PLAYWORLD: BENCHMARKING WORLD MODELS WITH AGENT PLAYERS OVER LONG-HORIZON OBJECTIVES

Anonymous authors
Paper under double-blind review

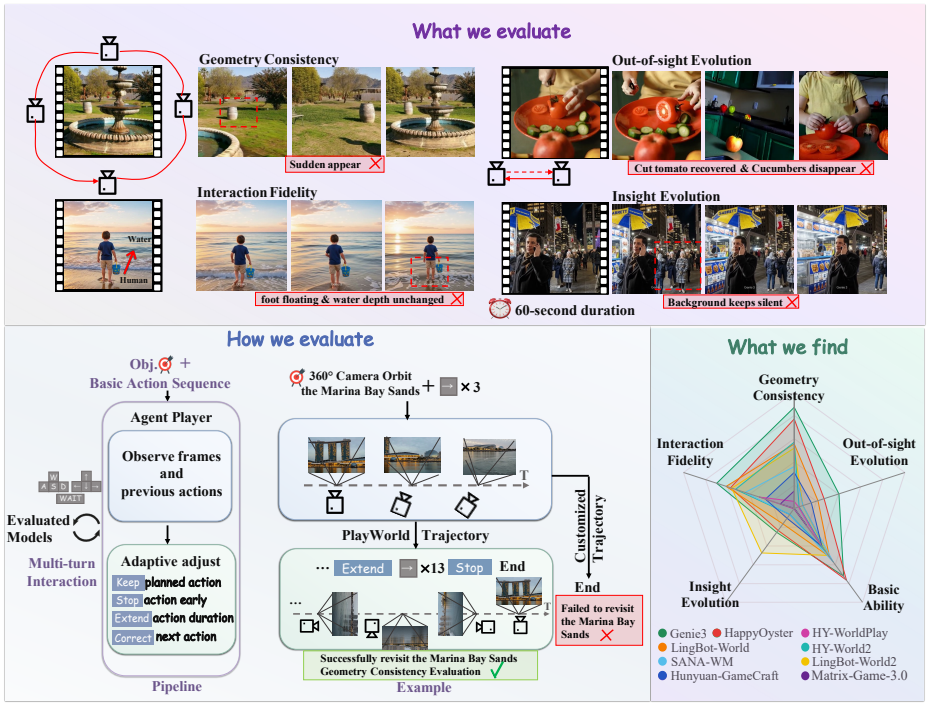

Figure 1: **PlayWorld evaluates world models from the perspective of a human player.** We define long-horizon objectives such as "turning around 360 degrees" or "walking into the water", and use an Agent Player to interact with the world models. We then evaluate them from multiple perspectives, including geometry consistency, interaction fidelity, and persistent state evolution.

ABSTRACT

Video world models simulate future states conditioned on current observations and user actions. Recent systems have demonstrated impressive video consistency and action controllability over long sequences. However, fairly comparing these interactive models remains challenging. In practice, a human player typically evaluates a world model by pursuing long-horizon objectives through interaction. For example, a user may turn around 360 degrees to see whether the environment remains consistent, or walk into the water and inspect whether realistic water ripples are generated. The action sequence required to achieve the same objective may vary substantially between models, making fixed action-conditioned evaluation unsuitable for cross-model comparison. To address this, we employ multi-modal Agent Players to interact with world models toward specified long-horizon objectives. Building on this paradigm, we introduce PlayWorld, a benchmark providing 171 scenarios, each with a specified objective. To evaluate performance thoroughly, we assess models along four core dimensions: geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution. In addition, we incorporate basic ability metrics for video quality and controllability. Experiments across nine state-of-the-art world models reveal that current models remain unreliable on long-horizon interactive objectives, particularly in maintaining spatial consistency and persistent state evolution. Code and data are provided in the anonymous supplementary material. More results are available on our project page (https://playworld-2026.github.io/project/PlayWorld/).

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.

1 INTRODUCTION

General video generation models have advanced rapidly in recent years, achieving remarkable visual quality and photorealism (OpenAI, 2024; Wan et al., 2025; Seedance et al., 2026). Video world models (Bruce et al., 2024; Google DeepMind, 2025; Alibaba Cloud Community, 2026) go a step further by simulating interactive environments that users can explore while maintaining causally consistent evolution across space and time. This turns them into open worlds that users can freely interact with.

As a rapidly growing body of work pursues this goal (Zhu et al., 2026b; Li et al., 2025; He et al., 2025; Wang et al., 2026; Gao et al., 2026; Zhu et al., 2026a; Sun et al., 2025b), evaluation has become an increasingly important challenge. General video-generation benchmarks such as VBench (Huang et al., 2024) assess aesthetics, temporal coherence, and text alignment, but do not measure whether a model responds correctly to user actions. World-model benchmarks further evaluate 3D consistency, memory consistency, interaction, and out-of-sight state evolution (Duan et al., 2025; Ma et al., 2026; Chen et al., 2026b; Xu et al., 2026b; Wu et al., 2026). However, these benchmarks typically rely on customized trajectories predefined for individual evaluation cases, as summarized in Tab. 1.

This evaluation paradigm does not fully match how human players judge interactive world models. In practice, a human user usually evaluates a world model by pursuing a high-level, long-horizon objective rather than by checking whether a predefined action trajectory is followed. For example, a customized trajectory may specify three $\rightarrow$ commands with the intent of rotating the camera by 360° and returning to the initial view of Marina Bay Sands, as illustrated in Fig. 1. Because action granularity varies across models, the same three commands may complete a full rotation in one rollout but only a partial turn in another. Marina Bay Sands may therefore leave and re-enter the view in some rollouts but fail to reappear in others, making the resulting geometry consistency scores incomparable. Such customized trajectories also cannot reliably complete more complex action patterns, such as orbiting a sculpture while maintaining a fixed viewing direction (Fig. 4), because the required action duration and stopping point differ across models. Moreover, web-based closed-source world models such as Genie 3 (Google DeepMind, 2025) and HappyOyster (Alibaba Cloud Community, 2026) are accessible only through a web interface and require direct human interaction, which is labor-intensive and prone to bias. These challenges call for an automated evaluation protocol that can interact with world models like a human player.

