SCENEMEMBENCH: A BENCHMARK FOR SCENE MEMORY IN INTERACTIVE VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video world models generate navigable worlds frame by frame, but a persistent world requires previously observed scenes to remain recoverable as users move away and return. Existing memory evaluations, however, often focus on seen-view revisits and full rollouts, leaving **cross-view scene recovery** underexplored while entangling memory with trajectory-following errors and accumulated generation degradation. We introduce SCENEMEMBENCH, a benchmark for scene memory in interactive video world models. It evaluates two complementary dimensions: **seen-view recall versus novel-view recovery**, and **controlled diagnosis versus end-to-end interaction**. We introduce Continued Rollout from a shared ground-truth observation history, alongside Full Rollout over complete model-generated interactions. SceneMemBench contains 600 controlled trajectories across 33 Unreal Engine scenes and 60 Full Rollout cases, covering eight representative open- and closed-source models. We evaluate scene recovery along subject consistency, appearance fidelity, and geometric consistency. Results show that **strong seen-view recall does not guarantee strong novel-view recovery**, while model profiles under controlled histories can differ markedly from those observed in full interaction. Together, these findings reveal complementary failure modes obscured by seen-view revisits or end-to-end rollouts alone. Our benchmark will be released at https://anonymous.4open.science/r/SceneMemBench.·

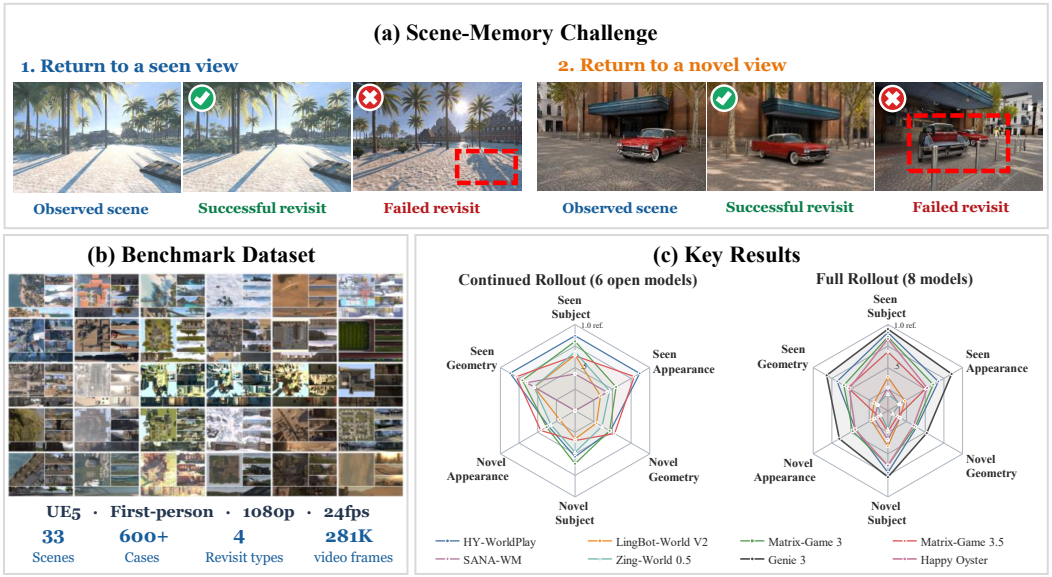


Figure 1: **SceneMemBench at a glance.** (a) Representative successful and failed revisits for seen- and novel-view memory tasks. (b) The benchmark contains 600 controlled trajectories across 33 scenes and 60 additional Full Rollout cases. (c) Dimension-wise evaluation results under Continued Rollout and Full Rollout protocols.

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

Recent video generation models have rapidly advanced in visual fidelity and temporal coherence (Kong et al., 2025; Team Wan et al., 2025; Team Seedance et al., 2026; MiniMax, 2026). Building on this progress, interactive video world models generate navigable environments conditioned on user actions or camera controls (Ball et al., 2025; Wang et al., 2026; Sun et al., 2026; Robbyant Team et al., 2026; Mao et al., 2025). As users move through these generated worlds, previously observed content may leave the field of view and later reappear, possibly from a different viewpoint. A persistent world therefore requires models not only to generate coherent video, but also to retain scene information and recover it consistently across interaction.

Recent benchmarks evaluate many aspects of interactive video world models, from geometric consistency (Duan et al., 2025) to multi-turn interaction over long-horizon objectives (Ying et al., 2026; Ding et al., 2026). Among them, memory-oriented benchmarks examine whether previously observed content persists across camera motion, disappearance and reappearance, and out-of-sight state evolution (Ye et al., 2026; Zhang et al., 2026; Ma et al., 2026; Chen et al., 2026a; Xu et al., 2026; Gu et al., 2026). While these benchmarks provide valuable evidence of memory failures, two questions remain insufficiently addressed: *whether remembered scene information remains usable from previously unseen viewpoints, and how scene memory can be diagnosed without letting model-specific errors accumulated before the revisit dominate the measurement.*

These two gaps reflect two recurring design choices in existing memory benchmarks. First, many protocols return the camera to a previously observed pose, which primarily tests seen-view recall rather than whether remembered scene information remains usable across viewpoints. For example, after observing an object from position A, the camera may move to B while facing away and then turn back toward the object from a novel viewpoint. The model should preserve the object's identity and surrounding spatial layout while producing changes in scale, perspective, and occlusion consistent with the new camera pose. Second, memory is often assessed only after a complete model-generated rollout. Differences in trajectory-following ability and generation stability can cause models to follow different paths and ultimately reach different positions and viewing directions. Trajectory drift and visual degradation complicate scene-memory diagnosis because pre-revisit histories differ across models.

To address these limitations, we introduce SCENEMEMBENCH, a benchmark for scene memory in interactive video world models. To extend memory evaluation beyond seen-view recall, SCENEMEMBENCH defines two complementary memory tasks: *Return to a Seen View*, which queries a camera pose present in the observation history, and *Return to a Novel View*, which revisits the same scene from a previously unseen pose. To reduce the influence of errors accumulated before the memory query, we introduce *Continued Rollout*, which provides models with the same ground-truth observation history for each case and begins generation of a short revisit segment from a controlled query boundary. We complement this diagnostic setting with *Full Rollout*, in which the model generates the complete trajectory and must preserve scene memory throughout the interaction. Together, these tasks and protocols evaluate scene recovery across viewpoints under both controlled histories and end-to-end model-generated rollouts.

Our contributions are summarized as follows:

- We introduce SCENEMEMBENCH, extending memory evaluation for interactive video world models beyond seen-view recall to novel-view recovery. Its two complementary tasks, *Return to a Seen View* and *Return to a Novel View*, assess scene recovery from previously observed and unseen viewpoints.
- We introduce Continued Rollout as a controlled diagnostic protocol that provides all models with a shared ground-truth history and begins generation from a fixed query boundary, reducing errors accumulated before the memory query. We complement it with Full Rollout, which evaluates scene recovery throughout complete model-generated interactions.
- We construct 600 controlled trajectories across 33 Unreal Engine scenes and 60 Full Rollout cases, and systematically evaluate eight representative open- and closed-source video world models. Our results show that strong seen-view recall does not guarantee strong novel-view recovery, while model profiles under controlled histories can differ markedly from those observed in full interaction.

Table 1: **Comparison of memory-oriented video world-model benchmarks.** ✓/△/– denote explicit/partial or related/unreported support. Txt./Img./Vid./Cam./Act./Obj. denote text/image/video/camera/action/objective.

Benchmark	Cases	Evaluation Input	GT Reference	Controlled Memory Eval.	Revisit Types				Rollout Length
					Reversed Rotation	Full Rotation	Forward and Look Back	New-Viewpoint Revisit	
MIND (Ye et al., 2026)	150	Txt. + Img./Vid. + Act.	✓	–	✓	–	–	–	370–4,906 frames
MBench (Zhang et al., 2026)	300	Txt. + Img. + Cam./Act.	–	△	✓	–	–	–	25 s
MemoBench (Chen et al., 2026a)	360	Txt. + Img. + Cam./Act.	✓	–	✓	–	✓	–	80–285 frames
WorldRoamBench (Xu et al., 2026)	211	Img. + Act.	–	△	✓	–	–	–	10–60 s
R2M-Bench (Gu et al., 2026)	300	Txt. + Img. + Act./Cam.	–	△	✓	–	–	–	≈ 481 frames
PlayWorld (Ding et al., 2026)	91	Txt. + Img. + Obj. + Act.	–	△	✓	✓	✓	△	10–60 s
SceneMemBench (Ours)	**660**	**Txt. + Img. + Vid. + Cam./Act.**	✓	✓	✓	✓	✓	✓	**2–60 s**

2 Related Work

Video Generation and Interactive Video World Models. Video generation has advanced from diffusion-based clip synthesis to increasingly capable video foundation models (Ho et al., 2022; Singer et al., 2023). Recent open models, including CogVideoX, LTX-Video, HunyuanVideo, and Wan, further improve generation quality, efficiency, and versatility (Yang et al., 2025; HaCohen et al., 2024; Kong et al., 2025; Team Wan et al., 2025). Building on these advances, interactive video world models extend generation conditioned on an initial prompt or frame to sustain interaction guided by successive user inputs, such as camera trajectories or actions (Wang et al., 2026; Qian et al., 2026; Zhu et al., 2026; Robbyant Team et al., 2026; Gao et al., 2026; Chen et al., 2026b; Ball et al., 2025; Alibaba Cloud Community, 2026). Such interaction requires models to follow user instructions while maintaining a consistent environment across multiple rounds. In particular, users may leave a scene and later revisit it from the same or a different viewpoint, requiring models to retain its appearance and geometry beyond the current view. Scene memory is therefore essential to persistent interactive environments, yet generating long, visually coherent videos alone does not establish this capability. This motivates explicit memory evaluation to determine whether interactive video world models can faithfully reconstruct previously observed scenes when revisited.

