MEMWORLD: EXPLORE, EDIT, AND REVISIT WITH VALIDITY-AWARE MULTIMODAL MEMORY

Anonymous authors
Paper under double-blind review

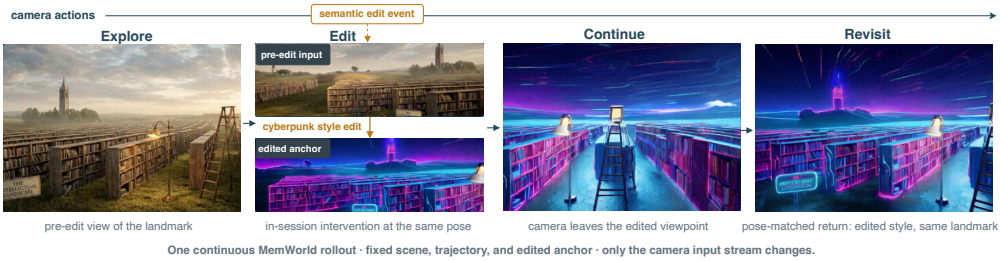

Figure 1: **Explore, edit, continue, and revisit within one evolving world.** All panels are real frames from a MemWorld rollout with a fixed scene, trajectory, and edited anchor. The user first explores the landmark, issues one global Cyberpunk-style intervention in session (the Edit panel shows the pre-edit input above and the edited anchor below), continues exploring away from the edited viewpoint, and finally returns to the pose-matched starting view: the world keeps its original structure while carrying the edited appearance.

ABSTRACT

Interactive video world models support two complementary interaction modes: users can *wander* through generated environments using camera actions or *direct* their evolution through semantic interventions. Existing models typically handle these modes separately. We unify them within one evolving session, enabling users to explore, edit, continue, and revisit. This setting exposes a memory challenge: limited context causes forgetting during long-horizon revisits, while an edit leaves pre-edit observations geometrically useful but visually obsolete. Reusing them as RGB can revert the edit, whereas discarding them loses scene structure. We introduce *MemWorld*, a memory-augmented video world model with validity-aware multimodal memory. It retrieves nonlocal observations through camera-frustum overlap and uses a version-conditioned RGB–depth policy to preserve current appearance in RGB while retaining obsolete observations as structural depth. Retrieved memory, recent history, and current latents are jointly processed by a camera-conditioned video DiT. We evaluate MemWorld on our proposed editing benchmark of 500 cases covering global appearance editing, continued navigation, and pose-matched revisitation, together with R2M-Bench and MBench. Among the evaluated learned systems, MemWorld achieves the highest R2M-Bench point estimate and the highest Composite point estimate for persistent editing and geometry, while remaining competitive on MBench. These results show that multimodal memory can unify wandering and directing while preserving long-term scene consistency and persistent world updates.

1 INTRODUCTION

Interactive video worlds expose two complementary forms of control. In *exploration*, a user moves through a generated environment and expects previously visited places to remain recoverable. In *direction*, a user changes the world through a semantic intervention, such as altering its visual style or environment. A practical world model should not force users to choose between these modes: it should support interleaved exploration and intervention within one evolving session. The desired

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.

interaction is simple to state— *explore, edit, continue, and revisit*—but difficult to sustain over a long autoregressive rollout.

Recent systems increasingly combine navigation with promptable events, timestamped text actions, and object-level controls (Parker-Holder et al., 2024; Song et al., 2026; Gu et al., 2026a; Wang et al., 2026a). However, accepting both navigation and semantic control does not ensure that an intervention becomes a persistent update to a revisitable world. Long-rollout methods improve local continuity but may forget observations outside the working window (Huang et al., 2026; Yang et al., 2025; Liu et al., 2026), while memory-augmented models retrieve nonlocal observations but usually assume that the retrieved content remains fully valid (Hong et al., 2025; Li et al., 2025; Sun et al., 2025; Wang et al., 2026c; Hu et al., 2026). This assumption fails after an edit because an old RGB observation may preserve useful identity and geometry while carrying obsolete appearance. Retaining it can revert the edit, whereas discarding it removes valuable scene evidence. PermaVid addresses this conflict through edit-aware RGB–depth filtering, but applies validity only to retrieved memory and supplies selected observations through the VACE reference-image interface, with only the latest frame as recent video context (Yang et al., 2026). It therefore lacks a unified, trained pathway for reconciling nonlocal memory, multi-frame recent history, and the evolving current state. Because fixed-clip editing cannot address future unobserved revisits (Molad et al., 2023; Qi et al., 2023; Geyer et al., 2024), a persistent editable world must retrieve relevant nonlocal evidence, preserve structure without reintroducing obsolete appearance, and propagate updates consistently across memory and recent context.

We address these requirements with *MemWorld*, a memory-augmented camera-conditioned video DiT that turns appearance interventions into persistent world-state updates. Its retrieve–validate–integrate design combines geometry-grounded retrieval of observations covering the current view with version-aware multimodal memory, which preserves authoritative appearance in RGB and reuses appearance-obsolete observations as structural depth. A shared self-attention pathway applies the same validity policy to nonlocal memory and recent history. Together, these components address retrieval, stale appearance, and fragmented context.

Concretely, each generated observation is stored with aligned RGB and depth latents, camera parameters, absolute time, and its world version. Camera-frustum overlap selects a fixed number of geometrically relevant nonlocal observations, while a FIFO buffer supplies recent history. After a global appearance edit, a version-conditioned rule represents current-version observations as RGB and earlier-version observations as aligned depth in both context sources. The resulting nonlocal memory, recent history, and current noisy latents are concatenated and processed through the native self-attention of the pretrained camera-conditioned DiT backbone (Team et al., 2026). Lightweight LoRA adapters are trained in two stages, first for robust RGB recall and then for mixed RGB–depth conditioning under imperfect self-generated context. At inference time, an external editor creates the edited anchor, and MemWorld then turns that anchor into the persistent world state rather than performing end-to-end text editing itself.

Figure 1 illustrates the interaction: the user explores a scene, observes a landmark, applies a global appearance edit, navigates away, and later revisits the location. The returned view retains the landmark structure and edited appearance, showing why obsolete pre-edit RGB memory must not revert the world state. Our appearance-intervention benchmark captures this setting through global style editing, continued navigation, and pose-matched revisitation. MemWorld preserves edited appearance, recognizable landmarks, and scene layout across revisits. A matched comparison with PermaVid shows that this benefit is not due to a stronger one-shot edit. Instead, it reflects better integration of the edited state into subsequent generation, reducing appearance reversion and structural drift. Strong R2M-Bench performance (Gu et al., 2026b) and competitive MBench results (Zhang et al., 2026) further show that edit-aware memory retains general revisit capability. Together, these results support editing as a persistent state transition rather than an isolated transformation.

Our contributions are summarized as follows:

- We introduce *MemWorld*, a memory-augmented video world model that unifies wandering and directing in a persistent explore–edit–continue–revisit process while preserving revisitable structure.
- We develop a validity-aware multimodal memory mechanism. Geometry-grounded retrieval selects relevant observations. Version-conditioned RGB–depth routing removes ob-

solete appearance while retaining structure across long-term memory and recent history. All contexts are integrated with the current state through the native self-attention pathway.

- We introduce a 500-case editing benchmark covering global appearance editing, continued navigation, and pose-matched revisitation. It evaluates edit adoption, long-term persistence, and structural consistency.

