PHYS-WORLDBENCH: A UNIFIED BENCHMARK FOR EVALUATING EMBODIED WORLD MODELS FROM PASSIVE PREDICTION TO ACTIVE INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied intelligence requires model to understand how actions transform the physical world. World models provide a natural foundation for this ability, yet their physical interaction capabilities remain insufficiently characterized. Existing benchmarks often evaluate passive prediction or action-conditioned modeling separately, lacking a unified protocol that connects world prediction, action-effect modeling, and real-world action generation. They also commonly organize evaluation around semantic tasks, making it difficult to cover the underlying mechanisms of physical interaction. We introduce **Phys-WorldBench**, a unified real-world benchmark that organizes tabletop manipulation through interaction skills and physical constraints. Its three-level protocol spans passive future prediction, action-condition modeling, and world action modeling within a shared physical interaction space. Built on 4,500 real-robot trajectories ($\sim$30 hours), Phys-WorldBench covers 50 atomic, compositional, and out-of-distribution tasks. Across representative world models and embodied policies, we observe substantial variation across the three levels, with persistent challenges on compositional and out-of-distribution tasks. Visually plausible prediction does not consistently translate into accurate action-effect modeling or robust real-world execution. Cross-level analysis further finds significant moderate associations between adjacent levels ($r = 0.34$–0.36 for L1–L2 and $r = 0.33$–0.38 for L2–L3, all $p < 0.05$ across Wan and Cosmos), while passive prediction remains weakly associated with real-robot progress. Together, these results suggest that predictive quality alone provides an incomplete measure of embodied world-model capability, motivating evaluation across the full progression from world prediction to physical interaction.

1 INTRODUCTION

Recent progress in world models has substantially advanced the ability of intelligent agents to predict and reason about environment dynamics (Zhu et al., 2024; Ding et al., 2025). As this paradigm moves toward embodied settings, world models are increasingly expected not only to predict how the physical world evolves, but also to model how specific actions change it and, ultimately, to generate actions for physical execution. This progression is reflected in a growing spectrum of embodied world models: passive models such as Wan 2.6 (Wan et al., 2025) and Veo (Wiedemer et al., 2025) forecast future observations from high-level instructions; action-conditioned models such as Ctrl-World (Guo et al., 2025) and DreamDojo (Gao et al., 2026) additionally predict outcomes conditioned on robot actions; and world action models such as FastWAM (Yuan et al., 2026) and Cosmos-Policy (Kim et al., 2026) directly generate executable actions from observations. These models therefore engage with different stages of the same physical interaction process, raising a fundamental question: *does capability in predicting the world reliably carry over to modeling action effects and, eventually, to effective physical interaction?*

At the core of embodied interaction is a common observation–action–transition process: an agent observes the environment, actions intervene on its state, and the physical world evolves into subsequent observations. The three model classes expose different interfaces to this process. Passive world models predict its outcomes without explicit action inputs; action-conditioned world models predict how supplied actions alter those outcomes; and world action models close the loop by generating the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Three-level evaluation protocol in Phys-WorldBench.** Level-1 evaluates passive world modeling by predicting future observations from the current observation and task instruction without explicit action inputs. Level-2 evaluates action-condition world modeling by predicting the physically realized future conditioned on a supplied action sequence, including both reference(GT) and counterfactual(CF) actions. Level-3 evaluates world action modeling by directly generating executable robot actions from observations and instructions.

actions themselves. Evaluating these capabilities within a shared interaction space therefore provides a natural way to study not only how well each capability performs, but also *how capability changes as models move from prediction toward physical interaction*.

Existing evaluation protocols provide only a partial view of this progression. First, passive prediction, action-conditioned modeling, and action generation are typically evaluated separately, making it difficult to determine how capabilities relate across stages or where they break down. Second, action-conditioned models are commonly assessed through rollout quality or downstream task success (Shang et al., 2026a), often with VLM-based judging (Bai et al., 2025). Such evaluations can conflate errors in action-condition prediction with evaluation or execution failures and rarely test whether predictions respond correctly when the supplied action is deliberately changed. Third, existing benchmarks are often simulator-based or organized around semantic task definitions rather than the underlying interaction skills and physical constraints, limiting systematic coverage of the mechanisms required for real-world manipulation. Together, these limitations motivate a unified real-world evaluation that connects prediction, action-condition world modeling, and execution within a shared physical interaction space.

To address these gaps, we introduce **Phys-WorldBench**, a unified real-world benchmark for evaluating embodied world models through physical interaction. Phys-WorldBench combines two complementary designs. First, an **atomic physical interaction space** organizes tabletop manipulation bottom-up by interaction skills and physical constraints rather than surface-level task semantics, enabling systematic construction of atomic, compositional, and out-of-distribution interactions. Second, a **three-level evaluation protocol** examines the progression from **Level 1 – Passive World Modeling**, to **Level 2 – Action-Condition World Modeling**, and **Level 3 – World Action Modeling** within the same task space. Level 2 further pairs reference(GT) and counterfactual(CF) actions with their physically executed futures, directly testing whether predicted outcomes respond to changes in action.

Built on 4,500 real-robot trajectories totaling approximately 30 hours, Phys-WorldBench contains 20 atomic tasks, 20 compositional tasks, and 10 compositional OOD tasks. The shared task and data foundation enables both level-specific diagnosis and direct analysis of how capabilities relate across levels. Experiments across representative world models and embodied policies reveal substantial variation across passive prediction, action-conditioned modeling, and closed-loop execution, with persistent degradation on compositional and OOD interactions. Cross-level analysis shows significant moderate associations between adjacent levels ($r = 0.34$–0.36 for L1–L2 and $r = 0.33$–0.38 for L2–L3 across backbone-matched Wan and Cosmos families), while passive prediction remains only weakly associated with real-robot progress. The interaction taxonomy further helps localize failure

Table 1: Comparison with existing embodied world-model benchmarks. "Comp. OOD" denotes unseen compositions of interaction skills. L1, L2, and L3 correspond to passive prediction, action-conditioned world modeling, and executable action generation, respectively. "Executed CF" indicates that counterfactual actions are paired with their physically realized futures.

Benchmark	Tasks & Data			Evaluation Capabilities				
	Data Source	Organization Design	Comp. OOD	Level 1 Passive Pred.	Level 2 Action-Condition	Level 2 Counterfactual	Level 2 Executed CF	Level 3 Action Gen.
WorldModelBench (Li et al., 2025a)	Real	Task-level	×	✓	×	×	×	×
EWMBench (Yue et al., 2025)	Real	Task-level	×	✓	×	×	×	×
WoW-World-Eval (Fan et al., 2026)	Real	Task-level	✓	✓	×	×	×	×
World-in-World (Zhang et al., 2025)	Sim	Task-level	×	×	*Indirect*	×	×	×
WorldArena (Shang et al., 2026a)	Sim	Task-level	×	✓	*Indirect*	×	×	×
WorldArena 2.0 (Shang et al., 2026b)	Sim + Real	Task-level	×	✓	*Indirect*	×	×	×
RoboWM-Bench (Jiang et al., 2026)	Real→Sim	Task-level	×	✓	×	×	×	×
RoboPhys-3D (Wang et al., 2026)	Sim	Task-level	×	✓	×	×	×	×
MiraBench (Yang et al., 2026)	Real	Task-level	×	✓	Direct	✓	×	×
Phys-WorldBench (Ours)	**Real**	**Interaction-level**	✓	✓	**Direct**	✓	✓	✓

modes: some compositional and OOD failures emerge already at the prediction stage, whereas deformation- and tool-mediated interactions can retain non-trivial predictive performance yet fail during execution. These results suggest that embodied world-model capability exhibits local continuity across adjacent stages but does not simply transfer end-to-end from prediction to physical interaction.

Our contributions are threefold:

- We introduce **Phys-WorldBench**, a unified real-world benchmark spanning passive world modeling, action-condition world modeling, and world action modeling within a shared task and data foundation. Evaluation across representative model families reveals significant moderate associations between adjacent levels, but only a weak relationship between passive prediction and real-robot progress, highlighting incomplete capability transfer from prediction to execution.
- We develop an atomic physical interaction space organized by interaction skills and physical constraints rather than task semantics, enabling systematic construction and diagnosis of atomic, compositional, and OOD tabletop manipulation tasks.
- We propose an interventional action-condition world evaluation protocol that pairs reference and counterfactual actions with their physically executed futures, enabling direct assessment of factual effect fidelity and robustness to action interventions.

2 Related Work

Embodied World Models World models have long been studied as a means for agents to internalize environment dynamics and support prediction and planning (Yang et al., 2023; Zhu et al., 2024; Ding et al., 2025), and have increasingly extended to embodied robotics (Park et al., 2023; Zhou et al., 2024). Recent systems expose three increasingly coupled capability interfaces. Passive world models, such as Wan 2.6 (Wan et al., 2025) and Veo (Wiedemer et al., 2025), predict future observations from high-level instructions; action-conditioned models, such as Ctrl-World (Guo et al., 2025) and DreamDojo (Gao et al., 2026), predict environment transitions conditioned on robot actions; and world action models, including Cosmos-Policy (Kim et al., 2026), DreamZero (Ye et al., 2026), FastWAM (Yuan et al., 2026), and GigaWorld-Policy, jointly leverage world dynamics for executable action generation. Unified video–action models such as UVA (Li et al., 2025b) and UWM (Zhu et al., 2025) further blur these boundaries by supporting combinations of video prediction, dynamics modeling, and policy learning within a single architecture. Despite this convergence, it remains unclear how predictive, action-effect, and execution capabilities are related, and whether strong performance at one stage is indicative of capability at another.

Embodied World Model Benchmarks Evaluation has similarly progressed from perceptual quality toward increasingly embodied measures. Earlier benchmarks emphasize visual fidelity, semantic alignment, physical adherence, and motion consistency (Meng et al., 2024; Li et al., 2025a; Yue et al., 2025), while WorldArena (Shang et al., 2026a) connects perceptual quality with functional utility through data generation, policy evaluation, and action planning. WorldArena 2.0 further expands

evaluation across modalities, interactive learning, and simulated and real robotic platforms, while World-in-World (Zhang et al., 2025) evaluates world models through closed-loop planning. More recent benchmarks target complementary aspects of physical grounding: RoboWM-Bench (Jiang et al., 2026) evaluates whether generated behaviors can be converted into executable robot actions, RoboPhys-3D (Wang et al., 2026) measures 3D- and execution-grounded consistency, and MiraBench (Yang et al., 2026) explicitly evaluates action-conditioned reliability, including action following and failure-inducing conditions.

These efforts substantially broaden embodied world-model evaluation, but focus on different slices of the prediction-to-action pipeline. As shown in Table 1, Phys-WorldBench instead evaluates passive prediction, action-condition world modeling, and world action modeling within the same real-robot task and data foundation, enabling direct analysis of capability relationships across levels. Its atomic physical interaction space supports systematic coverage and diagnosis of manipulation mechanisms, while paired reference and counterfactual executions provide physically realized targets for both action conditions, enabling direct evaluation of factual effect fidelity and robustness to action interventions.

3 PHYS-WORLDBENCH

3.1 THREE-LEVEL EVALUATION PROTOCOL

As illustrated in Fig. 1, Phys-WorldBench evaluates embodied world models at three levels within a shared real-robot interaction space. All three levels are grounded in the same underlying process:

$$(o_t,\, g) \;\rightarrow\; a_{t:t+T} \;\rightarrow\; o_{t+1:T}, \tag{1}$$

where o_t denotes the current observation, g the language-specified task goal, $a_{t:t+T}$ the robot action sequence, and $o_{t+1:T}$ the resulting future observations. The three levels probe different interfaces of this process by varying the available inputs and prediction target while keeping the task space and physical interactions shared, as summarized in Table 2.