Video World Model Memory Benchmarks. Existing benchmarks assess memory through revisits and the persistence or evolution of content outside the field of view (Ye et al., 2026; Zhang et al., 2026; Ma et al., 2026; Chen et al., 2026a). However, these evaluations raise two distinct questions: whether remembered content remains usable across viewpoints, and whether scene memory can be diagnosed without being dominated by rollout errors. Recovering a previously seen view alone does not establish that a model can reconstruct the same scene from an unseen viewpoint. Moreover, discrepancies at a revisit can reflect trajectory-following errors and accumulated generation degradation, complicating the assessment of memory. Prior work mitigates these confounds through observed-history conditioning, geometric alignment, temporal normalization, and adaptive action execution (Ye et al., 2026; Xu et al., 2026; Gu et al., 2026; Ding et al., 2026). SceneMemBench examines both questions by explicitly distinguishing seen-view recovery from novel-view reconstruction and evaluating each through short continuations from ground-truth histories and complete model-generated rollouts. The former reduces errors accumulated before the memory query, while the latter assesses scene recovery throughout the full interaction.

In Table 1, Cases counts memory-evaluation examples, and Evaluation Input lists input modalities supported by each benchmark's implementation. GT Reference denotes scoring against ground-truth videos; Controlled Memory Eval. denotes evaluation with a prescribed observation history that reduces pre-query rollout confounds. Revisit Types summarizes camera motions used to revisit observed scenes, while Rollout Length gives generated video lengths in seconds or frames.

3 SceneMemBench

SceneMemBench evaluates whether interactive video world models can recover scene content after it leaves the current view. As summarized in Fig. 2, the benchmark combines ground-truth data with two memory tasks, four revisit types, and two complementary evaluation protocols. We first

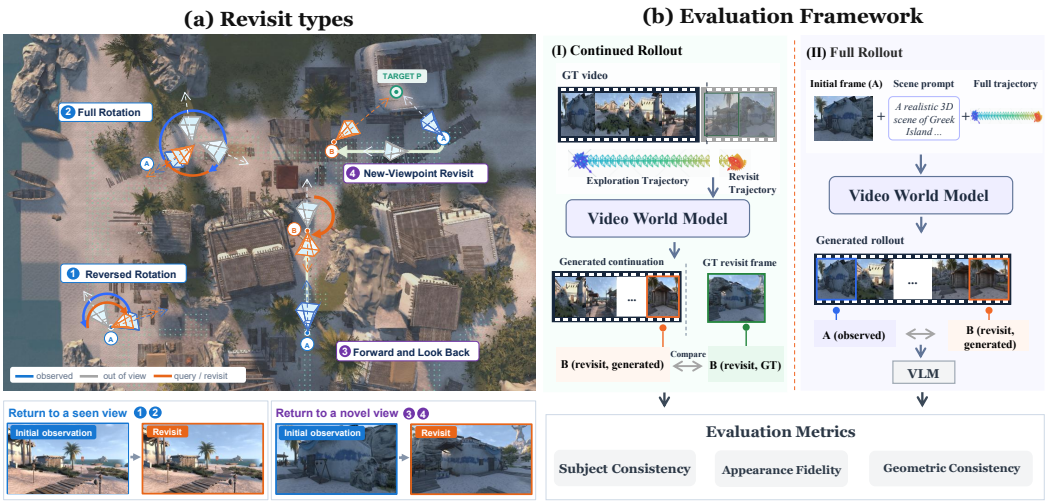


Figure 2: **Overview of SceneMemBench.** (a) Four revisit types instantiate two memory tasks: Reversed Rotation and Full Rotation return to seen views, whereas Forward and Look Back and New-Viewpoint Revisit query previously observed scene content from novel views. (b) The benchmark assesses these tasks under two complementary evaluation protocols. (I) Continued Rollout initializes the model with the prescribed GT history, generates the short query, and compares the output with held-out GT. (II) Full Rollout starts from the initial view, scene prompt, and reference actions, executes the complete interaction, and compares the revisit frame with the initial view. Each protocol reports Subject Consistency, Appearance Fidelity, and Geometric Consistency separately.

explain the data construction, then define the memory tasks and revisit types, and finally present the evaluation framework.

3.1 Dataset Construction

Even when asked to revisit the same scene, models generating a complete interaction may follow different paths, run for different lengths of time, and stop at different camera views. Their histories before the revisit can therefore differ, as can the visual degradation accumulated along the way. To compare scene memory after a common observation history, Continued Rollout initializes each model from the same recorded GT history and starts the short generated revisit at a fixed point in the trajectory. WorldMem and Memory Forcing have used GT histories to evaluate subsequent generation in Minecraft (Xiao et al., 2025; Huang et al., 2025), but comparable trajectory data for diverse, realistic scenes remain limited. We therefore record GT video trajectories in static rendered scenes with replayable camera motion, preserving the pre-query observations, camera poses, query boundary, and reference views needed for scoring.

We record trajectories in 33 Unreal Engine 5.6 scenes spanning urban and residential areas, historical ruins and stylized architecture, industrial and transportation facilities, indoor spaces, and natural terrain. Within each scene, we select navigable regions with recognizable structures that remain stable across viewpoints and construct deterministic rotational and translational revisits. The resulting 600 controlled trajectories record rendered video, frame-wise camera poses, a scene prompt, a reference action sequence, and an explicit history–query boundary. For Full Rollout, the benchmark additionally provides 60 image-anchored cases, each comprising an initial image, a scene prompt, and a reference action sequence. Appendix A provides detailed capture settings and dataset statistics; the memory queries supported by these data are defined next.

3.2 Scene-Memory Tasks and Revisit Types

We define two scene-memory tasks by relating the scored query view to the complete observation history, as illustrated in Fig. 2(a). *Return to a Seen View* asks whether a model can recover a view that has already appeared in the history, whereas *Return to a Novel View* asks it to reconstruct previously

observed scene content from a camera pose absent from that history. The distinction therefore separates view recall from reconstruction across camera poses while keeping the underlying scene fixed.

Two rotational revisit types instantiate Return to a Seen View. In *Reversed Rotation*, the camera turns away from its initial view and then retraces the rotational path in the opposite direction. In *Full Rotation*, the camera completes a continuous 360° in-place rotation: for Continued Rollout, the first 270° forms the Context, and the remaining 90° forms the query that closes the loop. Although their intervening motions differ, both revisit types culminate in a view already represented in the history.

Two further revisit types instantiate Return to a Novel View. In *Forward and Look Back*, the camera moves from position A to position B and turns by 180° at B; the Continued Rollout Query is the final 90° of this turn and does not translate back to A. In *New-Viewpoint Revisit*, the camera observes a target anchor P from A, moves to B while P leaves the field of view, and then turns until P reappears from the new position and orientation. In both types, novelty is defined by the constructed trajectory rather than a universal threshold on translation, rotation, or view overlap. Together, the four revisit types provide the shared scene-memory queries evaluated under the two protocols described next.

3.3 Evaluation Framework

We assess scene recovery under two complementary evaluation protocols. *Continued Rollout* evaluates a short revisit generated from a prescribed GT history, reducing errors accumulated before the query. *Full Rollout* evaluates recovery after a complete model-generated interaction, retaining the effects of action execution and long-horizon generation stability. Both assess *Subject Consistency*, *Appearance Fidelity*, and *Geometric Consistency*, but use different evidence and scoring procedures. Their scores are reported separately, without combining them or directly comparing their numerical values.

3.3.1 Continued Rollout

Evaluation Procedure. For each case, every model receives the same complete GT Context through its native history or memory state. It then generates the short Query from the prescribed boundary, using camera controls adapted to its native interface. We score the final decoded frame against a reference from the GT Query, which is withheld from generation. Initialization therefore supplies the complete observation history rather than a single conditioning image. Appendix B describes the native initialization and controls.

Reference selection accounts for residual camera-control errors that can otherwise obscure scene recovery (Fig. 11). For *Return to a Seen View*, we keep the generated final frame fixed and use image matching to select a GT reference with a better-matching viewpoint only within a local neighborhood of the GT endpoint. This allows an incomplete return turn while excluding references near the start of the continuation, so a static output cannot gain an inflated score by matching its starting view. For *Return to a Novel View*, substantial content differences can make viewpoint matching unreliable; we use the final GT Query frame without search. All six metrics use the same generated/reference pair, and all models follow the same rule within each task. Appendix B.3 gives the search bounds, matching algorithm, and fallback rule.

Metrics. We measure three aspects of recovery from this frame pair. *Subject Consistency* assesses agreement in scene content using DINOv2 CLS cosine similarity (Oquab et al., 2024). *Appearance Fidelity* uses LPIPS, SSIM, and PSNR to measure perceptual and image-level differences (Zhang et al., 2018; Wang et al., 2004). For *Geometric Consistency*, Depth Anything 3 estimates depth independently for each image (Lin et al., 2025); SI-AbsRel measures depth agreement after scale fitting, while normal cosine measures agreement between the derived surface orientations. These geometry measures are monocular-depth proxies, not metric 3D reconstruction accuracy. Appendix C details metric formulas and preprocessing.

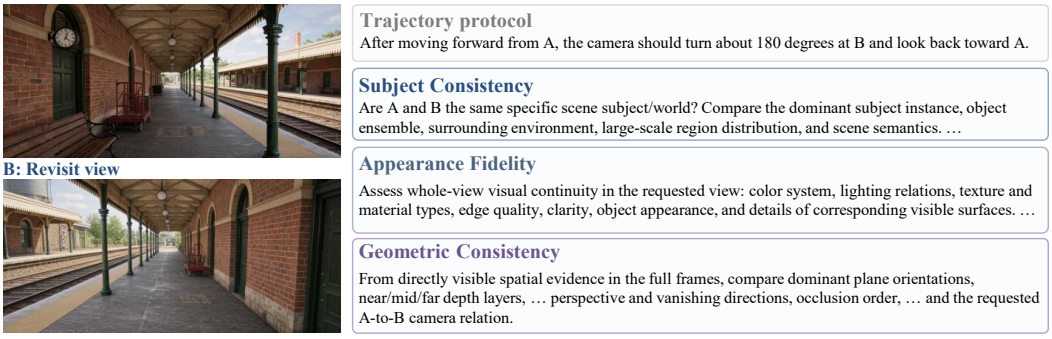


Figure 3: **Two-frame VQA for Full Rollout.** The initial view A and revisit view B are scored from 0 to 1 under the required view relation using three independent dimension questions.

