WORLDMEMBENCH: A COMPREHENSIVE BENCHMARK FOR MEMORY OF VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As video models enable longer and more realistic interactive experiences, maintaining a world consistent with its history becomes essential. This requires both preserving previously observed content and ensuring coherent world evolution over time. World memory is a critical capability to evaluate alongside visual fidelity and video duration. We introduce WorldMemBench, a benchmark for systematically evaluating world memory in video world models. Across 260 test cases, it assesses static memory through the retention of human identity and appearance, object appearance and geometry, and environmental content and geometry, and dynamic memory through event progression and state persistence. Using controlled camera departures and multi-revisit trajectories, the benchmark evaluates static content against fixed initial references and dynamic events against expected state transitions. It combines appearance and geometry metrics with event-specific assessments, and reports video quality and camera-following accuracy as complementary diagnostics. Evaluation of nine representative models shows that strong static retention does not imply reliable dynamic state evolution, while repeated revisits reveal memory degradation beyond the first return. Human studies support the alignment between our automatic evaluations and human judgments. WorldMemBench provides a systematic testbed for diagnosing memory failures and guiding the development of persistent, temporally coherent video worlds.

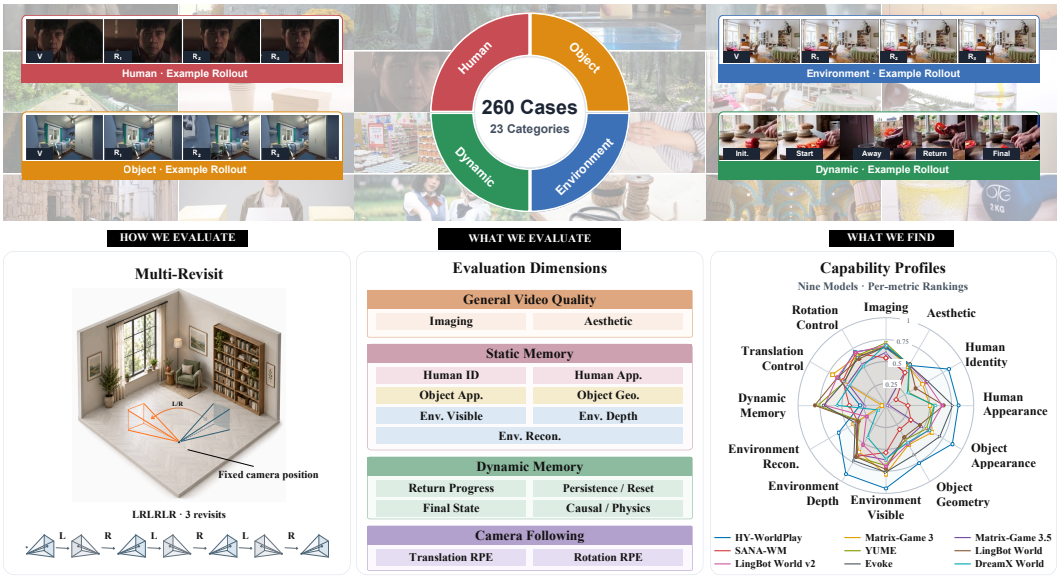

Figure 1: **Overview of WorldMemBench**. Top: case coverage across humans, objects, environments, and dynamic events. Bottom left: controlled departure-and-revisit evaluation compares each return against a fixed initial reference and averages scores across revisits for each memory metric. Bottom right: capability profiles of nine video world models.

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 advances in autoregressive video generation have enabled longer, more coherent video sequences Huang et al. (2025b); Zhu et al. (2026b); Zhao et al. (2026). In parallel, camera-controlled methods have improved viewpoint and trajectory control Bai et al. (2025). Together, these developments support video world models that enable sustained, interactive exploration of generated environments Mao et al. (2025); Gao et al. (2026); Yin et al. (2026); Sun et al. (2025).

Such exploration repeatedly moves content into and out of view, requiring generated worlds to remain consistent with past observations and events. We define **world memory** as the ability to preserve established content and sustain coherent state evolution across interruptions in observation. Static content, including human identities, object appearance and geometry, and scene layouts, should remain consistent across revisits. Meanwhile, ongoing events should progress coherently while out of view, with their expected consequences reflected upon return. Visual quality and local temporal coherence alone cannot establish these capabilities.

Recent benchmarks have begun to assess memory-related failures Zhang et al. (2026); Chen et al. (2026); Gu et al. (2026); Xu et al. (2026a). Building on these efforts, an important question is how memory behaves throughout continued exploration. A successful return demonstrates recovery at one moment, but provides only partial evidence of reliability over subsequent interactions. Repeated revisits help determine whether recovered content remains stable, whether evolving states remain consistent with prior events, and how these behaviors respond to changes in exploration direction. Tab. 1 compares benchmark coverage and highlights our joint assessment of memory persistence over extended exploration and repeated revisits, off-screen state evolution, and physical consistency.

To this end, we introduce **WorldMemBench**, a benchmark for systematically evaluating memory in video world models. It comprises 260 cases across 23 semantic categories, including 200 static-memory cases and 60 dynamic-memory cases as in Fig. 1. Our evaluation covers four dimensions. **Static Memory** assesses retention of human identity and appearance, object appearance and geometry, and environmental content and geometry. **Dynamic Memory** evaluates event progression, state persistence, and causal and physical plausibility. **General Video Quality** and **Camera Following** serve as complementary diagnostics of perceptual quality and adherence to prescribed exploration trajectories. Reporting these diagnostics separately helps interpret memory performance when rendering quality or camera accuracy differs across models.

In WorldMemBench, we use controlled trajectories for leaving and revisiting observed content, including translation loops, repeated rotations about the same axis, and rotations that switch axes between revisits. Each static revisit is compared with the same initial reference to track deviations across successive returns. Dynamic cases specify an initial state, an action, and an expected next state. Case-specific checklists assess event onset and whether subsequent observations reflect its expected progression and consequences. Dynamic memory is thus evaluated through coherent state changes rather than similarity to an unchanged initial frame.

Our evaluation of nine video world models reveals that strong static retention does not necessarily imply reliable dynamic evolution. Repeated revisits expose cumulative degradation, while switching rotation axes challenges content recovery. Human studies show agreement between our automatic metrics and human judgments. We use WorldMemBench to examine how context access affects memory, comparing training-free policies while holding the model and generation settings fixed. Combining recent context with view-relevant history improves most evaluated memory metrics and reduces latency relative to full-history access. These results demonstrate the value of WorldMemBench for characterizing memory limitations and evaluating context-access strategies.

Our contributions are summarized as follows:

- **A benchmark for world memory.** We introduce WorldMemBench, a comprehensive benchmark for the structured evaluation of memory capabilities in video world models.
- **A diagnostic evaluation protocol.** We develop a multi-revisit evaluation protocol to assess static and dynamic memory alongside camera controllability.
- **An empirical characterization of world memory.** We evaluate nine video world models, revealing distinct memory profiles, cumulative degradation across revisits, and sensitivity to exploration direction and context access.

Table 1: **Comparison with world-model evaluation benchmarks.** Input denotes model paradigm: C (camera-driven), A (action-driven), and T (language-driven). Gray icons indicate unevaluated paradigms. ✓: explicit coverage; △: partial coverage; ×: not reported.

Benchmark	Input	Length (s)	Camera control	Visual quality	Long-horizon revisit	State evolution	Physical consistency	Multi-revisit
WorldScore Duan et al. (2025)	C T	∼2–10	✓	✓	×	×	×	×
MIND Ye et al. (2026)	A	∼5–10	✓	✓	△	×	×	×
WorldMark Xu et al. (2026b)	C A T	20–60	✓	✓	✓	×	×	×
WBench Ying et al. (2026)	C A T	∼10–45	✓	✓	✓	△	△	×
MBench Zhang et al. (2026)	C A T	∼25	✓	×	✓	✓	✓	×
MemoBench Chen et al. (2026)	C A T	∼4–6	✓	✓	×	✓	✓	×
WorldRoamBench Xu et al. (2026a)	C A T	10–60	✓	✓	✓	△	△	×
WorldExam Yang et al. (2026)	C A T	∼3–12	✓	✓	△	△	△	×
R2M-Bench Gu et al. (2026)	C A	∼30	×	×	✓	×	×	×
WorldMemBench	C A T	∼20–35	✓	✓	✓	✓	✓	✓

2 Related Work

2.1 Video World Models

Self Forcing Huang et al. (2025b) and subsequent advances Zhu et al. (2026b); Zhao et al. (2026) have improved the efficiency and stability of autoregressive video generation. Together with progress in camera and action control, these developments have made long-horizon interactive generation increasingly practical, enabling a growing class of video world models Yin et al. (2026); Sun et al. (2025); Team et al. (2026b); Gao et al. (2026); Wang et al. (2026); Qian et al. (2026); Zhu et al. (2026a). Beyond generation quality and responsiveness, these models increasingly emphasize spatial memory for maintaining consistency over extended interactions. Relevant approaches include long-context modeling Team et al. (2026b), historical context reconstruction Sun et al. (2025), and field-of-view-based frame retrieval Yu et al. (2025), which help preserve previously observed scene content during exploration.

2.2 Evaluation of Video World Models

Video generation is commonly evaluated using FID Heusel et al. (2017) and FVD Unterthiner et al. (2018) for distributional similarity, CLIP-based scores Hessel et al. (2021) for semantic alignment, and PSNR, SSIM, and LPIPS Wang et al. (2004); Zhang et al. (2018) for visual fidelity when aligned references are available. Beyond these metrics, benchmarks such as VBench Huang et al. (2024), VBench-2.0 Zheng et al. (2025), and WorldScore Duan et al. (2025) provide assessments of visual quality, controllability, text alignment, and physical plausibility. However, performance on these dimensions alone does not establish whether a model can reliably retain and update world information over time. Recent efforts extend evaluation toward world-model capabilities: WBench Ying et al. (2026) examines multi-turn interaction and consistency, while WorldMark Xu et al. (2026b) standardizes control across heterogeneous model interfaces. WorldModelBench Li et al. (2025) assesses instruction following and physical plausibility, and WorldExam Yang et al. (2026) evaluates scene-conditioned reactions and goal-directed behaviors. Complementing these evaluation frameworks, our benchmark focuses on world memory: the ability to preserve previously observed scene information and track changes in world states over time.

3 WorldMemBench Construction

In this section, we describe the construction of WorldMemBench, including its data composition and camera trajectory design. The benchmark evaluates world memory through controlled departure-and-revisit rollouts, probing whether previously observed content and evolving states remain consistent across periods outside the camera's field of view.

