
WorldArena 2.0: Extending Embodied World Model Benchmarking on Modality, Functionality and Platform

Yu Shang^{1*†} Yinzhou Tang^{1*} Yiding Ma^{1*} Zhuohang Li^{1*} Lei Jin^{1*}
 Weikang Su^{1*} Xin Jin Zhaolu Wang Ziyu Wang¹ Xin Zhang¹
 Haisheng Su² Weizhen He³ Wei Wu¹ Haoyi Duan⁴ Gordon Wetzstein⁴
 Xihui Liu⁵ Dhruv Shah⁶ Zhaoxiang Zhang⁷ Zhibo Chen⁸ Jun Zhu¹
 Yonghong Tian⁹ Tat-Seng Chua¹⁰ Wenwu Zhu¹ Chen Gao¹ Yong Li^{1‡}

¹Tsinghua University ²Shanghai Jiao Tong University ³Zhejiang University
⁴Stanford University ⁵The University of Hong Kong ⁶Princeton University
⁷Chinese Academy of Sciences ⁸University of Science and Technology of China
⁹Peking University ¹⁰National University of Singapore

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 benchmark is available at: <https://world-arena.ai>.

1 Introduction

World models [1, 2, 3] 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

*Equal contribution.

†Project lead.

‡Corresponding author: liyong07@tsinghua.edu.cn

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 [4, 5, 6]. 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 [7], WorldEval [8], World-in-World [9], WoW-World-Eval [9] and WorldArena [10]. Among these, WorldArena [10] 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 a stronger connection between predictive accuracy and task-level effectiveness, making it a valuable foundation for evaluating world models in embodied intelligence.

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 [11] or static policy evaluation [8], 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 [12], 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 tactile prediction into benchmark evaluation, WorldArena 2.0 assesses whether world models can capture physical interaction dynamics that are more closely aligned with real robotic systems. Second, WorldArena 2.0 broadens functional evaluation by utilizing world models as online interactive RL environments. Rather than merely evaluating frozen policies, our framework uses world models as interactive proxies to train embodied agents. 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 directly evaluates whether learned dynamics can improve agent behavior, rather than merely predict 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 simulation environments, RoboTwin and LIBERO, we introduce real-world evaluation on the AgileX Split-Type ALOHA platform. We evaluate two real-world tasks: pour water and wipe the table, which test the complex manipulation abilities of embodied world models. Together, these extensions establish WorldArena 2.0 as a rigorous benchmark capable of assessing world models under increasingly realistic perception and deployment conditions. Our main contributions are threefold:

- We introduce WorldArena 2.0, a comprehensively upgraded benchmark that expands embodied world model evaluation across three dimensions: modality, functionality, and platform.
- We pioneer standardized evaluation protocols that incorporate visuotactile modality and evaluate world models as online interactive environments for reinforcement learning.
- We conduct extensive experiments on 12 embodied world models across simulated and real-world robotic platforms, revealing consistent trends between simulation and reality while highlighting a substantial sim-to-real usability gap.

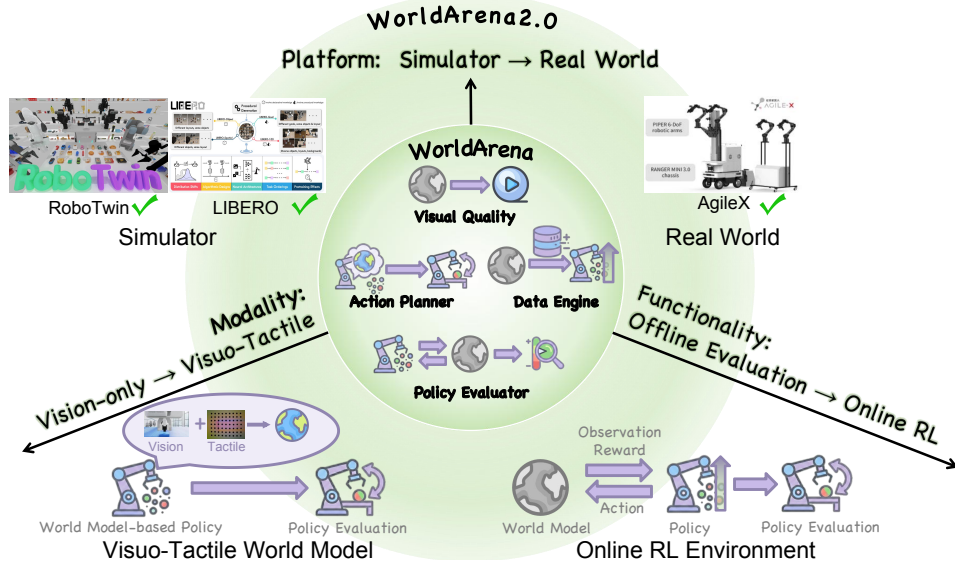


Figure 1: Overview of the extension from WorldArena to WorldArena 2.0 along three dimensions: modality, functionality, and platform.