3.3.2 Full Rollout

Evaluation Procedure. Each model generates a complete interaction from an initial image, a scene prompt, and a reference action sequence through its native control interface. We compare the first decoded frame A with the revisit frame B, fixed before scoring. Since the revisit may change the viewpoint, recovery is assessed under the view relation specified by the revisit type. Novel-view cases revisit overlapping regions of the same scene, so shared structures and their spatial relations provide evidence of recovery across viewpoints. Gemini 3.8 Flash receives this required view relation and the two frames for dimension scoring; GT novel-view images are not provided.

A separate VQA check screens for recognizable camera motion before scoring. Only passing videos are scored, but the check does not require the prescribed direction, exact angle, or trajectory completion. It therefore screens for motion presence rather than verifying a successful revisit. Intermediate frames are used only for this check, not for the three dimension scores. Appendix D.1 gives the sampling settings, judge inputs, and manual review of the evaluated videos.

Metrics. Three independent VQA questions assess recovery from the frame pair. *Subject Consistency* asks whether the frames depict the same specific scene. *Appearance Fidelity* assesses colors, materials, and visual details, while *Geometric Consistency* assesses layout, depth ordering, perspective, and occlusion under the required view relation. For novel-view queries, changes consistent with the requested viewpoint are allowed. A wrong viewing direction does not automatically negate Subject Consistency, but recognizing the same scene alone does not establish appearance or geometry recovery from the required view. For each dimension, the judge directly assigns a score from 0 to 1, allowing partial correspondence to receive partial credit. We average these scores over cases in each revisit type and task, reporting the three dimensions separately. Figure 3 illustrates the input pair, required view relation, and dimension questions; Appendix D gives the scoring guidance and prompt details.

4 Experiments

Our experiments assess the memory capabilities of video world models through their ability to recover previously observed scenes when revisiting seen or novel viewpoints under both Continued Rollout and Full Rollout protocols. We first describe the experimental setup (Section 4.1). We then present the main results (Section 4.2), comparing scene recovery under a shared ground-truth history in Continued Rollout with recovery over complete interaction sequences in Full Rollout, and examining failure cases across Subject Consistency, Appearance Fidelity, and Geometric Consistency.

4.1 Experimental Setup

Testing Models. We evaluate six open-weight video world models—HY-WorldPlay, LingBot-World v2, Matrix-Game 3, Matrix-Game 3.5, SANA-WM Streaming, and Zing-World 0.5—in both evaluation protocols, and additionally evaluate the closed Genie 3 and Happy Oyster web services

Table 2: **SceneMemBench results under Continued Rollout and Full Rollout.** Entries are per-model task means. **Bold** and underline mark the best and second-best means within each task and protocol, not statistical significance. Continued Rollout paired intervals: Appendix E.3.

Continued Rollout

	Subject Consistency	Appearance Fidelity			Geometric Consistency	
Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal Cos.↑
Return to a Seen View						
HY-WorldPlay	**0.9290**	**0.2775**	0.4339	**16.8042**	**0.2057**	**0.8227**
LingBot-World V2	0.7749	0.5128	0.3286	13.2291	0.3998	0.7021
Matrix-Game 3	0.8840	0.4182	0.3631	14.8309	0.2946	0.7554
Matrix-Game 3.5	0.7690	0.3501	**0.4367**	16.3734	0.2678	0.7979
SANA-WM Streaming	0.6245	0.4665	0.3200	14.4377	0.3638	0.7522
Zing-World 0.5	0.8007	0.4688	0.3549	14.1265	0.4544	0.7298
Return to a Novel View						
HY-WorldPlay	0.6987	0.5401	0.3410	13.5762	0.4622	0.6809
LingBot-World V2	0.5569	0.5834	0.3026	12.4137	0.6246	0.6369
Matrix-Game 3	**0.7543**	**0.5124**	0.3394	13.6729	0.5130	0.7004
Matrix-Game 3.5	0.5694	0.5141	**0.3674**	**14.3759**	**0.4368**	**0.7079**
SANA-WM Streaming	0.3318	0.6699	0.2387	11.5343	0.8050	0.5325
Zing-World 0.5	0.6633	0.5630	0.3349	13.4154	0.6102	0.6795

Full Rollout

	Return to a Seen View			Return to a Novel View		
Model	Subject↑	Appearance↑	Geometry↑	Subject↑	Appearance↑	Geometry↑
Closed-source models						
Genie 3	**0.947**	**0.853**	**0.820**	**0.770**	**0.650**	**0.520**
Happy Oyster	0.833	0.527	0.517	0.637	0.437	0.343
Open-source models						
HY-WorldPlay	0.907	0.677	0.680	0.717	0.433	0.383
LingBot-World-v2	0.390	0.227	0.183	0.413	0.223	0.163
Matrix-Game 3	0.857	0.607	0.590	0.633	0.473	0.367
Matrix-Game 3.5	0.683	0.503	0.533	0.243	0.130	0.130
SANA-WM	0.267	0.150	0.133	0.337	0.183	0.160
Zing-World 0.5	0.193	0.163	0.143	0.213	0.093	0.067

in Full Rollout. The six open-weight models use the same checkpoints across protocols; model and inference details are provided in Appendix B.

Evaluation Settings. Continued Rollout uses all 600 controlled cases per model with fixed context/query boundaries. Full Rollout uses all 60 image-anchored cases, shared across all eight models with 15 per revisit type. The two protocols yield 3,600 and 480 results, respectively. Execution details appear in Appendices B and D. To support interpretation of the benchmark scores, we examine history dependence, uncertainty in model comparisons, and evaluator reliability. We first check continuation quality, motion response, and scene recovery on two Reversed Rotation cases (Appendix B.4). History interventions on one case further support the models' dependence on early observations (Appendix B.4.2). We report paired confidence intervals to quantify uncertainty in score differences between models (Appendix E.3; Table 16). Human assessments support the VLM as an auxiliary evaluator of relative preferences, although agreement is lower on some novel-view tasks (Appendix D.2).

4.2 Main Results

Table 2 summarizes SceneMemBench results for Return to a Seen View and Return to a Novel View under both Continued Rollout and Full Rollout evaluations. The Continued Rollout block compares six open-weight models using six metrics spanning Subject Consistency, Appearance Fidelity, and Geometric Consistency. The Full Rollout block extends the comparison to two close models and reports scores for the same three dimensions after complete interaction sequences. Appendix E.1 examines how using fixed ground-truth (GT) endpoint references affects model rankings in Return to a Seen View. Figure 4 illustrates novel-view failures, including replaced landmarks and spurious occluders despite a familiar overall appearance. These examples show that visual resemblance alone does not establish the recovery of scene content and spatial structure.



Figure 4: **Continued Rollout examples across all four revisit types.** To highlight Subject Consistency in Seen views and Geometric Consistency in Novel views, badges show DINOv2 and SI-AbsRel/normal cosine, respectively; green bold values and red boxes mark the best displayed results and visible errors.



Figure 5: **Full Rollout examples across all four revisit types.** Rows compare the initial observation with four models' fixed-endpoint revisits. S/A/G are per-case Gemini 3.8 Flash scores for Subject Consistency, Appearance Fidelity, and Geometric Consistency.

Continued Rollout compares models with different memory designs. HY-WorldPlay and Matrix-Game 3 retrieve frames by camera geometry and view overlap; HY-WorldPlay also uses temporal reframing (Sun et al., 2026; Wang et al., 2026). Matrix-Game 3.5 retrieves and aligns geometry-aware spatial patches (Qian et al., 2026). LingBot-World v2 manages KV caches dynamically (Gao et al., 2026); Zing-World 0.5 uses a bounded cache with a prefix sink and sliding window (Chen et al., 2026b); SANA-WM combines Gated DeltaNet states and softmax attention (Zhu et al., 2026). Under the same GT history, HY-WorldPlay leads five Seen means; Matrix-Game 3 leads Novel Subject Consistency and LPIPS by mean score, and Matrix-Game 3.5 leads the other four Novel means. Paired confidence intervals support several of these metric-specific advantages; full comparisons for both protocols appear in Appendix E.3 and Table 16.

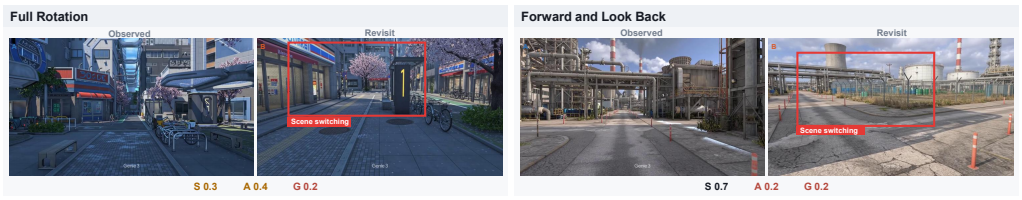


Figure 6: **Selected Genie 3 failures under Full Rollout.** The two examples pair the initial observation with its fixed-endpoint revisit; red boxes mark scene switching. S/A/G are per-case Gemini 3.8 Flash scores for Subject Consistency, Appearance Fidelity, and Geometric Consistency.

During Full Rollout, trajectory-following errors and accumulated generation degradation may obscure the strengths of different memory designs in recovering scene content, appearance, and spatial structure. Matrix-Game 3.5 provides a clear example of the difference between the two settings: it achieves the best mean values on both Novel geometry proxies under Continued Rollout, yet obtains low Geometric Consistency scores under Full Rollout and exhibits structural errors (Figure 5). Its performance illustrates why recovery from a shared GT history should be examined alongside complete interaction.

Although Genie 3 achieves the highest mean Full Rollout scores across all the evaluation dimensions (Table 2), we still find some failure cases as shown in Figure 6 which show significant differences between initial and final frames in both scene content and layout. These failures illustrate that recognizable scene elements can coexist with substantial discrepancies in appearance and spatial structure, underscoring the need to evaluate Subject Consistency jointly with Appearance Fidelity and Geometric Consistency.