To address this challenge, we introduce PlayWorld, which evaluates world models using long-horizon objectives. PlayWorld uses Agent Player, a multi-modal agent that simulates a human player. Each evaluation case provides a basic action sequence, which serves as a shared initial reference across world models. Given this reference, the Agent Player observes the generated frames and action history at each step and adaptively adjusts execution through *Keep*, *Stop*, *Extend*, *Correct*, or *End* decisions. These decisions adapt the number and duration of actions to each model's response, enabling fairer comparison under the same objectives. PlayWorld further introduces a VQA rubric verifier for open-ended interactive scenarios, covering four core world-model capabilities: geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution. The overall evaluation framework assesses model behavior over rollouts lasting approximately 10 to 60 seconds using continuous W/A/S/D, $\uparrow/\downarrow/\leftarrow/\rightarrow$, and WAIT controls. The benchmark contains 171 human-annotated cases evaluated across nine representative world models, revealing the capability boundaries of current systems.

2 RELATED WORK

World models. Video world models are rapidly developing across diverse application domains (Bruce et al., 2024; He et al., 2025). In autonomous driving, world models predict how traffic scenes change in response to ego motion, vehicle controls, and surrounding agents. GAIA-1 (Hu et al., 2023), DriveDreamer (Wang et al., 2023), DrivingWorld (Hu et al., 2024), and Vista (Gao et al., 2024) generate future observations for scenario simulation, planning, and policy evaluation. In robotics, world models instead predict future visual observations conditioned on actions, context, or task instructions. IRASim (Zhu et al., 2024), Cosmos (NVIDIA et al., 2025), RoboScape (Shang et al., 2025), and LVP (Chen et al., 2026a) use these predictions to support robot learning, planning, and control. Beyond these domain-specific applications, general-purpose interactive world models generate visual environments continuously in response to user actions rather than producing a fixed video from a text or image prompt. Representative systems include Matrix-Game (He et al., 2025; Wang et al., 2026), YUME (Mao et al., 2025a;b), WorldPlay (Sun et al., 2025b), LingBot-World (Robbyant Team

Table 1: **Comparison of world-model evaluation benchmarks.** We compare input formats, closed-loop adaptation, long-horizon revisitation, evaluated capabilities, and rollout duration. Objective denotes a scene-grounded high-level goal, while 🎥 and 🎮 denote camera and action inputs. For PlayWorld, the Objective is accompanied by a basic action sequence that serves only as the Agent Player's initial reference; State Evolution includes visible and out-of-sight evolution. ✓, ✗, and ∼ indicate full, no, and partial coverage, respectively.

Benchmark	Input	Closed-Loop Adaptation	Long-horizon Revisit	Geometry Consistency	State Evolution	Interaction Fidelity	Time Scale / Video
WorldScore (Duan et al., 2025)	Text + Image + 🎥	✗	✗	✓	✗	✗	∼2–10 s
WorldMark (Xu et al., 2026b)	Image + 🎮	✗	∼	✓	✗	✗	20 / 40 / 60 s
WBench (Ying et al., 2026)	Text + Image + 🎥/🎮	✗	∼	✓	∼	✓	10–45 s
WorldRoamBench (Xu et al., 2026a)	Image + 🎮	✗	∼	✓	∼	✓	10–60 s
MemoBench (Chen et al., 2026b)	Text + Image + 🎥	✗	✗	✓	✓	✗	4–6 s
Omni-WorldBench (Wu et al., 2026)	Text + Image + 🎥	✗	✗	∼	✓	✓	3–6 s
PlayWorld (Ours)	Text + Image + Objective	✓	✓	✓	✓	✓	10–60 s

et al., 2026), LingBot-World2 (Gao et al., 2026), SANA-WM (Zhu et al., 2026a), HY-World (Tencent Hunyuan, 2025), HY-World2 (Team HY-World et al., 2026), and Hunyuan-GameCraft (Li et al., 2025). These models progressively generate the environment in response to movement keys, mouse or gamepad controls, camera trajectories, or text instructions. Closed-source systems such as Genie 3 (Google DeepMind, 2025) and HappyOyster (Alibaba Cloud Community, 2026) provide similar interaction through user-facing web interfaces, although their implementation details remain unavailable. These systems use different architectures and control formats but share the goal of generating worlds that respond continuously and coherently to user actions. Open-ended exploration and adventure scenarios place users in direct control of the generated environment. We therefore consider them a fundamental setting.

Evaluation of world models. Existing world-model benchmarks evaluate several complementary capabilities. VBench (Huang et al., 2024) and T2V-CompBench (Sun et al., 2025a) measure perceptual quality, temporal stability, and compositional alignment, but do not assess responses to user actions. WorldSimBench (Qin et al., 2024) evaluates perceptual quality and downstream embodied-control performance, without adapting evaluation controls according to generated observations. WorldScore (Duan et al., 2025) and WorldMark (Xu et al., 2026b) assess camera control and multi-view geometry under predefined camera trajectories or action sequences. Their results may conflate geometric inconsistency with trajectory failure when identical controls produce different movement magnitudes across models. MIND (Ye et al., 2026), MemoBench (Chen et al., 2026b), and STEVO-Bench (Ma et al., 2026) evaluate memory and out-of-sight evolution through revisitation, visible–disappear–reappear processes, and lookaway controls. These protocols follow fixed schedules that cannot account for model-specific action granularity and response speed. Omni-WorldBench (Wu et al., 2026) broadens evaluation to interaction, physics, and memory, while WBench (Ying et al., 2026) and WorldRoamBench (Xu et al., 2026a) introduce predefined multi-turn navigation patterns. However, these patterns cannot be flexibly composed into diverse long-horizon trajectories or adjusted online when the generated observation deviates from the expected state. Overall, existing benchmarks predominantly drive world models with predefined low-level controls. Because models differ in action granularity and response dynamics, the same controls may fail to reach comparable evaluation states. PlayWorld specifies a shared long-horizon objective and uses a multi-modal Agent Player to adapt action execution online. This closed-loop protocol preserves a consistent evaluation intent across models while supporting more diverse long-horizon interaction.

3 PLAYWORLD

To better simulate how human players evaluate world models, PlayWorld introduces a multi-modal Agent Player that interacts with each evaluated model and adaptively adjusts action execution. The PlayWorld benchmark organizes long-horizon objectives across diverse scenario types. Its VQA rubric verifier assesses four core world-model capabilities: geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution. We also combine basic quality metrics for complementary evaluation.

