chunyang Zhu, Junhao Hu, et al. {RLinf}: Flexible and efficient {Large-Scale} reinforcement learning via {Macro-to-Micro} flow transformation. In *20th USENIX Symposium on Operating Systems Design and Implementation (OSDI 26)*, pp. 829–846, 2026.

Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, Katherine Driggs-Campbell, and Ismini Lourentzou. Vtam: Video-tactile-action models for complex physical interaction beyond vlas. *arXiv preprint arXiv:2603.23481*, 2026a.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b.

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.

Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Peihong Wang, Liangzhi Shi, Yuqing Xie, et al. Rlinf-vla: A unified and efficient framework for reinforcement learning of vision-language-action models. *arXiv preprint arXiv:2510.06710*, 2025.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016. URL https://arxiv.org/abs/2304.13705.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 62557–62583, 2024a.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024b.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024a.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. *arXiv preprint arXiv:2511.09515*, 2025.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024b.

A Experimental Platforms and Tasks

A.1 RoboTwin 2.0

RoboTwin 2.0 (Chen et al., 2025) is a scalable bimanual simulation environment comprising 731 objects across 147 categories, featuring extensive domain randomization (e.g., scene clutter, lighting, textures, table height, and language instructions). We evaluate two tasks that benchmark the world models' ability to simulate rich interaction dynamics: *Adjust Bottle* and *Click Bell*. *Adjust Bottle* requires the robot to reorient and position a bottle, testing fine-grained pose adjustment and grasp stability prediction. *Click Bell* involves pressing a bell to produce an audible click, requiring the world model to accurately capture contact-rich dynamics and precise force application. We disable domain randomization during RL checkpoint evaluation and evaluate policies within the selected task distribution. The systematic randomization also exposes visual and spatial distribution shifts, making RoboTwin 2.0 a stress test for robustness beyond memorized backgrounds or object poses.

A.2 LIBERO

LIBERO (Liu et al., 2023) provides 130 language-conditioned single-arm manipulation tasks built upon Robosuite (Zhu et al., 2020), utilizing procedural generation to isolate spatial, object, goal, and mixed knowledge shifts. We select *Turn on the Stove* as our evaluation task. This task enables a fine-grained assessment of a world model's ability to capture both object-relational and articulated-body dynamics. We include LIBERO because it decouples different types of knowledge transfer, allowing us to diagnose precisely which aspects of environment dynamics a world model struggles to learn. This structured diagnostic capability is absent in purely randomization-based benchmarks and is critical for understanding a model's failure modes.

A.3 AgileX Split-Type ALOHA

AgileX Split-Type ALOHA (AgileX Robotics, 2025) serves as the real-world robot testbed. This platform features a master-follower teleoperation architecture built on the RANGER MINI 3.0 chassis, equipped with PiPER 6-DoF lightweight robotic arms. For the cross-platform evaluation in Sec 4.3, we use two real-world tasks, *Pour Water*, which tests the world model's ability to predict fluid dynamics and deformable material behaviors under gravity; and *Wipe Table*, which evaluates long-horizon contact sequences and friction modeling across extended surface trajectories, with ten trials per condition. Separate physical studies for the tactile and WM-RL extensions use their own tasks and protocols, described in Appendices B.4 and C.10, respectively. Our physical evaluation uses small object-reset displacements. Physical trials retain natural lighting and background variation, providing a direct check of sim-to-real performance. The ALOHA platform brings the evaluation to physical reality, where the world model must contend with unmodeled effects such as sensor noise, variable friction, and imperfect actuation.

A.4 Real-World Task Definitions

The Pour Water and Wipe Table tasks are illustrated in Figure 5.

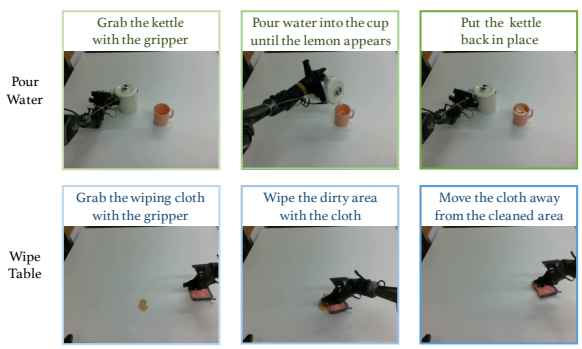


Figure 5: Illustration of the two real-world embodied tasks.

B Supplementary Details for the Modality Extension

B.1 Visuotactile Evaluation Setup

The evaluation uses the official UniVTAC dataset and the modular pipeline in Figure 2. A tactile VAE encodes tactile deformation sequences into the video latent space; a two-stream world model jointly predicts visual and tactile observations; an action diffusion head converts the predicted latents into actions. The simulated tasks are *Insert HDMI* and *Lift Bottle*.

B.2 Conditioning and Decoder Interface

The world model receives visual and tactile context together with task text. The action decoder receives past states and actions and predicted visuotactile latents. Decoders are trained separately for each world model under common training settings. Past actions supplied to the decoder do not imply that the generator is conditioned on the executed future action chunk. Downstream success therefore measures the complete world-model–decoder system.

B.3 Simulated Visuotactile Evaluation Protocol

Each visuotactile model is trained for 100 episodes. Tactile prediction quality is evaluated with PSNR and SSIM. Functional evaluation uses ten trials per model and task, with each trial capped at 200 environment steps. Table 1 reports the point estimates; confidence intervals are provided below.

As an additional temporal diagnostic, contact-timing coverage is 97.2% for Wan 2.2, 95.3% for Vidar, and 87.6% for Genie Envisioner. This alignment with task success demonstrates that contact-timing precision provides a critical dynamic diagnostic beyond static pixel fidelity, directly capturing the key phases of contact inception and release during manipulation.

B.4 Real-Robot Tactile Ablation

The comparison evaluates FastWAM (Yuan et al., 2026b) with and without the tactile branch while holding the remaining training and evaluation settings fixed. Each variant is trained for 200 episodes. Xense tactile sensing (Xense Robotics, 2026) is used for two tasks, Chip and Wiping, with 20 evaluation trials per task, action chunk 25, a maximum of 600 steps, and no human intervention. We evaluate tactile input through task-level success rates in this FastWAM ablation.

Table 4: Real-robot tactile evaluation. Success rates (%) for FastWAM with and without tactile input.