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 [2, 3, 13]. Existing approaches fall into three broad categories. The first learns compact latent dynamics for control and reinforcement learning, exemplified by the Dreamer family [14, 15], which improves sample efficiency in simulation. The second leverages large-scale generative models [16, 17, 18] to perform action-conditioned video prediction [19, 20, 21, 22], generating future visual observations conditioned on agent actions and language instructions via diffusion [23, 24, 25] or autoregressive architectures [26]. The third focuses on enhancing physical and interaction fidelity through physics-informed objectives and joint visual-action modeling [27, 28, 29, 30]. 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. EWMs can produce plausible-looking rollouts yet violate basic physical rules and accumulate errors over long horizons. Limited human demonstration data further reduces robustness under closed-loop sequential interactions. Thus, evaluating EWMs requires more than generative quality: models should be assessed on their ability to capture multimodal dynamics that are physically grounded and useful for embodied decision-making. This motivates benchmarks incorporating richer sensory modalities, interactive and closed-loop evaluation, and real robotic platforms alongside simulation.

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 [5, 31] 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., EWM-Bench [6], 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 [7], which translates generated videos into control signals via inverse dynamics, or WorldEval [8] and WoW-World-Eval [9], which test policy evaluation or action planning. The third type provides unified frameworks, exemplified by WorldArena [10], which jointly measures perceptual quality and multiple functional aspects, including data engine, policy evaluation, and action planning.

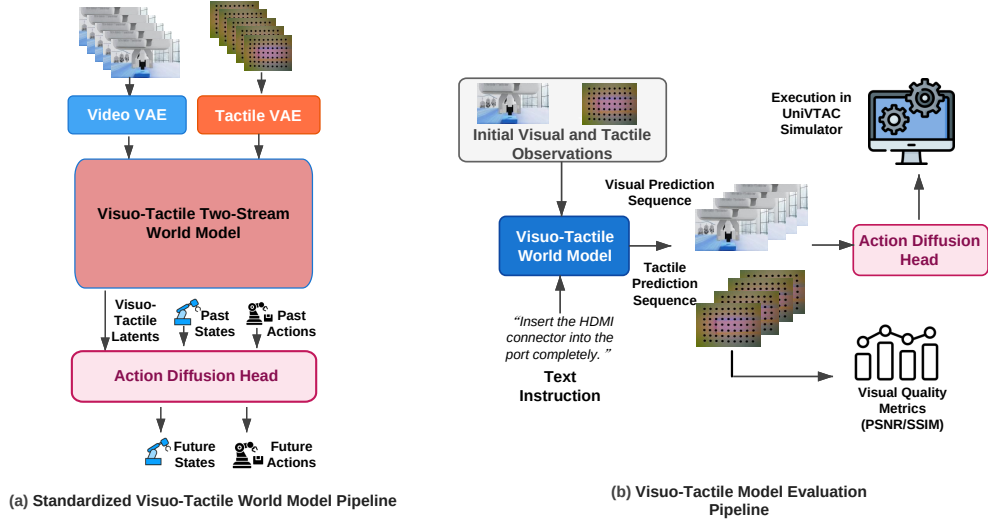


Figure 2: The standardized visuotactile world model architecture design (a) and the visuotactile model evaluation pipeline based on the UniVTAC simulator (b).

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, enabling interactive reinforcement learning, 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 [10] established a unified benchmark for embodied world models by jointly evaluating *perceptual quality* and *functional utility*. Specifically, it assessed open-loop visual prediction through 16 metrics spanning six dimensions, including visual quality, motion quality, content consistency, physics adherence, 3D accuracy, and controllability. It further evaluated world models in three embodied roles: as data engines for synthetic data generation, as policy evaluators for proxy-based policy assessment, and as action planners for closed-loop task execution. WorldArena provided an important first step toward benchmark standardization, moving evaluation beyond video realism alone toward functional utility. However, its scope remained centered on visual simulation and a restricted set of downstream uses, including data engine, policy evaluation and action planning.