3.1 AGENT PLAYER DESIGN

The Agent Player consists of a replaceable multi-modal agent model (*e.g.*, Claude or Gemini) and an agent interface. The agent model observes generated frames and execution history, then decides

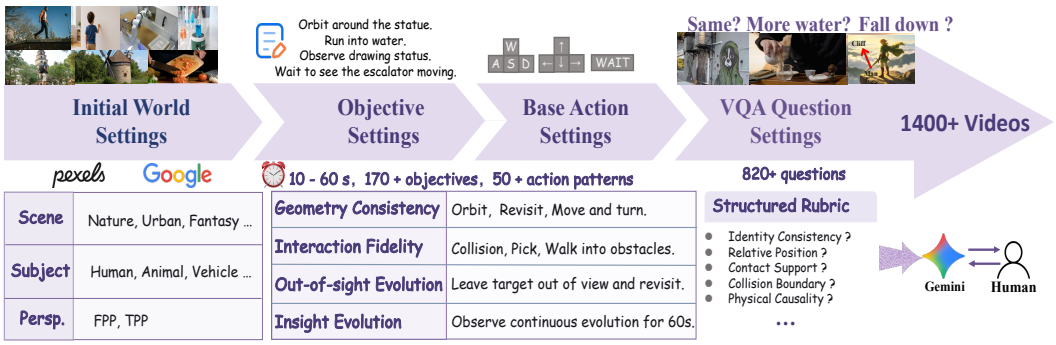

Figure 2: **Benchmark construction pipeline of PlayWorld.** Starting from diverse initial worlds, annotators define long-horizon objectives based on scenario-specific attributes, write basic action sequences, and construct sample-specific VQA rubrics. The resulting benchmark spans over 170 human-annotated cases and 50 action patterns over 10–60-second rollouts, with more than 820 questions applied to over 1,400 interactive videos.

how to adapt action execution toward the long-horizon objective. The agent interface executes these decisions on the evaluated world model, captures the resulting observation, and returns it to the agent model, forming a closed interaction loop.

Agent model decision. A straightforward way to instantiate the Agent Player is to let the agent plan the trajectory from scratch and interact with the world model without any human-provided action input. However, because each decision requires the agent model to reason over the current observation, fully online planning introduces substantial decision latency and slows action execution. It may also produce highly different trajectories across world models with different capabilities, making cross-model comparison less consistent. We therefore use a human-annotated basic action sequence as a shared initial reference. This preserves a consistent evaluation intent across models, reduces the burden of open-ended sequential planning, and still allows observation-conditioned adaptation. As shown on the left of Fig. 1, the Agent Player receives the long-horizon objective and the shared basic action sequence. At each interaction step, the agent model observes the generated frame, the previously executed actions, the scene description, and the objective. Based on this context, it decides whether to keep, stop, extend, correct, or end the current rollout. *Keep* retains the planned action, continuing the current hold or advancing to the next action as scheduled. *Stop* terminates the active action early when the intended visual state has been reached. *Extend* increases the action hold duration when additional movement is required. *Correct* revises or skips the next planned action when the observed state no longer matches the reference sequence. *End* terminates the evaluation case only when the Agent Player determines that the long-horizon objective and its required observation have been completed. For interaction fidelity cases involving an obstacle, reaching the obstacle does not trigger *End*; the Agent Player continues the forward action so that the rollout reveals whether the subject collides with the obstacle realistically or penetrates it.

Agent interface execution. The agent interface directly interacts with the evaluated world model. It receives an initial frame and a human-annotated basic action sequence expressed using W/A/S/D, ↑/↓/←/→, and WAIT, and translates these actions into the world model's native controls. At each step, it executes the active action, captures the resulting generated frame, returns the frame to the agent model, and applies the returned decision to update subsequent execution. This process continues in a closed loop until the Agent Player returns *End* or the 40-step interaction budget is exhausted. For web-based world models, the agent interface uses browser automation to dispatch controls and capture generated frames directly from the model interface. Further details are provided in the supplementary material.

3.2 PlayWorld Benchmark Overview

Benchmark construction. To support long-horizon evaluation, we construct the PlayWorld benchmark around three core elements: an initial world, a scene-grounded objective, and a basic action sequence. As shown in Fig. 2, we first collect candidate images from Pexels (Pexels, 2026) and Google Images (Google, 2026b), covering natural, urban, and fantasy environments with various subjects such as humans, animals, and vehicles. Annotators then screen the images for clear scene structure and select first-person or third-person views according to the target capability and scene

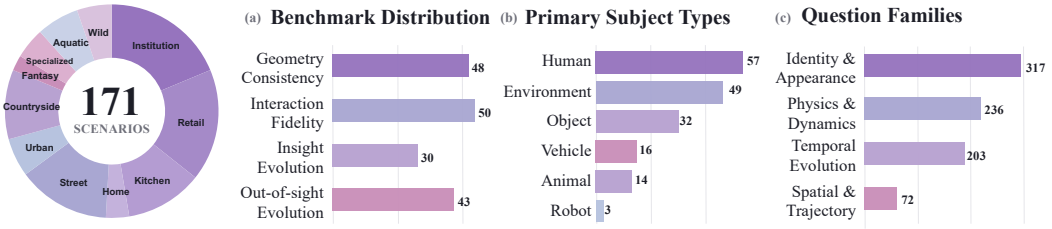

Figure 3: **Composition of PlayWorld.** The benchmark comprises over 170 human-annotated cases across diverse scenario categories. The cases span four evaluation dimensions and diverse subject types, with dimension-specific VQA questions tailored to their evaluation criteria.

content. For each selected image, Gemini 3.1 Pro (Google, 2026a) generates an environment caption, which is verified and revised by human annotators. Annotators further define a long-horizon objective based on scenario-specific attributes, specify a visually verifiable completion condition, and write a basic action sequence. Finally, annotators construct sample-specific VQA rubrics for the target evaluation dimension. All cases are manually curated to ensure that the expected outcome is observable and suitable for dimension-specific evaluation.

Benchmark composition. Fig. 3 summarizes the composition of PlayWorld. The benchmark comprises over 170 human-annotated cases covering 50 action patterns, ranging from a single action to combinations of up to five actions and producing 10–60-second rollouts. Evaluating nine world models yields over 1,400 interactive videos, which are assessed using more than 820 task-conditioned VQA questions. The cases span four evaluation dimensions and cover six primary subject types. The VQA questions are organized into four evidence families: identity and appearance, physics and dynamics, temporal evolution, and spatial and trajectory reasoning. Fig. 4 provides representative examples for each dimension. Geometry consistency cases emphasize revisitable visual anchors such as lettering and buildings with distinctive shapes; interaction fidelity cases include walking into water, fire, cliffs, or grass; out-of-sight evolution cases emphasize continued changes across disappear–reappear intervals, such as a person peeling an apple, making dessert, or painting a picture; and insight evolution cases focus on long visible processes without an explicit action trajectory, such as shoppers queuing for checkout or a chef plating food.