Method	Chip	Wiping	Mean
FastWAM	15	20	17.5
FastWAM+Tactile	45	30	37.5

FastWAM achieves 15% and 20% success on Chip and Wiping, respectively; adding tactile input yields 45% and 30% (Table 4).

B.5 Statistical Uncertainty

For binary task outcomes with recoverable integer counts, we calculate two-sided 95% Wilson score intervals (Wilson, 1927). With $\hat{p} = \frac{k}{n}$ and $z = 1.96$, the endpoints are

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}.$$

These intervals summarize trial-level uncertainty conditional on the evaluated policy and an independent-binomial approximation. They do not quantify seed variation, paired-method effects, or checkpoint-selection bias. The counts below are reconstructed from task success rates using 10 trials per UniVTAC condition and 20 per real tactile condition.

Table 5: 95% Wilson intervals (percent) for UniVTAC task success, with ten trials per task.

UniVTAC model ($n = 10$/task)	Insert HDMI	Lift Bottle
ACT	20 [5.7, 51.0]	80 [49.0, 94.3]
Vidar	70 [39.7, 89.2]	0 [0.0, 27.8]
Genie Envisioner	0 [0.0, 27.8]	0 [0.0, 27.8]
Wan 2.2	100 [72.2, 100.0]	0 [0.0, 27.8]

Table 6: Real tactile evaluation counts and 95% Wilson confidence intervals, corresponding to the point estimates in Table 4. The intervals describe each condition separately and do not constitute a test of the difference between conditions.

Method	Task	Successes/trials	Success (%; 95% CI)
FastWAM	Chip	3/20	15 [5.2, 36.0]
FastWAM	Wiping	4/20	20 [8.1, 41.6]
FastWAM+Tactile	Chip	9/20	45 [25.8, 65.8]
FastWAM+Tactile	Wiping	6/20	30 [14.5, 51.9]

C Supplementary Details for the Functionality Extension

C.1 WM-RL Pipeline and Training Data

To evaluate world models as RL training environments for embodied policies, we construct a unified end-to-end pipeline based on RLinf (Zang et al., 2025; Yu et al., 2026). We select task *Adjust Bottle* and *Click Bell* from RoboTwin 2.0 (Chen et al., 2025), with $\pi_{0.5}$ (Intelligence et al., 2025) as the base embodied policy. We collect 3,000 simulator trajectories containing state, action, and visual-observation sequences using two policies with different performance levels and the simulator's native action planner; the same dataset is used to adapt the world and reward models. We then initialize policy optimization by fine-tuning $\pi_{0.5}$ with SFT on 1,000 high-quality expert trajectories generated by the planner. During RL optimization, the initialized policy interacts with each learned environment, and GRPO (Shao et al., 2024) updates the policy under a fixed iteration budget. Finally, selected checkpoints are executed in the ground-truth RoboTwin 2.0 simulator, where the simulator task-success predicate determines task success.

C.2 Formalism and Component Interfaces

We model the underlying task as a finite-horizon POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, P, \Omega, R, \rho_0)$, where $s_t \in \mathcal{S}$ is hidden state, $o_t \sim \Omega(\cdot \mid s_t)$ is an observation, $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ is a transition, $r_t \sim R(\cdot \mid s_t, a_t)$ is the environment reward, and ρ_0 is the initial-state distribution. In the RL experiments, P and R denote dynamics and rewards. The video world model does not expose s_t; $\hat{P}_\phi$ is therefore a predictive kernel over future observations conditioned on the component-specific interaction history, rather than a parameterization of the latent-state kernel P. With instruction g and history $h_t = (g, o_0, a_0, \ldots, o_t)$, component context maps c_π, c_ϕ, c_ψ make the shorthand in Eq. 1 explicit:

$$a_t \sim \pi_\theta(\cdot \mid c_\pi(h_t)), \tag{2}$$

$$\hat{o}_{t+1} \sim \hat{P}_\phi(\cdot \mid c_\phi(h_t), a_t), \tag{3}$$

$$\hat{r}_t = \hat{R}_\psi(c_\psi(h_t, a_t, \hat{o}_{t+1})). \tag{4}$$

The history is updated by appending a_t and $\hat{o}_{t+1}$. Each context map selects the inputs its component actually supports; no full-history conditioning or Markov observation assumption is implied. The policy receives the head-camera observation, proprioceptive state, and task instruction. One model interaction uses an action chunk and predicts its corresponding future observations. Proxy and similarity rewards score a current image, whereas the VLM reward scores visual sequence context. For each sampled group of $G = 4$ rollouts with reward scores q_i, GRPO (Shao et al., 2024) forms group-relative advantages $\hat{A}_i = \frac{q_i - \bar{q}}{\mathrm{std}(q) + \epsilon}$ and uses them in the clipped policy objective, where ϵ is a numerical stabilizer. Optimization uses an actor-only loss without a critic model.

C.3 GRPO Optimization and Checkpoint Protocol

The policy is $\pi_{0.5}$ (3.3B), initialized by SFT on 1,000 planner-generated expert trajectories. Only the action expert (approximately 0.3B parameters) is updated; the VLM backbone remains frozen. To preserve the pre-trained temporal inductive bias of each foundation model, we adhere to each architecture's native action-chunk and temporal interface rather than enforcing an artificial uniform chunking that could degrade latent dynamics. Consequently, we standardize on a uniform optimization iteration budget to reflect practical model-based adaptation regimes (Appendix C.7).

Table 7: Shared actor-only GRPO settings for WM-based policy optimization. Rollout settings vary by world model.

Item	**Setting**
Optimization	GRPO, group size 4; token-mean loss aggregation
Clipping	Policy-ratio clip low/high $0.20/0.28$; dual clip $c = 3.0$
Regularization	KL coefficient 0; entropy bonus 0
Advantage	Group-relative normalization within groups of four
Optimizer	AdamW (Loshchilov & Hutter, 2019); weight decay 0.01; gradient clipping at 1.0
Action head	Flow matching (Lipman et al., 2022); 5 denoising steps; flow-SDE noise 0.3; 14-D absolute joint actions
Policy input	Head-camera observations, task instruction, and proprioceptive state
Checkpoint evaluation	Fixed-seed RoboTwin episodes; no domain randomization
Selection	Best success observed on these evaluation episodes
Training protocol & seed	Fixed iteration budget; standardized single initialization seed (App. C.7)

Following standard model-based RL evaluation conventions, checkpoints are periodically validated on fixed, held-out RoboTwin environments disjoint from the training trajectories. We report the peak policy performance achieved within the fixed training budget, establishing a standardized measure of the optimization upper bound enabled by each learned simulator.