WorldArena 2.0 builds on this foundation and extends WorldArena along three coordinated axes. First, along the *modality* axis, it moves from vision-only to visuotactile modalities, enabling assessment of whether learned dynamics capture contact-aware physical interactions that are essential for real-world manipulation. Second, along the *functionality* axis, it goes beyond offline downstream uses and evaluates world models as interactive environments for online reinforcement learning, testing whether imagined rollouts can support policy improvement. Third, along the *platform* axis, it extends evaluation from simulator-only settings to cross-embodiment sim-to-real testbeds, allowing benchmark results to reflect comprehensive deployment conditions. Together, these extensions transform WorldArena from a unified simulation benchmark into a more realistic and deployment-oriented evaluation protocol for embodied world models.

3.2 Modality Extension: Evaluating Visuotactile World Models

While contemporary world models [23, 24, 32, 33] 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 [34, 35, 36]. 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 [9, 37] 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 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 [23, 24, 16, 32] with plug-in tactile components rather than altering their underlying architectures. We emphasize that our evaluation scope centers exclusively on world models, which predict future video observations, as distinct from world action models [27, 38] that directly predict future actions. This extension is based on UniVTAC [12] simulator. 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 [39, 40, 41, 26]. However, existing approaches often rely on system-level designs without standardized interfaces, making it difficult to isolate the contributions of individual components and compare their impact on final performance.

Motivated by this, we propose a standardized framework (Figure 3) that integrates world models into a closed-loop RL pipeline for policy training and evaluation. We first formalize real-world interactions as a partially observable Markov decision process $\mathcal{M} = (\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho_0)$, where $\mathcal{O}$ is the observation space, $\mathcal{A}$ the action space, $\mathcal{P}(o_{t+1}|o_t, a_t)$ the true environment transition kernel, $\mathcal{R}(o_t, a_t)$ the real-world reward function, $\gamma \in [0, 1)$ the discount factor, and ρ_0 the initial observation distribution. With their increasing modeling capacity, world models can approximate $\mathcal{P}$ using a parameterized model $\hat{\mathcal{P}}^\phi$, effectively acting as a proxy for the real environment. The framework comprises four core components—world model environment, reward model, policy model, and optimization module—which together enable closed-loop RL training and systematic evaluation of policy performance.

World Model Environment: Parameterized by ϕ , it approximates the real-world transition dynamics as a conditional distribution $\hat{\mathcal{P}}_\theta(o_{t+1}|o_t, a_t)$, which takes the current observation o_t and policy action a_t as input, and outputs the predicted next observation $o_{t+1} \sim \hat{\mathcal{P}}_\theta(\cdot|o_t, a_t)$.

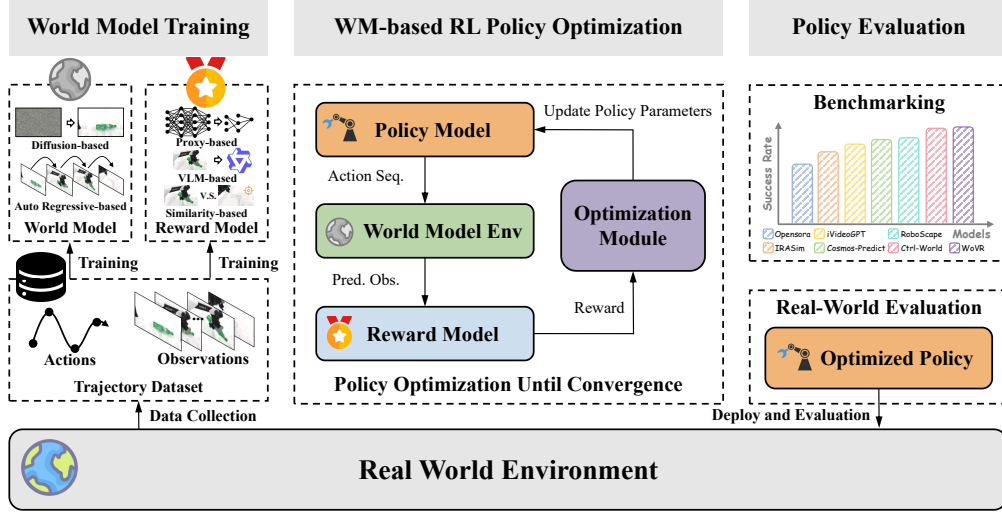


Figure 3: WorldArena 2.0 framework for leveraging world models as RL environments, comprising three stages: world model training, RL policy optimization using the world model, and policy evaluation.