2 Related Work

2.1 Exploration-Centric Interactive Video World Models

Interactive video world models add action or camera control to modern video generators, enabling navigation in learned games and open-domain environments (Bruce et al., 2024; Alonso et al., 2024; Sun et al., 2025; Wang et al., 2026c; Team et al., 2026). Long-horizon methods reduce autoregressive degradation through generated-context training, efficient attention, distillation, compressed caches, or state-space models (Huang et al., 2026; Liu et al., 2026; Yang et al., 2025; Chen et al., 2026; Yu et al., 2025; Li et al., 2026; Zhu et al., 2026), but continuity does not guarantee recall once a location leaves the active context. Persistent-memory systems address this limitation by compressing history into recurrent states, tokens, object slots, or positional representations (Chen et al., 2026; Yu et al., 2025; Li et al., 2026; Kim et al., 2026; Dou et al., 2026; Wu et al., 2026), or by retrieving directly addressable past observations (Hu et al., 2026; Hong et al., 2025; Peng et al., 2026). Geometry-aware retrieval further uses pose, field-of-view overlap, landmarks, or spatial representations to recover relevant history (Sun et al., 2025; Wang et al., 2026c; Li et al., 2025; Wang et al., 2026b). These mechanisms extend temporal reach, but largely assume that recalled observations remain valid in a stationary world. Once the user redirects scene appearance, retrieval must determine not only *where* to recall but *whether* the recalled content remains authoritative. MemWorld makes this validity explicit, preserving current appearance while reusing obsolete observations as structural evidence.

2.2 Semantic Intervention and Mixed-Control Worlds

A parallel line broadens control beyond camera motion. Text-guided video editing modifies an observed clip through source adaptation, spatiotemporal conditioning, inversion, attention, or cross-frame features (Molad et al., 2023; Qi et al., 2023; Geyer et al., 2024). Recent world models bring such semantic control into interactive generation. The Genie series supports real-time interactive world generation (Parker-Holder et al., 2024), while World in World unifies exploration and video editing in a frozen model (Song et al., 2026). WorldCraft adds object-trajectory control (Gu et al., 2026a), whereas WorldDirector combines semantic motion orchestration with persistent dynamic-object memory (Wang et al., 2026a). These systems show that navigation and semantic control can coexist, but do not establish whether an appearance intervention remains authoritative after the camera leaves and old memory is retrieved. This motivates MemWorld to treat semantic intervention as a persistent world-state update rather than a current-view transformation.

2.3 Persistent State Under Semantic Direction

Bringing these lines together, semantic direction updates an interactive world whose unseen future views may reappear much later, unlike fixed-clip manipulation of observed frames (Molad et al., 2023; Qi et al., 2023; Geyer et al., 2024). Existing memory methods access earlier observations through recurrent states, slots, retrieval, or geometry-aware matching (Kim et al., 2026; Dou et al., 2026; Wu et al., 2026; Hu et al., 2026; Hong et al., 2025; Li et al., 2025; Wang et al., 2026b), but generally treat stored content as immutable. Semantic direction therefore creates a conflict. RGB recall preserves identity and detail but may restore pre-intervention appearance, while discarding it loses geometry. PermaVid uses field-of-view retrieval and version-aware RGB–depth filtering, supplying selected images as VACE references while retaining only the latest frame as recent context (Yang et al., 2026). MemWorld instead applies one validity rule to nonlocal memory and multi-latent recent history, then jointly processes them with current noisy latents through native self-attention. Its shared RGB, depth, and mixed-context pathway is explicitly trained to make semantic direction persistent across navigation and revisitation.

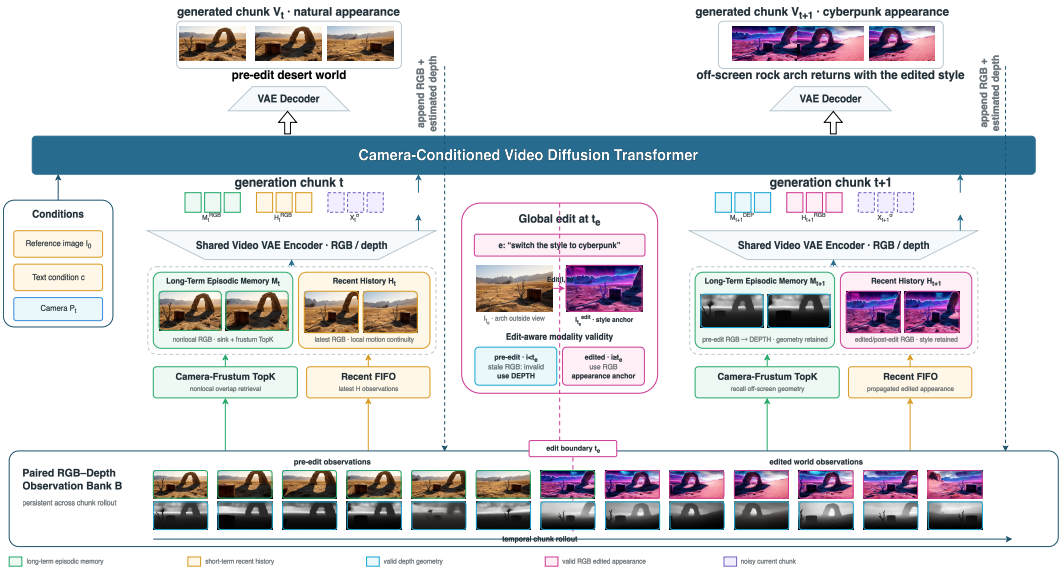

Figure 2: **MemWorld architecture and edit-aware rollout.** Camera-frustum TopK retrieves nonlocal memory, a FIFO selector supplies recent history, and the video DiT jointly processes both with the current chunk. After a global appearance edit, pre-edit RGB is routed as aligned depth while edited and post-edit observations remain RGB. Generated RGB and depth are appended to the observation bank.

3 METHOD

3.1 FRAMEWORK OVERVIEW: RETRIEVE, VALIDATE, AND INTEGRATE

We consider an interactive video world model initialized from a reference image I_0, a text condition c, and a camera trajectory $\mathcal{P}_{1:T} = \{(\mathbf{P}_i, \mathbf{K}_i)\}_{i=1}^{T}$, where $\mathbf{P}_i$ and $\mathbf{K}_i$ are the camera extrinsics and intrinsics at latent time i. The model generates $\hat{\mathbf{V}} = \{\hat{I}_i\}_{i=0}^{T}$ in temporal chunks. Most steps contain only camera actions; at time t_e, a global appearance intervention inserts an anchor created by an external editor and advances the world version. MemWorld models the persistence of this state update: subsequent generation must recover nonlocal scene evidence without allowing pre-edit appearance to overwrite it.

At generation step t, we maintain an observation bank

$$\mathcal{B}_{t-1} = \left\{ (\mathbf{z}_i^{\mathrm{rgb}}, \mathbf{z}_i^{\mathrm{dep}}, \mathbf{P}_i, \mathbf{K}_i, i, q_i) \right\}_{i<t}, \tag{1}$$

where $\mathbf{z}_i^{\mathrm{rgb}}$ and $\mathbf{z}_i^{\mathrm{dep}}$ are VAE latents of the generated RGB observation and its aligned depth visualization, and q_i is its world version. RGB preserves appearance and identity, whereas depth provides structural evidence less tied to appearance; both use the same video VAE and conditioning interface.