C.4 Reward Designs

C.4.1 Proxy Reward

The proxy uses a ResNet-18 visual backbone (He et al., 2016), frozen T5-base cross-attention (Raffel et al., 2020), and an MLP head, trained on the same 3,000-trajectory simulator dataset used to adapt the world models.

C.4.2 VLM Reward

The VLM is Qwen3.5-9B (Qwen Team, 2026), applied zero-shot to visual sequence context and served through SGLang (Zheng et al., 2024a).

C.4.3 Similarity Reward

The similarity design uses LPIPS similarity (Zhang et al., 2018) between a predicted observation and a task-goal observation.

C.4.4 Simulator Baseline Reward

Simulator-based RL uses the native RoboTwin task reward. The proxy, VLM, and similarity comparisons apply to policy optimization in learned world-model environments; SFT uses no RL reward. All policies are evaluated with the simulator's task-success criterion.

The simulator's success predicate uses internal spatial states that are unavailable for a generated video. Consequently, simulator rewards cannot simply be attached to arbitrary model-generated transitions without a validated state-reconstruction mechanism. The simulator-trained baseline changes the training dynamics, so it is not an oracle-reward-only ablation with the generator held fixed. Ground-truth simulator success is instead the evaluation outcome for every trained policy.

C.5 Reward Sensitivity Analysis

The three learned reward designs differ in supervision and computational cost. Within each reward comparison, the same reward model is applied to all generators; this does not imply equal calibration error across their image distributions.

Table 8: Correlations between success rates under different reward designs. Pearson r measures linear agreement; Spearman ρ measures rank agreement. The full set contains nine methods (including SFT and simulator RL) on two tasks. For this pooled analysis, the common SFT and simulator-RL baseline scores are replicated across reward-design comparisons because these baselines do not depend on the learned reward family. These are descriptive summaries of success rates after training, not direct reward-calibration measurements.

Subset	Reward pair	r	ρ
All 18 points	Proxy–VLM	0.916	0.846
	Proxy–similarity	0.971	0.977
	VLM–similarity	0.952	0.845
Click Bell (9 methods)	Proxy–VLM	0.917	0.983
	Proxy–similarity	0.988	0.983
	VLM–similarity	0.952	0.967
Adjust Bottle (9 methods)	Proxy–VLM	0.940	0.850
	Proxy–similarity	0.915	0.900
	VLM–similarity	0.972	0.833

Pooling tasks can mix between-task and within-task differences, and including common SFT and high-performing simulator baselines can strengthen correlation. Task-specific coefficients therefore accompany the pooled analysis. Correlation does not demonstrate identical rankings, eliminate shared reward bias, or estimate uncertainty across independently trained policies. The similarity results are retained as sensitivity diagnostics only. The seven-world-model analysis in Table 9 provides the primary reward-sensitivity comparison because it excludes the replicated common baselines.

Table 9: Additional sensitivity analysis restricted to the seven world models (14 method–task points), excluding SFT and simulator RL. Computed from the displayed point estimates; it remains a pooled two-task descriptive analysis.

Reward pair	Pearson r	Spearman ρ
Proxy–VLM	0.768	0.706
Proxy–similarity	0.925	0.957
VLM–similarity	0.824	0.710

Even when isolating the seven world models and excluding the high-performing simulator baseline, the rankings maintain substantial rank concordance across different reward formulations (Table 9, Spearman ρ up to 0.957). This consistency verifies that the relative capability profile of each world model is robust to the choice of training supervision.

The correlations in Tables 8 and 9 are computed from the success rates in Table 2.

C.6 Rollout Horizon and Coverage

The relevant recursive horizon is the number of world-model generations, not just the number of low-level environment steps. For a configured step limit H and action chunk size K, the maximum number of model queries is $\lceil H/K \rceil$, with fewer queries when an episode terminates early. We match each policy action chunk to the corresponding world model's temporal interface; chunk sizes, step limits, and batch sizes therefore vary across models. Action chunking bounds recursive generation depth, where prediction errors can accumulate, but a configured query budget is not an empirical stability horizon. Reliability also depends on reward calibration and coverage of the states and actions visited during policy optimization.

We evaluate trained policies under ground-truth simulator dynamics using the simulator's task-success criterion. This separates evaluation from learned training rewards, although reward exploitation during training remains possible. Rollout stability, reward calibration, and on-policy coverage require further quantitative characterization.

C.7 Computational Cost

RL training uses two A800-80GB GPUs, one for the policy and one for the learned environment. Average per-GPU memory use is approximately 57 GB. Under the proxy reward, recorded per-iteration wall times range from 7.8 to 28.8 minutes across the profiled world models. These measurements characterize batched offline policy optimization rather than real-time robot control.

C.8 Policy Optimization Dynamics

Figure 6 illustrates the policy optimization trajectories on Click Bell across representative world-model environments. Across training steps, policies consistently exhibit stable reward progression, verifying that the learned dynamics reliably provide gradient signals for iterative policy improvement.

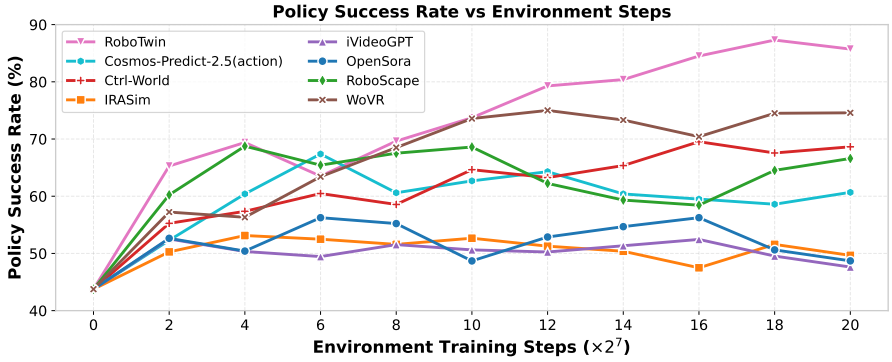


Figure 6: Checkpoint success rates on *Click Bell* under the proxy reward over training. Each curve shows one training seed.