Level-1: Passive World Modeling. Given (o_t, g), the model predicts future observations without access to explicit robot actions. This level measures whether the model can generate a plausible task-conditioned evolution of the physical scene.

Level-2: Action-Conditioned World Modeling. Given $(o_t, g, a_{t:t+T})$, the model predicts the future resulting from the supplied action sequence. For evaluation, reference and counterfactual trajectories are constructed from matched or closely aligned initial observations. Crucially, both action sequences are physically executed on the robot, so each prediction is evaluated against the future actually induced by that action rather than against the reference trajectory. Level-2 therefore tests whether a model captures how changes in action lead to corresponding changes in physical outcome.

Level-3: World Action Modeling. Given (o_t, g), the model directly generates executable actions $\hat{a}_{t:t+T}$ for accomplishing the task. This level evaluates whether the physical knowledge represented by the model can support closed-loop interaction with the real world. The full set of evaluated models is provided in Appendix A.

Because the three levels are evaluated over a shared physical interaction space, Phys-WorldBench supports not only level-specific diagnosis but also direct analysis of how passive prediction, action-conditioned prediction, and action generation relate to one another. This shared foundation distinguishes differences in model capability from differences in task domain and motivates the structured interaction space described next.

Table 2: Three evaluation levels in Phys-WorldBench. Each level probes a different interface of the same physical interaction process.

Level	Input	Output
1: Passive World Modeling	(o_t, g)	$\hat{o}_{t+1:T}$
2: Action-Conditioned World Modeling	$(o_t, g, a_{t:t+T})$	$\hat{o}_{t+1:T}$
3: World Action Modeling	(o_t, g)	$\hat{a}_{t:t+T}$

3.2 ATOMIC PHYSICAL INTERACTION SPACE

Phys-WorldBench focuses on fixed-base tabletop manipulation and defines an **Atomic Physical Interaction Space** around the physical transitions induced by robot actions. Rather than organizing

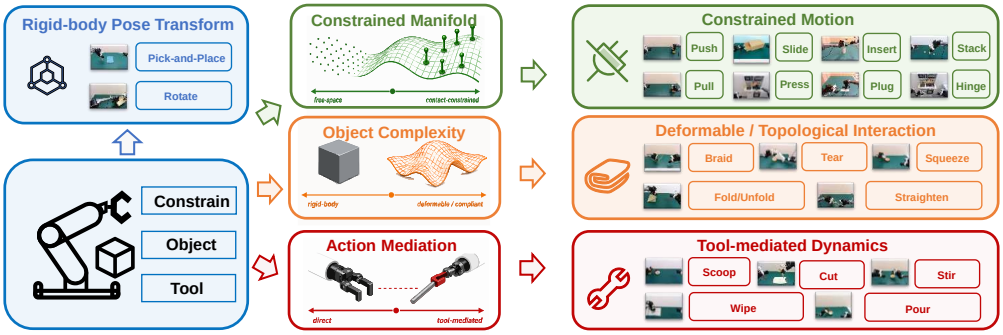


Figure 2: **Atomic Physical Interaction Space in Phys-WorldBench.** Tabletop interactions are organized according to constrained manifold, object complexity, and action mediation, yielding four categories: rigid-body pose transformation, constrained motion, deformable and topological interaction, and tool-mediated dynamics.

tasks primarily by semantic task names, we characterize interactions according to the physical structure among the robot, manipulated objects, tools, and environment, including contact, geometric constraints, deformation, and tool-mediated dynamics (Lynch & Park, 2017).

As shown in Fig. 2, the interaction space is organized along three factors—*constrained manifold*, *object complexity*, and *action mediation*—which give rise to four broad categories:

- **Rigid-body pose transformation**: direct manipulation of approximately rigid objects, such as pick-and-place and rotation.
- **Constrained motion**: manipulation governed by contact or geometric constraints, including push, pull, slide, insert, plug, stack, press, and articulated-object manipulation.
- **Deformable and topological interaction**: manipulation of non-rigid or topology-sensitive objects, including fold/unfold, straighten, braid, tear, and squeeze, requiring reasoning over high-dimensional shape and topology changes (Zhu et al., 2019).
- **Tool-mediated dynamics**: indirect manipulation in which object dynamics are mediated by a tool, such as scoop, stir, cut, wipe, and pour.

This interaction space supports systematic construction of atomic and compositional tasks and controlled evaluation of generalization to unseen compositions.

3.3 Task and Data Construction

Based on the atomic physical interaction space, Phys-WorldBench instantiates 50 real-robot tasks: 20 atomic tasks, 20 compositional tasks, and 10 **compositional OOD** tasks. Each atomic task isolates a dominant interaction skill, while compositional tasks combine multiple skills into ordered interaction sequences. The 20 in-distribution compositional tasks include 12 short-horizon tasks containing 2–3 skills and 8 long-horizon tasks containing 4–6 skills. The 10 compositional OOD tasks use interaction compositions held out from task-specific training, evaluating generalization beyond the compositions observed during fine-tuning.

For each of the 40 in-distribution atomic and compositional tasks, we collect 90 successful trajectories for training, together with 10 successful trajectories and 10 deliberately failed trajectories for evaluation. The successful evaluation trajectories provide factual/reference futures, while the failed trajectories are generated by executing alternative action interventions from matched or closely aligned initial states and provide physically realized counterfactual futures for Level-2 evaluation. For each of the 10 compositional OOD tasks, we collect 10 trajectories used only for evaluation.

In total, Phys-WorldBench contains 3,600 training trajectories and 900 evaluation trajectories: 400 successful in-distribution trajectories, 400 counterfactual failure trajectories, and 100 compositional-OOD trajectories. All trajectories contain synchronized visual observations, robot states, end-effector poses, language instructions, executed actions, and task outcomes. Dataset statistics, collection details, and the complete task list are provided in Fig. 5 and Appendix B.

Table 3: Level-1 results of passive world models on **Phys-WorldBench**.

Model	Generation Consistency		Visual Quality		Physics Adherence			Instruction Following	L1-PWS
	Background Consistency	Subject Consistency	MUSIQ Score	JEDi Similarity	Robot Arm Integrity	Object Plausibility	Interaction Quality	Task Alignment	
Veo 3.0 (API)	0.6595	0.7235	0.7734	0.3985	**0.5723**	**0.4185**	**0.4004**	**0.4691**	**0.5526**
Wan 2.6 (API)	**0.7578**	0.7529	**0.7972**	0.4456	0.4711	0.3659	0.3441	0.3648	0.5338
WoW-14B-2M	0.6439	0.6715	0.6934	**0.6564**	0.3560	0.3096	0.2936	0.3256	0.4945
ABot-PhysWorld	0.6315	0.6888	0.7495	0.6463	0.3752	0.3036	0.2852	0.2824	0.4905
Cosmos-Pred. 2.5-14B	0.6959	**0.7981**	0.6998	0.3081	0.4203	0.3432	0.3195	0.2992	0.4778
Cosmos-Pred. 2.5-2B	0.6394	0.7589	0.7296	0.2435	0.3987	0.3411	0.3127	0.3041	0.4602
Wan2.2-5B	0.5291	0.5509	0.7287	0.3718	0.2568	0.2528	0.2104	0.2152	0.3864
Gigaworld-0	0.4727	0.5530	0.7537	0.2300	0.2424	0.2552	0.2224	0.2504	0.3738
CogVideoX1.5-5B-I2V	0.2765	0.2757	0.5288	0.3411	0.2000	0.2064	0.2000	0.2008	0.2785

4 Level-1: Passive World Modeling

4.1 Experimental Setup

Setup. Level-1 evaluates passive world modeling under

$$(o_t, g) \rightarrow \hat{o}_{t+1:T}, \tag{2}$$

where models predict task-conditioned future observations without access to the executed robot actions. Because the future is underdetermined under this interface, we evaluate whether the generated evolution is visually coherent, physically plausible, and consistent with the task goal rather than requiring frame-exact reconstruction. All models are evaluated zero-shot on the shared atomic and compositional task space.

Target models. We evaluate Veo 3.0 (Google DeepMind, 2025), Wan 2.6 I2V, Wan2.2-5B (Wan et al., 2025), WoW-14B-2M (Fan et al., 2026), Cosmos Predict 2.5 Image2World (NVIDIA et al., 2026), Gigaworld-0 (Team et al., 2025), ABot-PhysWorld (Chen et al., 2026), and CogVideoX1.5 (Yang et al., 2025).

Metrics. We evaluate four complementary dimensions: *generation consistency*, *visual quality*, *physics adherence*, and *instruction following*. Generation consistency and visual quality measure temporal coherence and perceptual/distributional quality, while physics adherence and instruction following evaluate physical interaction plausibility and task consistency using a calibrated VLM judge. Detailed metric definitions, normalization, and VLM-judge calibration are provided in Appendices C and D. We average constituent metrics within each dimension and define the **Level-1 Phys-World Score (L1-PWS)** as

$$\text{L1-PWS} = \frac{S_{\text{cons}} + S_{\text{vis}} + S_{\text{phys}} + S_{\text{inst}}}{4}. \tag{3}$$

4.2 Comparison Results

Overall performance. Table 3 shows clear variation across models and evaluation dimensions. Veo 3.0 achieves the highest L1-PWS (0.5526), followed by Wan 2.6 (0.5338), while WoW-14B-2M performs best among the evaluated open-source models (0.4945). More importantly, strong visual or reference-based similarity does not necessarily imply physically meaningful task evolution: models with competitive generation-quality metrics can remain substantially weaker in physical interaction quality and task alignment. This gap indicates that visual coherence alone captures only part of passive world-modeling capability. Detailed model-wise comparisons are provided in Appendix E.1.

Interaction-category analysis. Performance further degrades for interactions involving non-rigid dynamics, tool-mediated effects, and longer compositional horizons, indicating that these settings remain particularly challenging for passive world models. Detailed category-level results are provided in Appendix E.2.

Takeaway. Level-1 evaluates whether a model can generate a plausible task-conditioned physical future without knowing which robot actions are executed. It therefore characterizes passive predictive capability, but cannot determine whether a model predicts the consequence of a *specific* action sequence. Level-2 addresses this stricter requirement by explicitly conditioning future prediction on executed robot actions.

Table 4: Detailed Level-2 results. Factual fidelity is computed from GT and CF samples. The three interventional-robustness components are higher-is-better percentile scores.

Model	Factual Effect Fidelity (GT)						Interventional Robustness (GT–CF)				L2-PWS
	Trajectory Accuracy	PSNR	V-JEPA Similarity	Robot-Arm Integrity	Object Consistency	Factual Fidelity	Trajectory Wass. Score	nDTW Mean Gap Score	nDTW Tail Gap Score	Robustness Score	
Cosmos-Pred. 2.5	0.2252	**0.5869**	**0.6096**	0.4168	0.4521	**0.4640**	**0.5627**	0.4287	0.4278	0.4731	0.4686
Motus	**0.2788**	0.4360	0.5784	**0.5119**	**0.5897**	0.4610	0.4389	**0.5774**	**0.5784**	**0.5316**	**0.4963**

5 Level-2: Action-Conditioned World Modeling

5.1 Experimental Setup

Target models. We evaluate two action-conditioned world models: **Cosmos-Predict 2.5 (action)** (NVIDIA et al., 2026) and **Motus** (Bi et al., 2025). Their world-model backbones belong to the same families as Cosmos-Policy and FastWAM, respectively, providing family-level pairings for the cross-level analysis in Section 7. These pairings do not control for differences in architecture, pretraining, or downstream training. Both models are fine-tuned on the **Phys-WorldBench** training split for 50,000 steps with a batch size of 8, retaining their respective optimization settings.

Metrics. Each prediction is evaluated against the recorded future corresponding to its input action sequence. We summarize performance using two complementary metrics.