The context contains geometrically retrieved long-term memory $\mathbf{M}_t$, recent history $\mathbf{H}_t$, and the noisy current latent $\mathbf{X}_t^{\sigma}$. MemWorld makes three corresponding decisions: camera-frustum retrieval determines *where* to recall from; a modality-validity rule determines *what* remains authoritative after an edit; and joint self-attention determines *how* memory, history, and the current window interact. This validity decision is rule-based rather than learned: it uses the edit boundary and stored world-version label. Figure 2 summarizes this retrieve–validate–integrate pipeline.

3.2 RETRIEVING WHERE THE USER RETURNS

Because appearance similarity is unreliable across viewpoints, we retrieve historical observations by camera-frustum overlap. Let $\mathcal{F}_i$ be the voxelized view frustum of historical latent i, and let $\mathcal{F}_t^q$

be the query frustum of the current generation window. We use the query-normalized intersection

$$s(i,t) = \frac{\mathrm{Vol}(\mathcal{F}_i \cap \mathcal{F}_t^q)}{\mathrm{Vol}(\mathcal{F}_t^q)}. \tag{2}$$

Unlike symmetric IoU, this score measures how much of the requested region a historical observation can explain, even when historical and query frusta have different volumes.

To complement rather than duplicate recent history, retrieval excludes the L_{recent} immediately preceding latent frames. The first observation is retained as a persistent identity anchor, and the remaining slots contain the highest-scoring nonlocal observations above an overlap threshold:

$$\mathcal{I}_t = \{0\} \cup \mathrm{TopK}_{\substack{i<t-L_{\mathrm{recent}},\ i\neq 0 \\ s(i,t)\geq\tau_s}} s(i,t). \tag{3}$$

This keeps the active context bounded while allowing any earlier observation to re-enter generation when it becomes geometrically relevant. Fixed context sizes, the retrieval diagnostic, and the leave-and-return trajectory construction used to train nonlocal recall are provided in Appendix B.

3.3 Integrating Memory, History, and the Present

Given retrieved indices $\mathcal{I}_t$, we concatenate episodic memory, recent history, and the noisy current window along time:

$$\mathbf{Z}_t = [\mathbf{M}_t \mid \mathbf{H}_t \mid \mathbf{X}_t^\sigma]. \tag{4}$$

For a clean current latent $\mathbf{X}_t^0$, we sample $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a scheduler timestep with noise level σ_t, and construct the flow-matching path and its constant-velocity target as

$$\mathbf{X}_t^\sigma = (1-\sigma_t)\mathbf{X}_t^0 + \sigma_t \boldsymbol{\epsilon}_t, \qquad \mathbf{v}_t^\star = \boldsymbol{\epsilon}_t - \mathbf{X}_t^0. \tag{5}$$

Memory and history remain clean conditions. Accordingly, each joint token j receives a token-wise diffusion timestep

$$\tau_j = \begin{cases} 0, & j \in \Omega_{\mathrm{M}} \cup \Omega_{\mathrm{H}}, \\ t, & j \in \Omega_{\mathrm{C}}, \end{cases} \tag{6}$$

where Ω_{M}, Ω_{H}, and Ω_{C} denote memory, history, and current-token regions.

Rather than introducing an auxiliary memory cross-attention module, the entire sequence passes through the native DiT blocks. Let $\mathbf{U}^{(\ell)}$ be the input to block ℓ. Its visual self-attention response is the sum of the pretrained RoPE branch and the parallel projective-camera branch of the backbone (Team et al., 2026):

$$\mathbf{A}^{(\ell)} = \mathrm{SA}^{(\ell)}_{\mathrm{RoPE}}\Big(\mathbf{U}^{(\ell)}; \mathbf{p}\Big) + \mathrm{SA}^{(\ell)}_{\mathrm{PRoPE}}\Big(\mathbf{U}^{(\ell)}; \widetilde{\mathcal{P}}, \mathcal{K}\Big). \tag{7}$$

Here $\mathbf{p}$ contains the original absolute latent indices, while $\widetilde{\mathcal{P}}$ and $\mathcal{K}$ collect the re-anchored extrinsics and intrinsics. The two responses are added before the shared time-modulated residual update; the block then applies its original text cross-attention and feed-forward layers. Consequently, memory, history, and current tokens interact through the same self-attention pathway, while PRoPE supplies camera geometry rather than a separate memory pathway.

Only the current region is supervised. The complete masked objective is

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{(\mathbf{X}^0, c, \mathcal{P})\sim\mathcal{D},\, t,\, \boldsymbol{\epsilon}} \left[\frac{1}{|\Omega_{\mathrm{C}}|} \sum_{j\in\Omega_{\mathrm{C}}} \left\| \mathbf{v}_\theta(\mathbf{Z}_t, \boldsymbol{\tau}, c, \widetilde{\mathcal{P}})_j - \mathbf{v}^\star_{t,j} \right\|_2^2 \right]. \tag{8}$$

The clean memory and history tokens therefore shape current denoising without contributing reconstruction targets of their own.

Every token receives projective camera encoding and retains its original absolute video index. Camera extrinsics are expressed relative to the first joint token, preserving relative geometry, while absolute coordinates prevent distant memories from being interpreted as temporally adjacent to the current window. The video DiT, text encoder, VAE, and PRoPE branch remain frozen; only rank-64 LoRA adapters on the native self-attention q, k, v, and output projections are optimized. Numerical camera processing, positional regularization, and adaptation details are provided in Appendix B.

3.4 Updating What Remains Valid

RGB is the most informative memory before an edit, but afterward it can restore obsolete appearance even when its geometry remains useful. Depth retains this reusable structure while weakening pressure toward the previous visual state. We use two training stages to first establish robust long-term recall and then teach the model to condition on whichever representation remains valid.

Stage I: robust RGB recall. The first stage uses RGB for both episodic memory and recent history, alternating among full-memory, history-only, and initial-image conditioning. This curriculum prevents the model from assuming that useful long-term evidence is always available. We also reinject residuals from partially denoised predictions into later memory and history, exposing the model to the imperfect self-generated context encountered during rollout. Sampling probabilities and the residual construction are specified in Appendix B.

Stage II: modality-flexible validity. Starting from the RGB-memory model, the second stage randomly replaces RGB slots in both memory and recent history with their aligned depth latents. Training on RGB, depth, and mixed contexts teaches the shared attention pathway to use appearance when it is valid and structural evidence when appearance is unavailable or stale, without modality-specific encoders or tokens. Mixing recent history is important because frames immediately preceding an edit can become invalid before they leave the short-term window.

Validity-aware rollout. Each generated RGB chunk and its estimated aligned depth are appended to the observation bank. At the edit boundary t_e, the edited RGB–depth pair replaces the latest bank entry and the world-version index advances. For the global appearance interventions considered here, the edited RGB anchor is paired with the pre-edit depth at the same pose; alternative anchor-depth policies are evaluated in Appendix B.2. Under a global appearance edit, observations from earlier versions retain useful geometry but no longer provide authoritative appearance:

$$\bar{\mathbf{z}}_i = \begin{cases} \mathbf{z}_i^{\mathrm{dep}}, & q_i < q_t, \\ \mathbf{z}_i^{\mathrm{rgb}}, & q_i = q_t, \end{cases} \tag{9}$$

where q_t is the current world version. The rule is applied to both retrieved memory and recent history before constructing Eq. 4. Consequently, obsolete RGB cannot override the edit, while its aligned depth still supports reconstruction when previously observed geometry re-enters the view.

4 Experiments

We evaluate MemWorld in four stages. Section 4.1 specifies the benchmarks, protocols, baselines, metrics, and implementation; Sec. 4.2 tests persistent appearance intervention, including a matched PermaVid comparison; Sec. 4.3 examines validity-aware context routing; and Sec. 4.4 measures edit-free revisit memory.