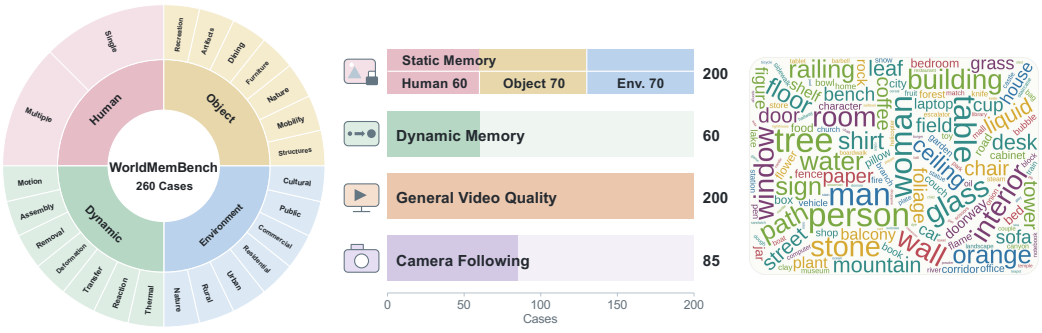

Figure 2: **Composition and diversity of WorldMemBench. (a)** Case taxonomy, with Human, Object, Environment, and Dynamic categories in the inner ring and subcategories in the outer ring. **(b)** Case counts for static memory, dynamic memory, general video quality, and camera-following evaluation. **(c)** Word cloud of entities, scenes, and actions in benchmark prompts.



Figure 3: **Evaluation camera trajectories in WorldMemBench**. Translation preserves the initial viewing direction and completes a square loop (green), while Rotation keeps the camera position fixed and applies left–right yaw (orange) or up–down pitch (blue).

3.1 Evaluation Dimensions

Inspired by hierarchical diagnostic evaluation Yang et al. (2026), we organize evaluation into four complementary dimensions. **Static Memory** assesses the retention of previously observed humans, objects, and environments upon revisitation, covering individual entities through global scene structure. **Dynamic Memory** examines whether events progress appropriately while out of view and whether their consequences persist upon return. Each case specifies an initial state, an action, and a next state, with assessment covering event progression, state persistence, final-state correctness, and causal and physical consistency. These two memory dimensions are complemented by **General Video Quality**, which measures imaging and aesthetic quality, and **Camera Following**, which evaluates adherence to the prescribed camera motion. Together, they characterize memory performance alongside the visual quality and control accuracy of the generated rollouts.

3.2 Data Composition

WorldMemBench comprises 260 core cases across four families: 60 Human, 70 Object, 70 Environment, and 60 Dynamic Memory cases. The first three families form the 200-case Static Memory subset. General Video Quality is evaluated on all static-memory rollouts, while Camera Following reuses 85 static-memory cases suitable for trajectory evaluation. As shown in Fig. 2, the four case families jointly cover 23 representative content categories. Human cases distinguish single- and multi-person scenes; Object and Environment cases span diverse target entities and spatial settings; and Dynamic Memory covers seven state-transition mechanisms. These 23 categories provide coverage across entity types, scene configurations, and dynamic processes.

3.3 CAMERA TRAJECTORIES

We construct translation and rotation trajectories to move the camera away from a previously observed region and subsequently return to it, as illustrated in Fig. 3. We denote the initial observation phase by V, the i-th departure by D_i, and the corresponding prescribed revisit by R_i, providing shared temporal anchors for memory evaluation. **Translation** maintains a fixed camera orientation and follows a Forward–Right–Backward–Left sequence, forming a square loop that returns to the initial position at R_1. **Rotation** keeps the camera position fixed while changing its viewing direction. Here, **L/R** denote left/right yaw rotations, and **U/D** denote upward/downward pitch rotations. Each **LR** or **UD** pair turns the camera away from and then back to its initial viewing direction, completing one departure-and-revisit cycle. Accordingly, **LRLR** and **LRLRLR** repeat the left–right cycle twice and three times, respectively. **LRUD** performs a left–right cycle followed by an up–down cycle, yielding two revisits across different rotation axes.

4 EVALUATION PROTOCOL AND METRICS

We describe the evaluation protocol and metrics for video quality, static memory, dynamic memory, and camera following.

4.1 MULTI-REVISIT PROTOCOL

For each rollout, we evaluate the revisits specified by the camera trajectory. At each revisit, the observed content is compared with a fixed reference from the initial observation phase. We compute scores for each target to track retention across returns. The overall score is obtained by averaging over revisits and then over targets:

$$S_u = \frac{1}{K}\sum_{k=1}^{K} s_{u,k}, \qquad S_{\text{rollout}} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} S_u, \tag{1}$$

where $\mathcal{U}$ is the set of targets, K is the number of prescribed revisits, and $s_{u,k}$ is the consistency score for target u at the k-th revisit. S_u is the average score for target u, and S_{rollout} is the overall rollout score. All prescribed revisits are included in K, including those without valid observations.

4.2 GENERAL VIDEO QUALITY

We assess visual quality by sampling every fourth frame of each rollout. **Imaging Quality** is measured using MUSIQ-SPAQ Ke et al. (2021), while **Aesthetic Quality** uses the LAION aesthetic predictor based on CLIP features Radford et al. (2021). Scores are averaged over sampled frames within each video, then equally across cases. Both metrics are normalized to $[0, 1]$ and reported separately, with higher values indicating better quality.

4.3 CAMERA FOLLOWING

Camera Following measures adherence to prescribed camera motion, supporting interactive exploration and helping identify viewpoint errors that can confound memory evaluation. We report translation and rotation relative pose errors (RPE). Let Q_t and $\widehat{Q}_t^{(e)}$ denote the target and estimated camera-to-world poses at frame t, where e identifies the pose estimator.

1) Translation, we normalize the target path to unit length and align the DA3 Lin et al. (2025) and ViPE Huang et al. (2025a) trajectories independently using $\mathrm{Sim}(3)$, removing differences in global coordinate frame and scale. Let $\Delta Q_t = Q_t^{-1} Q_{t+1}$ and $\Delta\widehat{Q}_t^{(e)} = (\widehat{Q}_t^{(e)})^{-1}\widehat{Q}_{t+1}^{(e)}$ denote adjacent target and estimated motions. Their translation disagreement over the full rollout is:

$$\mathrm{RPE}_{\text{trans}} = \frac{1}{2}\sum_{e\in\{\text{DA3,ViPE}\}}\sum_{t=0}^{T-2}\left\|\mathrm{trans}\Big((\Delta Q_t)^{-1}\Delta\widehat{Q}_t^{(e)}\Big)\right\|_2. \tag{2}$$

2) Rotation, only the DA3 orientations are used, without $\mathrm{Sim}(3)$ alignment. Let R_t and $\widehat{R}_t$ be the target and estimated rotations, with adjacent increments $\Delta R_t = R_t^\top R_{t+1}$ and $\Delta\widehat{R}_t = \widehat{R}_t^\top \widehat{R}_{t+1}$. We

normalize their accumulated angular disagreement by the target angular path:

$$\mathrm{RPE}_{\mathrm{rot}} = \frac{\sum_{t=0}^{T-2} \theta\Big((\Delta R_t)^\top \Delta \widehat{R}_t\Big)}{\sum_{t=0}^{T-2} \theta(\Delta R_t)}. \tag{3}$$

Here, T is the number of frames, $\mathrm{trans}(\cdot)$ extracts the translation component, and $\theta(\cdot)$ is the $\mathrm{SO}(3)$ geodesic angle. Both metrics are dimensionless and lower is better.

4.4 Static Memory

Static Memory evaluates the retention of previously observed world content across departures and revisits. All metrics use fixed references established during the initial observation phase, score each prescribed revisit independently, and aggregate the results using Eq. 1.

Human Memory. We evaluate each person identified in the conditioning frame. **1) Identity** uses ArcFace Deng et al. (2022) face embeddings. **2) Appearance** uses DINOv2 Oquab et al. (2024) features extracted within person masks to assess clothing and overall appearance. Both metrics compare valid revisit observations with initial references:

$$s_k = \frac{1}{|R_k|} \sum_{t \in R_k} \mathrm{clip}_{[0,1]} \left(\max_{r \in V} \frac{\phi(I_r)^\top \phi(I_t)}{\|\phi(I_r)\|_2 \, \|\phi(I_t)\|_2} \right), \tag{4}$$

where I_t is a valid observation of the person at frame t, ϕ is the corresponding encoder, and V and R_k denote the initial reference set and the k-th revisit set, respectively. The reference set V remains fixed throughout the rollout, and revisits without valid observations receive a score of zero. Identity and Appearance are aggregated separately using Eq. 1, with equal weight for each person. Both metrics lie in $[0, 1]$, with higher scores indicating better retention.

Object Memory. The target object is specified by a mask in the conditioning frame and located at each revisit using semantic object detection and video segmentation. **1) Appearance** follows the fixed-reference comparison in Eq. 4, with ϕ extracting DINOv2 features within object masks. **2) Geometry Consistency** compares object silhouettes after cropping each mask to its tight bounding box, resizing it while preserving its aspect ratio, and centering it on a common canvas:

$$s_k^{\mathrm{geo}} = \frac{1}{|R_k|} \sum_{t \in R_k} \max_{r \in V} \mathrm{IoU}(\mathcal{C}(M_r), \mathcal{C}(M_t)), \tag{5}$$

where M_t is the object mask at frame t and $\mathcal{C}(\cdot)$ denotes the normalization described above. This normalization removes image-plane translation and uniform scale differences while preserving silhouette changes. Both metrics lie in $[0, 1]$, with higher scores indicating better retention.

Environment Memory. We evaluate scene retention through appearance, depth ordering, and 3D structure. **1) Visible Memory** compares global DINOv2 features between the conditioning frame and each return. Within each prescribed rotational return window, we use the DA3 camera poses to select the frame whose estimated camera orientation is closest to that of the conditioning frame. Visible Memory is the mean of the clipped cosine similarities between the selected return frames and the conditioning frame. **2) Depth Spearman** computes the rank correlation between depth values at corresponding image locations in the initial and revisit views, measuring preservation of relative depth ordering independently of absolute scale. **3) Spherical Reconstruction Score** projects reference and revisit depth observations onto a shared spherical grid and compares reconstructed surfaces after global scale alignment, penalizing missing surfaces, additional surfaces, and distance discrepancies. Each angular bin stores the nearest log-range. Let H and G_k denote the reference and k-th revisit fields, while $\mathcal{C}_k$ and $\mathcal{U}_k$ denote the intersection and union of their occupied bins. We remove the global depth scale with $\delta_k = \mathrm{median}_{b \in \mathcal{C}_k}(H_b - G_{k,b})$. The three metrics are reported separately. Higher values indicate better retention for all three metrics.

$$S_{\mathrm{depth}} = \frac{1}{K} \sum_{k=1}^{K} \rho_{\mathrm{S}}(d_0, d_{r_k}), \qquad S_{\mathrm{rec}} = \frac{1}{K} \sum_{k=1}^{K} \frac{|\mathcal{C}_k|}{|\mathcal{U}_k|} \operatorname*{mean}_{b \in \mathcal{C}_k} \exp(-|H_b - (G_{k,b} + \delta_k)|). \tag{6}$$

Table 2: Quantitative comparison of video world models on WorldMemBench. ↑ / ↓ indicate higher / lower is better. The best and second-best results for each metric are shown in **bold** and underlined, respectively. App., Env., Recon., Trans., Rot., Prog., Persist., and Phys. denote Appearance, Environment, Reconstruction, Translation, Rotation, Progress, Persistence, and Physics, respectively.