Reward Model: Parameterized by ψ , it predicts the immediate reward based on the current observation and action, defined as $r^t = \mathcal{R}^\psi(o_t, a_t)$.

Policy Model: Parameterized by θ , it outputs the action distribution conditioned on the current observation, i.e., $a_t \sim \pi_\theta(\cdot|o_t)$, which recursively generates actions for the next timestep based on the predicted observations from the world model.

Optimization Module: It updates the policy parameters θ by maximizing the expected discounted return $\mathcal{J}(\theta)$, supporting any mainstream RL optimization algorithms.

The evaluation pipeline comprises three sequential stages of world model training, WM-based RL policy optimization, and optimized policy evaluation. Within this framework, the world model replaces the simulator in traditional RL settings, and the entire closed-loop pipeline is built on the recursive trajectory generation process:

$$o_0 \sim \rho_0, a_t \sim \mathcal{P}_\phi(\cdot|\hat{o}_t, a_t), \hat{r}_t = \hat{\mathcal{R}}_\psi(o_t, a_t). \quad (1)$$

The entire pipeline can be optimized via various optimization methods using the obtained reward signals.

In the evaluation pipeline, we first train the world model on the collected real-world data to learn the underlying environment dynamics. Formally, given a dataset $\mathcal{D} = \{(o_t, a_t, o_{t+1}, r_t)\}_{i=1}^N$ collected from real-world interactions, we optimize the world model parameters ϕ by minimizing a task-aligned loss function $\mathcal{L}_{\text{WM}}(\phi; \mathcal{D})$, which can be instantiated based on the specific world model architecture:

$$\phi^* = \text{argmin}_\phi \mathcal{L}_{\text{WM}}(\phi; \mathcal{D}). \quad (2)$$

Next, we perform policy optimization under the proposed framework until convergence. With the pre-trained world model and reward model fixed, we update the policy parameters θ by maximizing the expected discounted return $\mathcal{J}(\theta)$. For general policy gradient-based methods, the gradient is formally derived as:

$$\nabla \mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta, \hat{\mathcal{P}}_\phi} \left[\left(\sum_{t=0}^{T-1} \nabla \log \pi_\theta(a_t|o_t) \right) \hat{A}_t(\tau) \right], \quad (3)$$

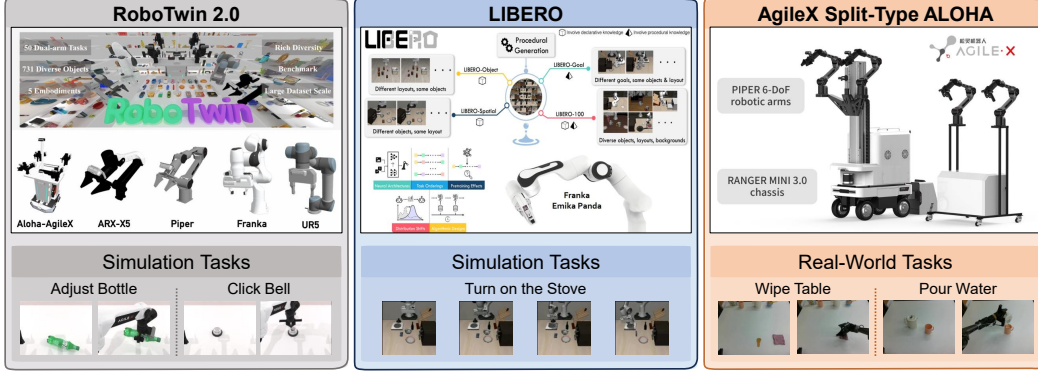


Figure 4: Overview of three test platforms in WorldArena 2.0: RoboTwin 2.0, LIBERO, and a real-world AgileX split-body ALOHA embodiment.

where $\hat{A}_t(\tau)$ denotes the advantage function estimator at timestep t , which can be instantiated differently based on the specific RL algorithm.

Finally, we deploy the optimized policy to interact with the real environment and measure its success rate. We evaluate the performance of a world model as an RL training environment by benchmarking the success rates of policies trained using different world models, where the success rate is defined as the ratio of successful task executions over total real-world rollouts.

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

Most existing benchmarks [6, 8] 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, which together form a graded pipeline from domain-randomized simulation to structured knowledge-transfer benchmarks to real-world physical execution. An overview of the six evaluation tasks across the three platforms is presented in Figure 4, with each platform detailed in Appendix A.