4.1 Experimental Setup

Benchmarks and protocols. Our proposed appearance-intervention benchmark pairs 100 input scenes with five target styles, yielding 500 cases. Each system first explores a scene, receives its own externally edited anchor, continues along a shared camera trajectory, and revisits pre-intervention viewpoints. All compared systems are evaluated on the full set. PermaVid is additionally evaluated under a matched protocol using the same scenes, styles, editor, actions, metric implementation, and common resized evaluation resolution. We also evaluate edit-free memory on R2M-Bench (Gu et al., 2026b) and MBench (Zhang et al., 2026). R2M-Bench uses approximately 481-frame rollouts on three return trajectories. MBench uses the official 547-item manifest from its causal, environment, human, and object subsets, with ten-chunk free rollouts of approximately 417 frames.

Baselines. On the appearance-intervention benchmark, we compare with LingBot-World-2.0, HY-WorldPlay, SANA-WM, Matrix-Game 3.0, and Lyra 2.0. We separately report a complete matched comparison with PermaVid, the closest mechanism-level prior. For R2M-Bench and MBench, we follow their public comparison sets and distinguish learned video world models from WonderWorld, an explicit-3D reference for revisitation.

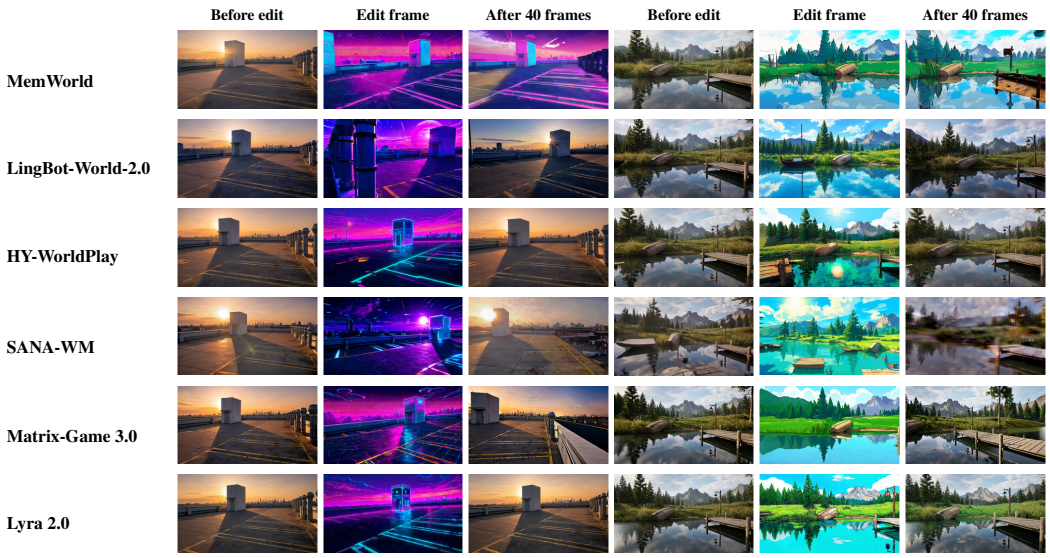

Figure 3: **Persistent appearance intervention.** Each row shows one method on the same two scene–style cases. Within each case, *Explore* is a pre-intervention observation, *Intervene* is the method-specific edited anchor, and *Revisit* is the later pose-matched return view.

Evaluation metrics. Style-fidelity and intervention-response metrics measure whether the anchor appearance is adopted, while retention measures whether it persists through the rollout. Pose-matched depth metrics evaluate structural consistency. XEdit compares a pre-intervention observation with its post-intervention revisit, whereas Post compares views within the updated world. We summarize the trade-off using the harmonic mean of a min–max-normalized Edit Axis and Consistency Axis. Because this Composite depends on the compared method set, we interpret it together with the raw metrics. R2M-Bench reports MemoryGain in each metric's native units and Normalized Memory Ratio (NMR) relative to the model's short-range dynamic range. MBench uses its official mean score.

Implementation. All results use the same MemWorld checkpoint trained with the two-stage procedure in Sec. 3.4, without benchmark-specific finetuning. Full protocols, metric definitions, confidence intervals, and per-metric results are provided in the supplementary material.

4.2 Persistent Appearance Intervention

The central question is whether an intervention becomes the current world state. After changing a visible view, the user should be able to leave and later revisit it without recovering obsolete appearance or losing the scene layout.

4.2.1 Qualitative Comparison

Figure 3 shows that MemWorld carries the target appearance into later revisits while preserving recognizable landmarks and scene layout. Competing systems often revert toward the earlier appearance, lose the edited style, or change the spatial organization during continued generation. The examples therefore test persistence rather than only the quality of the initial intervention.

4.2.2 Quantitative Comparison

Table 1 confirms the same pattern. MemWorld leads 11 of 14 base metrics in the full evaluation. Its CSD intervention response is 0.2164, nearly twice Lyra 2.0's 0.1104, and its CSD retention remains positive at 0.0484 while Lyra 2.0 falls to -0.0425. Lyra 2.0 has the highest XEdit DepthCorr, but this accompanies a substantially smaller appearance change. Within the updated world, MemWorld

Table 1: Compact persistent-intervention results. CSD and MegaStyle Δ measure intervention response, CSD Ret. measures late-versus-early retention, XEdit compares a pre-intervention view with its later revisit, and Post compares views within the updated world. All methods are evaluated on all 500 cases. Full metrics are reported in Supplementary Table 6.

Method	CSD $\Delta\uparrow$	MegaStyle $\Delta\uparrow$	CSD Ret. $\uparrow$	XEdit DepthCorr $\uparrow$	Post DepthCorr $\uparrow$	Composite $\uparrow$
LingBot-World-2.0	0.0441	0.0526	0.0111	0.7098	0.8432	0.217
HY-WorldPlay	-0.0016	-0.0036	<u>0.0222</u>	<u>0.9478</u>	0.9297	0.157
SANA-WM	-0.0239	-0.0078	-0.0304	0.8175	0.7596	0.097
Matrix-Game 3.0	0.0080	0.0015	0.0202	0.8350	0.8502	0.202
Lyra 2.0	<u>0.1104</u>	<u>0.0990</u>	-0.0425	**0.9612**	<u>0.9777</u>	<u>0.694</u>
MemWorld	**0.2164**	**0.1983**	**0.0484**	0.9273	**0.9860**	**0.965**

obtains the highest Post DepthCorr at 0.9860. The resulting Composite is 0.965, compared with 0.694 for the strongest baseline. Full metrics are reported in Supplementary Sec. C.1.

4.2.3 Comparison with PermaVid

PermaVid retrieves RGB and depth memories with field-of-view and version rules, then supplies selected images through VACE reference conditioning while using only the latest frame as recent video context. MemWorld instead applies one validity rule to both retrieved memory and a multi-latent recent-history window, then processes both with the current noisy chunk through native self-attention. This tests whether validity covers all visual context, not only retrieved memory.

Under the matched 500-case protocol, the two systems have comparable initial intervention response and immediate style fidelity. Their CSD intervention responses are 0.274 for MemWorld and 0.260 for PermaVid. The gap appears after continued generation. MemWorld retains a positive CSD change of 0.045, whereas PermaVid falls to -0.115, and it achieves a Post DepthCorr of 0.989 compared with 0.854. The retention and consistency differences have the same direction across all five styles, and their paired confidence intervals exclude zero. Because the systems use different backbones, this is a system-level comparison consistent with the joint-context design rather than a causal isolation of VACE conditioning versus native self-attention. Complete metrics and confidence intervals are reported in Supplementary Sec. C.2.