Metric	HY WorldPlay	Matrix Game-3	Matrix Game-3.5	SANA-WM	YUME	LingBot World	LingBot World-v2	Evoke	DreamX World
Video Quality									
Imaging ↑	0.659	0.680	0.603	0.542	**0.713**	0.671	0.683	0.670	0.686
Aesthetic ↑	0.527	0.510	0.429	0.434	0.533	**0.545**	0.534	0.533	0.537
Static Memory									
Human ID ↑	**0.828**	0.483	0.019	0.134	0.298	0.393	0.534	0.550	0.286
Human App. ↑	**0.830**	0.502	0.043	0.251	0.551	0.661	0.645	0.762	0.556
Object App. ↑	**0.879**	0.612	0.471	0.328	0.531	0.448	0.564	0.710	0.569
Object Geometry ↑	**0.761**	0.518	0.457	0.314	0.449	0.418	0.506	0.651	0.491
Env. Visible ↑	**0.947**	0.782	0.620	0.537	0.682	0.728	0.694	0.762	0.610
Env. Depth ↑	**0.905**	0.611	0.645	0.668	0.676	0.680	0.517	0.730	0.419
Env. Recon. ↑	**0.619**	0.423	0.390	0.367	0.347	0.364	0.253	0.400	0.105
Dynamic Memory									
Overall ↑	0.294	0.046	0.240	0.415	0.775	**0.808**	0.696	0.706	0.563
Return Prog. ↑	0.292	0.050	0.183	0.342	0.708	**0.775**	0.667	0.667	0.558
Persist./Reset ↑	0.308	0.050	0.233	0.458	0.817	**0.833**	0.733	0.717	0.600
Final State ↑	0.267	0.042	0.200	0.333	0.717	**0.750**	0.592	0.658	0.492
Causal/Phys. ↑	0.308	0.042	0.342	0.525	0.858	**0.875**	0.792	0.783	0.600
Camera Follow									
Trans. RPE ↓	0.609	**0.424**	1.018	0.771	0.853	0.575	0.565	0.838	1.139
Rot. RPE ↓	**0.417**	0.554	0.837	0.464	0.529	0.637	0.439	0.573	0.928

4.5 Dynamic Memory

Dynamic Memory evaluates whether an event continues to evolve while out of view and whether its consequences remain visible after return. Each case specifies an initial state, an action, and an expected next state. We use Qwen3.8-27B Qwen Team (2026) to generate candidate binary questions and select eight questions with fixed reference answers, covering four aspects with two questions each: return progress, persistence/reset, final state, and causal/physical consistency. During evaluation, Qwen3.8-27B first determines whether the target event visibly starts during the initial hold and then answers all eight questions from the ordered rollout evidence. The event-start decision acts as a gate: if the event does not start, all eight question scores are zero; otherwise, each answer receives one point if it matches the fixed reference answer and zero otherwise. Each sub-dimension score is the mean of the two gated scores for that aspect, while Overall is the mean of all eight gated scores. All scores lie in $[0, 1]$, with higher values indicating better performance. The complete prompts and event-start decomposition are provided in Appendix D.

5 Experiment

5.1 Setup

We evaluate nine models across all dimensions of WorldMemBench: HY-WorldPlay Sun et al. (2025), Matrix-Game 3 Wang et al. (2026), Matrix-Game 3.5 Qian et al. (2026), SANA-WM Zhu et al. (2026a), YUME Mao et al. (2025), LingBot-World Team et al. (2026b), LingBot-World-v2 Gao et al. (2026), Evoke Yin et al. (2026), and DreamX-World Team et al. (2026a). All models are evaluated under consistent benchmark conditions using a unified evaluation protocol, with each model configured according to the defaults provided in its official release. Model-specific resolutions, frame rates, temporal generation strategies, control adaptations, and inference settings are detailed in Appendix C.

5.2 Evaluation of World Models

Quantitative Results. Following the evaluation protocol, we evaluate videos generated by nine world models across four dimensions: General Video Quality, Static Memory, Dynamic Memory, and Camera Following. Tab. 2 reveals distinct strengths across these dimensions. HY-WorldPlay

leads all seven Static Memory metrics, demonstrating strong retention of human, object, and environmental content. However, its Dynamic Memory score is substantially lower than those of LingBot-World and YUME, which achieve the highest and second-highest scores, respectively. These results suggest that preserving static content and maintaining evolving states pose distinct challenges. Camera-following performance also varies: Matrix-Game 3 achieves the lowest translation error, while HY-WorldPlay achieves the lowest rotation error. Overall, no model leads across all four dimensions, highlighting the need to evaluate complementary capabilities separately.

Qualitative Results. To examine whether our metrics reflect observable memory behavior, we present three multi-revisit comparisons in Appendix E. In the Human, Object, and Environment cases shown in Figs. 5, 6, and 7, HY-WorldPlay most consistently preserves subject identity, object geometry, and scene layout across three revisits. These observations align with its first-place performance on all seven Static Memory metrics in Tab. 2, providing qualitative support for our quantitative evaluation of memory consistency across different content types and repeated revisits.

Human Evaluation. To complement the qualitative analysis, we assess how well our Static and Dynamic Memory metrics align with human judgments. For Static Memory (Tab. 6), blinded A/B agreement ranges from 78.5% to 92.5% across the six metrics, with model-level Spearman correlations of 0.988, 0.952, and 0.929 for Environment Visible, Human Appearance, and Human Identity, respectively. For Dynamic Memory (Tab. 7), the checklist judge achieves a Spearman correlation of 0.732 and a Pearson linear correlation coefficient (PLCC) of 0.808 for the Overall score across 50 instances. Complete annotation protocols, per-dimension results, and interfaces are provided in Appendix F. Together, these results support the validity of our metrics for assessing world memory.

5.3 Multi-Revisit Analysis

Fig. 4 shows how the seven Static Memory metrics and Camera Rotation RPE change across nine models from R_1 to R_3 under LRLRLR, revealing two main findings: **1) Multi-Revisit Evaluation Exposes Cumulative Memory Decay.** Nine-model averages degrade monotonically across revisits on all eight metrics: Object Appearance, Human Identity, and Object Geometry drop by 0.229, 0.217, and 0.205 from R_1 to R_3, while Environment Reconstruction Score decreases by 0.196. The trend also holds per model: on the Human, Object, Environment Visible, and Environment Reconstruction metrics, every model is weaker at R_3 than at R_1, and Environment Depth degrades for eight of the nine models. Content recovered on the first return therefore does not necessarily persist across later revisits; multi-revisit evaluation exposes cumulative memory decay that single-return evaluation cannot observe. **2) Cumulative decay stems from compounding viewpoint misalignment and autoregressive error propagation.** First, the nine-model mean Camera Rotation RPE grows from 0.430 to 0.502 and 0.530, so consecutive rotations make the returning viewpoint progressively harder to align with the initially observed region. Second, all nine models generate long videos autoregressively: each new segment conditions on previously generated content, so early appearance, structure, and geometry errors propagate into later conditioning and impose longer error-propagation chains on later revisits.

5.4 Cross-Axis Revisit Robustness

LRLR and LRUD share the first yaw revisit but differ at the second: LRLR repeats the yaw cycle, whereas LRUD returns through pitch. For each higher-is-better Static Memory metric, we define $d = R_1 - R_2$ and $\Delta_{\text{axis}} = d_{\text{LRUD}} - d_{\text{LRLR}}$. Across nine models, mean Δ_{axis} is positive for all seven metrics, and 50 of 63 model–metric pairs degrade more under LRUD; Environment Visible shows the largest additional degradation at 0.152. Switching exploration axes therefore introduces difficulty beyond same-axis multi-revisit evaluation. Trajectory-wise means and full per-model results are provided in Appendix G.Appendix G.1 reports departure/return coverage and evaluates both trends on tracking-supported and pose-qualified subsets.

5.5 Diagnostic Case Study: Training-Free Context Access

The main results reveal distinct performance profiles for static content recovery and dynamic state evolution. To examine how context access affects these capabilities, we compare three training-free policies on LingBot World v2, keeping the model and other generation settings fixed. FullKV

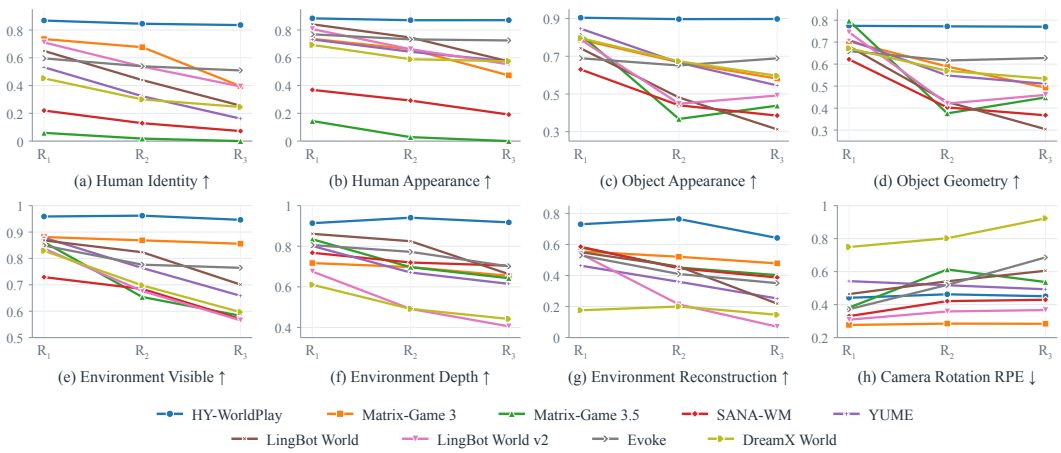

Figure 4: Per-model changes of the nine models across the three revisits of LRLRLR, four metrics per row. ↑/↓ indicate higher/lower is better; all panels keep the original value direction of each metric. The red Δ marks the largest single-model change from R_1 to R_3 on each metric.

Table 3: Controlled comparison of context-access policies on LingBot World v2. Best and second-best results are **bold** and underlined.