3.3 VQA Rubric Verifier

Automated metrics primarily capture pixel-level fidelity and low-level perceptual quality, but often fail to measure consistency under large viewpoint changes, assess interaction plausibility, or determine whether long-horizon objectives are completed. We therefore design a VQA rubric verifier, where Gemini 3.1 Pro (Google, 2026a) answers sample-specific Yes/No rubric questions for each generated rollout. Each rubric assigns greater importance to questions that are more relevant to the target capability and excludes questions that are not applicable to the case. Because instruction-following failures can make memory evaluation unreliable, as discussed in MemoBench (Chen et al., 2026b), the verifier uses dimension-specific validation as a gate before rubric scoring. The weighted Yes/No outcomes are aggregated and converted into a case-level score on a 1–5 scale (↑: higher is better), as reported in Tab. 2. The evaluation covers four dimensions:

Geometry consistency. To quantitatively assess long-horizon geometry consistency, the rubric considers whether the environment preserves its structure during camera movement and revisitation. We first verify *Trajectory Validity* by determining whether the rollout follows the intended path and reaches the required viewpoint. Once the rollout passes this validation, its sample-specific VQA rubric covers two aspects. *Scene Identity* checks the preservation of object identity, count, texture, material, and color, such as whether a wall mural preserves its color and shape after a 360-degree turn or whether a revisited sculpture remains unchanged. *Spatial Consistency* considers relative object positions and the overall scene layout, for example whether a red car that is initially to the right of a green car appears in the correct left-right order after the camera crosses the street and looks back.

Interaction fidelity. For interaction fidelity, we examine whether the intended interaction occurs and elicits a physically plausible response. Before rubric scoring, *Subject and Reachability* determines whether the controlled subject responds to the actions and reaches the intended interaction region. A completely static subject fails this validation because the intended interaction is never tested. The subsequent rubric covers *Contact and Collision*, *Motion and Causality*, and *Visual Response*. For



Figure 4: **Representative cases in PlayWorld.** Each case combines an initial world, a scene-grounded long-horizon objective, and a sample-specific VQA rubric. The expected trajectory is visualized alongside each case; for insight evolution, the camera remains stationary and observes the scene for 60 seconds.

instance, it asks whether the subject collides with a wall rather than passing through it and whether its motion follows plausible kinematics, such as maintaining grounded footsteps without floating. It also checks whether entering water produces appropriate physical feedback, such as visible ripples, splashes, or changes in apparent water depth around the subject.

Out-of-sight evolution. Out-of-sight evolution focuses on whether a target preserves its identity and continues to change while unobserved. We first use *Trajectory Validity* to determine whether the target leaves the field of view and subsequently reappears as required. For eligible rollouts, *Reappearance Consistency* asks whether the revealed target is the same entity observed before it disappeared. *Hidden-State Evolution* determines whether the target state progresses plausibly during the unobserved interval, such as whether an in-progress drawing gains additional content while it is out of sight. *Physical Causality* checks whether the revealed state can reasonably result from that hidden evolution, such as a person continuing to move upward on an escalator, blue ink spreading through clear water, or a cup becoming fuller as water is poured into it.

Insight evolution. Insight evolution focuses on continuously observed phenomena without a fixed action trajectory. The rubric evaluates a visible subject or process from three aspects. *Identity and State Progression* checks whether the relevant person or object remains present and whether its state progresses over time, rather than suddenly disappearing. *Motion and Physical Plausibility* identifies static behavior, repetitive but non-progressive actions, abrupt transitions, and implausible dynamics. For example, if a person is eating one of three buns, the rubric checks whether the bun being eaten gradually changes and whether the number or appearance of the remaining buns evolves consistently. Meanwhile, *Temporal Scene Consistency* detects unintended changes to the surrounding environment. Because this dimension uses stationary observation, it does not require trajectory validation.

3.4 Basic Ability Evaluation

We further use automatic metrics to test video quality and controllability. We choose Aesthetic Quality, Imaging Quality, Motion Smoothness, and Temporal Flickering from VBench (Huang et al., 2024), Temporal Consistency from Omni-WorldBench (Wu et al., 2026), Depth Stability from MemoBench (Chen et al., 2026b), and Subject Consistency from HyDRA (Chen et al., 2026c). These video quality metrics do not require ground-truth videos and mainly capture frame-level or inter-frame signals. All video quality metrics are normalized to $[0, 1]$ and reported as percentages. We then evaluate Action Controllability using Translation Pass Rate and Rotation Pass Rate, following WorldMark (Xu et al., 2026b). Camera poses are estimated by VGGT (Wang et al., 2025) from

Table 2: **Rubric-based VQA evaluation across four dimensions on the PlayWorld benchmark.** Scores range from 1 to 5 (↑: higher is better), and Validity is reported as a percentage. Overall is the unweighted mean across the four evaluation dimensions; Validity averages the three movement-dependent dimensions. **Bold**: best; underline: second best.

Model	Validity (%)	Geometry Consistency	Interaction Fidelity	Insight Evolution	Out-of-sight Evolution	Overall
Closed-source Models						
Genie 3	**87.1%**	**2.74**	**2.40**	1.51	**1.81**	**2.12**
HappyOyster	79.6%	2.54	2.15	1.47	1.54	1.92
Open-source Models						
LingBot-World	72.7%	2.11	2.23	1.33	1.43	1.78
LingBot-World2	78.8%	2.04	2.13	**1.95**	1.16	1.82
HY-World2	51.7%	2.14	2.06	1.13	1.09	1.61
SANA-WM	80.4%	1.72	1.89	1.13	1.16	1.48
Hunyuan-GameCraft	61.5%	1.62	1.52	1.21	1.31	1.42
HY-WorldPlay	41.6%	1.12	1.63	1.01	1.08	1.21
Matrix-Game-3.0	68.2%	1.30	1.25	1.00	1.00	1.14

uniformly sampled frames, while target trajectories are defined by the executed actions, including online adjustments made by the Agent Player. A rollout passes the translation or rotation criterion when the corresponding error falls below an acceptable threshold. Both pass rates are computed over rollouts with valid estimates and reported as percentages, where higher values indicate better controllability. Tab. 6 reports all nine higher-is-better metrics together with a Basic Ability Score, which converts their average rank into a percentage.