C.9 Future-Information Ablation and Attribution

To test whether downstream action prediction genuinely uses the world-model future rather than relying only on history inputs, we perform an ablation on *Adjust Bottle* using the per-WM action decoder. Providing ground-truth future observations yields 66% success, compared with 32% without future inputs and 33% with shuffled futures. This confirms that temporally meaningful future information contributes substantially to downstream action prediction. This substantial performance gap (66% with ground-truth futures vs. 33% with shuffled futures) confirms the strong architectural sensitivity of the action decoder to temporal dynamics, ensuring that downstream execution effectively leverages the physical fidelity of predicted rollouts.

C.10 Real-Robot WM-RL Pilot

On AgileX ALOHA Lift Bottle, ten trials per policy produce progress summaries of 10% (SFT), 25% (WoVR-trained), and 30% (Ctrl-World-trained). To capture fine-grained physical execution beyond coarse binary success, the real-world transfer study evaluates multi-stage task progress on Lift Bottle (tracking reaching, grasping, and lifting phases). Policies trained in world models achieve notable progress gains over SFT (25% for WoVR and 30% for Ctrl-World vs. 10% for SFT), demonstrating meaningful transfer of learned interaction primitives to real hardware.

D SUPPLEMENTARY DETAILS FOR THE PLATFORM EXTENSION

D.1 DATA ENGINE AND ACTION PLANNER PROTOCOLS

Table 3 evaluates world models as data engines and action planners across RoboTwin 2.0, LIBERO, and AgileX ALOHA. The functional configurations use visual context and task text for future generation; the action decoder uses the generated future and its history inputs. Decoder parameters are trained separately per world model with common training settings. Task success measures the complete generator–decoder system.

D.2 PHYSICAL EVALUATION PROTOCOL

The Data Engine and Action Planner study uses ten trials per task/cell, chunk 45, at most 500 steps, and small object-reset displacements. The protocol permits only emergency-stop intervention. Failures include inaccurate grasping and incomplete stain removal. To ensure cross-hardware extensibility, the benchmark framework is built with modular abstraction layers, providing native interface support for additional robotic arms such as the Franka Research 3 (FR3) alongside the primary AgileX setup.

D.3 EXTENDED REAL-ROBOT TASKS AND QUALITATIVE OBSERVATIONS

Task coverage and capabilities. The extended suite includes six vision-only tasks: *Clean Table*, *Instruction-Following Cleanup*, *Fold Cardboard Box*, *Fold Shirt*, *Pour Water*, and *Wipe Table*. Three visuotactile tasks complement these activities: *Insert Two-Prong Plug*, *Peel Cucumber*, and *Pick Potato Chip*. Together, they probe instruction-conditioned object selection, deformable-object manipulation, sequential assembly, surface interaction, fragile-object handling, sustained tool contact, and precise insertion. Figure 7 illustrates the procedures using chronological demonstration frames, rather than model-evaluation rollouts. The original two-task illustration in Figure 5 remains associated with the controlled cross-platform study.

Execution interface. The interface separates a remotely hosted policy, a scheduling hub, and a robot-side worker. Policies receive camera observations, robot states, and task instructions, and return action chunks for execution. Visuotactile configurations additionally expose tactile images and force/torque observations. The interface supports AgileX dual-arm joint control and Franka single-arm end-effector control; this support does not imply that every system was evaluated on both embodiments. Episodes are subject to a step budget, and successful completion requires avoiding safety interventions.

Qualitative observations. Available evaluations indicate uneven physical competence across tasks, with precision-demanding cardboard-box folding remaining particularly difficult. Even systems that sometimes completed box folding performed less reliably on this task than on wiping or garment folding. Box folding requires ordered manipulation of multiple panels and accurate edge alignment, making it a useful probe of fine manipulation. However, aggregate outcomes alone cannot determine whether failures originate in perception, planning, or contact control. Because the evaluated systems differ in training and implementation, these observations characterize task difficulty rather than establish a controlled model ranking. Demonstration frames illustrate the required procedures and should not be interpreted as evidence of an evaluated model's success.

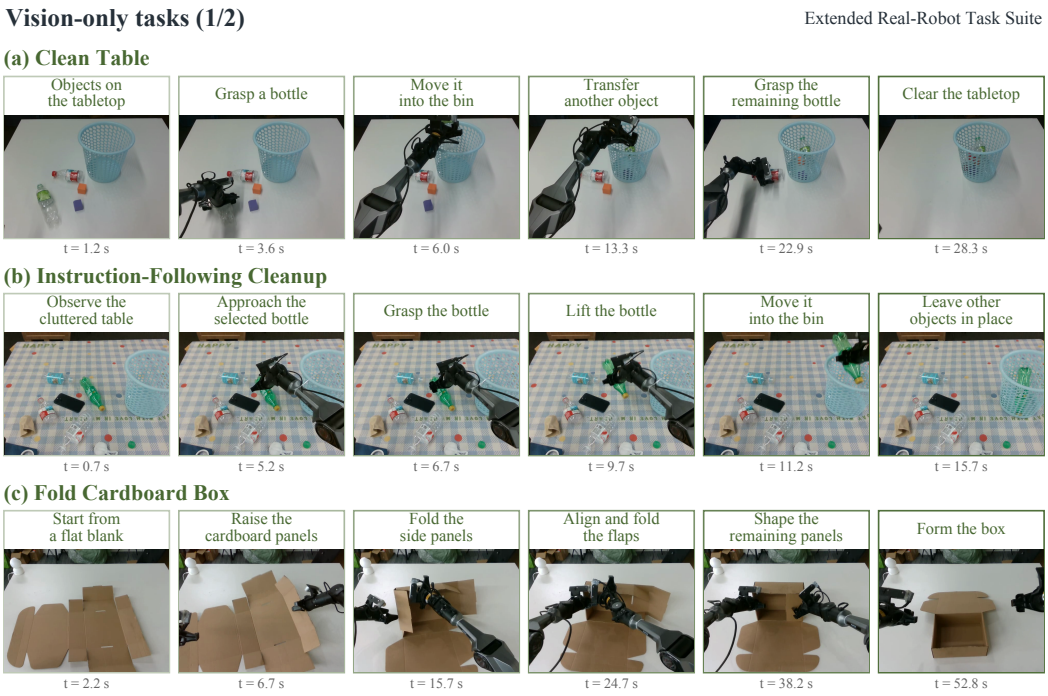


Figure 7: **Extended real-robot task suite (1/3): visual clearing and folding.** Tasks (a–c) show table clearing, instruction-following cleanup, and cardboard-box folding. Six chronological frames are shown per task. Timestamps indicate elapsed video time and are not aligned across tasks. Demonstration frames are selected and proportionally resized without generative editing; they illustrate procedures rather than model-evaluation outcomes. The remaining visual and visuotactile tasks appear on the following two pages.