Factual effect fidelity ($\mathcal{F}$) measures prediction quality with GT and CF samples by equally averaging trajectory accuracy, normalized PSNR (Gonzalez & Woods, 2018), V-JEPA similarity (Assran et al., 2025), and physics adherence, which combines VLM-rated robot-arm integrity and object consistency.

Interventional robustness ($\mathcal{R}$) measures how prediction quality changes from GT to CF actions through three trajectory-based statistics: the Wasserstein distance between trajectory-accuracy distributions, the mean nDTW drop, and the increase in the low-nDTW fraction. These statistics are converted to higher-is-better percentile-rank scores and averaged to obtain $\mathcal{R}$. We combine the two metrics with equal weight:

$$\text{L2-PWS} = \frac{\mathcal{F} + \mathcal{R}}{2}. \tag{4}$$

Model-level scores are averaged equally across tasks. Detailed definitions and aggregation rules are provided in Appendix C.

Evaluation protocol. Level-2 evaluates action-conditioned causal reasoning over 1-second prediction horizons, aligned with typical world model control frequency. From each trajectory, we extract three 1s clips at **Early (16.7%)**, **Mid (50.0%)**, and **Late (83.3%)** temporal anchors. At each anchor, models receive the initial observation o_t and a 1s action chunk resampled to their specific FPS requirements, and must predict the future frames under different action conditions: GT actions following the original goal-achieving trajectory, and CF actions that deviate from task goals while remaining physically plausible. We conduct three independent trials per clip with predictions sampled at 5Hz, and report performance averaged across all trials and anchors.

5.2 Comparison Results

Motus achieves a higher overall L2-PWS than Cosmos-Predict 2.5 (0.4963 vs. 0.4686; Table 4), with similar factual fidelity but higher interventional robustness.

Factual effect fidelity. The two models obtain nearly identical factual fidelity (Cosmos: 0.4640; Motus: 0.4610). Cosmos achieves higher PSNR and V-JEPA similarity, indicating closer visual agreement with the recorded videos. Motus achieves better trajectory accuracy and stronger robot and object consistency.

Interventional robustness. Motus achieves higher aggregate robustness (0.5316 vs. 0.4731), which accounts for its overall advantage. Its higher nDTW mean-gap and tail-gap scores indicate smaller average alignment degradation and a smaller increase in poorly aligned predictions when actions change from GT to CF. Cosmos achieves a higher Wasserstein-based score, indicating a smaller change in the trajectory-accuracy distribution, but trails Motus on both nDTW-based components.

Table 5: **Real-robot Level-3 results** over 50 tasks (500 rollouts per model). SR is success rate (%), and Prog. is normalized progress (0–100). Bold denotes the best result in each metric.

Task Split	# Tasks	$\pi_{0.5}$		Cosmos-Policy		Cosmos-Policy (2.5)		FastWAM	
		SR	Prog.	SR	Prog.	SR	Prog.	SR	Prog.
Atomic	20	**45.0**	**72.0**	14.0	32.5	21.0	40.3	24.0	44.2
Compositional	20	**15.0**	**40.2**	2.5	20.0	5.5	24.9	7.0	26.9
OOD	10	**5.0**	**49.0**	0.0	23.8	0.0	28.2	0.0	29.6
All tasks	**50**	**25.0**	**54.7**	6.6	25.8	10.6	31.7	12.4	34.4

6 Level-3: World Action Modeling Evaluation

6.1 Experimental Setup

Target models. We evaluate four world action models (WAMs): Cosmos-Policy (NVIDIA et al., 2026), Cosmos-Policy (2.5), Motus (Bi et al., 2025), and FastWAM (Yuan et al., 2026), together with the VLA baseline $\pi_{0.5}$ (Intelligence et al., 2025). The two Cosmos variants are initialized from Cosmos-Predict2-2B and Cosmos-Predict2.5-2B, respectively, while FastWAM uses Wan2.2-5B. These Cosmos and Wan backbone families also appear in Levels 1–2, enabling the cross-level analysis in Section 7. All models are trained on the same **Phys-WorldBench** split and evaluated under the same conditions; implementation details are provided in Appendix F.3.

Evaluation protocol. We evaluate each model with closed-loop real-robot rollouts on 50 tasks, using 10 trials per task (500 rollouts per model) under the same multi-view observations and scene-reset protocol. We report task-averaged success rate (SR, %) and normalized progress (Prog., 0–100), where progress is derived from a 0–5 partial-completion score. Per-task results, protocol details, and uncertainty analyses are provided in Appendix F.

6.2 Comparison Results

Policy performance. $\pi_{0.5}$ achieves the highest success rate and progress on every split (Table 5). Its overall success rate is 25.0%, compared with 12.4% for FastWAM, the strongest WAM policy. FastWAM also leads the WAMs in overall progress (34.4), followed by Motus (33.1), Cosmos-Policy (2.5) (31.7), and Cosmos-Policy (25.8). Cosmos-Policy (2.5) and Motus achieve similar overall success rates (10.6% and 10.4%, respectively).

Backbone comparisons. Within the Cosmos family, the 2.5 variant improves overall success from 6.6% to 10.6% and progress from 25.8 to 31.7. Across families, however, FastWAM outperforms both Cosmos policies despite its Wan backbone scoring lower at Level-1. The Level-1 model ranking therefore does not carry over directly to closed-loop policy performance. Because these policies also differ in architecture and training, the comparison does not isolate the effect of backbone quality.

Composition and generalization. Success rates decline sharply on compositional and OOD tasks for all evaluated action model. $\pi_{0.5}$ drops from 45.0% on atomic tasks to 15.0% on compositional tasks and 5.0% on OOD tasks; FastWAM drops from 24.0% to 7.0% and 0.0%. All four WAM policies record zero successes in their 100 OOD rollouts. Nevertheless, their OOD progress scores range from 23.8 to 29.6, indicating partial completion without reaching the final goal. These results identify reliable completion of composed and unfamiliar tasks as a shared limitation of the evaluated policies.

7 Cross-Level Analysis

We examine whether tasks with higher prediction quality also yield better real-robot execution. We compare task-level scores within two backbone families: Wan2.2, Motus, and FastWAM at Levels 1–3, respectively; and Cosmos-Predict 2.5, its action-conditioned variant, and Cosmos-Policy (2.5). The analysis uses in-domain tasks with valid scores for each compared pair. Levels 1 and 2 use their respective PWS scores, while Level-3 uses real-robot progress normalized to $[0, 1]$.

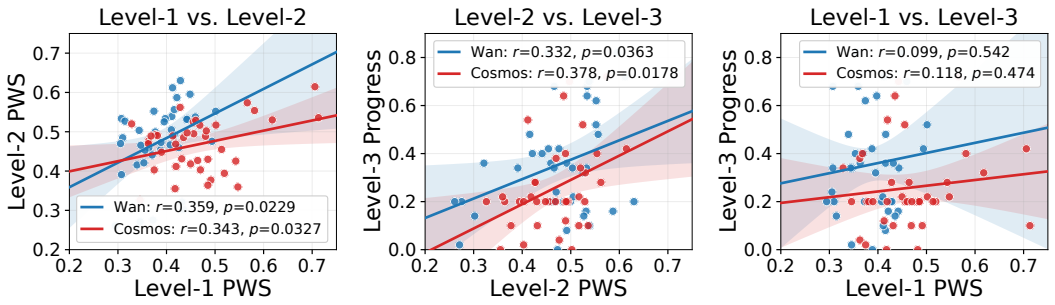


Figure 3: **Cross-level task correlations.** Each point represents an in-domain task within the Wan or Cosmos family. Levels 1 and 2 use PWS; Level-3 uses real-robot progress normalized to $[0, 1]$. Legends report Pearson's r and p-values. Shaded regions denote 95% confidence interval.

Cross-level performance relation. L1 and L2 are moderately correlated for both model family Wan ($r = 0.359$, $p = 0.0229$) and Cosmos ($r = 0.343$, $p = 0.0327$), reflecting their shared requirement for physically coherent future prediction. However, L2 additionally tests whether the predicted future is consistent with the specified actions, capturing action–effect fidelity beyond passive plausibility.

Importantly, L2 also correlates with real-robot progress (Wan: $r = 0.332$, $p = 0.0363$; Cosmos: $r = 0.378$, $p = 0.0178$), whereas L1 does not (Wan: $r = 0.099$, $p = 0.542$; Cosmos: $r = 0.118$, $p = 0.474$). This pattern suggests that **action-conditioned prediction provides a stronger bridge from world prediction to physical execution**, while passive prediction alone is an insufficient proxy for control. The three levels therefore capture complementary capabilities: physical plausibility (L1), action–effect fidelity (L2), and closed-loop execution (L3).

Skill-level transfer. Fig. 4 further reveals substantial skill- and model-dependent gaps across levels. Although both families obtain relatively strong L1/L2 scores on many skills, these capabilities often fail to translate into execution; for example, Pour, Scoop, Braid, and Plug retain moderate-to-high L2 scores but achieve near-zero L3 success. In contrast, Stir and Slide show more consistent transfer from action-conditioned prediction to real-world execution across both families. Notably, Cosmos is often stronger at L1/L2 yet does not consistently outperform Wan at L3, showing that improved predictive modeling alone does not means stronger action control.

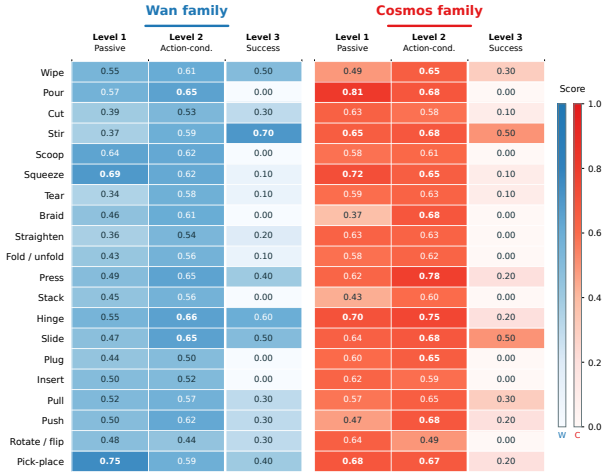


Figure 4: **Cross-level performance across skills.**

8 Conclusion

We introduced Phys-WorldBench, a unified real-world benchmark for evaluating embodied world models across passive prediction, action-conditioned reasoning, and real-world action generation. Grounded in an atomic physical interaction space, it enables systematic diagnosis across mechanism-based skills, compositional tasks, and compositional-OOD tasks. Our experiments show that current models remain limited in action-effect understanding and robust real-world execution, despite producing visually plausible predictions. Cross-level analysis further reveals model-dependent transfer and skill-specific gaps between modeling and execution. These findings call for evaluating embodied world models through physical interaction, beyond passive prediction or final task success alone.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improving the clarity and presentation of the manuscript. In addition, generative AI models were used as VLM-based judges for selected evaluation metrics, as described in the Appendix D. The authors reviewed all AI-assisted manuscript revisions and designed and validated the VLM-based evaluation procedure. The authors take full responsibility for the final content, claims, analyses, and reported results.

ETHICS STATEMENT.

This work focuses on evaluating and improving embodied AI systems through robot interaction. All data used in our experiments are collected from robot manipulation in controlled simulation or laboratory environments and do not involve human subjects, personally identifiable information, or sensitive personal data. The benchmark and proposed models are intended for research on physical reasoning, world modeling, and robot learning. As with other embodied AI systems, deployment in real-world environments may introduce safety risks due to model prediction or control errors; therefore, all real-robot experiments are conducted under controlled conditions with appropriate operational safeguards and human supervision.

REPRODUCIBILITY STATEMENT.