4.3 Ablation of Validity-Aware Context Routing

We ask whether validity must cover both long-term memory and recent history. On 100 paired scene–style cases, all variants use the same model, edited anchor, trajectory, seed, and inference configuration. We compare retaining stale RGB, converting only episodic memory to depth, and converting both episodic memory and recent history. The pre-intervention rollouts are not pixel-identical across variants, so this study is a controlled diagnostic rather than a strict causal estimate.

Keeping stale RGB preserves geometry with a Consistency Axis of 1.0000, but it nearly prevents the intervention from propagating. Its Edit Axis is only 0.1089 and its Composite is 0.1964. Converting episodic memory alone raises the Edit Axis to 0.3569 and the Composite to 0.5240 while retaining a Consistency Axis of 0.9856. This indicates that stale RGB in recent history can still override the update. Converting both context sources reaches an Edit Axis of 1.0000 and a Composite of 0.9859 while retaining a Consistency Axis of 0.9722. The full result therefore trades a modest consistency change for a much larger gain in valid appearance propagation rather than indiscriminately suppressing memory. The full confidence intervals, base metrics, and per-style breakdown are reported in Supplementary Sec. B.1. The same subsection also compares reset-and-re-anchor alternatives, while edited-anchor depth policies are reported in Supplementary Sec. B.2.

4.4 General Memory Evaluation

Finally, we test whether updateable memory preserves general long-horizon recall when no appearance intervention is issued.

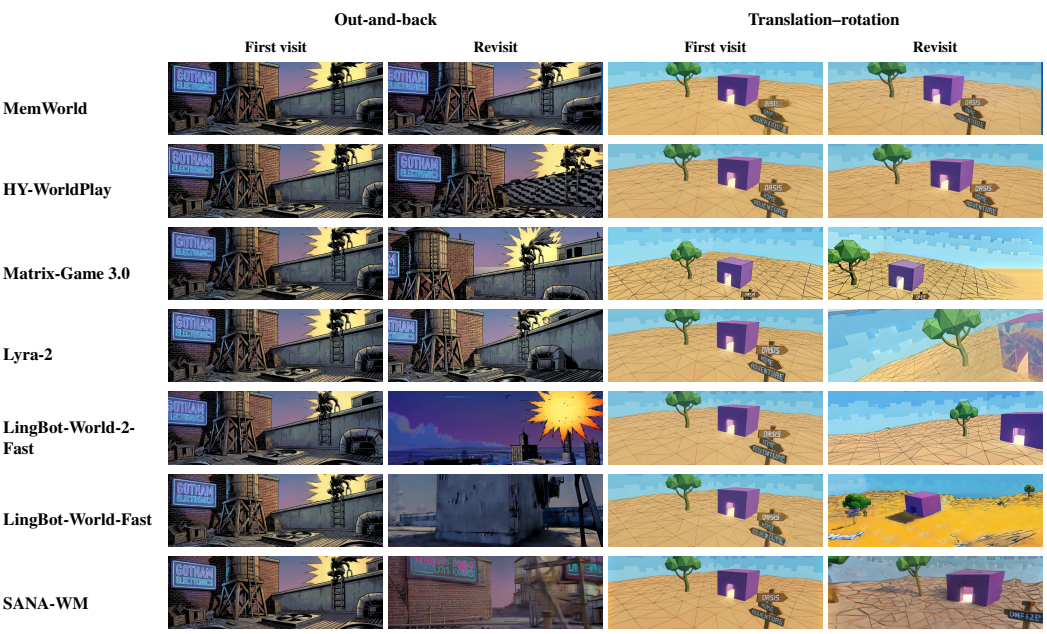

Figure 4: **Qualitative comparison of long-horizon revisit memory.** Each row compares one video world model. The four columns show the first visit and pose-matched revisit for two representative return trajectories. MemWorld preserves scene identity, landmarks, and spatial organization across both returns, whereas competing methods often replace, omit, or rearrange previously observed content.

Figure 4 shows the long-horizon memory advantage directly. On the out-and-back trajectory, MemWorld recovers the same facade, sign, water tower, and street layout. On the harder translation–rotation return, it preserves the distinctive tree, structure, and sign under the changed viewpoint. Competing models frequently produce a plausible but different place, lose landmarks, or alter their spatial relations. Camera-frustum retrieval brings geometrically relevant nonlocal observations back into context, while joint memory–history–current attention uses them to anchor the revisit.

Among the seven learned video world models, MemWorld ranks first on R2M-Bench with an Overall NMR of 0.706, compared with 0.485 for the strongest learned baseline, HY-WorldPlay. It also leads 9 of 11 per-metric NMRs. The explicit-3D reference WonderWorld reaches 1.510, showing that exact 3D reconstruction remains a strong upper reference for revisitation. Full results are reported in Supplementary Sec. C.4.

On MBench, MemWorld ranks second among the seven learned models with a mean of 68.48, compared with 69.57 for HY-WorldPlay. It is strongest in object texture and rendering light, while human identity and interaction state remain limitations. This result shows that the validity-aware memory design retains competitive general video-memory performance rather than specializing only to appearance intervention. Full results are reported in Supplementary Sec. C.6.

5 Conclusion

Interactive video world models must treat edits as persistent state updates across explore–edit–continue–revisit sessions. We introduce MemWorld, a memory-augmented camera-conditioned video DiT combining frustum retrieval, appearance validity, obsolete-RGB depth, and joint attention over memory, history, and generation. MemWorld preserves edits while maintaining structure. Matched PermaVid results and routing ablations support the context-wide validity design; leading R2M-Bench and competitive MBench results show that the model retains general long-horizon memory. Validity-aware multimodal memory therefore enables explorable, persistently directed worlds.

AI USE STATEMENT

We used generative AI tools only for tasks in the recommended-disclosure category: editing text to improve readability, drafting and revising prose, reorganizing the narrative and suggesting document structure, and creating or modifying draft figures and illustrative placeholder images during writing. We did not use generative AI tools for any task in the required-disclosure category: we did not use them to generate datasets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design methodology or experiments, implement methods, clean or reformat data, perform qualitative analysis, or interpret results. All datasets, the model and its implementation, experimental protocols, and reported results were produced and verified by the authors. We have reviewed all AI-assisted text and figures for accuracy and originality. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a video world model and does not involve experiments on human subjects beyond a voluntary human evaluation in which annotators rated the visual quality and consistency of generated videos; no personally identifiable or sensitive information was collected from annotators. Model training and evaluation use publicly available research datasets (e.g., SpatialVID, RealEstate10K, Sekai, and DL3DV) under their respective terms, together with game recordings and internally rendered synthetic engine data; real-world footage is used solely for non-commercial research. We acknowledge that controllable generative world models could be misused to produce misleading or deceptive visual content, and we share our analysis to support research on reliable and persistent video generation rather than such misuse. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results in the main paper and appendix. The two-stage training configuration and data composition are described in Appendix A; additional method details, including the memory-validity policy, edited-anchor depth handling, context and retrieval design, and modality mixing, are given in Appendix B. Full implementation and evaluation settings for persistent editing, R2M-Bench, and MBench are reported in Appendices C.1, C.4, and C.6, with metric definitions in Appendix C.7 and the exact evaluation prompt in Appendix C.8.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Taiye Chen, Zihan Ding, Anjian Li, Christina Zhang, Zeqi Xiao, Yisen Wang, and Chi Jin. Recurrent autoregressive diffusion: Global memory meets local attention. In *European Conference on Computer Vision*, pp. 492–511. Springer, 2026.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. Slotmemory: Object-centric kv memory for streaming long-video generation. *arXiv preprint arXiv:2605.31033*, 2026.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Bohai Gu, Taiyi Wu, Yueyang Yuan, Jian Liu, Xiaocheng Lu, Dazhao Du, Jie Zhang, Jinxiang Lai, Shuai Yang, Xiaotong Zhao, et al. Worldcraft: From camera navigation to object manipulation in interactive video world models. *arXiv preprint arXiv:2605.25077*, 2026a.