5 Conclusion

We introduced SceneMemBench, a benchmark for evaluating scene memory in interactive video world models through seen-view recall and novel-view recovery. Its two tasks, *Return to a Seen View* and *Return to a Novel View*, assess whether models can recover previously observed scene content with previously both seen and unobserved camera poses. We introduce Continued Rollout, which supports controlled diagnosis from a ground-truth observation history, while Full Rollout evaluates revisit scenes after the complete model-generated interactions. Across eight representative models, our evaluation shows that strong seen-view recall does not guarantee strong novel-view reconstruction, and that model profiles under controlled histories can differ markedly from those observed in full interaction. These findings underscore the need to evaluate both recovery tasks under both protocols: each exposes complementary strengths and failure modes that a single task or protocol can obscure.

AI Use Statement

Large language models assisted with writing, translation, and result analysis and interpretation; coding agents assisted with code development. The authors reviewed these outputs and tested the assisted code, and take full responsibility for the paper. Models evaluated and used as judges are described in Sections 3.3.2 and 4.1.

Ethics Statement

The benchmark uses rendered and generated scenes, with human preference assessments to examine evaluator agreement. VLM judgments can be biased; reported memory scores do not establish safe or reliable deployment in real environments.

Reproducibility Statement

Sections 3–4 describe tasks, protocols, and experimental settings. Appendices A–E provide dataset statistics, model adaptations, metric formulas, scoring prompts, and validation results. The anonymous project repository is available at https://anonymous.4open.science/r/SceneMemBench.

References

Alibaba Cloud Community. Alibaba launches HappyOyster, a world model product for real-time immersive creation and interaction. Alibaba Cloud Community blog, April 2026. URL https://www.alibabacloud.com/blog/alibaba-launches-happyoyster-a-world-model-product-for-real-time-immersive-creation-and-interaction_603048. Published April 20, 2026; accessed September 26, 2026.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. Google DeepMind blog, August 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

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, 2026a. URL https://arxiv.org/abs/2606.27537.

Mingyang Chen, Shengdong Chen, Xiaoxiao Fu, Bosheng Gong, Haoyuan Guo, Bowen Li, Jiawen Li, Kejun Li, Tianpeng Li, Yin Liu, Haoze Sun, Zeyang Tian, Meng Wang, Xinmiao Wu, Jiangqiao Yan, and Zining Zhao. Zing-0.5: Toward playable worlds with real-time joint action and text control, 2026b. URL https://arxiv.org/abs/2609.17909.

Kaixin Ding, Xi Chen, Minghong Cai, Zhiyuan Xu, Yiyang Wang, Yuxiang Lu, Junyi Li, Shuyang Chen, Yuan Gao, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Playworld: Benchmarking world models with agent players over long-horizon objectives, 2026. URL https://arxiv.org/abs/2608.13552.

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 (ICCV)*, pp. 27713–27724, October 2025.

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.

Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2m-bench: Evaluating revisit memory via relative consistency in interactive video world models, 2026. URL https://arxiv.org/abs/2608.27328.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. doi: 10.52202/068431-0628. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft, 2025. URL https://arxiv.org/abs/2510.03198.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models, 2026. URL https://arxiv.org/abs/2603.13215.

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, 2025. URL https://arxiv.org/abs/2507.17744.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities, 2026. URL https://www.minimax.io/blog/minimax-h3. Accessed: September 19, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory, 2026. URL https://arxiv.org/abs/2608.29910.

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, 2026. URL https://arxiv.org/abs/2601.20540.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GfSwkDSr8J.

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. URL https://arxiv.org/abs/2604.14148.

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, 2025. URL https://arxiv.org/abs/2503.20314.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395.

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, 2026. URL https://arxiv.org/abs/2604.08995.

Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 49632–49652. Curran Associates, Inc., 2025. doi: 10.52202/085713-1659. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/470629a47e2d65ce0606c40055df5d26-Paper-Conference.pdf.

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, 2026. URL https://arxiv.org/abs/2606.31672.

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.

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, 2026. URL https://arxiv.org/abs/2602.08025.

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, 2026. URL https://arxiv.org/abs/2605.25874.

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

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. Mbench: A comprehensive benchmark on memory capability for video world models, 2026. URL https://arxiv.org/abs/2606.00793.

Fangwei Zhong, Kui Wu, Churan Wang, Hao Chen, Hai Ci, Zhoujun Li, and Yizhou Wang. Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5769–5779, October 2025.

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, 2026. URL https://arxiv.org/abs/2605.15178.

A Dataset and Trajectory Design Details

A.1 Dataset Coverage and Capture

The capture and trajectory specifications below define the observation histories and revisit queries used in the benchmark. SceneMemBench contains 600 controlled trajectories across 33 scenes and 60 image-anchored cases, covering indoor, urban, industrial, and natural environments. We use UnrealZoo (Zhong et al., 2025) as the capture platform for the 600 GT video trajectories. Tables 3 and 4 summarize capture specifications, annotations, and per-scene case counts. The image-anchored cases overlap the controlled corpus in scenes and initial views. Capture fixes lighting and rendering settings, synchronizes RGB and camera poses, and checks trajectory motion and video integrity.

Table 3: **Dataset statistics.** Trajectory statistics cover Context start through Query completion, excluding subsequent travel back to the starting position.

Dataset detail	Value
Total cases	660: 600 controlled trajectories + 60 Full Rollout cases
Controlled trajectories	
Case length	168–877 frames (mean 468.3); 7.0–36.5 seconds (mean 19.5)
Total observations	280,973 RGB frames with synchronized camera poses
Capture format	1920×1080, 24 FPS, 90° horizontal FOV
Annotations	Camera poses, actions, trajectory phases, and Context/Query boundaries
Full Rollout cases	
Case contents	60 initial images, each paired with a scene prompt and a reference action sequence
Evaluation cases	60 shared cases; 15 per revisit type

Table 4: **Scene coverage and case counts: 33 scenes and 600 cases.**

Scene	Cases	Scene	Cases
AbandonedDistrict	30	Map_ChemicalPlant_1	29
AncientRuins	29	Medieval_Daytime	15
Brass_Gardens	27	MiddleEast	2
Brass_Palace	29	ModularBuilding	6
Brass_Streets	32	ModularNeighborhood	19
Cabin_Lake	18	ModularSciFiVillage	5
ChineseWaterTown_Ver1	26	ModularVictorianCity	11
ContainerYard_Night	29	PostSoviet_Village	7
CourtYard	25	RussianWinterTownDemo01	12
Demonstration_Castle	23	SnowMap	10
DesertMap	18	Stadium	4
Desert_ruins	29	SuburbNeighborhood_Day_dooropen	6
Greek_Island	35	SunsetMap	1
IndustrialArea	28	Tokyo	15
InteriorDemo_NEW	21	UltimateFarming	4
JapanTrainStation_Optimised	22	Venice	6
KoreanPalace	27		

A.2 Trajectory Construction and Query Boundaries

Table 5 summarizes motion, Context coverage, and Query onset for all four revisit types. A denotes the initial camera position, B the displaced position, and P the revisited target. Boundaries are fixed before inference. Durations are computed from captured frame counts and frame rate, including holds within each interval. Translation distance is horizontal path length, excluding travel back after the revisit.

Table 5: **Collection design of the four revisit types. Time gives Context / Query duration ranges in seconds; distance is horizontal translation path length in meters.**

Type (cases)	Motion and Context	Query onset	Time (s)	Dist. (m)
Reversed Rotation (100)	Turn 90° or 180° in place, then reverse; Context includes the outward turn and return motion before the final segment.	Final approximately 90° of the reverse turn.	3.50–8.46 / 3.50–3.54	0
Full Rotation (100)	Rotate 360° in place; Context covers the first 270°.	At 270°, before the final 90°.	8.46 / 3.54	0
Forward and Look Back (200)	Observe a full turn at A, move straight to B; Context includes the first half of the turnaround at B.	Final 90° of the 180° turnaround at B.	17.21–19.29 / 2.54	18–26
New-Viewpoint Revisit (200)	Observe a full turn at A, move to B, then turn toward P; Context precedes the final turn segment. $\angle APB$ bins: 45°, 90°, 135°.	Final turn toward P at B, with at most 90° remaining.	18.04–33.00 / 2.58–3.58	18–28

New-Viewpoint Revisit uses horizontal $\angle APB$ design bins of 45°, 90°, and 135°, containing 48, 87, and 65 cases, respectively. These are target bins; realized angles vary with navigable paths and target positions. Reversed Rotation includes 80 yaw and 20 pitch trajectories.

B Continued Rollout Implementation and Validation

B.1 Model Selection and History Initialization

To compare recovery after a shared observation history, we supply the same GT Context through each model's native memory mechanism. GT Context is encoded into each model's native history representation, then used for continuation through its cache, retrieval, or history-replay mechanism. Table 6 summarizes history initialization and use. The adapters retain native encoding, denoising, and history-update mechanisms, incorporating newly generated content into the history. Context is resampled to the native resolution and frame rate; the final source frame is preserved and repeated only when required by VAE or chunk cadence. Camera motion uses the model's supported pose or action interface.

Table 6: **Model-native GT-history initialization and continuation.**

Model	GT Context initialization	History use during continuation
HY-WorldPlay	Encode history latents, conditions, and poses.	Select anchors, recent history, and FOV-retrieved blocks; rebuild the vision cache.
Matrix-Game 3	Build a history-latent bank and camera trajectory.	Retrieve latents by FOV overlap and condition on the recent overlap segment.
Matrix-Game 3.5	Teacher-force GT chunks into rolling KV, patch memory, and dynamic context.	Combine rolling KV, spatial patch queries, and dynamic context.
SANA-WM Streaming	Store GT latents, poses, and refiner history.	Replay the history prefix to rebuild state and generate a new suffix each call.
Zing-World 0.5	VAE-encode GT; fill KV with clean latent chunks and actions.	Continue with the same KV cache and write back new latents.
LingBot-World v2	Streaming-VAE encode GT; fill KV in native four-latent chunks.	Continue using sink and local-window caches.

B.2 Inference Settings

Across both protocols, the six open-weight models share released distilled checkpoints, seed 42, and native output resolutions and frame rates. Experiments use six NVIDIA A800 80 GB GPUs.