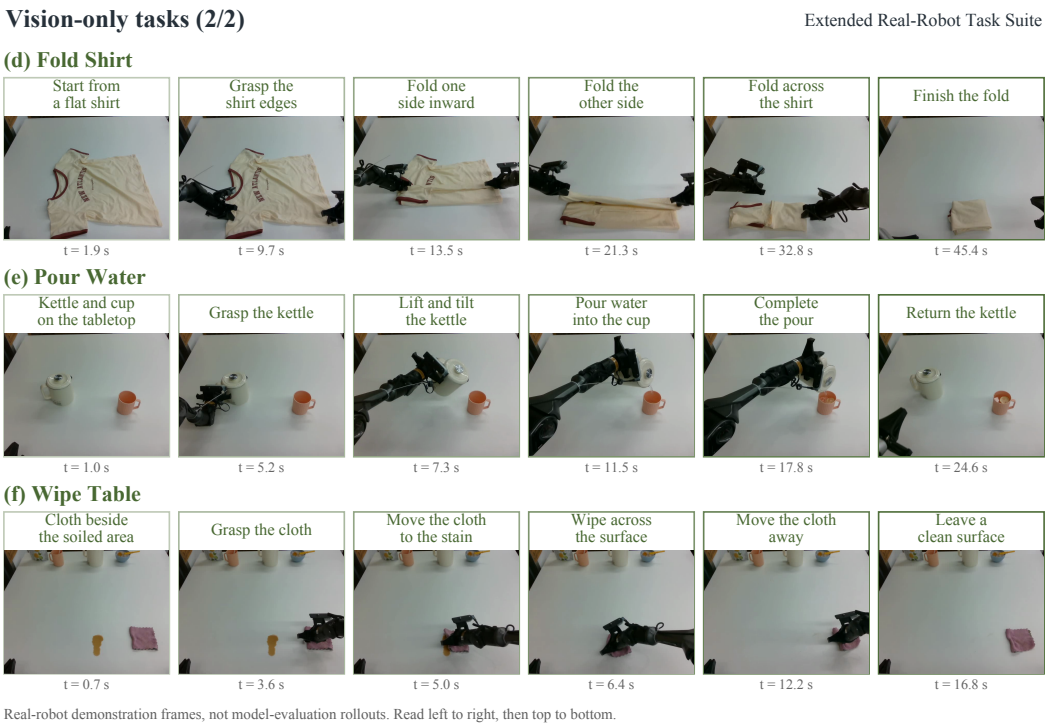


Figure 7: **Extended real-robot task suite (2/3): garment folding, pouring, and wiping.** Tasks (d–f) show shirt folding, water pouring, and table wiping, with six chronological frames per task, read left to right. All panels are vision-only task demonstrations, not outcomes attributed to a submitted model.

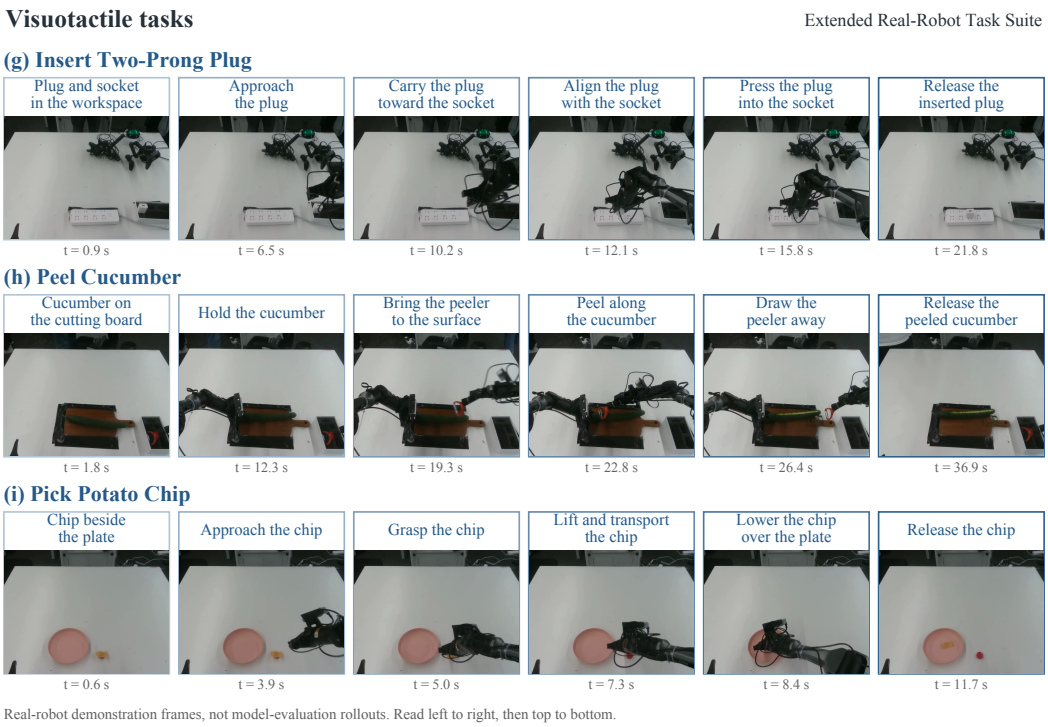


Figure 7: **Extended real-robot task suite (3/3): visuotactile manipulation.** Tasks (g–i) show plug insertion, cucumber peeling, and potato-chip transfer, with six chronological frames per task. The panels show the external RGB view of these visuotactile tasks, not tactile sensor images. These demonstrations illustrate task procedures and do not establish model success rates.

D.4 Physical Execution Cost

In the action-planner setting, one generation-and-decoding cycle takes approximately 3.5–21 seconds per replan. The real-robot Data Engine and Action Planner experiments use action chunks of 45, approximately 1.5 seconds of inference, a 2-second control cycle, and 22 GB of GPU memory on A800 or 4090 hardware. These physical-pipeline measurements are not directly comparable with the batched RL timings.