We provide detailed descriptions of the dataset construction, task splits, annotation procedures, model architectures, training settings, evaluation metrics, and experimental protocols in the main paper and supplementary material. For benchmark evaluation, we specify the construction of atomic, compositional, and out-of-distribution tasks, as well as the protocols for passive prediction, action-conditioned prediction, and real-robot action generation. We also document the implementation details required to reproduce the reported offline, simulation, and real-robot experiments, including model inputs, optimization settings, evaluation procedures, and metric computation. Code, evaluation scripts, and permissible benchmark resources will be released to facilitate reproducibility and future research.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang

Su, and Jun Zhu. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment. *CoRR*, abs/2603.23376, 2026. doi: 10.48550/ARXIV.2603.23376. URL https://doi.org/10.48550/arXiv.2603.23376.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 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.

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.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.

Google DeepMind. Veo: a text-to-video generation system. Technical report, Google DeepMind, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

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.

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.

Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, et al. Robowm-bench: A benchmark for evaluating world models in robotic manipulation. *arXiv preprint arXiv:2604.19092*, 2026.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025b.

Ge Ya Luo, Gian Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond fvd: Enhanced evaluation metrics for video generation quality, 2024. URL https://arxiv.org/abs/2410.05203.

Kevin M Lynch and Frank C Park. *Modern robotics*. Cambridge University Press, 2017.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2026.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Yu Qi, Xinyi Xu, Ziyu Guo, Siyuan Ma, et al. Mme-cof-pro: Evaluating reasoning coherence in video generative models with text and visual hints. In *European Conference on Computer Vision*, 2026a.

Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, et al. Bear: Dissecting embodied abilities in multimodal language models through skill-level evaluation and diagnosis. In *Proceedings of the International Conference on Machine Learning*, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

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

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, et al. Worldarena 2.0: Extending embodied world model benchmarking on modality, functionality and platform. *arXiv preprint arXiv:2605.17912*, 2026b.

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. Gigaworld-0: World models as data engine to empower embodied AI. *CoRR*, abs/2511.19861, 2025. doi: 10.48550/ARXIV.2511.19861. URL https://doi.org/10.48550/arXiv.2511.19861.

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.

Tianyi Wang, Jiazhou Chen, Yiming Xu, Xiangyu Li, Tianyi Zeng, Chih-Hsien Chou, Ning Lu, Liang Peng, Junfeng Jiao, and Christian Claudel. Robophys-3d: A comprehensive embodied world model evaluation via 3d reconstruction. 2026.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. Mirabench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

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.

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.

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.

Jiahan 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.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

Jihong Zhu, Benjamin Navarro, Robin Passama, Philippe Fraisse, André Crosnier, and Andrea Cherubini. Robotic manipulation planning for shaping deformable linear objects withenvironmental contacts. *IEEE Robotics and Automation Letters*, 5(1):16–23, 2019.

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.

Table 6: Models evaluated in **Phys-WorldBench**, organized by type and evaluation level.

Model	Type	Eval Mode	Level	Fine-tuned
Wan 2.6	WM	API	L1	No
Wan 2.2	WM	Code	L1	No
Veo 3.0	WM	API	L1	No
WoW-14B-2M	WM	Code	L1	No
Cosmos-Pred. 2.5	WM	Code	L1	No
ABot-Physworld	WM	Code	L1	No
Gigaworld-0	WM	Code	L1	No
CogVideoX1.5	WM	Code	L1	No
Motus	AWM	Code	L2	Yes
Cosmos-Pred. 2.5 (action)	AWM	Code	L2	Yes
$\pi_{0.5}$	VLA	Code	L3	Yes
FastWAM	WAM	Code	L3	Yes
Cosmos-Policy	WAM	Code	L3	Yes

A Details of Evaluated Models

In this section, we provide a detailed list of the models evaluated in **Phys-WorldBench**, as shown in Table 6. We categorize the models based on their architectural goals and evaluation levels to ensure a comprehensive assessment of physical world understanding and generation.

- **World Models (WM) - Level 1:** These models focus on high-fidelity video generation and physical consistency. We evaluate state-of-the-art diffusion-based models such as Wan 2.6, Veo 3.0, and Cosmos-Pred. 2.5. The evaluation primarily focuses on their ability to maintain structural integrity and motion causality in open-domain scenarios.
- **Actionable World Models (AWM) - Level 2:** These models are designed to predict future states conditioned on specific actions. Models like Ctrl-World and the action-conditioned version of Cosmos-Pred. 2.5 are assessed on their forward-prediction accuracy and how well they simulate the physical consequences of robotic interventions.
- **Vision-Language-Action (VLA) & Policy Models - Level 3:** This level involves end-to-end decision-making. We include $\pi_{0.5}$, FastWAM, and Cosmos-Policy to investigate how world-model-based pre-training or internal representations translate into concrete robotic control tasks and physical reasoning.

The evaluation modes vary between direct **API** calls for closed-source models and **Code**-based local inference for open-source frameworks, ensuring that each model is tested under its optimal configuration.

B Details of Dataset

B.1 Task Composition and Dataset Splits

Phys-WorldBench contains 50 real-robot tasks: 20 atomic tasks, 20 compositional tasks, and 10 compositional out-of-distribution (OOD) tasks. Atomic tasks isolate a dominant interaction skill, while compositional tasks combine skills into ordered sequences, including 12 short-horizon tasks (2–3 skills) and 8 long-horizon tasks (4–6 skills).

Table 7 summarizes the dataset split. For each of the 40 atomic and compositional tasks, we collect 90 successful training trajectories, 10 successful evaluation trajectories, and 10 deliberately failed evaluation trajectories. Level-2 and Level-3 fine-tuning uses only the 3,600 successful training trajectories. All evaluation trajectories, including the 400 deliberate failures, are held out from fine-tuning. The 10 compositional OOD tasks contribute 100 successful evaluation trajectories and no training data.

Table 7: **Dataset splits.** The same 40 tasks provide successful training trajectories and successful/failed evaluation trajectories. Ten additional compositional OOD tasks are held out from fine-tuning.

Split	Tasks	Traj./task	Total
Training: successful		90	3,600
Evaluation: successful (GT)	40	10	400
Evaluation: deliberate failure (CF)		10	400
Evaluation: compositional OOD	10	10	100
Total	50 unique	–	4,500

Scope of compositional OOD. The OOD split evaluates unseen skill–object pairings and combinations of previously observed atomic skills. All constituent skills belong to the same 20-skill vocabulary, and all four interaction categories are represented during training. Objects may be familiar or previously unseen, depending on the task. For example, held-out tasks use a spoon to move blocks or press coffee grounds, extending the combinations of skills, tools, and target objects observed during training. The robot platform, camera configuration, and laboratory environment remain the same. This split targets generalization to novel task combinations and does not systematically vary lighting, viewpoint, background, or embodiment.

The CF split evaluates prediction under deliberately unsuccessful executions of tasks seen during fine-tuning. It therefore tests a different generalization setting from compositional OOD. The compositional OOD split contains successful trajectories only and is not used to compute GT–CF robustness statistics.

B.2 Counterfactual Trajectory Collection

Action conditions. GT denotes successful reference executions, and CF denotes deliberately unsuccessful alternative executions. Both conditions consist of physically executed robot trajectories with recorded actions and corresponding observed futures. CF executions retain the original task instruction while intentionally taking actions that diverge from its goal.

Collection procedure. GT and CF trajectories are collected through the same human teleoperation pipeline. Collection comprises the following steps:

1. **Scene preparation.** The operator prepares the task-specific objects and nominal initial configuration on the same robot platform. Evaluation trials include variations in initial object positions.
2. **Successful execution.** The operator completes the task while recording synchronized visual observations, robot states, end-effector poses, actions, and the outcome.
3. **Alternative execution.** Using the same task instruction, objects, and nominal initial configuration, the operator deliberately fails at a task-critical step through physically executable robot motions. The alternative actions and their consequences are recorded with the same pipeline.

GT and CF trials are acquired in separate collection batches. CF actions are generated through human teleoperation rather than post-hoc editing of recorded actions or synthetic noise injection. Each alternative execution consequently provides a recorded physical response to its action sequence. The interventions are task-dependent, with no common perturbation magnitude imposed across skills.

Initial-state matching. GT and CF executions share the task instruction, object identities, and nominal initial configuration. Physical resets introduce small placement and appearance differences between collection batches. Matching therefore refers to the initial task configuration; the remaining visual variation is quantified in Appendix B.3.

Action-matched prediction targets. For each evaluation clip, the model receives that clip's initial observation and recorded action sequence. Its prediction is compared with the future recorded in

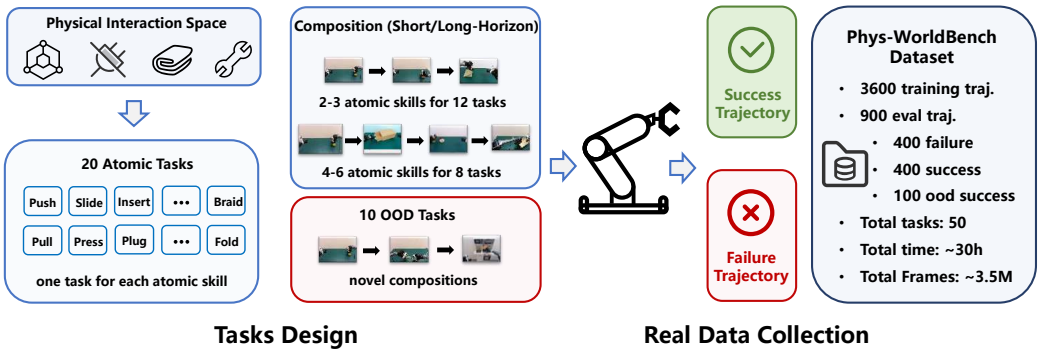


Figure 5: **Task design and dataset composition.** Phys-WorldBench comprises 20 atomic tasks, 20 compositional tasks (12 short-horizon and 8 long-horizon), and 10 compositional OOD tasks. The 40 atomic and compositional tasks provide 3,600 successful training trajectories and 800 evaluation trajectories with successful and deliberately failed executions. The 10 held-out tasks provide 100 additional successful evaluation trajectories.

the same execution. A CF prediction is therefore evaluated against its observed CF future, and an accurate prediction of an unsuccessful execution receives a high prediction-quality score. Later clips from GT and CF trajectories may start from different states following their preceding actions. Comparisons across these clips quantify prediction performance under the two execution conditions.

B.3 Validation of the CF Evaluation Set

Visual alignment. We quantify initial visual alignment using brightness-normalized first-frame structural similarity (SSIM). Mean SSIM between test-GT and test-CF observations is 0.722, compared with 0.820 between test-GT and same-task training observations (Table 8). These statistics characterize visual variation across independently collected batches under the same nominal task configuration. Each prediction is evaluated against the recorded future of its own execution.

Recorded action statistics. We characterize GT and CF motion using consecutive-frame end-effector translation and rotation increments. Mean translation increments are 3.8 mm for GT and 5.3 mm for CF; mean rotation increments are 0.0186 rad and 0.0238 rad, respectively. Using the corresponding GT 99.5th percentile as a reference threshold, 1.25% of CF translation increments and 0.60% of CF rotation increments exceed this threshold. Both conditions have complete sensor logs. These results show that alternative executions have motion magnitudes of the same order as successful executions while preserving complete recordings of the resulting physical interactions.

B.4 Dataset Statistics

End-effector Motion Statistics. Fig. 6 reports both per-frame and cumulative end-effector motion statistics for the four interaction categories. The per-frame translation $\|\Delta p\|$ and rotation $\Delta\theta$ are generally comparable across categories, suggesting that the categories are not separated by simple instantaneous motion-scale differences. However, the cumulative statistics show more pronounced category-specific patterns. Deformable and topological interactions accumulate larger translational motion because they often require repeated local adjustments over flexible or topology-sensitive objects. Tool-mediated dynamics exhibit broader rotational motion, since the robot must frequently orient and reorient tools before applying them to target objects. Constrained motion has smaller cumulative displacement, consistent with tasks that emphasize alignment, contact constraints, and precise local control. Rigid-body pose transformation shows moderate cumulative motion, corresponding to direct object relocation or reorientation in relatively low-dimensional pose space. These results indicate that the dataset maintains comparable low-level control scales while covering distinct long-horizon motion requirements across physical interaction types.