B.3 Evaluation Reference Selection

Residual camera-control errors can shift the generated revisit away from the prescribed GT endpoint. Local reference selection aims to reduce this view mismatch when evaluating recovery of a seen view. For Return to a Seen View, candidate GT frames r_i are restricted to a local angular window around the GT endpoint: $\pm 30^\circ$ for Full Rotation, and 30° on the approach side for Reversed Rotation, allowing an incomplete return turn. The generated endpoint g is compared with these candidates. Mutual SIFT matching and homography RANSAC identify inliers; C_g and C_{r_i} are their convex-hull coverage in the respective images. We maximize

$$S_{\text{sel}}(g, r_i) = N_{\text{inlier}}(g, r_i) \min(C_g, C_{r_i}) . \tag{1}$$

If the selected pair has fewer than 12 inliers or a minimum coverage below 0.01, we use the fixed endpoint reference. Return to a Novel View always uses the final decoded GT Query frame, without content-adaptive search. Matching does not verify the generated camera pose. All six metrics share the frame pair, and all models share each task's rule. No subsequent warping, pose alignment, temporal interpolation, or content-adaptive cropping is applied; the selection score is not a reported memory metric.

B.4 Validation of Continued Rollout

We check whether initialization from GT history supports usable, action-responsive continuations and recovery of earlier scene content. For this check, we compare native Full Rollout returns (A) with GT-history continuations (B) for six models on two Reversed Rotation cases, assessing quality, motion, and scene recovery.

B.4.1 Generation Quality and Motion Response

We score the first 2 generated seconds, excluding GT history. Quality uses MUSIQ-SPAQ means and percentile 10 (P10); smoothness is one minus normalized AMT-S interpolation error, both at a maximum edge of 512 pixels and native frame rates. Flow at 8 Hz measures the fraction of moving pairs and the fraction of moving pairs with correct direction. Table 7 equally averages both cases: continuation quality is broadly comparable to native generation, and all six models attain 100% on both motion measures, supporting usable, action-responsive short continuations.



Figure 7: **Six-model comparison on one Reversed Rotation case.** Each model has a Full Rollout row followed by a Continued Rollout row. Columns show the initial frame, the frame before return control, and the final output; in Continued Rollout, the first two columns are the early and actual last GT-history frames. The platform reappears in all six continuations. Endpoints are fixed last frames, not similarity-selected; rollout lengths and histories differ. The table uses two cases and the first 2 seconds, whereas this figure shows one case at its final endpoints.

Table 7: **Generation quality and motion response on two Reversed Rotation cases.** A: native Full Rollout return; B: GT-history continuation. Scores cover the first 2 seconds and equally average the two cases. Quality uses raw MUSIQ-SPAQ scores, not percentages.

Model	Quality mean A / B ↑	Quality P10 A / B ↑	Smoothness A / B ↑	B effective motion ↑	B direction accuracy ↑
HY-WorldPlay	78.94 / 79.00	77.81 / 78.15	0.9898 / 0.9855	100%	100%
LingBot-World v2	78.29 / 77.66	76.32 / 75.95	0.9613 / 0.9437	100%	100%
Matrix-Game 3	73.01 / 78.33	70.68 / 76.62	0.9783 / 0.9622	100%	100%
Matrix-Game 3.5	78.02 / 77.02	77.08 / 74.28	0.9711 / 0.9596	100%	100%
SANA-WM Streaming	69.22 / 71.29	62.47 / 59.56	0.9798 / 0.9705	100%	100%
Zing-World 0.5	78.46 / 77.99	78.02 / 77.17	0.9845 / 0.9755	100%	100%

Table 8: **Target-region scores under history interventions.** One Reversed Rotation case, scored against the fixed, unmodified GT endpoint. Top: six-model interventions. Bottom: a shared occluded initial frame; Full retains the original history after that frame.

	ROI DINOv2 ↑			ROI LPIPS ↓		
Model	Full	Recent	Replaced	Full	Recent	Replaced
HY-WorldPlay	0.7171	0.4149	0.5775	0.6110	0.6341	0.6733
LingBot-World v2	0.8579	0.4325	0.6196	0.6129	0.6811	0.5932
Matrix-Game 3	0.7128	0.3192	0.3312	0.6305	0.6470	0.8084
Matrix-Game 3.5	0.8361	0.3809	-0.0342	0.3310	0.6924	0.8213
SANA-WM Streaming	0.3806	0.4191	0.1366	0.5930	0.6484	0.6798
Zing-World 0.5	0.3545	0.1409	0.4089	0.6113	0.7714	0.7498
Model	Full	Target	Control	Full	Target	Control
HY-WorldPlay	0.6998	0.5775	0.7229	0.6205	0.6733	0.6182
Matrix-Game 3.5	0.7941	-0.0342	0.8028	0.4597	0.8213	0.4609

B.4.2 Recovery of Earlier Scene Structure

Figure 7 compares all six models on one snow-scene case. The previously observed checkerboard platform is outside the view when continuation begins and reappears in all six returns. History-selection records collected during generation show that HY-WorldPlay and Matrix-Game 3 use GT-history latents retrieved by field-of-view (FOV) overlap, while Matrix-Game 3.5 performs patch queries and selects GT-sourced context. The remaining models use the cache or history-replay mechanisms in Table 6. In a Full Rotation example, HY-WorldPlay retrieves a historical frame from the 0° view when returning to the initial viewpoint after a 360° rotation.

Sensitivity to earlier history. To test whether recovered content depends on earlier observations, we intervene on the history supplied to the models. On one representative Reversed Rotation case, we compare full history, genuine truncation of early history, and replacement of early target observations, while holding recent RGB context and Query controls fixed across conditions. Each condition rebuilds its states independently, with effective Query noise paired within each model. Target-region scores compare fixed generated endpoints with the unmodified GT endpoint. A two-model follow-up uses a shared target-occluded initial frame and an equal-area non-target occlusion control (Fig. 9).

Truncating early history worsens target-region LPIPS for all six models and lowers DINOv2 similarity for five (Table 8). Replacing early target observations also changes recovered content and scores (Fig. 8). LingBot-World v2 changes the distant wall and its opening; despite a slight improvement in ROI LPIPS, this content change and lower ROI DINOv2 support history sensitivity. The shared-initial-frame experiment with HY-WorldPlay and Matrix-Game 3.5 further shows that early evidence after the initial frame affects continuation. These results provide direct intervention evidence for history dependence in Continued Rollout, supporting its validity as a memory diagnostic on this case.



Figure 8: **Six-model history interventions on one Reversed Rotation case.** Columns show GT, Full history, Recent only, and Target replaced. Recent RGB inputs, Query controls, and effective Query noise are paired within each model. All conditions rebuild their states independently. Images are fixed generated endpoints; gray regions in the replacement condition can reflect the altered history rather than forgetting.

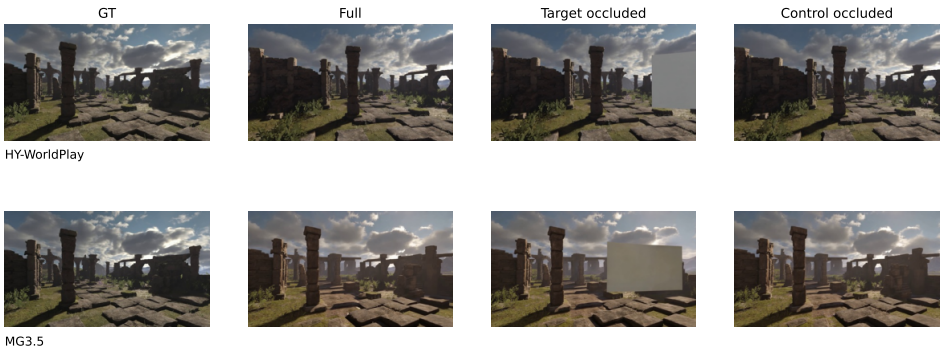


Figure 9: **History interventions with a shared initial frame.** All branches start from the same target-occluded image. Full provides original target evidence after that frame; Target occludes the early target region, while Control occludes an equal-area non-target region. The recovered opening in Full and Control, and gray region in Target, support a role for subsequent early observations.

C Evaluation Metric Formulas

The formulas and preprocessing below specify how Continued Rollout scores the three recovery dimensions from a shared image pair. All metrics compare the generated endpoint g with the GT reference r selected in Appendix B.3. Image metrics use Lanczos resizing and center cropping to $W \times H = 448 \times 252$, with RGB in $[0, 1]$.

C.1 Subject Consistency

Let f denote the final-layer DINOv2 CLS feature, using ImageNet channel-mean and standard-deviation input normalization (Oquab et al., 2024).

$$S_{\text{subj}}(g, r) = \frac{f(g)^{\mathsf{T}} f(r)}{\|f(g)\|_2 \|f(r)\|_2}. \tag{2}$$

C.2 Appearance Fidelity

LPIPS uses AlexNet features and learned channel weights (Zhang et al., 2018):

$$D_{\text{LPIPS}}(g, r) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{f}_l(g)_{hw} - \hat{f}_l(r)_{hw} \right) \right\|_2^2. \tag{3}$$

Here $\hat{f}_l$ is the channel-normalized feature at layer l, H_l, W_l are its spatial dimensions, w_l contains channel weights, and $\odot$ is elementwise multiplication. Network inputs are mapped from $[0, 1]$ to $[-1, 1]$. SSIM averages local-window scores over positions and RGB channels (Wang et al., 2004):

$$\text{SSIM}(g, r) = \text{mean}_{u,c} \frac{(2\mu_g \mu_r + C_1)(2\sigma_{gr} + C_2)}{(\mu_g^2 + \mu_r^2 + C_1)(\sigma_g^2 + \sigma_r^2 + C_2)}. \tag{4}$$

The window means, variances, and covariance are $\mu, \sigma^2, \sigma_{gr}$; u, c index window positions and channels. We use an 11×11 Gaussian window with standard deviation 1.5, $C_1 = 0.01^2$, and $C_2 = 0.03^2$, retaining scikit-image's sample-covariance and border-cropping conventions. PSNR is

$$\text{MSE} = \frac{1}{3HW} \sum_{u,c} (g_{uc} - r_{uc})^2, \qquad \text{PSNR} = -10 \log_{10} \max(\text{MSE}, 10^{-10}). \tag{5}$$