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 the world model directly predicts closed-loop action sequences, and task completion rate reflects its planning reliability. By running these protocols across RoboTwin, LIBERO, and the 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 [24], Wan2.2 [16], Genie Envisioner [32], and ACT [42] with a tactile tokenizer [12], as our baseline. For tactile prediction quality, we adopt PSNR and SSIM 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.

Table 1: Comparison of visuo-tactile world model task success rate on the UniVTAC simulator.

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

Table 2: Success rate of $\pi_{0.5}$ policy trained with different reward models in different tasks on RoboTwin 2.0.

Method	Proxy-based		VLM-based		Similarity-based	
	Click Bell	Adjust Bottle	Click Bell	Adjust Bottle	Click Bell	Adjust Bottle
SFT	43.75	55.08	43.75	55.08	43.75	55.08
Simulator-based RL	87.30	78.90	87.45	78.90	87.45	78.90
OpenSora	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)	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

Wan2.2 achieves higher tactile prediction quality than specialized embodied models, as general-purpose world models retain richer cross-modal knowledge priors that align more effectively with tactile modalities. Consequently, Wan2.2 attains **100%** success on *Insert HDMI*, confirming that superior tactile prediction capability directly translates to improved performance in fine-grained contact-rich manipulation. Conversely, *Lift Bottle* yields a counter-intuitive outcome where the ACT baseline achieves **80%** while all world models fall to **0%**, because this long-horizon task demands sustained force control where tactile information mainly serves high-frequency feedback, yet the long-horizon planning capability of current world models remains limited.

These results demonstrate that our standardized tactile injection pipeline equips world models with tactile comprehension capabilities, improving their ability to perceive the complete world and approaching the performance of existing tactile VLA models on contact-intensive manipulation tasks. Through the unified visuotactile evaluation framework that couples tactile prediction with downstream manipulation assessment, WorldArena 2.0 extends the capability boundary of existing world models, revealing their embodied competence in real-world physical interactions. This benchmark effectively exposes the capability gaps across different architectures, advancing world models toward more realistic and physically grounded intelligence.

4.2 World Model Evaluation as RL Environments

We evaluated the performance of the policy trained with different world model RL environments. The implementation is listed in Appendix B.1.

Overall performance. We evaluated 7 world models on two tasks, and compared their performance with the baseline SFT policy as well as the policy trained on the ground-truth RoboTwin2.0 simulator. The results are summarized in Table 2. The results demonstrate that, although policies trained with existing world models cannot yet outperform those trained on the simulator, the top-performing strategy has achieved comparable performance. Specifically, WoVR [43] achieves the best performance on the short-horizon Click Bell task, while Ctrl-World [21] yields state-of-the-art results on the long-horizon Adjust Bottle task. Furthermore, limited by their video generation quality, OpenSora [44], IRASim [45], and iVideoGPT [26] exhibit only marginal improvements in success rate. In contrast, Cosmos-Predict-2.5 (action) [46] and RoboScape [20] achieve favorable performance.

Analysis of different reward model design. Furthermore, considering that the reward model is also a core component in the RL training pipeline that guarantees the correctness of the policy optimization

Table 3: Task success rates (%) of world models as embodied data engines and action planners across three platforms.

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	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

direction, we also conduct systematic validation and comparison on the performance of three reward models with distinct architectures, which are specified as follows: the proxy-based reward model, which adopts ResNet as the backbone network and takes the current visual observation and task text instruction as inputs to achieve end-to-end immediate reward value prediction; the VLM-based reward model, which employs Qwen-3.5 [47] as the reward evaluator to perform fine-grained reward scoring based on the observation sequence and task instruction; and the visual Similarity-based reward model, which takes the feature similarity between the visual observation predicted by the world model at the current timestep and the target state observation of the task as the core metric for reward calculation. The results in Table 2 indicate that the proxy-based reward achieves the robustest performance. This is owing to the fact that the VLM-based reward is not finetuned on this task and the similarity-based reward highly depends on the observation prediction performance.

4.3 Cross-platform Evaluation and Analysis

Cross-platform video quality evaluation. Following the evaluation protocol established in WorldArena[10], 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 4, 5, 6, 7, 8, 9. A consistent pattern emerges across platforms. Commercial models such as Veo 3.1 and Wan 2.6 lead in visual quality, yet their advantage narrows on physics adherence, where embodied models such as CtrlWorld and IRASim achieve comparable or better trajectory accuracy. This indicates that visual fidelity alone is an insufficient proxy for dynamics modeling. Among embodied models, WoW and CtrlWorld consistently rank highly on physics adherence and content consistency.