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*, 2026b.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

Kunyang Li, Mubarak Shah, and Yuzhang Shang. Attend locally, remember linearly: Linear attention as cross-frame memory for autoregressive video diffusion. *arXiv preprint arXiv:2605.16579*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

J Parker-Holder, P Ball, J Bruce, V Dasagi, K Holsheimer, C Kaplanis, A Moufarek, G Scully, J Shar, J Shi, et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model*, 2024.

Zhan Peng, Jie Ma, Huiqiang Sun, Chong Gao, Zhijie Xue, Zhiyu Pan, Zhiguo Cao, Jun Liang, and Jing Li. Compression and retrieval: Implicit memory retrieval for video world models. *arXiv preprint arXiv:2606.23105*, 2026.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15886–15896. IEEE, 2023.

Chenxi Song, Yanming Yang, and Chi Zhang. World in world: Explore the world with world models. *arXiv preprint arXiv:2609.11548*, 2026.

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*, 2026.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, et al. Worlddirector: Building controllable world simulators with persistent dynamic memory. *arXiv preprint arXiv:2607.02517*, 2026a.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026b.

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*, 2026c.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Shuai Yang, Bingjie Gao, Ziwei Liu, Jiaqi Wang, Dahua Lin, and Tong Wu. Permavid: Consistent video generation across edits via disentangled context memory. *arXiv preprint arXiv:2606.16449*, 2026.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, et al. Mbench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026.

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*, 2026.

Appendix overview. Section A summarizes the two-stage training configuration. Section B provides additional method details, including memory-validity and reset-baseline ablations, edited-anchor depth policies, context and retrieval design, training curriculum, and modality mixing. Section C reports full implementation and evaluation results for persistent editing, R2M-Bench, and MBench, together with additional qualitative comparisons, a limitation diagnostic, metric definitions, and the evaluation prompt. Section D defines the R2M-Bench MemoryGain and normalized memory ratio scores.

A Training Configuration

Training data. Our training pool combines internally constructed UE videos, real-world videos, and game recordings. The real-world portion uses SpatialVID, RealEstate10K, Sekai, and DL3DV, with recovered camera trajectories normalized to a common coordinate convention. The game portion uses Sekai-Game and OmniWorld-Game, whose engine-exported poses are converted to the same convention. The UE subset was constructed internally at ByteDance by collecting controllable free-camera and character-driven trajectories, validating their motion and camera paths, and rendering accepted trajectories offline with synchronized frames, camera poses, actions, and metadata. Its manifest contains 102,661 samples; static filtering with $t_{\mathrm{lat,min}} = 21$ retains 34,470 15-second samples for training. Appendix B describes the additional leave-and-return construction used to train nonlocal recall.

We train MemWorld in two consecutive stages. Table 2 reports the complete configuration of the first stage, *Memory Depth+RGB*; the second stage is summarized below only through its differences to avoid duplicating unchanged settings.

Table 2: Training configuration for Stage I (*Memory Depth+RGB*).

Item	Stage I configuration
Base model	Wan2.2-TI2V-5B
UE subset	102,661 manifest samples; 34,470 15-second samples after static filtering with $t_{\mathrm{lat,min}} = 21$
Resolution	1280×720 source videos; 1280×704 during training
Frame rate	16 FPS
Chunk layout	15-latent active window (57 raw frames), comprising 5 history latents and 10 current latents (37 raw frames), with 5 additional memory latents
Modality mixing	Episodic memory uses depth with probability 0.5 and RGB otherwise; recent history remains RGB
Optimizer	Fused AdamW, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$, weight decay 10^{-3}, and gradient clipping at 0.05
Learning rate	10^{-4} with 200 optimizer updates of linear warmup, followed by a constant schedule
Batch size	Per-process batch size 1, global microbatch 16, gradient accumulation 2, and effective batch size 32
Training length	20,000 training steps
LoRA	Rank 64, alpha 64, scale 1; self-attention $q/k/v/o$ projections in all 30 DiT blocks (120 adapters), without dropout
Precision and parallelism	BF16, FSDP v2, gradient checkpointing, and no EMA
Hardware	16 NVIDIA H20 GPUs

Stage II, *Memory+History Depth+RGB (train 2)*, continues from the Stage-I model and retains the settings in Table 2, except for the conditioning mixture and stage duration. Depth substitution is independently enabled with probability 0.5 for both episodic memory and recent history, whereas Stage I applies it only to episodic memory and keeps history in RGB. Stage II is trained for a further 10,000 steps. Both stages use the same 16-GPU NVIDIA H20 configuration.

B SUPPLEMENTARY METHOD DETAILS

This section reports the memory-validity and reset-baseline analyses, edited-anchor depth policies, context configuration, training-trajectory construction, and adaptation details omitted from the main method for readability.

B.1 MEMORY-VALIDITY AND RESET-BASELINE ABLATION

We evaluate five inference-time policies on 100 paired cases comprising 20 scenes and five appearance edits (Ghibli, cyberpunk, watercolor, pixel art, and Van Gogh). The policies (i) keep stale RGB in both context sources, (ii) hard reset the context and re-anchor from the edited observation, (iii) reset and re-anchor while retaining the pose-indexed pre-edit depth bank, (iv) convert only episodic memory to depth, or (v) convert both episodic memory and recent history to depth. All variants use the same trained MemWorld checkpoint, edited anchor, camera trajectory, random seed, post-edit noise schedule, context budget, and inference configuration; only the edit-boundary handling of memory and history changes. Because each policy is rolled out separately, pre-edit generations are not guaranteed to be pixel-identical, so we treat the comparison as a paired diagnostic rather than a strict causal estimate.

Edit propagation is measured by the CSD, MegaStyle, Gram, and SigLIP-T changes from pre-edit to post-edit windows. DepthCorr measures structure across the edit boundary and within the post-edit rollout. We normalize these six quantities with bounds fixed from the six-method editing benchmark, form Edit and Consistency axes, and report their harmonic-mean Composite. For the three validity-routing policies, confidence intervals use 10,000 scene-clustered bootstrap resamples, retaining all five styles from each sampled scene and using the same resampling matrix for every policy. The two reset baselines are reported as point estimates because matched bootstrap intervals were not computed for these additional runs.

Table 3: **Memory-validity ablation over 100 matched scene–style cases.** The three validity-routing policies report means with 95% scene-clustered bootstrap confidence intervals from 10,000 resamples; the two reset baselines report point estimates only. The paired column reports control minus Full. Converting both memory and recent history to depth gives the strongest overall point-estimate balance.