C.3 Geometric Consistency

Depth Anything 3 estimates d^g, d^r independently, without joint-image inference or pose/scale alignment (Lin et al., 2025). The reference is resized to the generated image dimensions, with intrinsics K scaled accordingly. Define $\mathcal{V}_x = \{i : d_i^x \text{ is finite and } d_i^x > 10^{-4}\}$ for $x \in \{g, r\}$, and $\mathcal{V} = \mathcal{V}_g \cap \mathcal{V}_r$. A global scale is fitted from the depth ratios on this intersection:

$$q_i = d_i^r / d_i^g, \qquad s = \underset{i \in \mathcal{V}:\, P_{10} \le q_i \le P_{90}}{\text{median}} q_i. \tag{6}$$

P_{10}, P_{90} are percentiles 10 and 90 of the ratios; fewer than 10 ratios are used without trimming. Scale-invariant AbsRel assigns unit error to reference-valid pixels with invalid generated depth:

$$E_{\text{SI}} = \frac{\sum_{i \in \mathcal{V}} |d_i^r - s d_i^g| / d_i^r + |\mathcal{V}_r \setminus \mathcal{V}|}{|\mathcal{V}_r|}. \tag{7}$$

The depth score is undefined when $\mathcal{V}$ is empty. Camera-space normals use back-projection and central differences:

$$X_{uv} = d_{uv} K^{-1} [u, v, 1]^{\mathsf{T}}, \qquad \tilde{n}_{uv} = (X_{u+1,v} - X_{u-1,v}) \times (X_{u,v+1} - X_{u,v-1}), \tag{8}$$

$$S_{\text{n}} = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} (n_i^r)^{\mathsf{T}} n_i^g. \tag{9}$$

Normalize $\tilde{n}$ and orient it toward the camera to obtain the unit normal n. $\mathcal{N}$ contains only interior pixels where, in both images, the center and four difference neighbors have valid depth, the normal is finite, and $\|\tilde{n}\|_2 > 10^{-8}$. An empty set yields an undefined score. These metrics measure agreement between monocular depth predictions, not ground-truth 3D reconstruction accuracy.

Table 9: **Native Full Rollout output formats.** The six open-weight models also participate in Continued Rollout using the same checkpoints. These frame rates describe Full Rollout outputs only.

Model	Access	Resolution	FPS
HY-WorldPlay	Open weights	848×480	24
LingBot-World v2	Open weights	832×464	16
Matrix-Game 3	Open weights	1280×704	17
Matrix-Game 3.5	Open weights	1280×704	17
SANA-WM Streaming	Open weights	1280×704	16
Zing-World 0.5	Open weights	1280×704	24
Genie 3	Web service	1280×704	20
Happy Oyster	Web service	1296×720	24

D Full Rollout VLM Evaluation

To make Full Rollout scores interpretable, we describe the visual evidence and scoring instructions supplied to the VLM, then assess its agreement with human preferences. The evaluated videos are generated under the native-interface conditions below. Full Rollout adapts action counts and durations to each model's native interface; realized paths and endpoints may differ. The six open-weight models use the shared settings in Appendix B.2. Genie 3 and Happy Oyster use native web interfaces without controllable backend versions or seeds. Table 9 lists output formats.

D.1 VQA Prompts and Scoring Details

Judge and visual evidence. Gemini 3.8 Flash evaluates all 480 Full Rollout outputs with fixed scoring prompts. Before scoring, a separate VQA check examines 12 uniformly time-sampled intermediate frames for recognizable camera motion. All 480 videos passed. This check does not require a prescribed direction, exact angle, or trajectory completion. Separately, we visually reviewed each of the 480 videos and found no static outputs or outputs limited to a small camera excursion followed by a return to nearly the initial view. This is an observation about the evaluated videos, not a minimum-motion guarantee of the VQA prompt. For dimension scoring, the evaluator receives the initial frame A, the revisit frame B fixed before scoring, and their side-by-side composite; no intermediate video frames or GT novel-view images are supplied. Novel-view cases target overlapping regions of the same scene. In Forward and Look Back, the initial and return views observe this region from opposing directions after a 180° turn. Shared objects and scene structures provide the basis for judging spatial relations across these views; missing or inconsistent structures are recovery errors. Subject Consistency, Appearance Fidelity, and Geometric Consistency are queried independently, without passing answers between dimensions. Dimension scores are used directly, without human editing or programmatic caps. We average 15 cases per revisit type and 30 per Task, reporting the three dimensions separately.

Unscored camera-motion check prompt.

> This is a separate unscored camera-motion check, NOT a memory or image-quality assessment. Inspect all 12 chronological intermediate frames. Decide whether recognizable camera movement through/viewing a scene is present (rotation, translation, or mixed motion). Do not require an exact angle, prescribed direction, complete loop, return to the initial view, or successful completion of the requested trajectory. Partial motion and overshoot are acceptable. Do not reject because subjects, appearance or geometry drift, or because the final view is wrong: those are separate endpoint assessments. Static images, only object animation, exposure flicker, or cuts with no discernible camera movement do not establish camera motion. Use uncertain when evidence cannot distinguish these. The trajectory label is context, not evidence. Sparse frames cannot certify full-video continuity. Cite visible changes and frame numbers. Do not assign S/A/G or numeric scores.

Endpoint scoring prompts. The twelve questions share the instructions and scale below. Each revisit type has three dimension questions with a common view requirement.

Common instructions

Use only visible evidence in the complete frames A and B; the side-by-side image repeats them. Do not assume successful motion or infer off-screen content. Allow legitimate viewpoint, parallax, scale, and occlusion changes. Assess the coverage of specific correspondence and genuine conflicts. Give partial credit for supported regions, but not a high scene-level score for one landmark or shared style. Score each dimension independently; report no conflict when none is visible.

Shared dimension definitions

Subject Consistency assesses the same specific world through its main subjects and environment. Appearance Fidelity assesses colors, lighting, materials, textures, clarity, and detail on corresponding surfaces in the requested view. Geometric Consistency assesses planes, depth ordering, boundaries, perspective, occlusion, scale, and spatial relations independently of appearance.

Scoring scale

Assign a score from 0.0 to 1.0 in steps of 0.1.

1.0 Nearly all major scene information is strongly supported without contradiction.

0.9 Broad strong agreement, with only minor differences.

0.8 Most important information is supported, with limited gaps.

0.7 The core scene and several major regions agree, with noticeable differences or incomplete coverage.

0.6 Scene-level agreement predominates, with substantial unsupported or conflicting content.

0.5 Scene-level agreement and disagreement or missing evidence are roughly balanced.

0.4 One substantial region matches, but most of the view differs or lacks support.

0.3 Specific partial correspondence is insufficient to support the whole view.

0.2 Only limited local structural or visual correspondence.

0.1 Only category, color, or style is similar.

0.0 No meaningful specific correspondence, or the main scene evidence conflicts.

1. Reversed Rotation

After leaving A, the camera should return to A's viewing direction: a same-view revisit.

Q1. Subject Consistency

After the reverse turn, does B recover the same specific subjects and surrounding scene as A? Allow small residual pose differences; one recognizable landmark is insufficient.

Q2. Appearance Fidelity

Do surfaces visible again after the reverse turn preserve their materials, colors, lighting, and detail? Distinguish small alignment changes from blur, texture melting, or repainting.

Q3. Geometric Consistency

Do planes, depth ordering, boundaries, and relative positions agree with A? Allow residual pose and edge-cropping differences without treating them as structural replacement.

2. Full Rotation

After a 360-degree turn, B should return to A's initial viewing direction: a same-view revisit.

Q4. Subject Consistency

After the full turn, do the subjects and environment identify the same specific world? A wrong final viewing direction does not alone imply a different world.

Q5. Appearance Fidelity

Has the initial direction been recovered with matching visible appearance? A clear side view or approximately 180-degree view of the same location does not restore the requested view.

Q6. Geometric Consistency

Do subject orientation, left/right layout, dominant planes, and vanishing directions support closure to the initial view? Recognizing the location alone does not demonstrate geometric closure.

3. Forward and Look Back

After moving forward from A, the camera should turn about 180 degrees at B and look back toward A.

Q7. Subject Consistency

Does the view from B show the same specific path and connected environment seen from A? Shared corridor, road, or architectural style alone is insufficient.

Q8. Appearance Fidelity

Do the same path or connected structures retain their appearance when viewed backward? Allow new foreground after forward travel; a clear forward or side view does not satisfy the look-back.

Q9. Geometric Consistency

Do path boundaries, connected openings, depth directions, and left/right relations support forward motion followed by an approximately 180-degree turn? Allow new foreground and elements lost through distance or occlusion.

4. New-Viewpoint Revisit

B should be a meaningful new view of the same specific 3D world as A; pixel alignment is not required.

Q10. Subject Consistency

Does B show the same specific subject and directly visible environmental anchors under the new viewpoint? An isolated recognizable object without verifiable surroundings supports limited scene identity, typically 0.2–0.4; this is a rubric anchor, not a hard cap.

Q11. Appearance Fidelity

Do corresponding surfaces preserve appearance with sufficient surrounding scene correspondence? An isolated matching object with unverifiable surroundings typically supports 0.2–0.4; contradictory surroundings merit less. These are rubric anchors, not hard caps.

Q12. Geometric Consistency

Can subject-ground-background adjacency, support surfaces, occlusion, and perspective follow one plausible camera translation and turn? Require direct non-reflected anchors and continuity between shared regions and newly revealed content.

D.2 Agreement with Human Preferences

We assess whether the VLM's relative preferences agree with human judgments of scene recovery, using paired comparisons of the evaluated outputs.