Config	Human		Object		Environment			Dynamic Memory				Inference Speed	
	ID ↑	App. ↑	App. ↑	Geometry ↑	Visible ↑	Depth ↑	Recon. ↑	Return ↑	Persist. ↑	Final ↑	Causal ↑	Latency (s) ↓	Speedup ↑
FullKV	0.534	0.645	**0.564**	**0.506**	**0.694**	0.517	0.253	0.667	0.733	0.592	0.792	322.16	1.00×
FoV	**0.547**	**0.663**	0.531	0.485	0.687	**0.550**	**0.287**	**0.725**	**0.792**	**0.675**	**0.800**	262.69	1.23×
Sliding Windows	0.392	0.540	0.426	0.392	0.610	0.501	0.235	0.667	0.717	0.600	0.775	**260.42**	**1.24×**

provides access to the complete history. Sliding Windows uses recent context within a fixed budget, while FoV combines recent context with view-relevant historical chunks under the same budget. FoV achieves the best scores on eight of the eleven capability metrics and outperforms Sliding Windows on all eleven. Compared with FullKV, it scores higher on all four Dynamic Memory sub-metrics and both Human metrics. FullKV, however, retains an advantage in Object Appearance, Object Geometry, and Environment Visible, suggesting that the benefits of context selection vary across memory dimensions. Both budgeted policies also reduce inference time. Across 660 matched videos, mean end-to-end latency is 322.2 s for FullKV, compared with 262.7 s for FoV and 260.4 s for Sliding Windows. These correspond to speedups of $1.23\times$ and $1.24\times$, respectively. Overall, the results suggest that combining recent context with view-relevant history offers a favorable balance between memory performance and efficiency. Context budgets, evaluation sets, timing protocols, and further analysis are provided in Appendix H.

6 Conclusion

In this paper, we introduce WorldMemBench, a benchmark for evaluating world memory in video world models through controlled departure-and-revisit rollouts. Across 260 cases, it separately assesses the retention of previously observed content, event evolution while out of view, general video quality, and adherence to prescribed camera trajectories. Evaluation of nine representative models shows that no model consistently leads across all dimensions and that strong static retention does not necessarily imply reliable dynamic state evolution. Multi-revisit evaluation reveals cumulative memory degradation beyond the first return, while cross-axis revisits expose additional sensitivity to changes in exploration direction. Human studies further support the alignment between automatic assessments and human judgments. With the model and generation settings held fixed, a controlled context-access study reveals how different temporal context policies affect memory performance and inference efficiency. Together, these findings establish WorldMemBench as both an evaluation benchmark and a diagnostic tool for improving world memory. We hope that this benchmark will help advance video world models toward reliable simulators that support long-term interaction, learning, and planning in persistent virtual environments.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

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. In *European Conference on Computer Vision (ECCV)*, 2026.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3087709. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

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

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. *arXiv preprint arXiv:2607.07534*, 2026.

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. *arXiv preprint arXiv:2608.27328*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception, 2025a. URL https://arxiv.org/abs/2508.10934.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, et al. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

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. *arXiv preprint arXiv:2511.10647*, 2025.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

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, 2021. URL https://arxiv.org/abs/2103.00020.

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. *arXiv preprint arXiv:2512.14614*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

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. *arXiv preprint arXiv:2601.20540*, 2026b.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 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.

Yuxue Yang, Shuyao Shang, Jiahe Wang, Zitong Zhou, Liang Tan, Junhan Zeng, Ruizhi Li, Junyan Li, Yu Liu, Xiao Yang, Yong Li, Jun Zhu, Hongsheng Li, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. Worldexam: Benchmarking world models from apparent appearance to inherent reactivity. *arXiv preprint arXiv:2608.02603*, 2026.

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. *arXiv preprint arXiv:2602.08025*, 2026.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world, 2026. URL https://arxiv.org/abs/2608.13546.

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.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval, 2025. URL https://arxiv.org/abs/2506.03141.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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. *arXiv preprint arXiv:2606.00793*, 2026.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 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. *arXiv preprint arXiv:2605.15178*, 2026a.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026b.

A Overview of the Appendix

In this section, we briefly introduce the contents of the appendix. We begin with a statement on LLM usage during manuscript preparation. In Sec. C, we detail the inference settings and output video specifications of the evaluated models. In Sec. D, we describe the Dynamic Memory assessment protocol and provide the corresponding prompt template. In Sec. E, we present qualitative comparisons across Human, Object, and Environment cases. In Sec. F, we describe the human evaluation protocols and annotation interfaces, and report the agreement between automatic assessments and human judgments. In Sec. G, we provide trajectory-wise results and cross-axis comparisons. In Sec. H, we analyze how context-access policies affect static memory retention and dynamic state progression. Finally, in Sec. I, we discuss the limitations of WorldMemBench and directions for future work.

B LLM Usage Statement

During the development of this work, LLMs were used primarily to identify and correct grammatical errors and occasionally as rewriting tools to improve the clarity and readability of the text.

C Inference Settings

All formal inference runs on machines equipped with eight NVIDIA H20-3e GPUs. Every model uses frozen random seeds and inference configurations, so that no setting varies across cases or trajectories.

HY-WorldPlay. HY-WorldPlay consumes per-frame OpenCV camera-to-world poses; we extract one latent camera anchor every four pixel frames and generate its native pose condition with fixed camera intrinsics. Formal inference uses seed 1, 4-step distilled AR sampling, and eight-card model parallelism, with prompt rewriting and super-resolution disabled.

Matrix-Game 3. The native interface of Matrix-Game 3 is jointly controlled by keyboard translation and mouse rotation. We compile the normalized camera trajectory into per-frame keyboard/mouse conditions. The model runs 12 iterations with seed 42 and 3-step AR sampling: the first segment generates 57 frames and each subsequent iteration adds 40 frames, producing 497 frames in total; inference uses seven GPUs for diffusion and one GPU for asynchronous MG-LightVAE decoding.

Matrix-Game 3.5. Matrix-Game 3.5 constructs its native PRoPE condition from per-case camera intrinsics and camera-to-world poses. Formal inference adopts the released six-block distilled configuration, where each block corresponds to 84 camera poses, with a three-step schedule of $[1000, 667, 333]$ and seed 3407. The six blocks share the same content prompt and complete a continuous 505-frame rollout through the generated history and dynamic context.

SANA-WM. SANA-WM builds chunk-level Plücker conditions from per-case camera intrinsics and poses, and generates videos through a streaming pipeline consisting of stage-1, a refiner, and a causal VAE. Stage-1 uses 4-step self-forcing sampling, and both stage-1 and the refiner treat three latent frames as one temporal block; formal inference fixes seed 42 and outputs 481 frames.

YUME. YUME does not directly consume numerical camera poses. We convert the normalized trajectory into segmented camera captions composed of Left, Right, Up, Down, Hold, and translation directions; each caption stage corresponds to 29 frames, and the MemoryBench content prompt is appended verbatim to every stage. Formal inference disables additional prompt refinement and uses seed 43 with 4-step Euler sampling; LRLR and LRUD produce 493 frames, while LRLRLR, which contains three revisits, produces 551 frames.

LingBot World. LingBot World interpolates per-frame camera-to-world poses to the latent rate and converts them into framewise relative poses and Plücker conditions; camera intrinsics are fixed on the model's input canvas. We set the chunk size to four latent frames so that the 493-frame target trajectory is generated in full, and run eight-card causal AR inference with seed 1 and FullKV attention.

LingBot World v2. LingBot World v2 uses the same pose, intrinsics, and four-latent-frame chunk protocol as LingBot World. The nine-model main results use seed 1 and FullKV attention; FoV and

Table 4: Video specifications and sampling configurations of the formal inference artifacts. Resolution and FPS are the models' native outputs; Frames / Duration reports the actual frame count of each output video together with its duration; Seed and Sampling give the sampling configuration of formal inference; Inference Setting only distinguishes explicit FullKV from the models' built-in memory modes.

Model	Resolution	FPS	Frames / Duration	Seed	Sampling	Inference Setting
HY-WorldPlay	848×480	24	493 / 20.5 s	1	4-step AR	Built-in
Matrix-Game 3	1280×704	17	497 / 29.2 s	42	3-step AR ×12	Built-in
Matrix-Game 3.5	1280×704	16	505 / 31.6 s	3407	3-step, six-block	Built-in
SANA-WM	1280×704	16	481 / 30.1 s	42	4-step self-forcing	FullKV
YUME	1280×704	16	493–551 / 30.8–34.4 s	43	4-step Euler	Built-in
LingBot World	832×464	16	493 / 30.8 s	1	causal AR	FullKV
LingBot World v2	832×464	16	493 / 30.8 s	1	causal AR	FullKV
Evoke	640×384	24	501 / 20.9 s	44	3-step windowed AR	Built-in
DreamX World	1280×704	16	501 / 31.3 s	42	distilled AR	FullKV

Sliding Windows in Sec. 5.5 are used only for the context-mechanism comparison with all other conditions frozen, and do not change the inference configuration behind the formal main table.

Evoke. Evoke uses per-case camera intrinsics and a 501-frame absolute camera-to-world trajectory. The camera trajectory undergoes no additional smoothing or segmented motion cropping: absolute poses drive geometric warping, and latent-rate relative Plücker poses serve as the camera condition. Formal inference fixes seed 44, adopts 3-step CFG-free windowed AR generation, and keeps 21 static pose samples at the end to cover the final return view.

DreamX World. DreamX World performs camera forcing with chunk-relative PRoPE conditions. Formal inference fixes seed 42, decodes 126 latent frames into 501 pixel frames, and retains the full generation history with FullKV attention.

D Dynamic Memory Evaluation Details

D.1 Assessment Protocol

Each rollout is sampled at a fixed stride of 8 frames, always including the first and last frames. The sampled frames are provided to Qwen3.8-27B in chronological order, together with the checklist and the planned camera-phase ranges (initial hold, departures, revisits, and final stationary hold). These ranges allow the assessment model to locate the evidence window for each question without access to the generation prompt. The model first determines whether the central event visibly starts during the initial hold and then answers all eight questions. If the model determines that the event has not visibly started, all eight question scores are set to zero. Otherwise, each question receives one point if the model's answer matches the frozen reference answer. We set `max_tokens` to 4096 for the assessment model.

D.2 Event Initiation and Conditional Dynamic Scores

We separate the event-start prerequisite from subsequent checklist performance using the same 60 videos per model. The event-start rate p_{start} is the fraction of videos judged to show concrete event progress during the initial hold. The conditional score S_{cond} averages the eight-question checklist score only over those videos. Because videos with a negative gate receive zero, the reported Overall satisfies $S_{\mathrm{gated}} = p_{\mathrm{start}} S_{\mathrm{cond}}$.

HY-WorldPlay has a conditional score of 0.928 on 19 started events, but its 31.7% start rate yields a gated Overall of 0.294. LingBot World starts 56 of 60 events and obtains a conditional score of 0.866, yielding 0.808 overall. Thus, differences in the headline Dynamic Memory scores reflect both event initiation and subsequent checklist performance. Matrix-Game 3's conditional score of 0.917 is based on only three videos. We retain the gated score as the end-to-end benchmark result and use the decomposition to distinguish these sources of failure. The conditional scores remain judge-based evaluations of different successful-start subsets, not isolated measures of memory.