Table 8: **GT–CF alignment and motion statistics.** SSIM uses brightness-normalized first frames. Motion values are consecutive-frame end-effector increments. Exceedance rates use the corresponding GT 99.5th-percentile increment as the threshold.

Initial visual alignment		
Comparison	Mean SSIM	
Test-GT vs. test-CF	0.722	
Test-GT vs. same-task training	0.820	
Recorded action statistics		
Metric	GT	CF
Complete sensor logs	100%	100%
Mean translation increment	3.8 mm	5.3 mm
Mean rotation increment	0.0186 rad	0.0238 rad
Translation above GT 99.5th percentile	–	1.25%
Rotation above GT 99.5th percentile	–	0.60%

Temporal Distributions. We provide additional statistics to characterize the duration distributions of Phys-WorldBench. Fig. 7 shows the trajectory duration distribution across task types, where the x-axis is truncated at 80 seconds for visualization. The full set contains 4,500 episodes at a 30 Hz equivalent frame rate, with an average duration of 22.28 seconds and a median duration of 17.47 seconds. Trajectory durations range from 3.00 seconds to 261.23 seconds, with a standard deviation of 16.51 seconds. This indicates that most trajectories are relatively short manipulation episodes, while the dataset also preserves a long tail of extended compositional interactions. Atomic tasks are generally shorter and more concentrated, whereas compositional tasks show a broader temporal distribution due to multi-step interaction sequences. Compositional OOD tasks have relatively compact durations in this dataset.

Instruction Vocabulary Distributions. We further visualize the language diversity of task instructions using noun and verb word clouds. The noun distribution covers a wide range of everyday objects, containers, tools, object parts, and spatial references, such as *handle*, *bottle*, *drawer*, *cup*, *rope*, *center*, and *rightmost*. This suggests that the instructions require models to ground actions not only on object identities, but also on object parts, spatial locations, and relational descriptions. The verb distribution is dominated by manipulation-oriented interactions such as *pick*, *grasp*, *insert*, *close*, *open*, *stir*, *push*, and *align*, reflecting the interaction-centric nature of Phys-WorldBench. These word clouds summarize the object references and manipulation verbs present in the task instructions.

C Details of Metrics

To provide a comprehensive evaluation of the generated videos, we categorize our metrics into two levels: Level-1 for long-horizon open-ended generation and Level-2 for short-horizon action-conditioned dynamics. Base video metrics are normalized to $[0, 1]$ using the calibration described below; heterogeneous Level-2 causal penalties are aligned with empirical percentile ranks before aggregation.

C.1 Level-1: Open-Ended Generation Metrics

Level-1 metrics assess the model's ability to generate stable, high-quality, and instruction-compliant video sequences.

- **Generation Consistency ($\mathcal{M}_{consist}$)** This dimension evaluates the temporal persistence of both the workspace environment and the robotic agent. To prevent the common "static-frame shortcut" where models generate motionless sequences to inflate consistency metrics, we implement a motion-regularized alignment approach:
  - *Background Consistency ($\mathcal{M}_{bg}$)*: We utilize **CLIP-ViT-L/14** (Radford et al., 2021) to extract global scene descriptors $\psi_t = \text{CLIP}(I_t)$. The raw stability score is computed as the average cosine similarity of the current frame relative to both the initial frame and the immediate

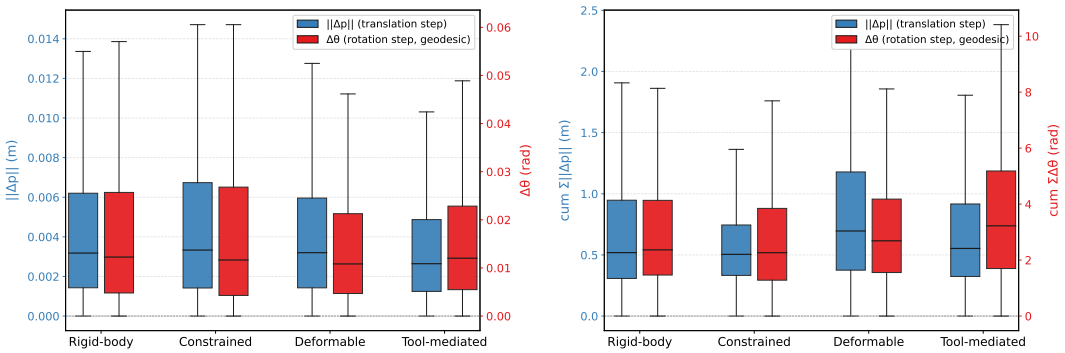


Figure 6: **End-effector motion statistics across interaction categories.**

predecessor:

$$\mathcal{M}_{bg_raw} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{\text{sim}(\psi_t,\psi_1)+\text{sim}(\psi_t,\psi_{t-1})}{2} \tag{5}$$

The final score is obtained by applying min-max normalization to $\mathcal{M}_{bg_raw}$ using the interval [0.8365, 0.9774].

– *Subject Consistency ($\mathcal{M}_{subj}$)*: To monitor the structural integrity of the robotic arm, we extract latent topological features $\phi_t = \text{DINO}(I_t)$ using **DINO-v2** (Caron et al., 2021). The raw integrity score $\mathcal{M}_{subj_raw}$ follows an analogous formulation to $\mathcal{M}_{bg_raw}$:

$$\mathcal{M}_{subj_raw} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{\text{sim}(\phi_t,\phi_1)+\text{sim}(\phi_t,\phi_{t-1})}{2} \tag{6}$$

We normalize the raw score using the interval [0.7297, 0.9916].

- **Visual Quality ($\mathcal{M}_{visual}$)** This dimension assesses the perceptual fidelity and the spatio-temporal distribution alignment of the generated sequences.

  – *Image Fidelity ($\mathcal{M}_{img}$)*: We adopt the **MUSIQ** model (Ke et al., 2021) to evaluate technical distortions in a no-reference setting. For a video V, the score is the average of frame-level predictions:

$$\mathcal{M}_{img_raw} = \frac{1}{T}\sum_{t=1}^{T}\zeta_{musiq}(I_t) \tag{7}$$

  We normalize $\mathcal{M}_{img_raw}$ using the 1st and 99th percentile interval of [0.6151, 0.7977].

  – *JEPA Similarity ($\mathcal{M}_{JEPA}$)*: We employ a similarity-form of the **JEDi** metric (Luo et al., 2024) via a **V-JEPA** encoder (Bardes et al., 2024). The MMD is estimated using a second-order polynomial kernel $\kappa(x,y) = (\langle x,y\rangle + c)^2$, and converted into a similarity measure via an exponential mapping:

$$\mathcal{M}_{JEPA} = \exp\left(-\lambda \cdot \text{MMD}^2_{poly}(\mathcal{G},\mathcal{R})\right) \tag{8}$$

  where $\lambda = 40$ is a scaling factor.

- **Physics Adherence ($\mathcal{M}_{phys}$)** We implement a **VLM-as-a-Judge** framework using **Gemini 2.5 Pro** (Comanici et al., 2025), selected via the human–VLM alignment study in Appendix D.3. Following a hyper-critical auditing philosophy, the judge evaluates the videos across three granular physical dimensions:

  – *Robot Arm Integrity*: Audits structural persistence and kinematic realism, penalizing disintegration and human-hand hallucinations.

  – *Object Consistency*: Evaluates 3D volume persistence and targets "telekinesis" and "melting" artifacts.

  – *Interaction Quality*: Scrutinizes force transfer causality, penalizing "action-at-a-distance" and interpenetration.

Table 9: **Per-task real-robot evaluation across 50 tasks.** Each task is evaluated over ten trials. Opening and closing are displayed separately for slide and hinge, with five trials per direction. **SR** reports successful trials out of evaluated trials. **Prog.** reports cumulative progress, with each trial scored from 0 to 5. Directional rows are combined into their parent tasks before aggregation, and aggregate progress is normalized to 0–100. Best per-row entries are in **bold**.