Cross-platform task success rate. Beyond perceptual quality, we evaluate world models as embodied data engines and action planners on RoboTwin, LIBERO, and the real-world AgileX platform. Table 3 summarizes the task success rates. As a data engine, no world model matches real demonstration data, and the performance gap widens from simulation to real-world tasks. Real-world evaluation remains the most challenging setting, with only a few models achieving non-zero success rates, all far below practical deployment requirements. These results indicate that a substantial gap remains between simulation-trained world models and the demands of real-world robotic task execution.

Cross-platform performance correlation. We analyze cross-platform ranking correlations for both perceptual quality and task success, with pairwise results shown in Figures 6, 7, 8, and 9. For perceptual quality, visual quality, motion quality, physics adherence, and 3D accuracy show strong correlations across platforms, suggesting that low-level fidelity and geometric reasoning transfer relatively well. By contrast, content consistency and controllability exhibit weaker correlations, indicating greater domain sensitivity in semantic and instruction-level alignment. This gap becomes more pronounced in functional evaluation: task success correlates positively between the two simulators but drops greatly when compared with real-world performance. These results reveal a clear sim-to-real gap, showing that simulation performance—whether perceptual or functional—is not a reliable proxy for real-world deployment and that physical evaluation remains indispensable.

5 Conclusion and Future Work

We introduce WorldArena 2.0, a comprehensive benchmark built on WorldArena that extends the evaluation of embodied world models across three key dimensions: modality, functionality, and platform. It incorporates visuotactile sensory inputs, interactive reinforcement learning tasks, and real-world robotic testbeds. Experiments across 12 state-of-the-art models reveal substantial sim-to-real gaps and identify critical areas for improving embodied world model design. Looking ahead, we plan to expand the benchmark to include additional sensory modalities, increase task complexity and diversity, and explore more challenging real-world scenarios, enabling a deeper understanding of the capabilities and limitations of embodied world models.

References

- [1] 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.
- [2] 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. 2026.
- [3] 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.
- [4] 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*, 2025.
- [5] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldsore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.
- [6] 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.
- [7] 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.
- [8] 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*, 2025.
- [9] 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.
- [10] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyong 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.
- [11] Jiahao Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, et al. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025.
- [12] 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*, 2026.
- [13] 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*, 2025.

- [14] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*.
- [15] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.
- [16] 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.
- [17] 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. *arXiv preprint arXiv:2408.06072*, 2024.
- [18] 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*, 2024.
- [19] Jinwei Gu. Cosmos world foundation models for physical ai. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pages 39–39, 2025.
- [20] 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.
- [21] 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.
- [22] 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*, 2026.
- [23] 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.
- [24] 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.
- [25] 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*, 2024.
- [26] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideopt: Interactive videopts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.
- [31] 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.

- [32] 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.
- [33] 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.
- [34] 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*, 2026.
- [35] Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026.
- [36] Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.
- [37] 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.
- [38] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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*, 2024.
- [45] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025.

- [46] 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.
- [47] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.
- [48] Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Liangzhi Shi, Yuqing Xie, Zhexuan Xu, et al. Rlinf-vla: A unified and efficient framework for vla+ rl training. *arXiv preprint arXiv:2510.06710*, 2025.
- [49] Tianxing Chen, Zanzin 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.
- [50] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. 2025.
- [51] Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, et al. Playworld: Learning robot world models from autonomous play. *arXiv preprint arXiv:2603.09030*, 2026.
- [52] 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.

A Platform Introduction

RoboTwin 2.0 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 select RoboTwin 2.0 for its systematic domain randomization, which exposes world models to severe visual and spatial distribution shifts. A model that performs well here must learn generalizable visual and physical priors rather than memorizing specific background textures or object poses, making this platform a strong stress test for out-of-distribution generalization.

LIBERO provides 130 language-conditioned single-arm manipulation tasks built upon Robosuite, 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.

AgileX Split-Type ALOHA serves as our 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. We evaluate 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. All real-world evaluations inherently incorporate natural lighting and background variations, providing a direct and rigorous sim-to-real performance metric. 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. Any model that achieves high scores in simulation but fails here reveals a fundamental sim-to-real gap, making this platform the definitive arbiter of real-world deployability.

Together, these three platforms span the key axes of cross-embodiment evaluation: RoboTwin 2.0 probes robustness to visual and spatial diversity through systematic randomization, LIBERO enables fine-grained diagnosis of specific knowledge transfer capabilities, and the AgileX ALOHA platform serves as the final physical reality check. By covering domain-randomized simulation, structured diagnostic benchmarks, and real-world execution, this tripartite design captures the full spectrum of challenges that a world model must overcome to be considered truly general, ensuring that improvements in one setting are validated against the others and that performance rankings reflect genuine cross-platform competence rather than narrow specialization.