Table 5: Dynamic Memory decomposition from the frozen 60-case evaluation. Started is the number of judge-positive event-start decisions. Conditional averages only the Started subset.

Model	Started	Start rate (%)	Gated Overall	Conditional
HY-WorldPlay	19/60	31.7	0.294	0.928
Matrix-Game 3	3/60	5.0	0.046	0.917
Matrix-Game 3.5	23/60	38.3	0.240	0.625
SANA-WM	35/60	58.3	0.415	0.711
YUME	54/60	90.0	0.775	0.861
LingBot World	56/60	93.3	0.808	0.866
LingBot World v2	53/60	88.3	0.696	0.788
Evoke	49/60	81.7	0.706	0.865
DreamX World	39/60	65.0	0.563	0.865

D.3 Assessment Prompt

Below, we present the prompt template used for Dynamic Memory assessment.

Dynamic Memory Assessment Prompt

You are judging one generated Dynamic Memory LR-hold video using eight case-specific questions. The attached images are sampled in chronological order. The target generation prompt and expected answers are deliberately hidden.

Use only visible evidence; do not complete a plausible story from the question wording. Question wording describes the intended evaluation target, but it is not visual evidence. For a change claim, cite visible before-and-after images that show a concrete difference; otherwise answer no. If the required subject or object disappears, is replaced, or changes into a different scene, do not treat its absence as successful completion. In the reason, describe only concrete evidence visible in the cited images and do not invent an unseen state or event.

The evaluator has assigned the images to the model's planned camera phases. Each range below is an inclusive pair of 1-based image numbers.

Planned phase image ranges: <planned phase ranges JSON: initial_hold, departures, revisits, final_stationary>

Useful comparison ranges: <compact comparison ranges JSON>

Phase guidance:
- returned_view means the returned_hold range, compared with the departure_baseline range when the question asks about change.
- returned_hold_sequence means the progression inside the returned_hold range.
- returned_and_later_visible_sequence means the returned_hold range through its final images.
- final_stationary_view means the final_stationary range.
- ordered_visible_sequence may use the full chronological image sequence.
- For a comparison or temporal claim, use visible before/after evidence when available.
- If an event never occurs, say that directly. Never invent a reset, reversal, worsening, disappearance, or change that is not visible.
- Interpret extent words relative to the stated baseline, not by absolute impression.
- Camera motion, framing, lighting, an action pose, or a replaced subject alone is not evidence of state progress.
- If the exact state is not visibly established, answer no.
- Cite only decisive 1-based image numbers.

Event-start prerequisite:
- First decide whether the central event described by the questions visibly starts during initial_hold, before the camera leaves the original view.
- event_started is yes only when at least two initial_hold images show concrete temporal progress by the principal subject or object.
- Mere pose, contact, camera motion, background motion, or a plausible intention is not an event start.
- If no relevant action or state change is visibly underway by the end of initial_hold, event_started is no.
- Even when event_started is no, answer all eight questions from visible evidence; the evaluator applies the zero-score gate.

Questions (expected answers are intentionally hidden): <checklist JSON: id, category, evidence_window, question for each of the eight frozen questions>

Output format. Return one JSON object with exactly two keys:

```
- event_started: {answer: "yes" or "no"; evidence_image_numbers: at least two decisive
1-based image numbers from initial_hold; reason: one concise sentence describing the
visible change, or the visible lack of change}.
- answers: exactly eight objects in the same ID order, each {id: the question ID;
answer: "yes" or "no"; evidence_image_numbers: a short list of decisive 1-based image
numbers; reason: one concise sentence describing only the cited visible evidence}.
Return JSON only.
```

E QUALITATIVE RESULTS

In this section, we present qualitative comparisons of the video world models evaluated in Sec. 5.2. Target regions in the Human and Object cases are highlighted with red boxes, and the corresponding metric scores are shown for each revisit. Across all three cases, HY-WorldPlay provides the most consistent multi-revisit recovery: it preserves the subject's identity and appearance, the wardrobe's appearance and geometry, and the room's visual content and spatial structure through R_3. This observation matches Tab. 2, where HY-WorldPlay leads all seven Static Memory metrics. Evoke also remains comparatively stable in the Human and Object examples, consistent with its second-best results on the four entity-level metrics, while other models show identity drift, target disappearance, or scene-layout changes at later revisits.

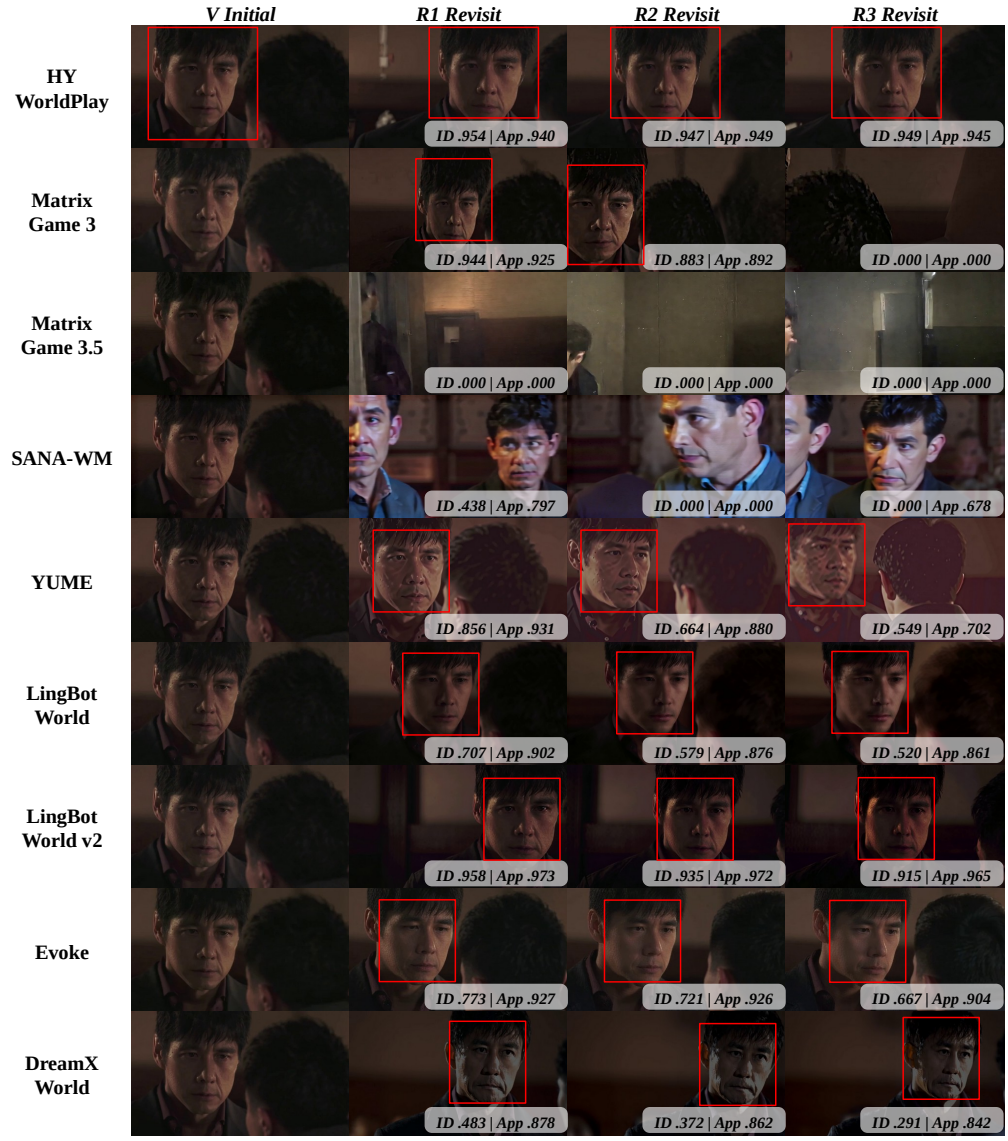

Figure 5: Qualitative comparison of video world models on a human case.

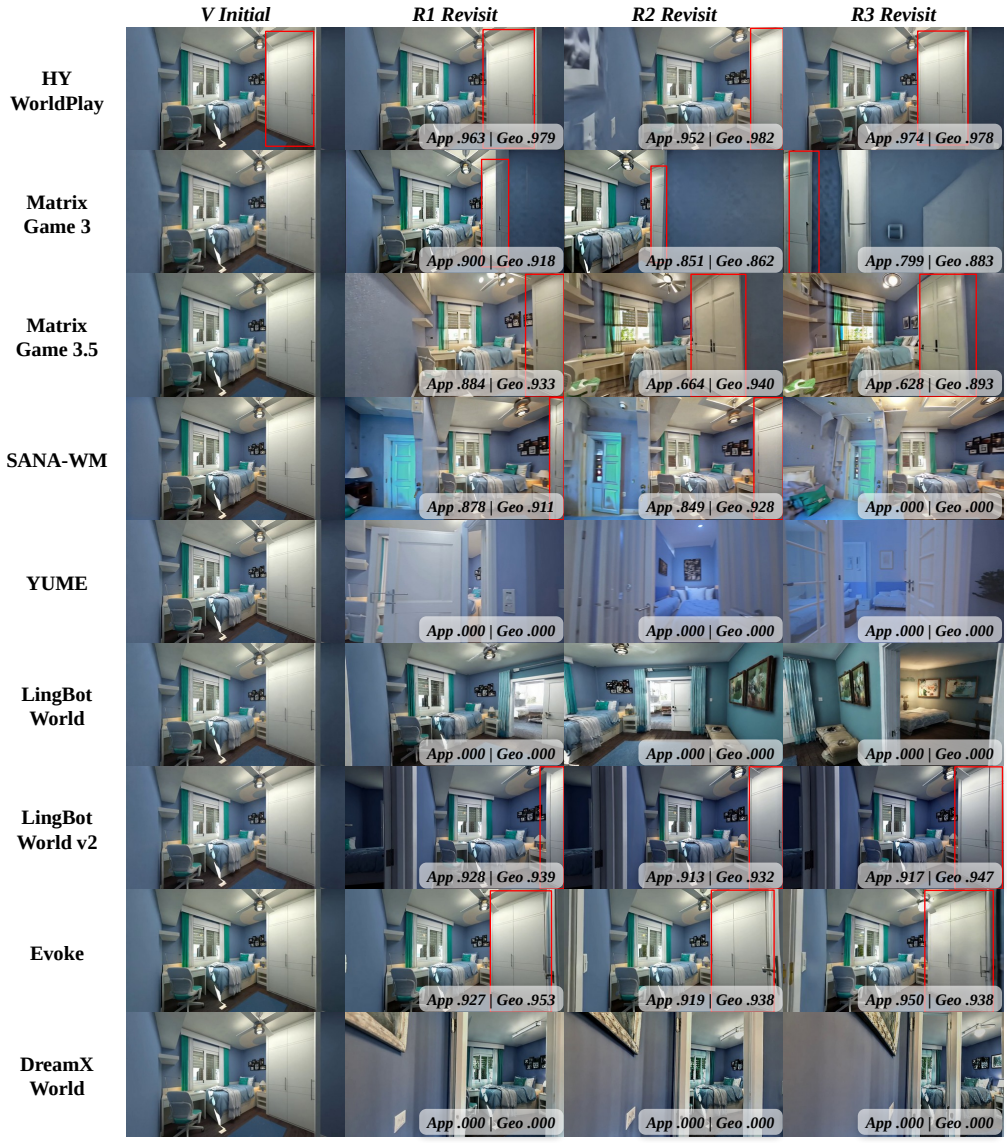

Figure 6: Qualitative comparison of video world models on an object case.