Skill	Instruction (abbrev.)	$\pi_{0.5}$		Cosmos-Policy		Cosmos-Policy (2.5)		FastWAM	
		SR	Prog.	SR	Prog.	SR	Prog.	SR	Prog.
pick-place	Pick yellow Lego, place into blue tray (L/R arm)	**8/10**	**45/50**	2/10	21/50	3/10	25/50	4/10	31/50
push	Push object on table (forward/backward/left/right)	**7/10**	**43/50**	2/10	14/50	3/10	19/50	3/10	24/50
pull	Grasp and pull object on table (4 directions)	**6/10**	**41/50**	3/10	20/50	3/10	22/50	3/10	16/50
rotate (flip)	Bimanual rotate object 45/90/135/180° CCW	**6/10**	**38/50**	0/10	11/50	1/10	16/50	3/10	21/50
insert	Bouquet hand-over and insertion into bottle	**3/10**	**41/50**	0/10	20/50	0/10	22/50	0/10	20/50
plug	Hand-over plug, insert into power-strip socket	0/10	**24/50**	0/10	10/50	0/10	12/50	0/10	10/50
slide (open)	Open the drawer (left/right arm)	**4/5**	**24/25**	3/5	18/25	3/5	20/25	3/5	18/25
slide (close)	Close the drawer (left/right arm)	**4/5**	**24/25**	2/5	16/25	3/5	19/25	2/5	18/25
open hinge	Insert gripper into recess and pull cabinet door open	**4/5**	**23/25**	1/5	11/25	2/5	15/25	3/5	21/25
close hinge	Push cabinet door closed with nearer gripper	**5/5**	**25/25**	2/5	13/25	3/5	17/25	3/5	20/25
stack	Stack all the blocks together	**3/10**	**33/50**	0/10	10/50	1/10	14/50	0/10	17/50
unfold	Unfold the folded towel	**2/10**	**21/50**	0/10	10/50	0/10	13/50	1/10	17/50
straighten	Bimanual pull rope ends apart until straight	**9/10**	**47/50**	0/10	14/50	1/10	18/50	2/10	17/50
braid	Braid three ropes right-to-left (×6 crossings)	**3/10**	**29/50**	0/10	10/50	0/10	12/50	0/10	10/50
tear	Tear tissue while holding paper roll steady	**2/10**	**38/50**	1/10	14/50	1/10	18/50	1/10	24/50
squeeze	Squeeze ketchup into bowl (left/right arm)	**4/10**	**33/50**	1/10	10/50	2/10	15/50	1/10	20/50
press	Press button on stove / range-hood (L/R gripper)	**6/10**	**36/50**	2/10	16/50	3/10	21/50	4/10	26/50
scoop	Scoop coffee powder from bowl into cup	**2/10**	**33/50**	0/10	10/50	1/10	14/50	0/10	10/50
stir	Stir liquid in bowl with stirring stick (L/R)	3/10	**41/50**	5/10	35/50	6/10	38/50	**7/10**	**41/50**
cut	Cut playdough with small spatula (L/R arm)	2/10	**36/50**	1/10	11/50	2/10	17/50	**3/10**	21/50
pour	Pour coffee grounds from cup into bowl (L/R)	0/10	**5/50**	0/10	**5/50**	0/10	**5/50**	0/10	**5/50**
wipe	Erase pen marks on the board	**7/10**	**40/50**	3/10	26/50	4/10	31/50	5/10	35/50
	Compositional Tasks (12 short-horizon; 8 long-horizon)								
slide→pp→slide	Open drawer, take out scoop, place on table, close drawer	**1/10**	**22/50**	0/10	17/50	0/10	18/50	0/10	17/50
pp→close (hinge)	Place glasses into case and close lid	**1/10**	**28/50**	0/10	10/50	0/10	12/50	**1/10**	12/50
pp→stack	Assemble hamburger (bread, lettuce, patty, bread)	**2/10**	**24/50**	0/10	10/50	1/10	14/50	1/10	18/50
pp→pour	Pour debris from plate into trash can (L/R)	1/10	31/50	**2/10**	**32/50**	**2/10**	**32/50**	**2/10**	**32/50**
pp→fold	Stack socks, then fold stack right-to-left	**9/10**	**47/50**	3/10	19/50	4/10	27/50	5/10	34/50
pp→insert	Grasp coin on box, insert into safe slot (L/R)	**3/10**	**32/50**	0/10	11/50	1/10	15/50	1/10	18/50
pp→rotate→insert	Remove test tube, shake, return to rack (L/R)	0/10	**22/50**	0/10	10/50	0/10	12/50	0/10	10/50
tool→rotate	Pry off bottle cap with bottle opener	0/10	**17/50**	0/10	10/50	0/10	12/50	0/10	10/50
slide→pp	Unzip bag, place clothes inside while holding open	0/10	**11/50**	0/10	10/50	0/10	11/50	0/10	10/50
pp→rotate	Stabilize bottle, unscrew cap CCW (L/R)	**4/10**	**27/50**	0/10	10/50	1/10	14/50	1/10	18/50
pp→press	Stamp seal within triangular region (L/R)	0/10	**14/50**	0/10	5/50	0/10	7/50	0/10	8/50
rot→pp→stack	Reposition cup, grasp handle, hang on rack	0/10	**7/50**	0/10	2/50	0/10	3/50	0/10	4/50
pp→scoop→stir	Scoop coffee, fill water, stir mixture, place stick	0/10	**2/50**	0/10	0/50	0/10	1/50	0/10	0/50
braid→str→ ···	Braid two ropes right-to-left (×4) and straighten	**2/10**	**29/50**	0/10	10/50	0/10	12/50	0/10	10/50
pp→rot→press→pp	Hand object to scanner, scan QR code, place back	0/10	**14/50**	0/10	5/50	0/10	7/50	0/10	7/50
fold (repeat)	Fold paperboard along creases into rectangular carton	0/10	**8/50**	0/10	0/50	0/10	3/50	0/10	7/50
pp (repeat)	Stabilize and open bag; place each fruit inside sequentially	**7/10**	**43/50**	0/10	27/50	2/10	31/50	3/10	34/50
pp→insert (repeat)	Sequentially place test tubes into rack column (L/R)	0/10	6/50	0/10	5/50	0/10	7/50	0/10	**8/50**
pp (rep.)→push	Arrange dominoes in a row, then knock them over	0/10	**15/50**	0/10	6/50	0/10	9/50	0/10	11/50
pp→open→ins→close→pp	Insert earbuds into case, close lid, place on table	0/10	**3/50**	0/10	1/50	0/10	2/50	0/10	1/50
	Compositional OOD Tasks (10)								
wipe	Wipe concave area in center of spoon with cloth	0/10	**23/50**	0/10	17/50	0/10	18/50	0/10	17/50
press	Press paper towel on circular watermark until absorbed	**1/10**	**26/50**	0/10	10/50	0/10	13/50	0/10	18/50
wipe	Wipe outer rim of bowl with paper towel	0/10	**21/50**	0/10	17/50	0/10	18/50	0/10	17/50
tool	Stir coffee grounds in bowl with chopsticks	0/10	**17/50**	0/10	10/50	0/10	12/50	0/10	10/50
tool	Move blocks left/right with a spoon	0/10	**29/50**	0/10	10/50	0/10	12/50	0/10	10/50
tool	Press coffee grounds with bottom of a spoon	0/10	**26/50**	0/10	15/50	0/10	16/50	0/10	15/50
tool	Insert fork downward into center of bowl of coffee	0/10	**30/50**	0/10	15/50	0/10	17/50	0/10	15/50
deform.+insertion	Insert spoon diagonally into bowl, handle up	**2/10**	**32/50**	0/10	10/50	0/10	14/50	0/10	20/50
press	Press hand sanitizer pump	0/10	**6/50**	0/10	5/50	0/10	**6/50**	0/10	5/50
deform.+pull	Separate two tangled charging cables without damage	**2/10**	**35/50**	0/10	10/50	0/10	15/50	0/10	21/50

- **Instruction Following ($\mathcal{M}_{inst}$)** This metric measures the semantic alignment between the generation and the provided text instruction.The judge evaluates whether the specified task is completed through intentional and physically sound movements. Detailed rubrics and prompts are provided in Appendix D.

C.2 Level-2: Action-Conditioned Dynamics Metrics

Level-2 evaluates the fidelity and robustness of short-horizon action-conditioned future prediction. Given an initial observation and an action sequence, the model predicts future visual frames, which are compared with the real video recorded under the same action sequence. We first measure factual prediction fidelity under reference (GT) actions from successful demonstrations. Accurate prediction under these actions alone, however, leaves open whether the model can predict physical effects when

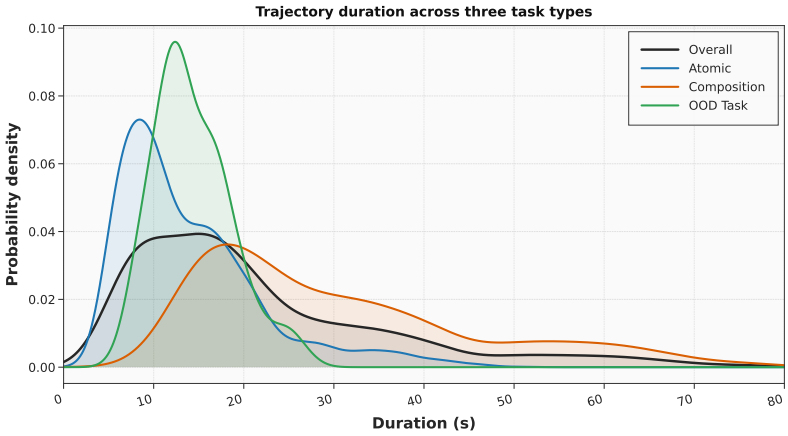


Figure 7: **Trajectory duration distribution across task types.** The x-axis is truncated at 80 seconds for visualization, while the full duration range spans from 3.00 to 261.23 seconds. Across all trajectories, the mean and median durations are 22.28 seconds and 17.47 seconds respectively, indicating that most trajectories are short manipulation episodes with a long tail of extended interactions.

actions depart from the successful behaviors used for fine-tuning. We therefore further evaluate counterfactual (CF) actions from deliberately unsuccessful executions, scoring each prediction against its corresponding recorded future. Comparing the GT and CF prediction-quality distributions then quantifies interventional robustness: how well the model maintains prediction fidelity under this change in the action distribution. The CF collection protocol is detailed in Appendix B.2.

Base prediction metrics. We measure pixel fidelity using normalized PSNR ($\mathcal{M}_{\text{pixel}}$), visual representation similarity using V-JEPA ($\mathcal{M}_{\text{latent}}$), spatial motion fidelity using trajectory accuracy ($\mathcal{M}_{\text{traj}}$), and temporal trajectory alignment using nDTW ($\mathcal{M}_{\text{ndtw}}$). Robot-arm integrity ($\mathcal{M}_{\text{arm}}$) and object consistency ($\mathcal{M}_{\text{obj}}$) are evaluated by the VLM judge. All metrics are oriented so that higher scores indicate better prediction quality. Visual features are obtained from the V-JEPA 2 ViT-g 384 encoder (Assran et al., 2025). For visual trajectiries, we use SAM3 (Carion et al., 2026) to detect the two grippers in the first frame and use their mask pixels nearest to the image center as tracking points. We then track these points throughout the video with CoTracker (Karaev et al., 2024) and normalize the resulting trajectories to $[0, 1]$. Ground-truth and generated videos are aligned by timestamps and processed with the same SAM3–CoTracker pipeline.

Factual effect fidelity. For each GT-action prediction, we average four dimensions: trajectory accuracy, pixel fidelity, latent similarity, and physics adherence:

$$\mathcal{F}_{\text{clip}} = \frac{1}{4}\left[\mathcal{M}^{\text{GT}}_{\text{traj}} + \mathcal{M}^{\text{GT}}_{\text{pixel}} + \mathcal{M}^{\text{GT}}_{\text{latent}} + \frac{\mathcal{M}^{\text{GT}}_{\text{arm}} + \mathcal{M}^{\text{GT}}_{\text{obj}}}{2}\right]. \tag{9}$$

The two VLM scores jointly form the physics-adherence dimension, giving each of the four dimensions equal weight. For model m and evaluation case c, we average GT-action predictions with valid evaluation metrics to obtain $\mathcal{F}_{m,c}$. The scoring procedure applies no additional success-based filtering, and CF predictions do not contribute to factual fidelity.

Interventional robustness. We evaluate whether prediction quality remains stable when the input action changes from the ground-truth (GT) action to a counterfactual (CF) action which leads to task unsuccessful. Each prediction is compared with the recorded future produced by executing that same action; thus, CF predictions are evaluated against CF futures, not against GT futures.

The computation is performed first at the case level. Let m index models, c index evaluation cases, and $a \in \{\text{GT}, \text{CF}\}$ denote the action condition. Each model–case pair contains multiple prediction samples collected across episodes, temporal clips, and repeated generations. We use $P^{a}_{\text{traj},m,c}$ and $P^{a}_{\text{ndtw},m,c}$ to denote the empirical distributions of valid trajectory-accuracy and nDTW scores, respectively.

We characterize the effect of the intervention using three complementary statistics:

$$D^{\text{traj}}_{\text{Wass},m,c} = W_1\big(P^{\text{GT}}_{\text{traj},m,c}, P^{\text{CF}}_{\text{traj},m,c}\big)\,, \tag{10}$$

$$D^{\text{ndtw}}_{\text{gap},m,c} = \mathbb{E}_{P^{\text{GT}}_{\text{ndtw},m,c}}[\mathcal{M}_{\text{ndtw}}] - \mathbb{E}_{P^{\text{CF}}_{\text{ndtw},m,c}}[\mathcal{M}_{\text{ndtw}}], \tag{11}$$

$$D^{\text{ndtw}}_{\text{tail},m,c} = \Pr_{P^{\text{CF}}_{\text{ndtw},m,c}}(\mathcal{M}_{\text{ndtw}} \le \tau_{m,c}) - \Pr_{P^{\text{GT}}_{\text{ndtw},m,c}}(\mathcal{M}_{\text{ndtw}} \le \tau_{m,c})\,. \tag{12}$$

These statistics capture different failure modes. The Wasserstein distance measures the overall discrepancy between the GT and CF trajectory-accuracy distributions, including changes in both location and shape. It is non-directional: a large value indicates a substantial distributional shift, but does not distinguish improvement from degradation. The signed mean gap measures the change in average temporal alignment. A positive value indicates that mean nDTW decreases under CF actions, whereas a negative value indicates an increase. The tail gap measures whether CF actions produce more severely low-scoring predictions than GT actions.

For each model–case pair, we define the low-quality region using the 20th percentile of its GT nDTW scores:

$$\tau_{m,c} = Q_{0.2}\big(P^{\text{GT}}_{\text{ndtw},m,c}\big)\,. \tag{13}$$

The same threshold $\tau_{m,c}$ is then applied to both the GT and CF distributions. Consequently, $D^{\text{ndtw}}_{\text{tail},m,c}$ measures the change in the frequency of low-quality predictions relative to the model's own GT baseline on the same case. For example, if approximately 20% of GT predictions and 45% of CF predictions fall at or below $\tau_{m,c}$, the tail gap is $0.45 - 0.20 = 0.25$. Positive values indicate an increase in low-quality predictions under CF actions, values near zero indicate a similar low-score frequency, and negative values indicate a decrease. A near-zero tail gap therefore indicates stability relative to the GT condition, but does not by itself imply high absolute prediction quality.