B Supplemental Description for World Model as RL Environments

B.1 Pipeline Implementation

To verify the capability of world models serving as the online RL environment for embodied policies, we construct a unified and extensible end-to-end experimental pipeline based on RLinf [48]. In this evaluation aspect, two representative robotic manipulation tasks, namely adjust bottle and click bell from the RoboTwin 2.0 [49] benchmark, are selected as the test tasks, with the $\pi_{0.5}$ [50] model adopted as the base embodied policy. To sufficiently characterize the dynamic characteristics of the manipulation task environment, we collect a total of 3000 trajectory samples containing complete state, action, and visual observation sequences using two embodied policies with distinct performance levels as well as the native action planner module of the simulator [51], and this dataset is simultaneously utilized for the training of both the world model and the reward model. In addition, based on 1000 high-quality expert trajectory samples generated by the action planner, we finetune the $\pi_{0.5}$ model with SFT to construct the initialized base policy for the subsequent RL optimization stage. During the RL policy optimization phase, the initialized base policy conducts online environmental

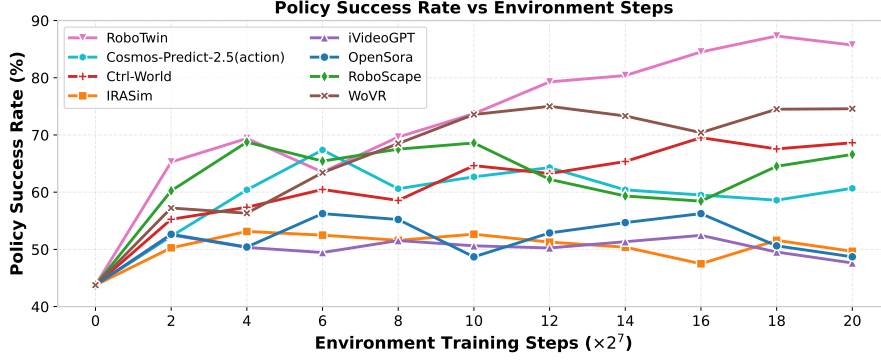


Figure 5: Relationship between environment training steps and the policy success rate in Click Bell task with Proxy-based reward model.

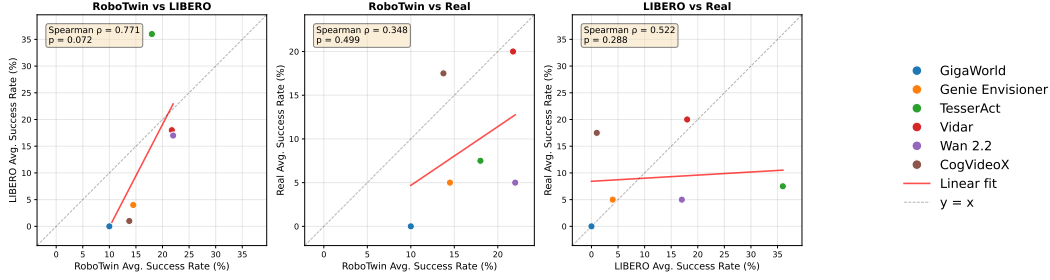


Figure 6: Cross-platform task success rate correlation between RoboTwin, LIBERO and the real-world robotic data.

interaction with world models of different configurations respectively, and policy iterative training is performed using the Group Relative Policy Optimization (GRPO) algorithm [52] until model convergence. Finally, the fully trained policies are subject to performance evaluation in the RoboTwin 2.0 simulation environment, with task success rate adopted as the core evaluation metric.

B.2 Analysis of world model and policy interaction step

In evaluating world models as RL environments, we have also analyzed the success rate curves of the policies with increasing the number of policy-environment interaction steps, with the results presented in Figure 5. The results indicate that nearly all models can guide policy updates to varying degrees as the number of interaction steps increases.