D.5 Functional Evaluation and Uncertainty

The physical cells in Table 3 use ten trials per condition. Table 10 maps the reported rates to 95% Wilson intervals using the formula in Appendix B.5. These intervals describe individual conditions and do not account for training variability or paired-method effects. Small samples and zero-success outcomes limit ranking precision.

Table 10: 95% Wilson intervals for the real-world success rates in Table 3, with ten trials per condition.

Real-world reported rate	Count	95% interval
0%	0/10	[0.0, 27.8]
10%	1/10	[1.8, 40.4]
20%	2/10	[5.7, 51.0]
30%	3/10	[10.8, 60.3]
40%	4/10	[16.8, 68.7]
50%	5/10	[23.7, 76.3]

D.6 Cross-Platform Functional Correlation

RoboTwin–LIBERO (Figure 8) shows a stronger positive association ($\rho = 0.771, p = 0.072$), whereas both simulator-to-real comparisons remain statistically uncertain (RoboTwin–Real: $0.348, 0.499$; LIBERO–Real: $0.522, 0.288$). The small model set and many zero-success physical evaluations make this estimate uncertain. We treat the correlation as descriptive and assess real-world utility through direct physical evaluation.

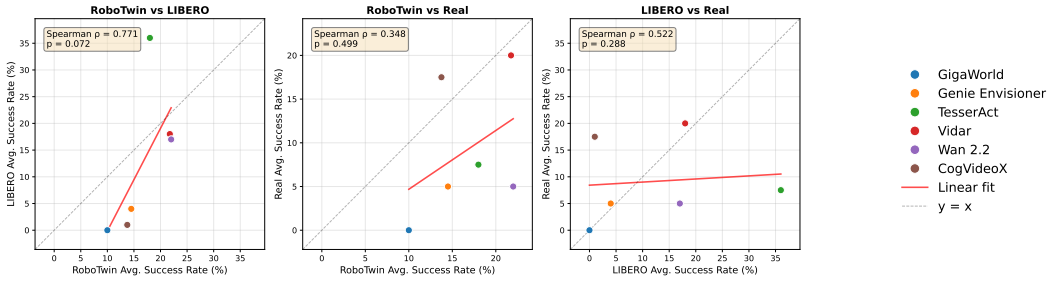


Figure 8: Cross-platform task success rate correlation between RoboTwin, LIBERO and the real-world robotic data.

D.7 Full Perceptual Evaluation Results

Perceptual tables (Table 11– 16) include both text- and action-conditioned configurations, explicitly including both Cosmos variants. Their output-only metrics describe the evaluated systems, but trajectory matching and controllability across heterogeneous conditioning are not controlled comparisons of transition fidelity. Accordingly, dynamics-quality comparisons should remain stratified by interface; functional success measures the complete generator–decoder system.

D.8 Cross-Platform Perceptual Correlations

Perceptual ranking agreement varies by dimension (Figures 9–11). Motion quality, physics adherence, and 3D accuracy provide complementary diagnostics; their ability to predict physical task success requires further validation.

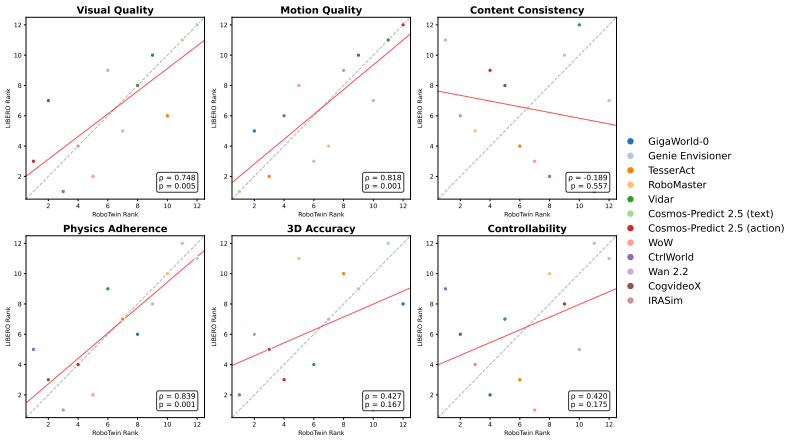


Figure 9: Cross-platform video quality correlation between RoboTwin and LIBERO.

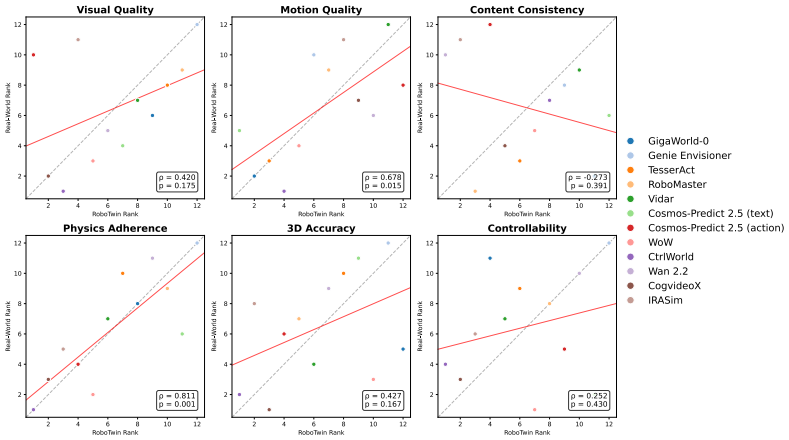


Figure 10: Cross-platform video quality correlation between RoboTwin and the real-world robotic data.

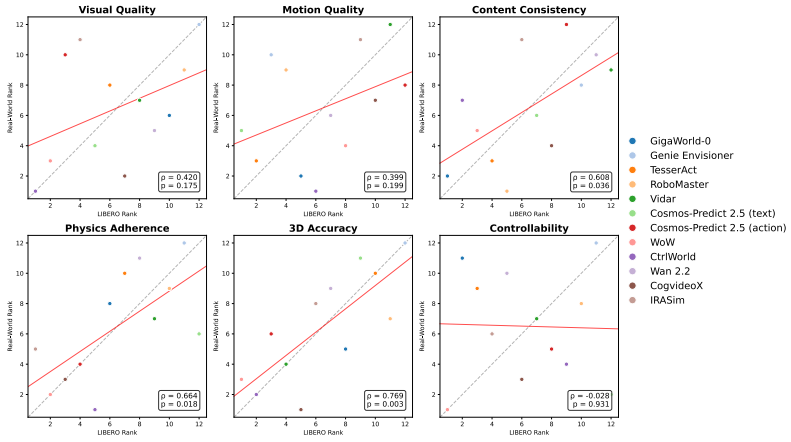


Figure 11: Cross-platform video quality correlation between LIBERO and the real-world robotic data.