We compute the 20th percentile using linear interpolation and include scores equal to the threshold in the low-quality region. Because of finite samples and tied scores, the empirical GT fraction at or below the threshold need not equal exactly 20%.

Task aggregation and score calibration. A benchmark task may consist of one or more evaluation cases. Let $\mathcal{C}_t$ denote the set of cases associated with task t. We first compute factual fidelity and the three robustness statistics separately for each model–case pair, and then average them across the constituent cases:

$$\mathcal{F}_{m,t} = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\mathcal{F}_{m,c}, \qquad D_{k,m,t} = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t} D_{k,m,c}, \tag{14}$$

where $k \in \{\text{Wass}, \text{gap}, \text{tail}\}$. In particular, each tail gap is computed using its own model–case threshold $\tau_{m,c}$ before case-level results are averaged. Samples from different cases are not pooled to estimate a shared task-level threshold.

The three task-level statistics have different numerical scales. We therefore convert each statistic into a higher-is-better score using an inverse empirical percentile rank:

$$\mathcal{S}_k(D_{k,m,t}) = 1 - \text{PercentileRank}_k(D_{k,m,t})\,. \tag{15}$$

For each statistic k, ranks are computed jointly across all available task–model rows in the calibration pool, rather than separately within each model. Average ranks are used for ties and mapped to $[0, 1]$. The signed mean and tail gaps retain their signs during ranking: larger positive gaps indicate greater degradation, whereas negative gaps indicate improvement. The Wasserstein distance is non-negative, with smaller values indicating greater distributional stability.

The task-level interventional robustness score is the equal-weight average of the three calibrated components:

$$\mathcal{R}_{m,t} = \frac{1}{3}\sum_{k\in\{\text{Wass},\text{gap},\text{tail}\}}\mathcal{S}_k(D_{k,m,t}). \tag{16}$$

Because percentile calibration is relative, these scores quantify robustness within the specified evaluation pool and may change if the included task–model rows change.

Level-2 Phys-World Score. We assign equal weight to factual prediction quality and robustness under action intervention:

$$\text{L2-PWS}_{m,t} = \frac{\mathcal{F}_{m,t} + \mathcal{R}_{m,t}}{2}. \tag{17}$$

Here, $\mathcal{F}_{m,t}$ measures prediction quality under GT actions, while $\mathcal{R}_{m,t}$ measures distributional stability, mean degradation, and the increase in low-quality predictions under CF actions. Model-level results are obtained by averaging task-level scores equally across benchmark tasks. Cross-level correlation analyses use one score per task after combining its constituent cases.

D VLM-AS-A-JUDGE EVALUATION PROTOCOLS

D.1 LEVEL-1

To ensure a rigorous and reproducible VLM-as-Judge evaluation, the Level-1 prompt is designed following the **Chain-of-Thought (CoT)** and **Role-Playing** paradigms. In addition to a unified scoring rubric, we use a category-aware prompt template that assigns a task-specific *core evaluation focus* according to the interaction category. The design logic is centered on three principles:

- **Hyper-Critical Persona**: We instruct the VLM to act as a "Hyper-Critical Robotics Inspector." This role encourages the judge to attend to physical inconsistencies, contact artifacts, and temporal glitches rather than only general visual appeal.
- **Hierarchical Rubrics**: Instead of a monolithic score, the prompt decomposes physical adherence into four dimensions: *Robot Arm Integrity*, *Object Consistency*, *Interaction Quality*, and *Task Alignment*. Each dimension is evaluated on a 1–5 Likert scale with explicit anchors.
- **Category-Specific Focus**: For each interaction category, the prompt provides a dedicated inspection focus, such as rigid-body pose continuity, constrained contact, deformation and topology changes, tool-mediated dynamics, or long-horizon object persistence.

You are a **HYPER-CRITICAL Robotics Inspector** specializing in `[CATEGORY]`.
Your goal is to inspect AI-generated videos, focusing heavily on the custom criteria defined in the **CORE EVALUATION FOCUS** section below.

SCORING PHILOSOPHY (STRICT):

- **Motion is Required**: A static video with no movement is a failure of generative intent.
- **Causality is Mandatory**: Objects must not move unless a physical force, such as the robot arm, acts upon them.
- **Task Specifics First**: The custom boundaries and target details specified under **CORE EVALUATION FOCUS** determine the strictness of the final scores.

EVALUATION DIMENSIONS & SCORING RUBRICS:
1. ROBOT_ARM_INTEGRITY (Structure & Motion Quality)

- **Score 1**: Static video; morphing into human limbs; structural disintegration.
- **Score 3**: Basic structure holds, but motion shows noticeable inconsistencies or unnatural trajectories.
- **Score 5**: Perfect mechanical consistency and smooth, purposeful kinematics aligned with the **CORE EVALUATION FOCUS**.

2. OBJECT_CONSISTENCY (Structural Persistence & Natural State)

- **Score 1**: Object disappears or transforms; object moves or changes while the robot arm is static.
- **Score 3**: Maintains basic shape, but exhibits minor perspective drift, local texture morphing, or volume fluctuation during manipulation.
- **Score 5**: Objects maintain coherent identity, volume/mass, material behavior, and task-relevant state changes. Rigid objects remain geometrically stable, while non-rigid objects may plausibly deform, flow, fold, or reconfigure under contact, gravity, or task dynamics.

3. INTERACTION_QUALITY (Contact Fidelity & Causality)

- **Score 1**: "Action at a distance"; object moves before robot contact; ghosting or passing through objects.
- **Score 3**: Plausible interaction, but shows contact defects, micro-clipping, or brief physics latency at active boundaries.

- **Score 5**: Perfect contact physics; movement is strictly driven by contact, passing all micro-scrutiny under **CORE EVALUATION FOCUS**.

4. TASK_ALIGNMENT (Instruction Following)
Instruction: `{instruction_text}`

- **Score 1**: No attempt at the task; video is static; completely wrong object interaction.
- **Score 3**: Attempts the instruction, but the path layout is uncoordinated or aborted midway.
- **Score 5**: Task completed successfully with clear, intentional, and physically sound movements fulfilling the exact instruction specification.

CORE EVALUATION FOCUS
The following focus is instantiated according to the interaction category:

- **Rigid-body Pose Transformation**: Inspect macroscopic progression, general clearance, forward advancement, gripper trajectory continuity, and boundary pixels where the gripper sweeps through the workspace.
- **Constrained Motion**: Inspect contact arrival, layout alignment, gripper approach to environmental constraints, phase transition, and structural continuity under constraint.
- **Deformation and Topological Interaction**: Inspect local micro-errors, pixel melting, tearing artifacts, material boundary behavior, texture blurring, mass loss, localized shape snapping, and glitching artifacts.
- **Tool-mediated Dynamics**: Inspect connection errors, independent slipping, floating tools, gripper–tool contact boundaries, spatial gaps, mesh clipping, tool-tip delay, and ghosting artifacts.
- **Short-Horizon / Long-Horizon**: Inspect persistent errors, object shrinking, fading memory artifacts, background/foreground consistency, object disappearance, spontaneous motion without contact, unexpected scaling, and tracking drift from start to finish.

OUTPUT FORMAT:
Please provide your evaluation in JSON format:

```
{
    "TASK_ALIGNMENT": {"score": 0, "reason": "..."},
    "INTERACTION_QUALITY": {"score": 0, "reason": "..."},
    "ROBOT_ARM_INTEGRITY": {"score": 0, "reason": "..."},
    "OBJECT_CONSISTENCY": {"score": 0, "reason": "..."}
}
```

D.2 LEVEL-2

The Level-2 prompt is specifically tailored for **Fine-Grained Spatio-Temporal Auditing**. Unlike Level-1, which focuses on long-range task logic, this prompt is designed to identify transient generative failures that occur during high-speed or short-duration robotic actions. The design philosophy incorporates:

- **Temporal Sensitivity**: Given the limited temporal context (1-second clips), the prompt directs the VLM to focus on *frame-level* sharpness. It emphasizes the detection of "motion-induced melting" and structural flickering that typically escape long-horizon metrics but are critical for downstream robot control.
- **The "Unreality Tax"**: We introduce a zero-tolerance policy for pixel-level degradation. This mechanism explicitly instructs the judge to cap scores at 2/5 if any high-frequency detail is lost or if edge definitions collapse, effectively penalizing models that use excessive motion blur to mask physical errors.
- **Mechanical Fidelity vs. Perceptual Quality**: The evaluation dimensions are narrowed down to *Robot Arm Integrity* and *Object Consistency*, with a specific focus on mechanical joint stability and 3D volume persistence. This ensures that the VLM ignores stylistic aesthetics and evaluates the clip purely as a physical world simulator.

You are an ELITE ROBOTICS EVALUATOR. You are auditing 1-second (5-frame) video clips that often suffer from severe generative degradation.

SCORING CRITERIA - THE "UNREALITY" TAX:

- **Blur & Noise**: If the video is "blurry" or "muddy" (loss of high-frequency detail), you MUST penalize both scores.
- **Structural Collapse**: If the arm or object loses its edge definition for even ONE frame, the maximum score is 2.
- **Industrial Reality**: A score of 5 is only for videos indistinguishable from a real 4K industrial camera.

EVALUATION DIMENSIONS:

1. ROBOT_ARM_INTEGRITY (Focus: Sharpness & Mechanical Structure)

- **Score 1**: Total collapse. The arm is a blurry "ghost," morphs into the table, or is completely static.
- **Score 2 (Severe Artifacts)**: The arm is visible but "muddy." Significant "melting" or structural disintegration occurs. The gripper's shape is lost in motion blur.
- **Score 3**: Structure is held, but edges are "soft" or "fuzzy." Noticeable texture swimming or frame-to-frame jitter.
- **Score 4**: Sharp and recognizable, but has minor non-physical micro-motions or slight aliasing.
- **Score 5**: Crystal clear, sharp edges, perfect mechanical joints, and smooth 1-second trajectory.

2. OBJECT_CONSISTENCY (Focus: 3D Stability & Scene Clarity)

- **Score 1**: Objects are unrecognizable blobs; they vanish or teleport; the background "warps" violently.
- **Score 2**: Significant "color bleeding" or "texture melting." Objects move telekinetically. The table surface looks like liquid.
- **Score 3**: Geometric shape is mostly stable, but the object looks "pasted on" due to inconsistent shadows or "fuzzy" boundaries.
- **Score 4**: Stable geometry and sharp edges, but exhibits subtle lighting flickering or shadow misalignment.
- **Score 5**: Perfect 3D volume, rock-solid textures, and realistic optical shadows/reflections.

TARGET ACTION: "Manipulate objects."

OUTPUT FORMAT (JSON ONLY):

```
{
"ROBOT_ARM_INTEGRITY": {"score": 0, "reason": "..."},
"OBJECT_CONSISTENCY": {"score": 0, "reason": "..."}
}
```

D.3 Human-VLM Alignment Study for VLM-as-a-Judge

To validate the reliability of our VLM-as-a-Judge protocol for evaluating world models, we conducted an expert human study to measure the alignment between automated VLM judges and human perception.

We curated an evaluation subset of 150 diverse robotic manipulation videos. Human experts annotated each video across our four proposed fine-grained dimensions (*Task Alignment*, *Interaction Quality*, *Robot Arm Integrity*, and *Object Consistency*) on a 1–5 Likert scale, yielding a total of 600 human annotations. We then evaluated three representative VLMs acting as judges: Gemini 2.5 Pro, GPT-5.2, and Qwen3-VL-8B. Alignment is measured using Pearson correlation (r) and Spearman rank correlation (ρ) between the VLM-generated scores and human annotations, computed at the aggregate per-task score level.