4 Experiments

4.1 Models and Evaluation Protocol

Evaluation models. We evaluate nine representative world models. Five models are evaluated through their web-based interfaces to reproduce the user-facing interaction setting: Genie 3 (Google DeepMind, 2025), LingBot-World (Robbyant Team et al., 2026), LingBot-World2 (Gao et al., 2026), HY-World2 (Team HY-World et al., 2026), and HappyOyster (Alibaba Cloud Community, 2026). Four models are executed locally through chunk-wise generation: SANA-WM (Zhu et al., 2026a), Hunyuan-GameCraft (Li et al., 2025), HY-WorldPlay (Sun et al., 2025b), and Matrix-Game-3.0 (Wang et al., 2026). All models are controlled by the same Agent Player, whose agent interface converts the decisions into each model's native input format. Implementation details and generation configurations are provided in the supplementary material.

Evaluation protocol. For each evaluation case, all models receive the same initial world, long-horizon objective, and basic action sequence. The same Agent Player observes the generated frames and adjusts action execution for every model. We evaluate the resulting rollouts, which last approximately 10–60 seconds, using the rubric verifier in Sec. 3.3 and the basic ability metrics in Sec. 3.4. The rubric verifier reports geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution, while the basic ability evaluation reports Video Quality and Action Controllability. We also report dimension-specific validation pass rates in Tab. 3. Specifically, geometry consistency is validated using Trajectory Validity, interaction fidelity using Subject and Reachability, and out-of-sight evolution using Trajectory Validity. These three validation criteria are evaluated independently. If a rollout does not pass the corresponding validation criterion, its score for that dimension is set to the minimum value of 1. The insight evolution dimension uses stationary observation and therefore does not require trajectory validation.

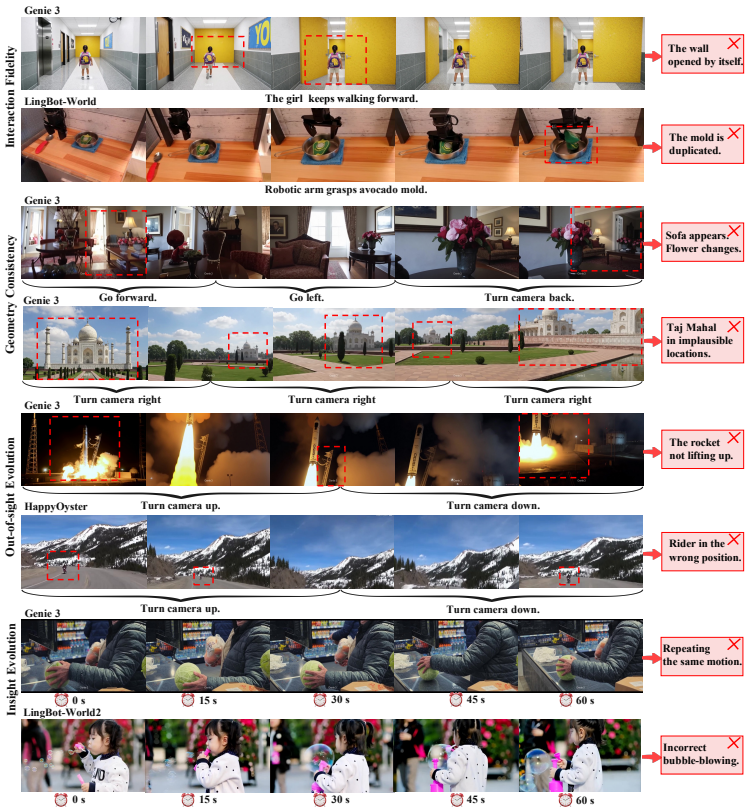

Figure 5: **Representative failures across four evaluation dimensions.** Current models can follow camera or action controls but struggle with coherent, realistic world dynamics.

4.2 Agent Player Analysis

Control strategy. Existing benchmarks typically execute a predefined trajectory for each evaluation case. We retain a basic action sequence as a shared initial reference but adapt its execution online to account for differences in action granularity across world models. Directly executing the basic action sequence without online adaptation (Preset Only) may cause the same controls to overshoot or undershoot the intended state. Agent Only instead predicts every action online from the latest generated observation and previously executed actions, but planning each action from scratch increases the sequential inference burden and slows trajectory execution. Preset + Agent combines both sources of guidance: it uses the basic action sequence as a reference while adapting its execution to the generated observations and model-specific control granularity. We compare these three strategies on the same 25 cases for Genie 3 and HappyOyster.

As shown in Tab. 4, Preset + Agent achieves the highest Trajectory Score and Human Preference on both world models, confirming the benefit of combining a shared action prior with online adaptation. Its Agent-modified Action Ratio remains within 10–20% on both models, showing that the Agent makes targeted modifications. By adapting action execution online, PlayWorld reaches the intended objective more reliably than executing a predefined trajectory alone.

Agent model selection. We compare three agent models under Preset + Agent on identical Genie 3 cases (Tab. 5). Their similar trajectory quality is consistent with adapting a basic action sequence rather than planning from scratch. Claude Sonnet leads in Trajectory Score, whereas Claude Haiku achieves the highest Human Preference and lowest measured Decision Latency. We select Haiku for its quality–efficiency balance and allow future upgrades. Latency remains dependent on provider infrastructure and network conditions.

4.3 Observations in World Models

Sustained world evolution remains the primary bottleneck. Tab. 2 shows that Genie 3 leads overall, while out-of-sight and insight evolution remain weak across models. Generated processes

Table 4: Comparison of trajectory control strategies on Genie 3 and HappyOyster.

Setting	*Genie 3*				*HappyOyster*			
	Trajectory Score ↑	Human Preference ↑	Majority Agreement	Agent-modified Action Ratio	Trajectory Score ↑	Human Preference ↑	Majority Agreement	Agent-modified Action Ratio
Preset Only	0.92	39.6%	88.0%	0.0%	1.00	40.8%	76.0%	0.0%
Agent Only	0.88	29.2%	84.0%	100.0%	0.68	24.4%	92.0%	100.0%
Preset + Agent	**1.08**	**65.6%**	86.0%	12.0%	**1.12**	**67.4%**	84.0%	14.9%

Table 5: Comparison of agent models on Genie 3 under Preset + Agent.