Paired assessment. Two annotators jointly assessed all 240 pairs, covering all 480 Full Rollout outputs, and produced final human consensus labels. Within each of the 60 trajectory–case groups, the eight model outputs were partitioned into four disjoint pairs, so each output appeared once. Pairing balanced model combinations and randomized left/right positions without consulting Gemini scores. The annotators viewed each candidate's initial and revisit frames and judged Subject Consistency, Appearance Fidelity, and Geometric Consistency separately: left preferred (L), right preferred (R), tie (T), or unable to judge (U). Human labels were locked before Gemini scores were revealed. Figure 10 illustrates three annotated pairs with their original frames and labels; Gemini scores are added for comparison.

Table 10: **Human–Gemini preference agreement by task and revisit type.** Cells give matches/valid pairs and percentages. U is excluded per dimension; ties agree only when both labels are ties. Task rows pool their two revisit types and are not independent samples.

Task / revisit type	S	A	G
Seen	97/120 (80.8%)	101/120 (84.2%)	100/120 (83.3%)
Novel	96/120 (80.0%)	84/120 (70.0%)	77/112 (68.8%)
Reversed Rotation	46/60 (76.7%)	49/60 (81.7%)	48/60 (80.0%)
Full Rotation	51/60 (85.0%)	52/60 (86.7%)	52/60 (86.7%)
Forward and Look Back	45/60 (75.0%)	35/60 (58.3%)	33/53 (62.3%)
New-Viewpoint Revisit	51/60 (85.0%)	49/60 (81.7%)	44/59 (74.6%)
All	193/240 (80.4%)	185/240 (77.1%)	177/232 (76.3%)

Agreement and scope. Gemini preferences are derived by comparing the two dimension scores, with equal scores treated as ties. Agreement requires an exact match with the human consensus L/R/T label; U labels are excluded for that dimension. Overall, VLM preferences agree with human consensus on 193/240 (80.4%) pairs for Subject Consistency, 185/240 (77.1%) for Appearance Fidelity, and 177/232 (76.3%) for Geometric Consistency, with eight geometry U labels excluded. These results support the VLM as an auxiliary evaluator of relative preferences. However, agreement is lower on some novel-view tasks (Table 10), indicating discrepancies between VLM and human judgments that limit its reliability as a cross-view consistency evaluator in these conditions. Future work could incorporate intermediate frames to provide richer visual evidence for cross-view correspondence.

E Additional Results and Confidence Intervals

We examine how the main results vary across revisit types, how sensitive they are to GT reference selection, and how uncertain the observed model differences are under resampling. The following tables disaggregate Table 2 by revisit type. Each entry is an arithmetic mean over the fixed cases for one model and revisit type. The main-table Task score equally averages its two revisit type means: 100 cases per Seen revisit type and 200 per Novel revisit type in Continued Rollout, and 15 per revisit type in Full Rollout.

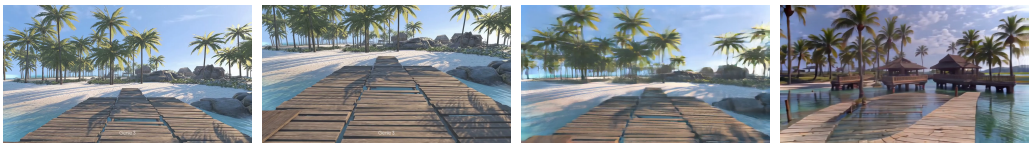
E.1 Continued Rollout by revisit type

Table 11: **Continued Rollout by revisit type: Return to a Seen View.** Arithmetic means over all scheduled cases; metric definitions follow Appendix C. PSNR is in dB.

Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal Cos.↑
Reversed Rotation (100 cases per model)						
HY-WorldPlay	0.9386	0.2268	0.4712	17.6024	0.1730	0.8519
LingBot-World v2	0.8134	0.5292	0.3185	13.2270	0.3876	0.6950
Matrix-Game 3	0.9014	0.3502	0.3887	15.6712	0.2338	0.7815
Matrix-Game 3.5	0.7869	0.3531	0.4325	16.2107	0.3071	0.7668
SANA-WM Streaming	0.7262	0.4413	0.3293	14.6233	0.3636	0.7596
Zing-World 0.5	0.8769	0.3844	0.3735	15.2095	0.3204	0.7794
Full Rotation (100 cases per model)						
HY-WorldPlay	0.9194	0.3281	0.3967	16.0060	0.2383	0.7935
LingBot-World v2	0.7364	0.4963	0.3386	13.2312	0.4119	0.7092
Matrix-Game 3	0.8666	0.4862	0.3376	13.9905	0.3554	0.7293
Matrix-Game 3.5	0.7511	0.3472	0.4409	16.5361	0.2286	0.8290
SANA-WM Streaming	0.5227	0.4916	0.3108	14.2522	0.3641	0.7448
Zing-World 0.5	0.7246	0.5532	0.3362	13.0436	0.5885	0.6802

In Tables 11 and 12, HY-WorldPlay has the highest Subject Consistency in both Seen revisit types, whereas Matrix-Game 3 leads this metric in both Novel revisit types. Matrix-Game 3.5 has the strongest Novel normal cosine in both revisit types, but its lowest SI-AbsRel is specific to New-Viewpoint Revisit: HY-WorldPlay has lower SI-AbsRel in Forward and Look Back.

(a) Reversed Rotation: agreement on all three dimensions



	Subject	Appearance	Geometry
Human preference	L	L	L
Gemini scores (left / right)	1.0 / 0.1	1.0 / 0.2	0.9 / 0.1
Gemini preference	L	L	L

(b) Full Rotation: agreement on all three dimensions



	Subject	Appearance	Geometry
Human preference	L	L	L
Gemini scores (left / right)	1.0 / 0.1	0.9 / 0.1	0.9 / 0.1
Gemini preference	L	L	L

(c) Forward and Look Back: disagreement on all three dimensions



	Subject	Appearance	Geometry
Human preference	R	R	R
Gemini scores (left / right)	1.0 / 1.0	0.8 / 0.3	0.9 / 0.9
Gemini preference	T	L	T

Figure 10: **Examples of human–Gemini preference assessment.** (a, b) Agreement on all three dimensions. (c) Disagreement on all three dimensions.

Sensitivity to the GT reference frame. We examine how GT reference selection affects model comparisons within Return to a Seen View. Table 13 evaluates the same 3,600 outputs using the final GT Query frame for both tasks, with the same metrics and aggregation as Table 2. Novel scores are unchanged because this task already uses fixed endpoint references.

Fixed references change the Seen rankings: Matrix-Game 3 leads DINOv2, and Matrix-Game 3.5 leads the other five means. Fixed GT endpoint comparisons reflect both scene-recovery errors and residual camera-control errors. Figure 11 shows that HY-WorldPlay preserves the main ruins, including the archway, steps, and walls, but its endpoint view more closely resembles an earlier GT frame along the return trajectory. Comparing directly with the prescribed endpoint also counts structure displacements and changes in visible extent caused by the rotational mismatch as image error. When assessing scene recovery, this error should not be attributed entirely to forgetting. Local reference matching therefore aims to reduce the influence of residual view mismatch. The fixed-reference analysis tests sensitivity to reference selection; it does not match history duration or other trajectory conditions between Seen and Novel.

Table 12: **Continued Rollout by revisit type: Return to a Novel View.** Arithmetic means over all scheduled cases; metric definitions follow Appendix C. PSNR is in dB.

Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal Cos.↑
Forward and Look Back (200 cases per model)						
HY-WorldPlay	0.6754	0.5323	0.3469	13.6646	0.4503	0.6874
LingBot-World v2	0.5374	0.5792	0.2938	12.5741	0.5802	0.6434
Matrix-Game 3	0.7392	0.5104	0.3412	13.5699	0.5281	0.6855
Matrix-Game 3.5	0.5199	0.5228	0.3634	14.1885	0.4663	0.6882
SANA-WM Streaming	0.3414	0.6707	0.2351	11.5193	0.8454	0.5235
Zing-World 0.5	0.6527	0.5753	0.3200	13.0158	0.6436	0.6659
New-Viewpoint Revisit (200 cases per model)						
HY-WorldPlay	0.7220	0.5478	0.3352	13.4877	0.4740	0.6745
LingBot-World v2	0.5765	0.5876	0.3115	12.2532	0.6691	0.6305
Matrix-Game 3	0.7695	0.5143	0.3376	13.7758	0.4979	0.7154
Matrix-Game 3.5	0.6190	0.5054	0.3713	14.5632	0.4073	0.7276
SANA-WM Streaming	0.3221	0.6690	0.2422	11.5493	0.7646	0.5416
Zing-World 0.5	0.6739	0.5508	0.3499	13.8150	0.5768	0.6931

Table 13: **Continued Rollout with fixed GT endpoint references.** The same 3,600 generated outputs are compared with the final GT Query frame in both tasks. Seen/Novel contain 200/400 cases per model, with equal revisit-type weights. Novel results are unchanged from Table 2. Bold marks the best means within each task.

Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal cos.↑
Return to a Seen View						
HY-WorldPlay	0.8468	0.5383	0.3360	13.2550	0.4665	0.6771
LingBot-World v2	0.7757	0.5085	0.3306	13.2874	0.3911	0.7078
Matrix-Game 3	**0.8808**	0.4158	0.3684	14.9269	0.2864	0.7624
Matrix-Game 3.5	0.7748	**0.3028**	**0.4821**	**17.3612**	**0.2419**	**0.8275**
SANA-WM Streaming	0.6274	0.4460	0.3270	14.6909	0.3392	0.7643
Zing-World 0.5	0.8010	0.4981	0.3452	13.8171	0.4601	0.7158
Return to a Novel View						
HY-WorldPlay	0.6987	0.5401	0.3410	13.5762	0.4622	0.6809
LingBot-World v2	0.5569	0.5834	0.3026	12.4137	0.6246	0.6369
Matrix-Game 3	**0.7543**	**0.5124**	0.3394	13.6729	0.5130	0.7004
Matrix-Game 3.5	0.5694	0.5141	**0.3674**	**14.3759**	**0.4368**	**0.7079**
SANA-WM Streaming	0.3318	0.6699	0.2387	11.5343	0.8050	0.5325
Zing-World 0.5	0.6633	0.5630	0.3349	13.4154	0.6102	0.6795



Figure 11: **Residual rotation mismatch in a Reversed Rotation case.** HY-WorldPlay's generated endpoint preserves recognizable ruins, while the matched GT frame shows closer image-space correspondence than the prescribed endpoint. The matched GT frame precedes the prescribed endpoint along the return trajectory, with a GT camera rotation difference of 24°.