E Extended Limitations and Responsible Use

E.1 Statistical Limitations

Our RL benchmarking protocol balances extensive model coverage against foundation-scale compute demands (Appendix C.7), utilizing fixed evaluation environments to benchmark optimization ceilings. While scaling up seed replicates and physical trials will provide narrower estimation bounds, the current multi-platform protocol successfully exposes systematic performance discrepancies across embodiments and reward specifications.

E.2 Modeling and Attribution

For decoder-based functional protocols, success reflects the complete generator–decoder system; WM-RL results similarly reflect the coupled learned-environment–policy pipeline. Reward sensitivity and rollout reliability are discussed in Appendices C.5 and C.6. Generalization beyond the evaluated tasks, scenes, and embodiments remains to be tested.

E.3 Computational Limitations

Computational cost varies substantially across world models and reward implementations (Appendix C.7). The learned environments support offline policy optimization; derived replanning frequencies must be distinguished from physical low-level control rates. Open-loop action chunks constrain rapid contact feedback.

E.4 Physical Safety and Supervision

The non-tactile physical protocol permits only emergency-stop intervention, and the tactile pilot uses no human intervention. Physical deployment should retain human supervision and emergency-stop access and report all interventions. Because imperfect dynamics and rewards can induce unsafe actions, simulator gains and perceptual scores do not justify unsupervised deployment.

F Broader Impact

WorldArena 2.0 evaluates embodied world models across modalities, functionalities, and platforms, exposing failures that can guide robust agents. Errors in learned dynamics can transfer to trained policies. Physical deployment therefore requires validation, intervention and safety-stop reporting, and human supervision.

Table 11: Video quality evaluation results across visual quality, motion quality and content consistency dimensions on Robotwin dataset.

Models	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Backgrd. Consistency	Photom. Consistency
GigaWorld-0	0.5041	0.3991	0.4413	0.6709	0.3118	0.7811	0.7303	0.8563	0.1756
Genie Envisioner	0.2305	0.3289	0.3340	0.6930	0.0855	0.6966	0.7760	0.9024	0.2006
TesserAct	0.3322	0.4590	0.4579	0.5150	0.2447	0.7579	0.8250	**0.9238**	0.2491
RoboMaster (RoboMaster Team, 2025)	0.3487	0.3842	0.2966	0.6124	0.1484	0.6940	0.8295	0.9123	0.3356
Vidar	0.4145	0.4068	0.5608	0.2767	0.1426	0.7973	0.7629	0.8300	0.2350
Cosmos-Predict 2.5 (text) (NVIDIA, 2025)	0.6668	0.4501	0.3126	0.5911	0.4302	0.7882	0.7488	0.8511	0.1383
Cosmos-Predict 2.5 (action)	0.4489	0.3576	0.9296	0.3994	0.0573	0.7100	0.8197	0.8894	0.3528
WoW	0.4587	0.3868	0.7440	0.4608	0.2706	0.7692	0.8161	0.9025	0.2170
CtrlWorld	0.3522	0.3893	0.9185	0.4257	0.3449	0.7377	**0.8411**	0.9057	0.1729
Wan 2.2	0.3884	0.3963	0.7575	0.4349	0.1269	0.7019	0.8388	0.9042	**0.4776**
CogvideoX	0.3582	0.3777	**0.9384**	0.3166	0.2189	0.7391	0.8083	0.8773	0.3580
IRASim	0.3489	0.3623	0.9330	0.4139	0.2083	0.7052	0.8312	0.9068	0.3522
Veo 3.1	0.6605	**0.4632**	0.5694	0.5450	0.1396	0.6989	0.7878	0.8710	0.3247
Wan 2.6	**0.6824**	0.4433	0.7229	**0.7421**	**0.4532**	**0.8539**	0.7517	0.8687	0.1904

Table 12: Video quality evaluation results across physics adherence, 3D accuracy and controllability dimensions on Robotwin dataset.

Models	Physics Adherence		3D Accuracy		Controllability		
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	Action Response Sensitivity
GigaWorld-0	0.5368	0.1552	0.6316	0.7596	0.6156	0.8591	0.1134
Genie Envisioner	0.2052	0.0679	0.8663	0.5284	0.2028	0.8544	0.0109
TesserAct	0.5800	0.1396	0.7159	0.7920	0.6152	0.8783	0.0311
RoboMaster	0.5364	0.1158	0.8335	0.7588	0.5772	0.8761	0.0352
Vidar	0.5348	0.1928	0.7872	0.7592	0.5912	0.8826	0.0819
Cosmos-Predict 2.5 (text)	0.3872	0.0816	0.7051	0.7964	0.2664	0.7733	**0.1418**
Cosmos-Predict 2.5(action)	0.5500	0.2945	0.8862	0.7644	0.5840	0.8879	0.0133
WoW	0.5564	0.2058	0.7283	0.7672	0.5692	0.8842	0.0434
CtrlWorld	0.6212	**0.4766**	0.9300	0.7960	0.7272	0.8912	0.0210
Wan 2.2	0.5184	0.1627	0.7768	0.7660	0.5376	0.8877	0.0512
CogvideoX	0.5940	0.3526	0.9097	0.7828	0.7268	**0.8977**	0.0076
IRASim	0.5656	0.3639	**0.9312**	0.7788	0.6604	0.8849	0.0526
Veo 3.1	**0.7872**	0.1231	0.7421	**0.8276**	**0.9328**	0.8607	0.0852
Wan 2.6	0.7280	0.1182	0.7144	0.8032	0.8536	0.8728	0.0992

Table 13: Video quality evaluation results across visual quality, motion quality and content consistency dimensions on Libero dataset.