Agent Model	Trajectory Score ↑	Human Preference ↑	Majority Agreement	Agent-modified Action Ratio	Decision Latency (s/call) ↓
Claude Haiku	1.08	**57.8%**	88.9%	12.0%	**3.83**
Claude Sonnet	**1.24**	45.3%	94.1%	12.4%	6.21
Gemini 3.1 Pro	1.08	46.5%	94.1%	12.6%	4.36

may freeze or reset, and occluded targets may return unchanged, misplaced, or replaced, exposing limited persistent state modeling.

Long-horizon revisitation exposes global spatial inconsistency. Fig. 5 shows that locally plausible frames accumulate into global contradictions: during orbit and revisit cases, landmarks are regenerated at incompatible positions rather than anchored to a persistent 3D world.

Interaction fidelity remains limited beyond simple collision responses. Models may handle simple collisions, yet still pass through obstacles, miss expected water responses, or fail to bind actions consistently to the controlled character.

Trajectory control and world-model capability can diverge. SANA-WM attains a relatively high validation pass rate (Tab. 3) but modest rubric scores, showing that reaching the intended region does not ensure memory, spatial consistency, or plausible interaction.

Automatic metrics only test the basic ability of interactive world models. High Video Quality or action-control scores do not guarantee completion of a long-horizon objective; the full basic-ability results are reported in Tab. 6.

4.4 Human Validation

We conduct a human evaluation to examine whether the VQA metrics align with human perception. Participants provide 600 valid pairwise judgments on video pairs evenly sampled across the four dimensions. We aggregate pairwise wins for each model within each dimension and normalize the resulting preference scores. Each point in Fig. 6 represents an evaluated model; the red line shows a linear fit, and the shaded region denotes its confidence band. Spearman's ρ measures rank agreement across the nine models for each evaluation dimension and the Overall score. The consistently positive correlations show that the VQA metrics closely preserve human preference rankings. Inter-rater agreement results are reported in Tab. 7. Further details are provided in Appendix E.

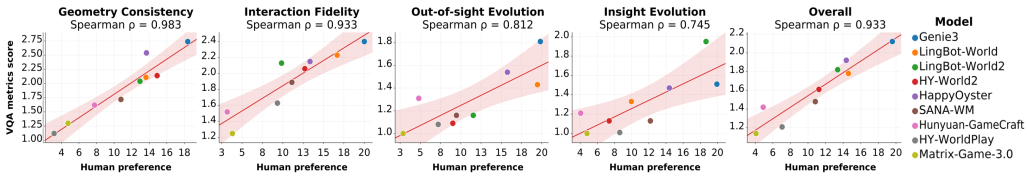

Figure 6: Alignment between human preferences and VQA metrics.

5 Conclusion

We introduce PlayWorld and use agent-as-player to adapt actions to each model's response, enabling fair long-horizon comparison of interactive world models. By exposing failures in geometry consistency, interaction fidelity and state evolution, we provides a reproducible testbed that guides future video world models beyond visual quality toward coherent, interactive world simulation.

AI USE STATEMENT

AI tools were used to assist with language polishing and improve the clarity of this manuscript. The authors reviewed and verified all AI-assisted edits and take full responsibility for the manuscript's accuracy, originality, and academic integrity.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code in the supplementary material. Implementation details, evaluation protocols, and agent interface settings are described in the main paper and the appendix.

REFERENCES

Alibaba Cloud Community. Alibaba launches happyoyster, a world model product for real-time immersive creation and interaction. https://www.alibabacloud.com/blog/alibaba-launches-happyoyster-a-world-model-product-for-real-time-immersive-creation-and-interaction_603048, April 2026. Alibaba Cloud Community blog post.

Anthropic. Claude haiku 4.5. `https://platform.claude.com/docs/en/about-claude/models/overview`, 2025. Claude model documentation, model ID: claude-haiku-4-5-20251001.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments, February 2024.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Caiyi Zhang, Peihao Li, Kiwhan Song, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large Video Planner Enables Generalizable Robot Control, May 2026a.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiyue Li, Alan Yuille, Paul Pu Liang, and Yilun Du. MemoBench: Benchmarking World Modeling in Dynamically Changing Environments. *arXiv preprint arXiv:2606.27537*, 2026b.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of Sight but Not Out of Mind: Hybrid Memory for Dynamic Video World Models, 2026c.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A Unified Evaluation Benchmark for World Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions, 2026. URL `https://arxiv.org/abs/2607.07534`.

Google. Gemini 3.1 pro. `https://ai.google.dev/gemini-api/docs/models`, 2026a. Gemini API model documentation.

Google. Google images. `https://images.google.com/`, 2026b. Accessed July 2026.

Google DeepMind. Genie 3: A new frontier for world models. `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source real-time and streaming interactive world model, August 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. DrivingWorld: Constructing world model for autonomous driving via video GPT. *arXiv preprint arXiv:2412.19505*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition, June 2025.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of Sight, Out of Mind? Evaluating State Evolution in Video World Models, March 2026.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A Text-Controlled Interactive World Generation Model, December 2025a.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An Interactive World Generation Model, July 2025b.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos World Foundation Model Platform for Physical AI, July 2025.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, March 2024.

Pexels. Pexels: Free stock photos and videos. `https://www.pexels.com/`, 2026. Accessed: 2026-05-03.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. WorldSimBench: Towards Video Generation Models as World Simulators, October 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL `http://proceedings.mlr.press/v139/radford21a.html`.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 779–788. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.91. URL https://doi.org/10.1109/CVPR.2016.91.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing Open-source World Models, January 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing Video Generation for World Complexity, 2026.

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

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, 2025a. doi: 10.1109/CVPR52734.2025.00787.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling, 2025b. URL https://arxiv.org/abs/2512.14614.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds, 2026. URL https://arxiv.org/abs/2604.14268.

Tencent Hunyuan. HY-World 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. https://3d-models.hunyuan.tencent.com/world/world1_5/HYWorld_1.5_Tech_Report.pdf, 2025. Technical report.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai

Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, April 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00499. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-Time and Streaming Interactive World Model with Long-Horizon Memory, April 2026.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-WorldBench: Towards a Comprehensive Interaction-Centric Evaluation for World Models, March 2026.

Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, Mu Xu, Qi Fan, Yang Gao, Yong Li, and Baoquan Chen. WorldRoamBench: An open-world benchmark for long-horizon stability of interactive world models. *arXiv preprint arXiv:2606.31672*, 2026a.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Yongtao Ge, and Kaipeng Zhang. WorldMark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024. doi: 10.52202/079017-0688.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking Memory Consistency and Action Control in World Models, February 2026.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. *arXiv preprint arXiv:2406.14540*, 2024.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer, 2026a. URL https://arxiv.org/abs/2605.15178.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General Interactive World Model with Autoregressive Denoising, January 2026b.