Figure 7: Qualitative comparison of video world models on an environment case.

F HUMAN EVALUATION

Static Memory protocol. We conduct a blinded A/B preference study on whether the six Static Memory metrics reproduce human judgments of long-term memory retention. In total, 504 blinded pairs are each judged by three annotators as A-wins, Tie, or B-wins, with model identities and order hidden, yielding 1,512 judgements (252 per metric). Pairwise agreement between the automatic metric and the human majority-vote direction serves as the primary measure; we further compare model-level win rates (Win/Tie/Loss as $1/0.5/0$) via Spearman ρ.

Pairwise agreement compares the automatic preference with a decisive human majority. Majority ties and three-way A/Tie/B splits are excluded from this denominator and reported separately. For example, Human Identity has 65 agreements among 76 decisive pairs, giving 85.5%; one majority tie and seven three-way splits account for the remaining eight pairs. Model-level Spearman ρ instead uses all raw votes to compute human win rates (Win/Tie/Loss as $1/0.5/0$), compared with automatic win rates on the same 84 tasks.

All six metrics reach 78.5%–92.5% pairwise agreement with human majority votes (Tab. 6). At the model-ranking level, Environment Visible, Human Appearance, and Human Identity attain $\rho = 0.988$, 0.952, and 0.929, confirming that the automatic metrics capture human-perceivable differences in long-term memory retention.

Table 6: Agreement between automatic Static Memory metrics and blinded human A/B preferences. Pairwise Agreement measures the per-pair directional consistency between an automatic metric and the human majority vote; model-level Spearman ρ is computed over the human and automatic win rates of the eight models.

Metric	Agree/Decisive	Tie/Split	Agreement (%)	Spearman ρ
Human Identity	65/76	1/7	85.5	0.929
Human Appearance	71/80	2/2	88.8	0.952
Object Geometry	62/79	0/5	78.5	0.606
Object Appearance	67/80	2/2	83.8	0.699
Environment Visible	74/80	1/3	92.5	0.988
Environment Depth	63/77	4/3	81.8	0.301

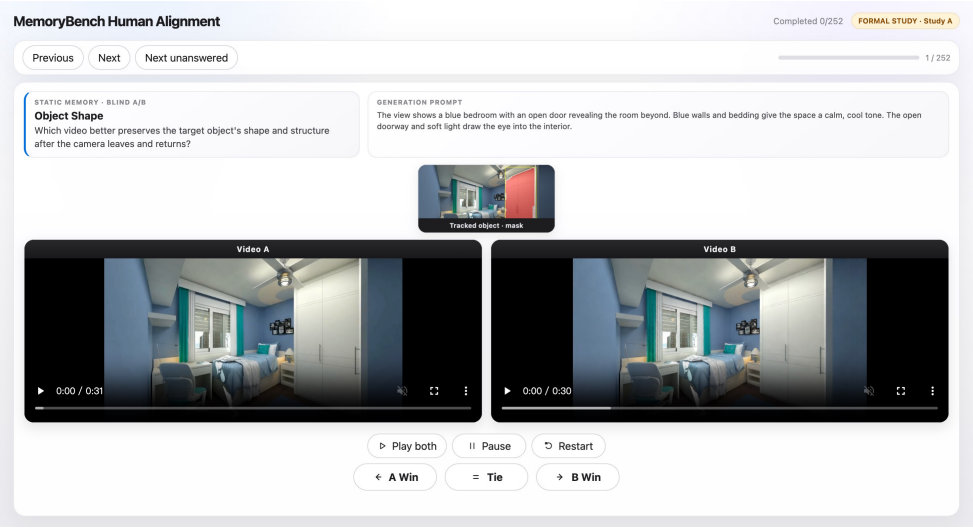

Figure 8: Blinded A/B preference annotation interface for Static Memory.

Dynamic Memory protocol. We validate the Dynamic Memory VLM judge on 50 evaluation instances with 400 checklist items. Three annotators independently answer the eight binary questions of each instance under the same event gate as the judge: if the target event has not started, all items of that instance are labeled 0. Each item thus receives three judgements (1,200 in total), whose

majority vote forms the human reference label; human and VLM scores are the respective fractions of satisfied checklist items per instance.

The judge attains an Overall Spearman $\rho = 0.732$ and PLCC of 0.808 against human evaluation (Tab. 7). Across the four sub-dimensions, ρ spans 0.401–0.684 and PLCC spans 0.491–0.683, with the highest agreement on Return Progress ($\rho = 0.684$, PLCC $= 0.683$). These results validate checklist-based Dynamic Memory evaluation as an effective and scalable automatic protocol.

Table 7: Human alignment of the Dynamic Memory checklist evaluation. The per-item majority vote of three annotators forms the human reference labels; Spearman ρ and PLCC are computed over the checklist satisfaction scores of the 50 instances.

Task	Checklist Items	Judgements	Spearman $\rho \uparrow$	PLCC $\uparrow$
Return Progress	100	300	0.684	0.683
Persistence / Reset	100	300	0.582	0.612
Final State	100	300	0.624	0.641
Causal / Physics	100	300	0.401	0.491
Overall	400	1200	0.732	0.808

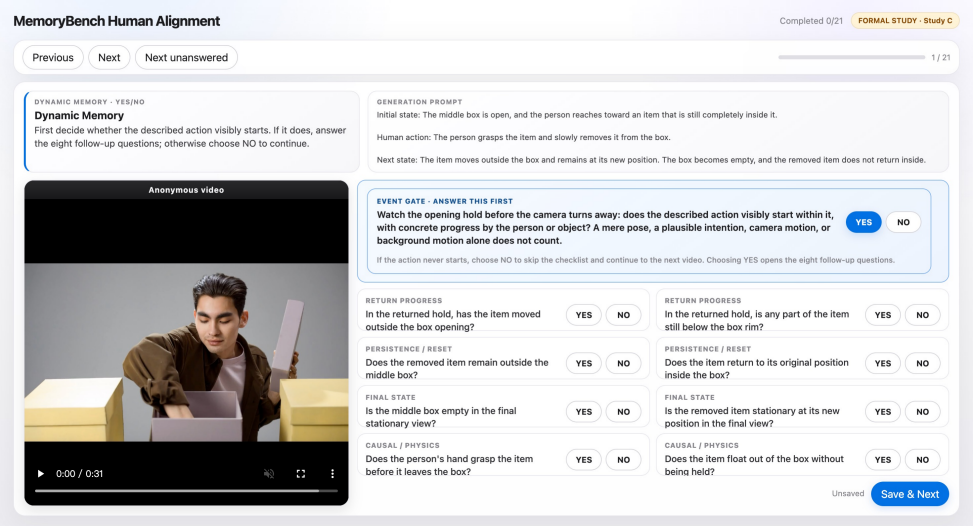

Figure 9: Event-gated checklist annotation interface for Dynamic Memory.

G Trajectory-Wise and Cross-Axis Results

In this section, we supplement the aggregate main results with per-model results for LRLR, LRLRLR, and LRUD, followed by the LRLR–LRUD comparison underlying the cross-axis analysis in Sec. 5.4. Environment Reconstruction is reported as a higher-is-better score throughout.

Trajectory-wise results. Tab. 8 shows the results of all nine models on the three rotation trajectories. HY-WorldPlay achieves the highest scores on all seven Static Memory metrics under LRLR and LRLRLR, and on six under LRUD, where Evoke leads in Human Appearance. Meanwhile, YUME achieves the highest Imaging scores across all three trajectories, indicating that strong visual quality does not necessarily coincide with the strongest memory retention.

Average cross-axis degradation. Tab. 9 summarizes the nine-model mean scores at the two revisits, with $d = R_1 - R_2$ measuring revisit-to-revisit degradation and $\Delta_{\text{axis}} = d_{\text{LRUD}} - d_{\text{LRLR}}$ measuring the excess degradation associated with switching rotation axes. Both trajectories show lower mean scores at the second revisit across all seven metrics, with larger drops under LRUD. The largest excess degradation occurs in Environment Visible ($+0.152$), followed by Object Appearance ($+0.127$) and Object Geometry ($+0.120$).

Table 8: Per-model results on the three formal Rotation trajectories. Best and second-best results within each trajectory are bold and underlined.

Metric	HY-WorldPlay	Matrix-Game 3	Matrix-Game 3.5	SANA-WM	YUME	LingBot World	LingBot World v2	Evoke	DreamX World
(a) LRLR									
Imaging ↑	0.668	0.686	0.613	0.554	**0.712**	0.677	0.698	0.673	0.683
Aesthetic ↑	0.538	0.522	0.446	0.451	0.539	**0.556**	0.553	0.541	0.530
Human ID ↑	**0.863**	0.577	0.015	0.150	0.295	0.389	0.559	0.546	0.267
Human App. ↑	**0.889**	0.597	0.041	0.282	0.529	0.643	0.702	0.783	0.522
Object App. ↑	**0.925**	0.647	0.510	0.311	0.498	0.442	0.626	0.744	0.486
Object Geometry ↑	**0.847**	0.547	0.501	0.303	0.411	0.422	0.559	0.689	0.421
Env. Visible ↑	**0.953**	0.818	0.677	0.529	0.691	0.787	0.700	0.746	0.563
Env. Depth ↑	**0.920**	0.640	0.666	0.663	0.689	0.743	0.515	0.697	0.378
Env. Recon. ↑	**0.657**	0.460	0.411	0.342	0.378	0.440	0.281	0.404	0.073
Rot. RPE ↓	**0.366**	0.373	0.656	0.429	0.506	0.586	0.371	0.556	0.961
(b) LRLRLR									
Imaging ↑	0.687	0.688	0.606	0.511	**0.711**	0.672	0.690	0.670	0.691
Aesthetic ↑	**0.559**	0.522	0.445	0.432	0.539	0.554	0.550	0.539	0.549
Human ID ↑	**0.850**	0.602	0.026	0.140	0.339	0.448	0.547	0.548	0.333
Human App. ↑	**0.875**	0.624	0.057	0.284	0.650	0.720	0.673	0.742	0.619
Object App. ↑	**0.900**	0.679	0.540	0.485	0.686	0.512	0.576	0.676	0.688
Object Geometry ↑	**0.772**	0.595	0.540	0.464	0.590	0.471	0.542	0.634	0.592
Env. Visible ↑	**0.956**	0.868	0.699	0.661	0.767	0.797	0.693	0.797	0.708
Env. Depth ↑	**0.924**	0.689	0.724	0.731	0.695	0.783	0.524	0.759	0.515
Env. Recon. ↑	**0.712**	0.518	0.478	0.473	0.358	0.409	0.274	0.430	0.174
Rot. RPE ↓	0.453	**0.282**	0.511	0.393	0.532	0.538	0.348	0.541	0.825
(c) LRUD									
Imaging ↑	0.621	0.667	0.592	0.559	**0.715**	0.664	0.662	0.667	0.686
Aesthetic ↑	0.486	0.487	0.395	0.419	0.521	0.527	0.500	0.518	**0.531**
Human ID ↑	**0.771**	0.272	0.015	0.110	0.259	0.343	0.494	0.557	0.259
Human App. ↑	0.727	0.285	0.030	0.188	0.473	0.619	0.560	**0.763**	0.528
Object App. ↑	**0.814**	0.510	0.363	0.186	0.409	0.390	0.491	0.710	0.534
Object Geometry ↑	**0.664**	0.413	0.329	0.175	0.345	0.362	0.417	0.628	0.459
Env. Visible ↑	**0.933**	0.661	0.483	0.421	0.589	0.601	0.690	0.743	0.559
Env. Depth ↑	**0.871**	0.504	0.545	0.610	0.644	0.514	0.513	0.734	0.363
Env. Recon. ↑	**0.488**	0.292	0.282	0.287	0.303	0.242	0.205	0.367	0.066
Rot. RPE ↓	**0.433**	1.008	1.343	0.569	0.549	0.786	0.598	0.622	0.997