Variant	Edit Axis ↑	Consistency Axis ↑	Composite ↑	Paired Composite Δ vs. Full ↑
Keep stale RGB	0.1089 $[0.0860, 0.1376]$	1.0000 $[0.9520, 1.0000]$	0.1964 $[0.1584, 0.2410]$	-0.7894 $[-0.8363, -0.7267]$
Hard reset + re-anchor	0.7600	0.6900	0.7233	-0.2626
Geometry-aware reset + re-anchor	0.9000	0.8600	0.8795	-0.1064
RGB→Depth, memory only	0.3569 $[0.2902, 0.4296]$	0.9856 $[0.9367, 1.0000]$	0.5240 $[0.4477, 0.5980]$	-0.4618 $[-0.5401, -0.3864]$
Full: memory + history	**1.0000** $[0.9915, 1.0000]$	0.9722 $[0.9141, 1.0000]$	**0.9859** $[0.9542, 1.0000]$	–

Table 4: **Base metrics for the validity-routing policies.** All values are means with 95% scene-clustered bootstrap confidence intervals. The four appearance deltas measure propagation of the edited style; XEdit and Post DepthCorr measure structural consistency across the edit boundary and within the post-edit rollout, respectively.

Variant	CSD Δ ↑	MegaStyle Δ ↑	Gram Δ ↑	SigLIP-T Δ ↑	XEdit DepthCorr ↑	Post DepthCorr ↑
Keep stale RGB	0.0040 $[-0.0051, 0.0139]$	-0.0047 $[-0.0174, 0.0079]$	-0.0001 $[-0.0283, 0.0275]$	0.0020 $[-0.0018, 0.0059]$	0.9614 $[0.9382, 0.9800]$	0.9889 $[0.9832, 0.9936]$
RGB→Depth, memory only	0.0481 $[0.0280, 0.0698]$	0.0400 $[0.0152, 0.0669]$	0.0059 $[-0.0254, 0.0378]$	0.0364 $[0.0297, 0.0426]$	0.9540 $[0.9306, 0.9735]$	0.9894 $[0.9839, 0.9938]$
Full: memory + history	**0.2744** $[0.2378, 0.3100]$	**0.2354** $[0.1913, 0.2790]$	**0.3945** $[0.3319, 0.4552]$	**0.0861** $[0.0797, 0.0927]$	0.9472 $[0.9201, 0.9708]$	0.9891 $[0.9828, 0.9943]$

Table 3 shows that retaining stale RGB yields an Edit Axis of 0.1089 and a Composite of 0.1964. A hard reset improves edit adoption to 0.7600 but reduces the Consistency Axis to 0.6900, giving a Composite of 0.7233. Retaining the pose-indexed depth bank during reset raises these point estimates to 0.9000, 0.8600, and 0.8795, respectively. Converting episodic memory alone reaches a Composite of 0.5240 because stale recent-history RGB can still override the update. Converting both context sources achieves the best overall point-estimate balance: an Edit Axis of 1.0000, a Consistency Axis of 0.9722 $[0.9141, 1.0000]$, and a Composite of 0.9859 $[0.9542, 1.0000]$. Relative to geometry-aware reset, its point estimates improve CSD Δ by 0.0214, MegaStyle Δ by 0.0194,

XEdit DepthCorr by 0.0292, Post DepthCorr by 0.0051, and Composite by 0.1064. These reset-baseline gaps are descriptive and should not be interpreted as significance claims.

The base metrics in Table 4 explain this gap for the three validity-routing policies. Under stale RGB, all four appearance deltas remain near zero despite an XEdit DepthCorr of 0.9614; this high value reflects failure to update appearance and does not by itself imply a successful edit. Memory-only conversion gives intermediate appearance deltas, whereas converting both context sources gives the strongest response. For the additional reset runs, hard reset gives CSD $\Delta = 0.2210$, MegaStyle $\Delta = 0.1810$, XEdit DepthCorr $= 0.8210$, and Post DepthCorr $= 0.9630$; geometry-aware reset improves these values to 0.2530, 0.2160, 0.9180, and 0.9840. The separately generated pre-edit prefixes remain the main limitation of this diagnostic.

B.2 Edited-Anchor Depth Policy

We next isolate the depth paired with the edited RGB anchor on the same 100-case subset. All variants use the same MemWorld checkpoint, edited RGB anchor, camera trajectory, random seed, post-edit noise schedule, context budget, full validity-aware routing policy, and inference configuration; only the edited-anchor depth tensor differs. Whenever depth is estimated, we use the same frozen estimator, checkpoint, input resolution, depth representation, and preprocessing pipeline used throughout the experiments. No ground-truth or renderer depth is used. As above, each policy is rolled out separately, so the results are paired diagnostics reported as point estimates.

We compare reusing the pre-edit depth at the anchor pose, estimating depth directly from the edited RGB, scale-and-shift aligning that estimate to the pre-edit depth, and supplying the edited anchor without depth. The first policy is the default for the global appearance edits studied in this paper.

Table 5: **Edited-anchor depth policies over 100 matched scene–style cases.** All entries are point estimates. Anchor-aligned RMSE and edge F1 compare the conditioning depth tensor with the pose-matched pre-edit depth. The perfect alignment scores for depth reuse hold by construction and do not measure absolute depth accuracy.

Policy	Aligned RMSE ↓	Edge F1 ↑	CSD Δ ↑	MegaStyle Δ ↑	XEdit DepthCorr ↑	Post DepthCorr ↑	Composite ↑
Reuse pre-edit depth	0.000	1.000	0.275	0.236	0.950	0.990	**0.987**
Raw depth re-estimation	0.142	0.914	0.266	0.226	0.927	0.982	0.958
Re-estimation + scale/shift alignment	0.061	0.914	0.273	0.234	0.946	0.989	0.984
Edited anchor without depth	–	–	0.252	0.216	0.906	0.975	0.914

Table 5 shows that reusing pre-edit depth gives the highest Composite point estimate, 0.987. Raw re-estimation reaches 0.958, while scale-and-shift alignment recovers most of the gap and reaches 0.984. Omitting anchor depth remains functional but is weakest at 0.914. The zero aligned RMSE and unit edge F1 for depth reuse hold by construction because that policy copies the reference tensor; they are not claims of absolute depth accuracy. The rollout metrics are comparative diagnostics under the shared evaluation protocol, and the observed gaps should not be interpreted as significance claims.

These results support reusing the pose-matched depth for the appearance-only edits studied in this paper, where scene geometry is assumed unchanged. However, this ablation does not cover edits that alter scene geometry, and the findings should not be taken as prescriptive depth guidance for that setting. When geometry changes, pre-edit depth may itself become obsolete, and the memory bank's structural prior could conflict with the intended modification. Investigating depth policies under geometry-altering edits remains a limitation and a direction for future work.

B.3 Memory Context and Retrieval Configuration

The episodic memory contains five latent observations, including the persistent first-frame sink. We use five recent-history latents and predict ten current latents per regular chunk. The active DiT context therefore remains fixed even as the addressable observation bank grows.

To ensure that training requires nonlocal evidence, we synthesize leave-and-return trajectories with reverse-playback tails. The anchor selector searches for windows with enough geometrically over-

lapping historical frames and falls back to the best available history when an exact revisit is unavailable. A temporal-gap bias favors larger separation between a retrieved observation and the current window, preventing episodic memory from collapsing into a duplicate of recent history.

B.4 Context Curriculum and Residual Reinjection

During Stage I, the full-memory, history-only, and initial-image conditioning regimes are sampled with probabilities 0.8, 0.1, and 0.1, respectively. The latter two regimes match the absence of useful episodic evidence and the initialization of a chunk-wise rollout.