Code can be found in the Supplementary Materials.
Dataset can be found at the anonymous website:
https://huggingface.co/datasets/WorldModelBenchmark/playworld-bench-anonymous
More videos can be found at the anonymous website:
https://playworld-2026.github.io/project/PlayWorld/

A Implementation Details

Closed-loop execution. Every evaluation case provides all models with the same initial frame, long-horizon objective, and human-annotated basic action sequence. The agent interface executes the current action and returns the latest generated frame to the Agent Player, which observes the frame, the remaining basic action sequence, and the recent action history before returning *Keep*, *Stop*, *Extend*, *Correct*, or *End*. This closed loop continues until the Agent Player returns *End* or the 40-step interaction budget is exhausted, producing rollouts of approximately 10–60 seconds. We use Claude Haiku 4.5 (Anthropic, 2025) as the agent model for all reported benchmark experiments.

Model execution. Genie 3, LingBot-World, LingBot-World2, HY-World2, and HappyOyster are evaluated through their user-facing web interfaces, whereas SANA-WM, Hunyuan-GameCraft, HY-WorldPlay, and Matrix-Game-3.0 are evaluated locally through chunk-wise generation and model-specific adapters. All settings use the same action definitions and record the executed actions and generated observations. HY-World2 requires an input image that supports global scene modeling. Cases whose initial images do not satisfy this requirement are skipped and excluded from its score aggregation. Hunyuan-GameCraft requires a non-empty action input for every generated chunk; consequently, insight evolution cases that require prolonged passive observation or WAIT-only evolution cannot always be evaluated under their intended protocol.

Basic ability metrics. Video Quality includes seven higher-is-better metrics. Aesthetic Quality, Imaging Quality, Motion Smoothness, and Temporal Flickering follow VBench (Huang et al., 2024); Temporal Consistency follows Omni-WorldBench (Wu et al., 2026); Depth Stability uses Depth Anything V2 (Yang et al., 2024) following MemoBench (Chen et al., 2026b); and Subject Consistency is adapted from HyDRA's $\mathrm{DSC}_{\mathrm{ctx}}$ (Chen et al., 2026c). Subject Consistency detects dynamic subjects with YOLO (Redmon et al., 2016), extracts CLIP (Radford et al., 2021) features from the subject crops, and measures their cross-window similarity. When no corresponding dynamic subject is detected in both windows, Subject Consistency is reported as N/A.

For Action Controllability, VGGT (Wang et al., 2025) estimates camera poses from uniformly sampled frames following WorldMark (Xu et al., 2026b). The executed action sequence, including Agent Player adjustments, defines the target trajectory. A rollout passes Translation when its target-normalized Translation Error is below 0.3 and passes Rotation when its mean geodesic Rotation Error is below 45°. Each pass rate is computed only over rollouts with a valid estimate for the corresponding metric. Basic Ability Score ranks models independently on the seven Video Quality metrics and two Action Controllability pass rates, averages the nine ranks, and converts the result into a higher-is-better percentage.

VQA rubric verifier. Gemini 3.1 Pro (Google, 2026a) serves as the VQA rubric verifier. For geometry consistency, interaction fidelity, and out-of-sight evolution, dimension-specific validation first checks whether the rollout follows the trajectory required by the objective. Only valid rollouts proceed to rubric scoring; invalid rollouts receive the minimum dimension score of 1. Insight evolution uses stationary observation and therefore does not require trajectory validation.

To balance temporal coverage and spatial detail, the verifier receives two contact-sheet streams. The primary stream samples the rollout at 10 FPS, resizes each frame to 384×216, and groups every 25 frames into a 5×5 grid. The detail stream samples at 0.5 FPS, resizes each frame to 800×450, and groups every four frames into a 2×2 grid. Gemini receives both streams together with the case objective and its sample-specific VQA rubric. For each applicable question, it returns a binary answer $a_j \in \{0, 1\}$ and supporting evidence. Regular questions receive weight 1, the question expressing the

defining requirement receives weight 2, and N/A questions receive weight 0. The dimension score is

$$R_d = \frac{\sum_{j \in A_d} w_j a_j}{\sum_{j \in A_d} w_j}, \qquad S_d = 1 + 4R_d,$$

where A_d contains only applicable questions. The Overall score is the unweighted mean of geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution.

Radar-chart scales. In Fig. 1, the four rubric-based dimensions—geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution—are displayed on a range of 1–3, while Basic Ability is displayed on a range of 0–1. The axes therefore visualize performance within their respective ranges and should not be interpreted as sharing the same absolute scale.

B VQA Scoring Robustness

Gemini's outputs may vary across repeated calls even when the prompt and visual inputs are fixed. Repeating rubric verification for every rollout is costly at the full benchmark scale, so benchmark evaluations commonly rely on a single scoring pass under a fixed configuration. We likewise use one VQA scoring pass per rollout for the reported results and conduct an additional independent pass to assess robustness. Across the nine models, the mean per-model sample variance between the two scoring passes is 0.0112. This low variance indicates that a single fixed scoring pass provides stable aggregate results, although the two-pass analysis is not a precise estimate of scoring uncertainty. When the evaluation budget permits, we recommend averaging multiple independent VQA scoring passes to further reduce verifier variance.

C Agent Interface for Web-Based Models

Web-based world models do not expose a common local inference API. The agent interface therefore connects to an authenticated browser session, initializes the evaluation case, dispatches model-specific controls, monitors generation, and captures the resulting observation. The captured frame is returned to the Agent Player, and the interface executes the next control after *Keep*, *Stop*, *Extend*, or *Correct*, or terminates the episode after *End*. For reproducibility, it records the task identifier, initial frame, basic action sequence, executed controls, intermediate frames, Agent Player decisions, timing, termination condition, and final video.

Browser interaction. For each case, the agent interface opens the model interface, supplies the initial-world condition, selects the required perspective when this option is available, and waits until the generated world becomes interactive. During execution, controls are dispatched through the browser interface and the rendered world is captured after each step. The Agent Player decides only how the world should be controlled, while login, page navigation, submission, rendering-state checks, and output collection remain deterministic interface operations.

Example execution. Fig. 7 shows one HappyOyster case from caption entry to termination. The screenshots are direct crops from the recorded browser session; browser chrome, system controls, and assistant overlays are excluded without altering the generated world content.