E.2 Full Rollout by revisit type

In Tables 14 and 15, all six open-weight models score lower on Full Rotation than on Reversed Rotation in all three dimensions. Among Novel revisit types, Genie 3 retains high Subject Consistency in Forward and Look Back, but its geometry score is 0.300, compared with 0.740 in New-Viewpoint

Table 14: **Full Rollout by revisit type: Return to a Seen View.** Each entry averages 15 shared cases, scored by Gemini 3.8 Flash. S/A/G denote Subject Consistency, Appearance Fidelity, and Geometric Consistency; all scores lie in $[0, 1]$ and higher is better.

	Reversed Rotation			Full Rotation		
Model	S↑	A↑	G↑	S↑	A↑	G↑
Genie 3	0.980	0.900	0.887	0.913	0.807	0.753
Happy Oyster	0.793	0.573	0.567	0.873	0.480	0.467
HY-WorldPlay	1.000	0.973	0.973	0.813	0.380	0.387
LingBot-World v2	0.400	0.287	0.247	0.380	0.167	0.120
Matrix-Game 3	0.907	0.700	0.693	0.807	0.513	0.487
Matrix-Game 3.5	0.973	0.820	0.893	0.393	0.187	0.173
SANA-WM Streaming	0.413	0.233	0.227	0.120	0.067	0.040
Zing-World 0.5	0.333	0.287	0.273	0.053	0.040	0.013

Table 15: **Full Rollout by revisit type: Return to a Novel View.** Each entry averages 15 shared cases, scored by Gemini 3.8 Flash. S/A/G denote Subject Consistency, Appearance Fidelity, and Geometric Consistency; all scores lie in $[0, 1]$ and higher is better.

	Forward and Look Back			New-Viewpoint Revisit		
Model	S↑	A↑	G↑	S↑	A↑	G↑
Genie 3	0.800	0.440	0.300	0.740	0.860	0.740
Happy Oyster	0.660	0.287	0.193	0.613	0.587	0.493
HY-WorldPlay	0.813	0.333	0.287	0.620	0.533	0.480
LingBot-World v2	0.427	0.200	0.140	0.400	0.247	0.187
Matrix-Game 3	0.600	0.307	0.213	0.667	0.640	0.520
Matrix-Game 3.5	0.207	0.140	0.140	0.280	0.120	0.120
SANA-WM Streaming	0.267	0.147	0.127	0.407	0.220	0.193
Zing-World 0.5	0.180	0.100	0.087	0.247	0.087	0.047

Table 16: **Full Rollout paired differences between the two leading means.** Positive values favor the leader; pointwise 95% case-bootstrap intervals. HY: HY-WorldPlay; MG3: Matrix-Game 3. S/A/G denote the three evaluation dimensions.

Task	Dimension	Leader / runner-up	Advantage	95% interval
Seen	S	Genie 3 / HY	0.0400	$[-0.0300, 0.0933]$
Seen	A	Genie 3 / HY	0.1767	$[0.0833, 0.2600]$
Seen	G	Genie 3 / HY	0.1400	$[0.0333, 0.2367]$
Novel	S	Genie 3 / HY	0.0533	$[-0.0400, 0.1367]$
Novel	A	Genie 3 / MG3	0.1767	$[0.0667, 0.2900]$
Novel	G	Genie 3 / HY	0.1367	$[0.0067, 0.2667]$

Revisit. Matrix-Game 3.5 has low geometry scores in both Novel revisit types (0.140 and 0.120). These results reflect memory, action following, and generation stability together.

Paired confidence intervals. To quantify uncertainty in the observed Full Rollout mean differences, we resample cases while preserving model pairings. Table 16 compares the two highest observed task means per dimension. In each of 10,000 bootstrap replicates (seed 20260926), we resample 15 cases with replacement within each revisit type, keeping the eight models and three dimensions paired. Task means weight their two types equally. The 2.5 and 97.5 percentiles of leader-minus-runner-up differences give pointwise 95% intervals, unadjusted for selecting the leaders or multiple comparisons. These intervals cover case sampling for the existing outputs and scores, not scene clustering or variation across generation seeds and judge calls. Genie 3's Appearance Fidelity and Geometric Consistency advantages exclude zero in both tasks, whereas its Subject Consistency intervals include zero.

Table 17: **Continued Rollout means and 95% confidence intervals.** Each cell gives the task mean above its pointwise scene-cluster bootstrap interval; all 600 trajectories, including pitch, are used.

Return to a Seen View Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal cos.↑
	0.929	0.277	0.434	16.80	0.206	0.823
HY-WorldPlay	[0.917, 0.940]	[0.253, 0.300]	[0.385, 0.488]	[15.59, 18.10]	[0.168, 0.247]	[0.802, 0.846]
	0.775	0.513	0.329	13.23	0.400	0.702
LingBot-World v2	[0.726, 0.820]	[0.485, 0.539]	[0.290, 0.372]	[12.26, 14.30]	[0.340, 0.461]	[0.674, 0.731]
	0.884	0.418	0.363	14.83	0.295	0.755
Matrix-Game 3	[0.870, 0.897]	[0.396, 0.439]	[0.323, 0.410]	[13.70, 16.12]	[0.255, 0.335]	[0.731, 0.782]
	0.769	0.350	0.437	16.37	0.268	0.798
Matrix-Game 3.5	[0.746, 0.790]	[0.335, 0.366]	[0.397, 0.481]	[15.45, 17.37]	[0.219, 0.336]	[0.774, 0.821]
	0.624	0.466	0.320	14.44	0.364	0.752
SANA-WM Streaming	[0.590, 0.662]	[0.445, 0.487]	[0.278, 0.367]	[13.47, 15.46]	[0.267, 0.520]	[0.725, 0.782]
	0.801	0.469	0.355	14.13	0.454	0.730
Zing-World 0.5	[0.778, 0.824]	[0.445, 0.491]	[0.311, 0.405]	[13.10, 15.22]	[0.335, 0.598]	[0.705, 0.757]
Return to a Novel View Model	DINOv2↑	LPIPS↓	SSIM↑	PSNR↑	SI-AbsRel↓	Normal cos.↑
	0.699	0.540	0.341	13.58	0.462	0.681
HY-WorldPlay	[0.672, 0.725]	[0.521, 0.557]	[0.295, 0.390]	[12.47, 14.77]	[0.406, 0.525]	[0.653, 0.711]
	0.557	0.583	0.303	12.41	0.625	0.637
LingBot-World v2	[0.500, 0.613]	[0.558, 0.605]	[0.266, 0.343]	[11.40, 13.57]	[0.544, 0.705]	[0.606, 0.673]
	0.754	0.512	0.339	13.67	0.513	0.700
Matrix-Game 3	[0.726, 0.784]	[0.487, 0.536]	[0.294, 0.389]	[12.57, 14.85]	[0.421, 0.620]	[0.671, 0.731]
	0.569	0.514	0.367	14.38	0.437	0.708
Matrix-Game 3.5	[0.527, 0.614]	[0.493, 0.535]	[0.320, 0.415]	[13.38, 15.45]	[0.372, 0.507]	[0.678, 0.739]
	0.332	0.670	0.239	11.53	0.805	0.533
SANA-WM Streaming	[0.292, 0.375]	[0.657, 0.684]	[0.205, 0.274]	[10.77, 12.35]	[0.716, 0.894]	[0.505, 0.562]
	0.663	0.563	0.335	13.42	0.610	0.679
Zing-World 0.5	[0.625, 0.701]	[0.543, 0.581]	[0.290, 0.382]	[12.50, 14.41]	[0.524, 0.702]	[0.650, 0.711]

Table 18: **Paired differences between the two leading means.** Positive values favor the leader; intervals are pointwise 95%. HY: HY-WorldPlay; MG3/MG3.5: Matrix-Game 3/3.5.

Task	Metric	Leader / runner-up	Advantage	95% interval
Seen	DINOv2↑	HY / MG3	0.045035	[0.028900, 0.059344]
	LPIPS↓	HY / MG3.5	0.072669	[0.049055, 0.097578]
	SSIM↑	MG3.5 / HY	0.002767	[−0.020476, 0.023165]
	PSNR↑	HY / MG3.5	0.430867	[−0.093669, 0.980763]
	SI-AbsRel↓	HY / MG3.5	0.062154	[0.024301, 0.107078]
	Normal cosine↑	HY / MG3.5	0.024784	[0.006430, 0.042638]
Novel	DINOv2↑	MG3 / HY	0.055597	[0.032117, 0.080053]
	LPIPS↓	MG3 / MG3.5	0.001752	[−0.011911, 0.016540]
	SSIM↑	MG3.5 / HY	0.026324	[0.017300, 0.035062]
	PSNR↑	MG3.5 / MG3	0.702983	[0.343431, 1.045170]
	SI-AbsRel↓	MG3.5 / HY	0.025394	[−0.017210, 0.066247]
	Normal cosine↑	MG3.5 / MG3	0.007465	[−0.004238, 0.019892]

E.3 Paired Confidence Intervals for Continued Rollout

We quantify how Continued Rollout model means and their paired differences vary under scene resampling, keeping trajectories from each scene together.

Method. From 600 trajectories (3,600 outputs), 10,000 scene-cluster bootstrap replicates resample 29 Seen or 33 Novel scenes with replacement, retaining all their trajectories. Six models and all metrics share weights; revisit-type means are averaged equally. Percentiles 2.5 and 97.5 define 95% intervals. Tables 17 and 18 report model means and paired differences between the two highest-ranked observed means per metric. Differences favor the leader when positive: leader minus runner-up for higher-is-better metrics, reversed for lower-is-better metrics.

Main findings. For Seen, HY-WorldPlay's DINOv2, LPIPS, SI-AbsRel, and normal-cosine advantages exclude zero. For Novel, Matrix-Game 3's DINOv2 and Matrix-Game 3.5's SSIM and PSNR advantages exclude zero. Each comparison is against the respective mean runner-up; the remaining paired intervals include zero.