Table 9: Nine-model mean Static Memory scores at the two revisits of LRLR and LRUD. $d = R_1 - R_2$ and $\Delta_{\text{axis}} = d_{\text{LRUD}} - d_{\text{LRLR}}$; positive Δ_{axis} indicates greater degradation after switching from yaw to pitch.

Metric	LRLR			LRUD			Δ_{axis}
	R_1	R_2	d	R_1	R_2	d	
Human ID ↑	0.490	0.324	0.166	0.435	0.249	0.186	+0.020
Human App. ↑	0.623	0.486	0.137	0.556	0.371	0.185	+0.049
Object App. ↑	0.657	0.496	0.162	0.634	0.345	0.288	+0.127
Object Geometry ↑	0.588	0.457	0.132	0.547	0.296	0.251	+0.120
Env. Visible ↑	0.781	0.655	0.126	0.770	0.492	0.278	**+0.152**
Env. Depth ↑	0.703	0.611	0.091	0.688	0.489	0.199	+0.107
Env. Recon. ↑	0.454	0.312	0.141	0.386	0.177	0.209	+0.068

Per-model cross-axis comparison. Tab. 10 provides the complete per-model revisit scores and cross-axis differences. All nine models exhibit positive Δ_{axis} on Environment Visible, and eight of nine exhibit positive values on both Object Appearance and Object Geometry. The magnitude nevertheless varies substantially across models: LingBot World shows excess degradation of $+0.368$ on Environment Visible and $+0.465$ on Environment Depth, whereas Evoke shows $+0.002$ and -0.098, respectively. These results indicate that the average cross-axis trend masks substantial differences in sensitivity across models and metrics.

Table 10: Full per-model LRLR–LRUD cross-axis comparison. Panels (a) and (b) report the two revisit scores; panel (c) reports Δ_{axis}. All memory metrics are higher-is-better.

Metric	Slot	HY WorldPlay	Matrix Game-3	Matrix Game-3.5	SANA-WM	YUME	LingBot World	LingBot World-v2	Evoke	DreamX World	Mean
(a) LRLR: two yaw revisits											
Human ID ↑	$R1$	0.873	0.738	0.030	0.229	0.431	0.517	0.663	0.579	0.349	0.490
	$R2$	0.853	0.416	0.000	0.072	0.159	0.261	0.456	0.514	0.186	0.324
Human App. ↑	$R1$	0.894	0.719	0.076	0.395	0.620	0.734	0.785	0.799	0.582	0.623
	$R2$	0.884	0.476	0.006	0.168	0.438	0.553	0.618	0.767	0.463	0.486
Object App. ↑	$R1$	0.928	0.700	0.694	0.324	0.596	0.605	0.661	0.745	0.663	0.657
	$R2$	0.922	0.593	0.327	0.299	0.399	0.279	0.592	0.743	0.308	0.496
Object Geometry ↑	$R1$	0.859	0.589	0.665	0.304	0.483	0.553	0.603	0.696	0.543	0.588
	$R2$	0.835	0.506	0.337	0.302	0.339	0.291	0.516	0.683	0.300	0.457
Env. Visible ↑	$R1$	0.959	0.830	0.795	0.654	0.763	0.825	0.787	0.777	0.642	0.781
	$R2$	0.946	0.806	0.560	0.404	0.619	0.748	0.613	0.714	0.484	0.655
Env. Depth ↑	$R1$	0.937	0.674	0.753	0.731	0.726	0.781	0.594	0.730	0.397	0.703
	$R2$	0.903	0.606	0.580	0.596	0.652	0.704	0.435	0.664	0.360	0.611
Env. Recon. ↑	$R1$	0.723	0.478	0.490	0.484	0.450	0.482	0.419	0.443	0.111	0.454
	$R2$	0.591	0.442	0.331	0.200	0.306	0.397	0.142	0.364	0.035	0.312
(b) LRUD: yaw followed by pitch											
Human ID ↑	$R1$	0.840	0.364	0.031	0.205	0.427	0.504	0.633	0.580	0.337	0.435
	$R2$	0.702	0.180	0.000	0.016	0.091	0.181	0.356	0.535	0.182	0.249
Human App. ↑	$R1$	0.828	0.396	0.060	0.341	0.593	0.729	0.735	0.742	0.580	0.556
	$R2$	0.625	0.174	0.000	0.035	0.354	0.508	0.384	0.783	0.475	0.371
Object App. ↑	$R1$	0.876	0.667	0.633	0.310	0.598	0.593	0.634	0.740	0.651	0.634
	$R2$	0.751	0.353	0.093	0.062	0.219	0.186	0.347	0.681	0.417	0.345
Object Geometry ↑	$R1$	0.754	0.534	0.583	0.292	0.475	0.542	0.552	0.652	0.541	0.547
	$R2$	0.574	0.293	0.075	0.059	0.215	0.183	0.283	0.604	0.377	0.296
Env. Visible ↑	$R1$	0.951	0.829	0.767	0.598	0.764	0.823	0.780	0.775	0.645	0.770
	$R2$	0.914	0.493	0.200	0.244	0.414	0.378	0.600	0.710	0.473	0.492
Env. Depth ↑	$R1$	0.898	0.648	0.726	0.688	0.737	0.785	0.597	0.718	0.394	0.688
	$R2$	0.845	0.360	0.363	0.532	0.551	0.243	0.429	0.750	0.331	0.489
Env. Recon. ↑	$R1$	0.541	0.470	0.369	0.354	0.424	0.482	0.294	0.433	0.106	0.386
	$R2$	0.435	0.115	0.194	0.220	0.183	0.002	0.116	0.302	0.026	0.177
(c) Cross-axis excess degradation											
Human ID ↑	Δ_{axis}	+0.119	-0.138	+0.000	+0.031	+0.063	+0.068	+0.069	-0.020	-0.008	**+0.020**
Human App. ↑	Δ_{axis}	+0.193	-0.021	-0.010	+0.079	+0.058	+0.040	+0.185	-0.073	-0.013	**+0.049**
Object App. ↑	Δ_{axis}	+0.120	+0.207	+0.173	+0.223	+0.181	+0.082	+0.218	+0.057	-0.121	**+0.127**
Object Geometry ↑	Δ_{axis}	+0.156	+0.158	+0.181	+0.231	+0.117	+0.096	+0.182	+0.035	-0.080	**+0.120**
Env. Visible ↑	Δ_{axis}	+0.025	+0.312	+0.331	+0.104	+0.207	+0.368	+0.007	+0.002	+0.013	**+0.152**
Env. Depth ↑	Δ_{axis}	+0.020	+0.220	+0.191	+0.021	+0.113	+0.465	+0.008	-0.098	+0.026	**+0.107**
Env. Recon. ↑	Δ_{axis}	-0.027	+0.320	+0.017	-0.149	+0.097	+0.395	-0.099	+0.053	+0.004	**+0.068**

G.1 Memory Evaluation under Departure and Return Controls

Inaccurate camera returns can reduce memory scores even when scene content is preserved. We therefore test whether multi-revisit degradation and the cross-axis penalty persist among videos with evidence of departure and return. This analysis uses the existing evaluation results and leaves the main benchmark scores unchanged.

Controlling departure and return. For Human and Object, we retain videos in which the tracker provides departure evidence and valid observations after return for every initially evaluated target at every compared revisit. Targets that remain continuously visible do not satisfy this criterion. For Environment, DA3 poses must indicate a departure of at least 30° from the initial orientation and a return within 30° of it. Tab. 11 reports the resulting coverage. These criteria use tracking and estimated orientation as evidence of the executed trajectory; Reconstruction retains the same zero-score rule for invalid revisits as the main evaluation.

Comparisons within each model use a fixed set of cases: all three returns must qualify for LRLRLR, and both returns in both trajectories must qualify for LRLR–LRUD. We compute each case's revisit decrease or cross-axis difference before averaging over cases and then equally over models with qualifying cases. The unfiltered comparison uses the same models. Tab. 12 reports the retained sample sizes alongside the effects, since some model-specific subsets contain only one or two cases.

Multi-revisit degradation persists after controlling return accuracy. Among the 431 qualifying Environment model–case pairs across all nine models, Visible, Depth, and Reconstruction decrease from R_1 to R_3 by 0.152, 0.135, and 0.139, respectively. Visible and Reconstruction decline for all nine models, while Depth declines for eight. These decreases are smaller than the corresponding

Table 11: Tracking-supported target revisits and pose-qualified Environment returns. Entity columns are percentages averaged over targets, prescribed revisits, and videos; Object uses the shared appearance/geometry tracking evidence. Environment percentages use 490 slots per model. D: departure $\geq 30^\circ$; R_{30}: return error $\leq 30^\circ$; Q_{30}: departure $\geq 30^\circ$ and return error $\leq 30^\circ$. The last columns count retained Environment model–case units for multi-revisit (N_{multi}) and cross-axis (N_{cross}) comparisons.