Results and Conclusion: As shown in Table 10, **Gemini 2.5 Pro** achieves the highest overall alignment with human experts ($r = 0.744$, $\rho = 0.718$), outperforming GPT-5.2 and establishing a substantial lead over the open-weight Qwen3-VL-8B.

These results demonstrate that reliably evaluating physical interactions and causal effects in robotic videos requires the advanced reasoning capabilities of frontier models. This observation is also consistent with recent embodied multimodal model evaluation benchmarks, which leverage strong multimodal foundation models for skill-level assessment and diagnosis Qi et al. (2026b). Moreover,

Table 10: Agreement between VLM judges and human experts on 150 annotated videos. Best results are in **bold**.

Judge (Model)	N	Pearson r	Spearman ρ
Gemini-2.5-pro	150	**0.744**	**0.718**
GPT-5.2	150	0.671	0.673
Qwen3-VL-8B	150	0.500	0.494

recent video evaluation frameworks have adopted frontier VLMs as automated judges for assessing complex semantic and reasoning capabilities in generated videos Qi et al. (2026a).

Consequently, we adopt Gemini 2.5 Pro as the default automated judge for all evaluations in PhysWorldBench.

E Details of Evaluation

E.1 Additional Level-1 Model Analysis

Table 3 reveals complementary strengths across the evaluated passive world models. Veo 3.0 obtains the strongest robot-arm integrity, object plausibility, interaction quality, and task alignment. Wan 2.6 instead leads in background consistency and MUSIQ, while Cosmos Predict 2.5-14B achieves the highest subject consistency. Among the evaluated open-source models, WoW-14B-2M achieves the highest L1-PWS and the strongest JEDi similarity.

The detailed results further illustrate the mismatch between visual similarity and physically meaningful task evolution. WoW-14B-2M and ABot-PhysWorld achieve high JEDi similarity but remain weaker than Veo 3.0 and Wan 2.6 in object plausibility, interaction quality, and task alignment. Conversely, Veo 3.0 achieves the strongest aggregate performance despite not leading the reference-similarity metrics.

Within the Wan family, Wan 2.6 substantially improves over Wan2.2-5B across both generation- and interaction-oriented metrics. Task alignment increases from 0.2152 to 0.3648, while L1-PWS increases from 0.3864 to 0.5338. Cosmos Predict 2.5 exhibits a comparatively balanced profile, whereas Gigaworld-0 and CogVideoX1.5 show larger degradation in physics adherence and task alignment. Across the evaluated models, physical interaction quality and task alignment remain more challenging than basic visual consistency.

E.2 Level-1 Interaction-Category Analysis

To further examine where passive world models fail, we report category-level VLM-based evaluation results over the Atomic Physical Interaction Space and compositional tasks in Table 11.

Atomic interactions. Among atomic interactions, rigid-body pose transformation achieves the highest object plausibility (0.4522), while constrained motion performs best in robot-arm integrity (0.4189), interaction quality (0.3807), and task alignment (0.3663). Deformation and topological interaction is more challenging, particularly in object plausibility and interaction quality, while tool-mediated dynamics obtains the lowest overall scores among the atomic categories. This suggests that passive prediction becomes more difficult when outcomes depend on non-rigid state changes or indirect tool–object interactions.

Compositional interactions. Compositional tasks introduce an additional temporal challenge. Short-horizon compositions already show lower object plausibility, interaction quality, and task alignment than the strongest atomic categories, while long-horizon compositions further degrade across all four dimensions. In particular, task alignment decreases to 0.2443 and interaction quality to 0.2375 for long-horizon tasks. These results show that the interaction taxonomy exposes failure modes that are not visible from aggregate Level-1 scores alone, especially for complex physical mechanisms and extended interaction sequences.

Table 11: Category-level average VLM-based evaluation results for Level-1. Higher values indicate better performance.

Eval Level	Interaction Category	Robot Arm Integrity	Object Plausibility	Interaction Quality	Task Alignment
Atomic	Rigid-body Pose Transformation	0.3763	**0.4522**	0.3567	0.3400
	Constrained Motion	**0.4189**	0.4232	**0.3807**	**0.3663**
	Deformation and Topological Interaction	0.3804	0.3072	0.2848	0.3160
	Tool-mediated Dynamics	0.3476	0.2714	0.2497	0.2928
Compositional	Short-Horizon	0.3495	0.2565	0.2543	0.2748
	Long-Horizon	0.3322	0.2408	0.2375	0.2443

F Details of Real-Robot Evaluation

F.1 Robot Platform and Observation Setup

All real-world experiments are conducted on a dual-arm PiperX platform from AgileX Robotics. The two arms are mounted in a fixed tabletop workspace and observed by three RGB cameras: one overhead head camera and two wrist cameras mounted near the end-effectors. The head camera provides a global third-person view with a 110° field of view, while the wrist cameras provide close-up views of local interactions with a 160° field of view. All cameras operate at 30 FPS with a resolution of 640×360. Together, the three synchronized views capture both global scene context and fine-grained manipulation details.

F.2 Task Set and Evaluation Protocol

The real-robot benchmark contains 50 tasks: 20 atomic tasks, 20 compositional tasks, and 10 compositional OOD tasks. Atomic tasks follow the four interaction categories defined in Sec. 3.2. Compositional tasks combine atomic skills into multi-step procedures, while compositional OOD tasks contain held-out skill–object pairings or skill combinations. Full task instructions, trajectory-level annotations, and collection identifiers are provided with the released dataset.

Each task is evaluated over ten physical rollouts, resulting in 500 rollouts per model. For slide and hinge manipulation, opening and closing are treated as two directional conditions with five trials each. They are displayed separately in Table 9 but combined into a single task before task-level and split-level aggregation. Consequently, the 20 atomic tasks appear as 22 rows in the table.

Success and progress scoring. Each rollout receives a binary success label $z_i \in \{0, 1\}$ and a progress score $s_i \in [0, 5]$, where 5 denotes full task completion. For a task with n evaluated rollouts, we compute

$$\mathrm{SR} = \frac{100}{n} \sum_{i=1}^{n} z_i, \qquad \mathrm{Progress} = \frac{100}{5n} \sum_{i=1}^{n} s_i. \tag{18}$$

The per-task table reports raw success counts and cumulative progress, whereas split-level and overall results are normalized to 0–100. A ten-trial task therefore has a maximum cumulative progress of 50, and each five-trial directional condition has a maximum of 25. All aggregate results assign equal weight to each benchmark task.

F.3 Implementation Details

All models are trained on the same **Phys-WorldBench** training split. We reproduce $\pi_{0.5}$, Cosmos-Policy and FastWAM using their released implementations and adapt them to the benchmark interface. Camera streams and action interfaces are aligned across models, while each model retains its native action representation. All predicted actions are converted to the dual-arm end-effector pose representation required by the PiperX control interface before execution. For a fair comparison, we use no task-specific prompting, skill-specific controller tuning, or model-specific evaluation heuristics.

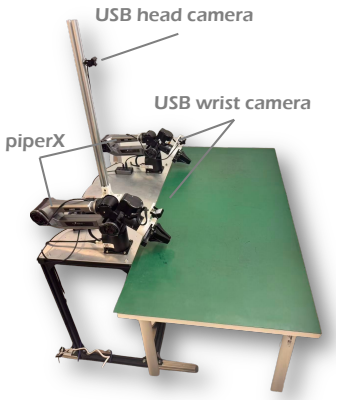


Figure 8: **Real-robot evaluation setup.** The dual-arm PiperX platform is observed by one overhead camera and two wrist-mounted cameras.

F.4 Per-Task Results

Table 9 reports the complete per-task results underlying the aggregate results in Table 5 of the main paper. We compare five policies: $\pi_{0.5}$, Cosmos-Policy, Cosmos-Policy initialized from Cosmos-Predict-2.5, Motus, and FastWAM. For each task, **SR** denotes successful rollouts out of the evaluated physical trials, and **Prog.** denotes cumulative progress.

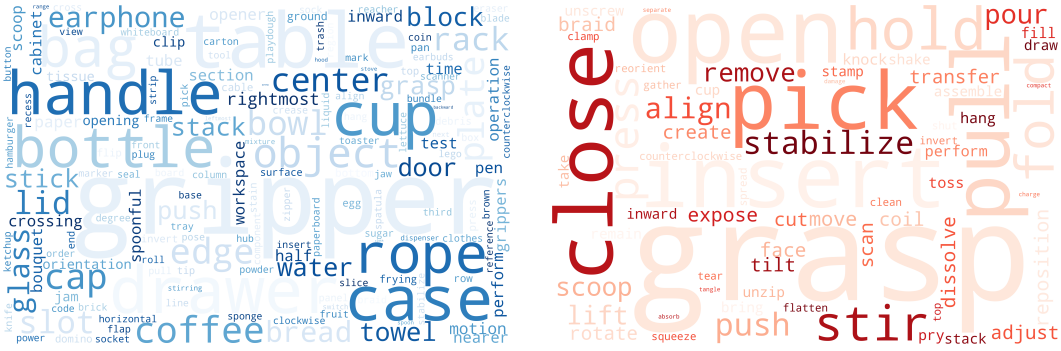
F.5 Uncertainty and Statistical Analysis

Confidence intervals. Each model is evaluated on 500 real-robot rollouts, with ten trials for each of the 50 tasks. We report Wilson 95% confidence intervals for aggregate success rates as a binomial reference. Because task difficulty varies substantially and trials from the same task share the same physical setup, we additionally report task-level bootstrap confidence intervals.

Specifically, we perform 20,000 bootstrap resamples by sampling tasks with replacement while retaining all ten trials of each selected task as one unit. For split-level estimates, tasks are resampled within the corresponding split; for overall estimates, tasks are resampled from all 50 tasks. The 2.5th and 97.5th percentiles of the bootstrap distribution define the reported 95% confidence intervals. These intervals capture variability across evaluated tasks, but do not capture variation across independent training runs.

Paired policy comparisons. For pairwise comparisons, we use the same resampled task indices for both models and compute the difference in mean task success rate. Relative to FastWAM, $\pi_{0.5}$ improves success by 12.6 percentage points (95% paired bootstrap CI: [7.6, 17.6]); relative to Cosmos-Policy, the difference is 18.4 points ([12.2, 25.0]). FastWAM exceeds Cosmos-Policy by 5.8 points ([3.4, 8.4]). All three paired intervals exclude zero on the evaluated task set.

Zero observed successes. The four WAM-based policies obtain no successful rollouts across the 100 compositional-OOD evaluations per model. Their task-bootstrap intervals therefore collapse to [0,0], which reflects the absence of observed successes in the evaluated tasks rather than certainty that the underlying success probability is zero. For reference, the corresponding Wilson 95% interval for 0 successes out of 100 trials is [0.0%, 3.7%].



(a) Nouns (b) Verbs

Figure 9: **Word clouds of language instructions.** Nouns cover diverse objects, object parts, tools, and spatial references, while verbs emphasize manipulation-oriented actions, indicating rich language grounding over both object-centric references and action-centric intents.

G Computational Resources

We summarize the computational resources required for our evaluation and fine-tuning experiments as follows:

- **Model Fine-tuning (Level 2 & 3):** For each model requiring fine-tuning (e.g., actionable world models and VLA policies), the training process was conducted on a cluster of **8**× **NVIDIA A100 GPUs**. On average, each training run took approximately **72 hours** to reach convergence.
- **Model Inference (Level 1):** For the evaluation of video generation and world modeling capabilities, inference for each model (e.g., Wan 2.6, Veo 3.0, Cosmos-Pred) was performed using a **single NVIDIA A100 GPU**. Depending on the model architecture and sampling complexity, the total inference time per model across the benchmark ranged from **12 to 24 hours**.

All experiments were conducted in a high-performance computing (HPC) environment with consistent software stacks (CUDA 12.4 and PyTorch 2.6) to ensure the stability of the benchmarks.