Models	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Backgrd. Consistency	Photom. Consistency
GigaWorld-0	0.3056	0.4860	0.1693	0.2387	0.0814	0.5886	**0.9082**	**0.9481**	0.7859
Genie Envisioner	0.3688	0.2518	0.0303	0.3330	0.2273	0.5776	0.6867	0.8670	0.2074
TesserAct	0.3779	**0.5249**	0.3524	**0.5060**	0.1433	0.6917	0.8659	0.9023	0.3294
RoboMaster	0.3530	0.3883	0.0468	0.3729	0.1263	0.5780	0.9080	0.9347	0.1964
Vidar	0.3071	0.4881	0.3725	0.0986	0.0342	0.5026	0.3607	0.3705	0.4354
Cosmos-Predict 2.5 (text)	**0.7406**	0.5014	0.0205	0.3174	**0.6573**	**0.9590**	0.6578	0.8210	0.3818
Cosmos-Predict 2.5 (action)	0.3642	0.5064	0.4205	0.1150	0.0374	0.4465	0.5550	0.5742	0.6496
WoW	0.3573	0.5061	0.4612	0.1972	0.0488	0.4773	0.7843	0.8184	**0.8629**
CtrlWorld	0.3530	0.4996	**0.6087**	0.2358	0.0698	0.5467	0.8396	0.8697	0.8227
Wan 2.2	0.3197	0.4753	0.2919	0.1341	0.0525	0.5597	0.3954	0.4253	0.4681
CogvideoX	0.3044	0.4614	0.4513	0.1334	0.0429	0.4793	0.5852	0.6036	0.6631
IRASim	0.3484	0.4843	0.4407	0.1364	0.0442	0.5012	0.5633	0.5941	0.7352

Table 14: Video quality evaluation results across physics adherence, 3D accuracy and controllability dimensions on Libero dataset.

Models	Physics Adherence		3D Accuracy		Controllability		
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	Action Response Sensitivity
GigaWorld-0	0.6225	0.3690	0.8272	0.8090	0.7281	0.9102	0.0100
Genie Envisioner	0.2000	0.1495	0.7520	0.4920	0.2000	0.8758	0.0187
TesserAct	0.6000	0.3381	0.7039	0.7960	0.6380	0.9029	0.0185
RoboMaster	0.4060	0.1984	0.7918	0.6500	0.3580	0.9021	0.0363
Vidar	0.4100	0.2810	0.9133	0.7940	0.3820	0.9179	0.0389
Cosmos-Predict 2.5 (text)	0.2340	0.0015	0.5853	**0.9400**	0.2000	0.7637	**0.1226**
Cosmos-Predict 2.5 (action)	0.5600	0.4552	0.9236	0.8020	0.4200	0.9131	0.0017
WoW	**0.6500**	0.5859	0.9381	0.8060	**0.7900**	0.9147	0.0121
CtrlWorld	0.5160	0.4824	**0.9431**	0.7920	0.4040	0.9212	0.0087
Wan 2.2	0.5360	0.2467	0.8901	0.7960	0.5180	0.9102	0.0364
CogvideoX	0.5200	0.6676	0.9267	0.7800	0.4820	0.9182	0.0177
IRASim	0.6200	**0.8427**	0.9184	0.7800	0.5620	**0.9240**	0.0071

Table 15: Video quality evaluation results across visual quality, motion quality and content consistency dimensions on the real-world robot.

Models	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Backgrd. Consistency	Photom. Consistency
GigaWorld-0	0.5976	0.4031	0.1592	0.4685	0.2706	0.6994	0.7667	0.9002	0.3251
Genie Envisioner	0.3302	0.2428	0.0041	0.2440	0.2031	0.4390	0.6691	0.8569	0.0776
TesserAct	0.6045	**0.4721**	0.0236	0.5263	0.2032	0.6883	0.7932	0.9070	0.2760
RoboMaster	0.6050	0.4258	0.0249	0.3556	0.0938	0.5299	**0.8137**	**0.9152**	0.3593
Vidar	0.6001	0.4570	0.0887	0.1510	0.0486	0.5224	0.4906	0.5472	**0.4149**
Cosmos-Predict 2.5 (text)	**0.6903**	0.4706	0.0473	0.3354	0.1844	0.6576	0.7320	0.8589	0.3299
Cosmos-Predict 2.5 (action)	0.5819	0.3266	0.1136	0.2738	0.1422	0.5700	0.4035	0.4715	0.1071
WoW	0.6733	0.4147	0.1839	0.4010	0.1535	0.6690	0.7780	0.9113	0.2692
CtrlWorld	0.6153	0.3540	**0.6621**	**0.5649**	**0.4963**	**0.7456**	0.7959	0.9093	0.0612
Wan 2.2	0.6267	0.4552	0.1047	0.2979	0.1935	0.6434	0.4964	0.5871	0.2339
CogvideoX	0.6145	0.3872	0.4909	0.3074	0.1202	0.5928	0.7768	0.8884	0.2940
IRASim	0.5175	0.3327	0.1056	0.1845	0.0914	0.5265	0.4458	0.5064	0.1858

Table 16: Video quality evaluation results across physics adherence, 3D accuracy and controllability dimensions on the real-world robot.

Models	Physics Adherence		3D Accuracy		Controllability		
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	Action Response Sensitivity
GigaWorld-0	0.7450	0.0815	0.9258	0.9238	0.4012	0.7816	0.0277
Genie Envisioner	0.2275	0.0039	0.6687	0.4838	0.2350	0.8209	0.0212
TesserAct	0.6025	0.0850	0.8300	0.8562	0.5175	0.8536	0.0247
RoboMaster	0.6812	0.0585	0.9164	0.8863	0.6987	0.8779	0.0343
Vidar	0.7162	0.1734	0.9468	0.9150	0.7100	0.9217	0.0399
Cosmos-Predict 2.5 (text)	**0.8600**	0.0445	0.5471	0.9450	0.9300	0.8850	**0.0574**
Cosmos-Predict 2.5 (action)	0.7200	0.2592	0.9703	0.8762	0.8012	0.8978	0.0088
WoW	0.8100	0.3603	0.9448	**0.9650**	**0.9887**	**0.9268**	0.0399
CtrlWorld	0.7387	**0.6865**	**0.9888**	0.9275	0.8838	0.8325	0.0059
Wan 2.2	0.5662	0.1164	0.8381	0.8637	0.4237	0.9092	0.0548
CogvideoX	0.7050	0.3107	0.9780	0.9425	0.8162	0.9081	0.0176
IRASim	0.6837	0.2621	0.9545	0.8475	0.7963	0.8873	0.0140