	Tracking support (%)			Environment pose (%)		Q_{30} paired cases	
Model	H.ID	H.App.	Object	D	R_{30}	N_{multi}	N_{cross}
HY-WorldPlay	80.8	58.6	65.2	99.2	99.4	68	67
Matrix-Game 3	56.7	45.1	31.7	99.4	91.2	69	25
Matrix-Game 3.5	3.5	6.9	61.2	100.0	91.0	62	47
SANA-WM	26.6	34.7	43.9	99.6	80.8	54	25
YUME	30.0	19.5	22.0	92.9	94.5	52	41
LingBot World	34.7	36.3	45.0	98.6	80.0	41	1
LingBot World v2	67.5	71.4	62.1	100.0	57.6	11	20
Evoke	84.1	77.6	78.3	99.6	95.3	65	52
DreamX World	27.6	19.5	15.6	76.9	58.4	9	3

Table 12: Paired revisit diagnostics. N counts retained model–case units (a cross-axis unit contains two videos); M counts contributing models. All cases uses the same M models as Qualified, without case filtering. Positive reports the number of models with a positive conditional effect. Tracking requires support for every target at every compared revisit; Pose uses the 30° criterion. Values are descriptive means, not significance tests or conditional model rankings.

Metric	N	M	All cases	Qualified	Positive	Criterion
(a) Multi-revisit decrease: $R_1 - R_3$						
Human Identity	125	8	0.237	0.110	8/8	Tracking
Human Appearance	93	7	0.162	0.058	5/7	Tracking
Object Appearance	155	9	0.229	0.158	9/9	Tracking
Object Geometry	155	9	0.205	0.101	9/9	Tracking
Environment Visible	431	9	0.161	0.152	9/9	Pose 30°
Environment Depth	431	9	0.138	0.135	8/9	Pose 30°
Environment Reconstruction	431	9	0.196	0.139	9/9	Pose 30°
(b) Cross-axis excess degradation: Δ_{axis}						
Human Identity	135	8	0.023	0.034	6/8	Tracking
Human Appearance	109	8	0.056	0.050	6/8	Tracking
Object Appearance	141	8	0.117	0.023	4/8	Tracking
Object Geometry	141	8	0.115	0.051	6/8	Tracking
Environment Visible	281	9	0.152	0.128	8/9	Pose 30°
Environment Depth	281	9	0.107	0.076	8/9	Pose 30°
Environment Reconstruction	281	9	0.068	0.056	5/9	Pose 30°

unfiltered values of 0.161, 0.138, and 0.196, but remain positive. The tracking-supported subsets likewise show mean decreases in Human Identity, Human Appearance, Object Appearance, and Object Geometry (0.110, 0.058, 0.158, and 0.101). Thus, content recovery deteriorates across revisits even among videos that satisfy the departure and return criteria.

The average cross-axis penalty persists, with greater variation across models. On 281 qualifying case pairs across all nine models, the mean additional degradation under LRUD is 0.128 for Environment Visible, 0.076 for Depth, and 0.056 for Reconstruction. The effect is positive for eight of nine models on Visible and Depth, but five of nine on Reconstruction. All four entity metrics also retain a positive mean cross-axis penalty; Object Appearance, however, has four positive and four negative model-level effects. Switching axes therefore remains harder on average in these subsets, although the effect is less consistent across models than multi-revisit degradation.

As stricter checks of the 30° analysis, we repeat the comparisons with return tolerances of 15° and 5°. Both mean effects remain positive for all three Environment metrics (Tab. 13). Tighter tolerances reduce coverage: the 5° cross-axis comparison contains only 30 case pairs from four models,

Table 13: Environment return-tolerance sensitivity, always requiring departure $\geq 30^\circ$ and fixed cases across compared revisits. N/M denotes retained model–case units / contributing models. Vis., Dep., and Rec. denote Visible, Depth, and Reconstruction. The retained case and model sets change with the tolerance.

	Multi-revisit decrease				Cross-axis excess degradation			
Tolerance	N/M	Vis.	Dep.	Rec.	N/M	Vis.	Dep.	Rec.
5°	195/8	0.092	0.057	0.116	30/4	0.094	0.158	0.050
15°	345/8	0.116	0.085	0.141	134/7	0.090	0.010	0.029
30°	431/9	0.152	0.135	0.139	281/9	0.128	0.076	0.056

so effect sizes across tolerances also reflect changes in the evaluated samples. Together, these results show that the degradation is not confined to videos with poorly executed returns. Residual differences in camera position, occlusion, and estimation accuracy can still affect the scores; the analysis supports the observed trends rather than attributing the remaining decrease exclusively to forgetting.

H Analysis of Context-Access Policies

In this section, we examine the capability differences induced by temporal KV-access policies in Section 5.5, detailing the controlled setup and the accompanying capability analysis.

Controlled Experimental Setup. We compare FullKV, Sliding Windows, and FoV on LingBot World v2 without retraining. We fix the checkpoint, conditioning first frame, text prompt, camera trajectory, random seed, number of sampling steps, generation length, and evaluators. The temporal KV-access policy during autoregressive generation is the only variable. FullKV allows the current chunk to access the entire historical KV context. Sliding Windows uses a fixed budget of 20 latents: 4 sink, 12 recent, and 4 current latents. FoV uses the same 20-latent budget, comprising 4 sink, 4 recent, 8 retrieved, and 4 current latents. The retrieved latents come from two historical chunks selected by current-to-history field-of-view overlap. The fixed-budget comparison therefore tests how allocating context to view-relevant history affects memory beyond access to recent observations.

Evaluation coverage and timing. All three policies use identical evaluation sets: 180 Human, 210 Object, 210 Environment, and 60 Dynamic Memory videos, totaling 660 matched videos per policy. Latency is the mean end-to-end time to generate a 493-frame video over this matched set. Speedup is the ratio of FullKV's mean latency to the corresponding policy's mean latency.

View-relevant context for static retention. Across 490 Environment revisit slots, FoV increases the proportion of returns that reach the target viewpoint from 57.6% to 65.3% and reduces the mean return rotation error from 31.0° to 26.7° relative to FullKV. These gains accompany improvements in Human Identity, Human Appearance, Environment Depth, and Environment Reconstruction, suggesting that view-relevant historical context supports both scene recovery and content retention. The deadlift example in Fig. 10 illustrates the difference in viewpoint recovery: FoV returns to a more complete view of the lifter and barbell, whereas FullKV retains a close-up dominated by the weight plates.

Context allocation for dynamic progression. Maintaining dynamic memory requires the model to preserve the latest event state while allowing it to evolve. Sliding Windows matches FullKV on Return Progress (0.667), improves Final State from 0.592 to 0.600, but scores lower on Persistence/Reset (0.717 vs. 0.733) and Causal/Physics (0.775 vs. 0.792). Prioritizing recent observations alone therefore does not consistently improve dynamic memory. FoV combines recent context with view-relevant historical evidence and achieves further gains over both baselines across all four submetrics. The erasing example in Fig. 10 also reveals differences in the consistency of task-relevant objects: Sliding Windows produces a visibly enlarged erasing object, while FoV maintains a compact eraser in contact with the paper during the returned sequence. Together, the quantitative comparison and these examples favor combining recent states with view-relevant historical context over a recent-only window.

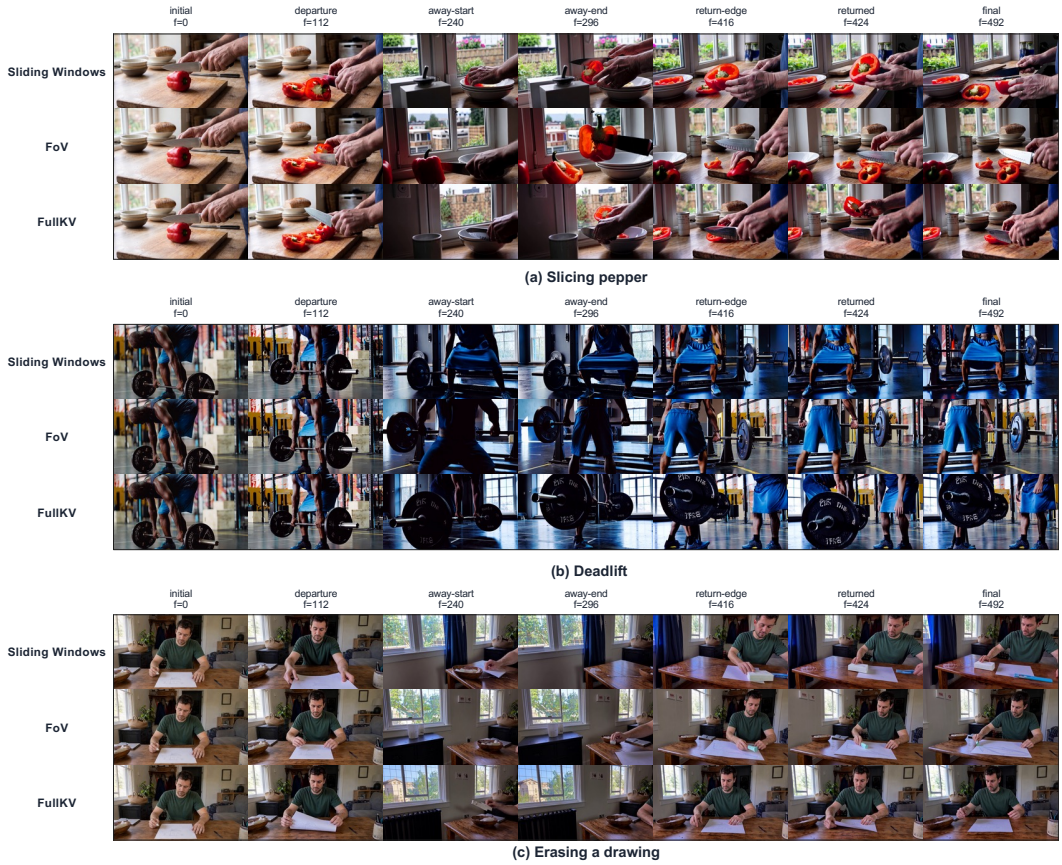

Figure 10: **World memory under different context-access policies in LingBot World v2.** The examples depict (a) slicing a pepper, (b) performing a deadlift, and (c) erasing a drawing. Within each example, rows show Sliding Windows, FoV, and FullKV from top to bottom. Columns present selected frames along the prescribed LR-Hold trajectory, spanning the initial observation, departure, and revisit. Camera phases and frame indices are indicated above.

I Limitation and Future Work

The current benchmark focuses on controlled translation and rotation trajectories over rollouts of several hundred frames. Longer horizons, free-form or compound traversal patterns, and embodied settings in which camera and object actions are jointly selected remain outside its present scope. Dynamic Memory additionally relies on frozen case-specific checklists evaluated by a VLM. Although the resulting scores show strong overall agreement with human judgments, the Causal/Physics dimension remains the most challenging sub-dimension. Extending the benchmark to richer event types, multi-agent interactions, and longer unobserved phases, together with stronger and more diverse evaluators, would further test the limits of persistent world generation. We hope WorldMem-Bench serves as a practical testbed for developing video world models that remember, evolve, and remain consistent over extended interaction.