Table 4: 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 Envioner	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	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)	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 5: 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
Vevo 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 6: 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 7: 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 8: 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 Envioner	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 9: 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 Envioner	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

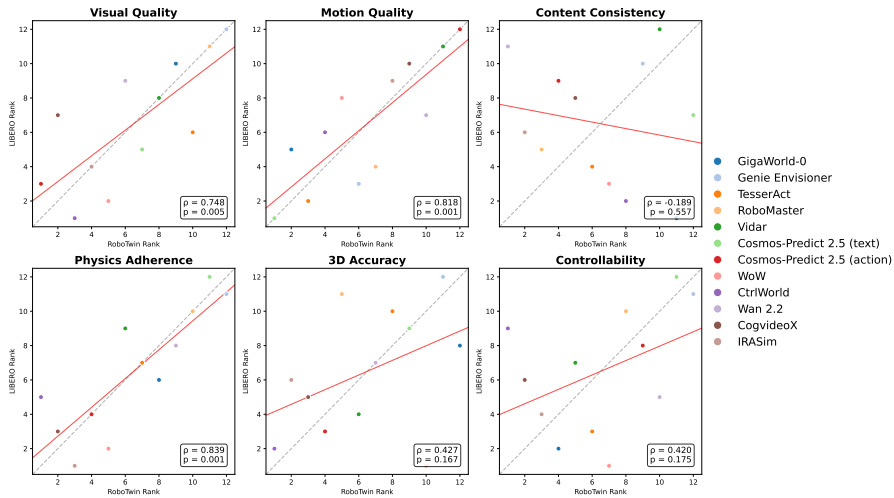


Figure 7: Cross-platform video quality correlation between RoboTwin and LIBERO.

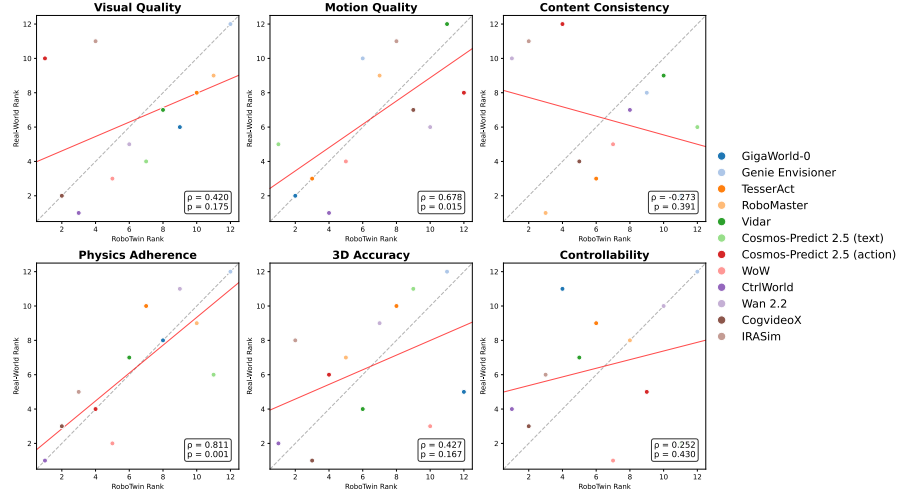


Figure 8: Cross-platform video quality correlation between RoboTwin and the real-world robotic data.

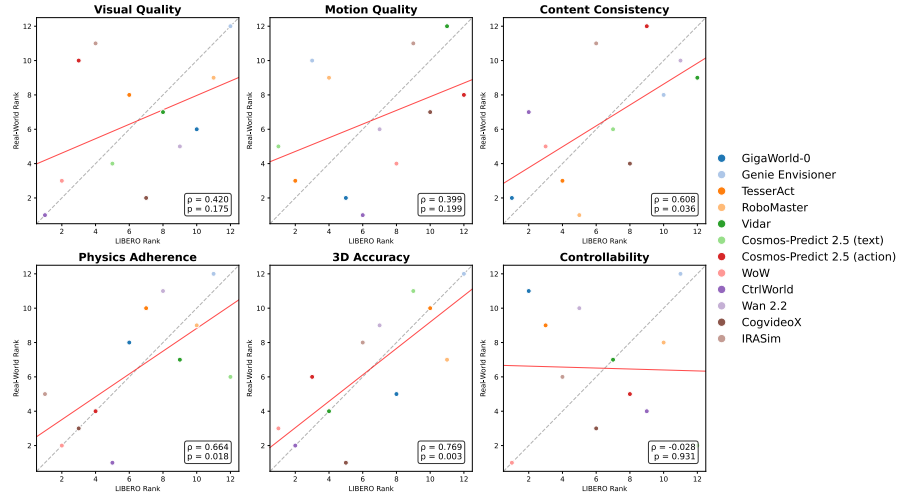


Figure 9: Cross-platform video quality correlation between LIBERO and the real-world robotic data.