D Additional Experimental Results

Agent Player analysis details. We compare Preset Only, Agent Only, and Preset + Agent on the same 25 cases for Genie 3 and HappyOyster (Tab. 4), and compare three agent models on the same Genie 3 cases using Preset + Agent (Tab. 5). For each comparison, we aggregate judgments from five independent raters with video-generation experience. Trajectory Score is assigned by Gemini 3.1 Pro on a 0–2 scale. Human Preference aggregates explicit pairwise preference judgments, assigning full credit to wins and half credit to ties. For the control-strategy comparison, Preset Only and Agent Only are each compared against Preset + Agent, whose rate pools comparisons against both baselines. For the agent model comparison, each agent model pools comparisons against the other agent models. The comparison order and left–right placement of the paired trajectories are randomized in the questionnaire. Majority Agreement is the percentage of pairwise comparisons for which more than



Figure 7: **Agent interface execution on a web-based world model.** (a) The agent interface enters the initial-world condition; (b) waits until the generated world becomes interactive; (c)–(d) dispatches controls and captures observations for the Agent Player's online decisions; and (e) terminates the episode while preserving the final video and execution record.

Table 3: **Trajectory-validation pass rates on the PlayWorld benchmark.** The insight evolution dimension uses stationary observation and therefore does not require trajectory validation and is marked as "–". Overall is the unweighted mean of the pass rates for geometry consistency, interaction fidelity, and out-of-sight evolution.

Model	Geometry Consistency	Interaction Fidelity	Insight Evolution	Out-of-sight Evolution	Overall
Closed-source Models					
Genie 3	**77.1%**	**93.5%**	–	90.7%	**87.1%**
HappyOyster	63.8%	**93.5%**	–	81.4%	79.6%
Open-source Models					
LingBot-World	58.3%	76.0%	–	83.7%	72.7%
LingBot-World2	58.3%	85.1%	–	**93.0%**	78.8%
HY-World2	47.9%	70.0%	–	37.2%	51.7%
SANA-WM	62.5%	88.0%	–	90.7%	80.4%
Hunyuan-GameCraft	25.0%	78.0%	–	81.4%	61.5%
HY-WorldPlay	14.6%	80.0%	–	30.2%	41.6%
Matrix-Game-3.0	37.5%	88.0%	–	79.1%	68.2%

half of the raters select the same outcome. Agent-modified Action Ratio is the macro-averaged per-case fraction of actions modified or inserted by the Agent Player. Decision Latency, reported only for the agent model comparison, is the average end-to-end waiting time per agent-model call.

E Human Annotation and Validation

Benchmark annotation. Human annotators screen candidate initial worlds, verify Gemini-generated captions, assign the viewpoint, write long-horizon objectives and basic action sequences, and construct sample-specific VQA rubrics. Annotators use observable evidence and remove questions whose answers cannot be determined from the generated rollout.

Table 6: **Basic ability of video world models on the PlayWorld benchmark.** Video Quality contains seven metrics; Action Controllability contains translation and rotation pass rates. Basic Ability Score converts the average rank across all nine metrics into a higher-is-better percentage. All values are percentages. Bold: best; underline: second best.

Model	Video Quality							Action Controllability		Basic Ability Score ↑
	Aes. ↑	Img. ↑	Mot. ↑	Flick. ↑	Temp. ↑	Depth ↑	Subj. ↑	Translation ↑	Rotation ↑	
Closed-source Models										
Genie 3	**52.00**	**75.22**	99.00	97.79	<u>98.64</u>	88.70	85.60	64.1	**50.6**	<u>72.2</u>
HappyOyster	49.66	73.66	**99.46**	<u>99.02</u>	**99.52**	<u>91.80</u>	<u>88.40</u>	58.0	<u>48.5</u>	**76.4**
Open-source Models										
LingBot-World	51.12	72.04	97.96	96.27	98.02	**92.20**	83.50	64.8	44.7	51.4
LingBot-World2	51.47	<u>73.98</u>	97.84	95.64	96.88	91.40	82.30	67.9	45.6	50.0
HY-World2	47.43	67.79	<u>99.12</u>	**99.46**	91.17	90.70	**89.80**	40.8	42.2	45.8
SANA-WM	<u>51.75</u>	72.59	98.92	96.48	98.20	84.40	85.90	<u>78.6</u>	42.2	56.9
Hunyuan-GameCraft	50.14	67.82	98.25	95.24	97.97	90.40	83.70	**89.6**	29.4	40.3
HY-WorldPlay	47.71	61.53	98.89	95.18	95.45	91.00	87.40	75.8	44.2	38.9
Matrix-Game-3.0	44.10	66.03	98.90	96.92	95.23	87.30	65.70	42.4	27.7	18.1

Table 7: **Inter-rater agreement in the human-alignment study.** All results are computed from the same 120 pairwise items and 600 judgments used in Fig. 6. Unanimous Agreement requires all five raters to select the same outcome, whereas Majority Agreement requires at least three raters to agree.

Dimension	Items	Judgments	Unanimous Agreement	Majority Agreement	Fleiss' κ
Geometry Consistency	30	150	33.3%	96.7%	0.461
Interaction Fidelity	30	150	16.7%	96.7%	0.323
Out-of-sight Evolution	30	150	23.3%	100.0%	0.419
Insight Evolution	30	150	43.3%	90.0%	0.483
Overall	120	600	29.2%	95.8%	0.434

Human-alignment study. Five participants independently evaluate the same 120 video pairs, comprising 30 comparisons for each of geometry consistency, interaction fidelity, out-of-sight evolution, and insight evolution. This yields 600 valid pairwise judgments, with no missing responses. For each comparison, participants view two videos generated for the same evaluation case, with model identities concealed using anonymous codes. They select the left video, the right video, or *Tie* when the two are difficult to distinguish.

Human preference assigns full credit to a win and half credit to a tie. Overall human preference pools comparisons across the four equally sampled dimensions. For visualization, we normalize the preference rates within each dimension so that the scores of the nine evaluated models sum to 100%.

At least three of the five raters agree on 95.8% of all pairwise items. Fleiss' κ is 0.434 overall and remains positive across all four dimensions, supporting the reliability of the aggregated human preference rankings. Figure 6 compares the normalized human preference scores with the corresponding rubric-based VQA scores. The Spearman correlations are $\rho = 0.933$ for Overall, $\rho = 0.983$ for geometry consistency, $\rho = 0.933$ for interaction fidelity, $\rho = 0.812$ for out-of-sight evolution, and $\rho = 0.745$ for insight evolution; all are statistically significant ($p < 0.05$).