To reduce the train–inference gap created by ground-truth conditioning frames, we maintain a per-rank FIFO error buffer $\mathcal{Q}$ inspired by residual reinjection for long-horizon world modeling (Wang et al., 2026c). After a training prediction at step s, we estimate the clean-latent error of each supervised current slice j as

$$\boldsymbol{\delta}_{s,j} = \mathrm{sg}\left[-\sigma_s \left(\hat{\mathbf{v}}_{s,j} - \mathbf{v}^{\star}_{s,j}\right)\right], \qquad j \in \Omega_{\mathrm{C}}, \tag{10}$$

where sg denotes stop-gradient and $\mathbf{v}^{\star}_{s,j} = \boldsymbol{\epsilon}_{s,j} - \mathbf{X}^0_{s,j}$. This follows from the flow path in Eq. 5: the implied clean estimate satisfies $\hat{\mathbf{X}}^0 - \mathbf{X}^0 = -\sigma_s(\hat{\mathbf{v}} - \mathbf{v}^{\star})$. We split these detached errors into individual latent slices, append them to $\mathcal{Q}$, and evict the oldest entries once its capacity of 512 slices is reached. Because insertion occurs after the prediction, a residual can perturb only later training examples.

When enough entries are available, we sample with replacement one stored slice for each clean conditioning slot, producing tensors $\boldsymbol{\Delta}_M \sim \mathrm{Sample}(\mathcal{Q}, |\Omega_{\mathrm{M}}|)$ and $\boldsymbol{\Delta}_H \sim \mathrm{Sample}(\mathcal{Q}, |\Omega_{\mathrm{H}}|)$. A single gate $b_s \sim \mathrm{Bernoulli}(p_{\mathrm{err}})$ controls whether the sampled errors are injected on that step:

$$\tilde{\mathbf{M}}_s = \mathbf{M}_s + b_s \gamma_m \boldsymbol{\Delta}_M, \qquad \tilde{\mathbf{H}}_s = \mathbf{H}_s + b_s \gamma_h \boldsymbol{\Delta}_H. \tag{11}$$

We use $p_{\mathrm{err}} = 0.9$ and $\gamma_m = \gamma_h = 1$. The buffer is local to each distributed rank and is not synchronized. Residuals affect only clean memory and recent-history conditions; noisy current tokens, their clean flow targets, and the real image-to-video anchor remain unmodified.

B.5 Modality Mixing and Adaptation Details

In Stage II, separate Bernoulli events enable depth substitution in episodic memory and recent history, each with probability 0.5. Once enabled, an event replaces all valid slots with probability 0.5 and otherwise replaces a nonempty random subset. This produces pure RGB, pure depth, and partially mixed contexts without explicit modality tokens.

All camera extrinsics are expressed relative to the first token of the joint context. Camera operations are performed in FP32 before casting back to the attention dtype. To reduce synchronized positional aliasing for distant memories, the temporal RoPE base is perturbed across attention heads:

$$\hat{\theta}_h = \theta_0(1 + \sigma_\theta \epsilon_h), \qquad \epsilon_h \in [-1, 1]. \tag{12}$$

The pretrained DiT and projective camera branch are frozen. We optimize rank-64 LoRA adapters attached to the self-attention q, k, v, and output projections.

C Implementation and Evaluation Details

This section reports the complete persistent-editing, R2M-Bench, and MBench tables moved from the main text, followed by the metric implementation and score definition needed to interpret them.

C.1 Full Persistent-Editing Results

Table 6 reports all style-fidelity, edit-response, retention, and pose-matched geometry metrics underlying the compact main-text comparison. The method-set-dependent Edit and Consistency axes should be read together with these raw metrics.

Table 6: Persistent-editing results over 100 scenes and five styles. Metrics are grouped by evaluation aspect in one table. XEdit pairs a pre-edit view with its post-edit revisit, whereas Post uses pairs entirely after the edit. Bold and underlined values denote the best and second-best result, respectively. All methods are evaluated on all 500 cases.

Aspect	Metric	MemWorld	LingBot-World-2.0	HY-WorldPlay	SANA-WM	Matrix-Game 3.0	Lyra 2.0
Post-edit style fidelity	CSD Post ↑	**0.7531**	0.5701	0.5141	0.3584	0.6012	0.6346
	MegaStyle Post ↑	**0.6855**	0.5050	0.4667	0.3229	0.5215	0.5486
	Gram Dist. Post ↓	**0.9234**	1.4410	1.4950	1.8672	1.2882	1.3361
	SigLIP-T Post ↑	**0.0571**	0.0146	0.0163	0.0472	0.0315	0.0321
Edit responsiveness	CSD Δ ↑	**0.2164**	0.0441	-0.0016	-0.0239	0.0080	0.1104
	MegaStyle Δ ↑	**0.1983**	0.0526	-0.0036	-0.0078	0.0015	0.0990
	Gram Δ ↑	**0.3030**	0.0609	-0.0308	-0.1329	-0.0231	0.1354
	SigLIP-T Δ ↑	**0.0567**	0.0101	0.0086	0.0207	0.0130	0.0305
Style retention	CSD Retention ↑	**0.0484**	0.0111	0.0222	-0.0304	0.0202	-0.0425
	MegaStyle Retention ↑	0.0095	-0.0052	**0.0141**	-0.0137	0.0028	-0.0465
Revisit consistency	XEdit DepthCorr ↑	0.9273	0.7098	0.9478	0.8175	0.8350	**0.9612**
	XEdit DepthSSIM ↑	0.8913	0.7765	0.8869	0.7997	0.8636	**0.9082**
	Post DepthCorr ↑	**0.9860**	0.8432	0.9297	0.7596	0.8502	0.9777
	Post DepthSSIM ↑	**0.9583**	0.8392	0.8728	0.7647	0.8597	0.9306
Overall	Edit Axis ↑	**1.000**	0.263	0.087	0.063	0.130	0.537
	Consistency Axis ↑	0.933	0.185	0.849	0.214	0.449	**0.982**
	Composite ↑	**0.965**	0.217	0.157	0.097	0.202	0.694

Additional qualitative comparisons. Figure 3 presents two representative cases; the six cases in Figure 5 broaden the comparison to painterly, watercolor-like, pixel-art, and neon appearance interventions in both indoor and outdoor environments. All panels use the same method ordering and the same three-stage protocol as the main-text figure. We omit case names and scene identifiers inside the figure so that the comparison emphasizes the visual trajectory from the original observation to the edited anchor and the later return view. From top to bottom, the paired cases cover a sparse outdoor arrangement and sculpture plaza, an indoor concourse and industrial yard, and a rooftop utility area and arched courtyard. Their distinctive structures provide visual anchors for assessing appearance retention and layout stability.

Across these cases, MemWorld most consistently carries the edited appearance into the revisit while retaining recognizable scene anchors. Competing methods often respond strongly at the intervention frame but then move back toward the pre-intervention rendering, lose the target palette, or alter prominent geometry. The comparison therefore tests persistence rather than the quality of a single edited frame.



Figure 5: **Additional persistent-appearance comparisons across six cases.** The top, middle, and bottom blocks each show two cases using the same Explore–Intervene–Revisit protocol and method order as Figure 3. Across diverse settings, MemWorld retains the intervened palette and rendering style while preserving the principal scene anchors. Several competing methods instead revert toward the original appearance, keep only a partial trace of the edit, or alter prominent geometry by the revisit.

C.2 Matched Comparison with PermaVid

We conduct a complete matched comparison on 500 scene–style cases (100 scenes and all five target styles). Both systems use the same Qwen-Image-Edit anchor generator, seed, WSLRRLL action sequence, temporal evaluation windows, metric implementation, and frozen normalization bounds. Before metric extraction, all outputs are resized to the same evaluation resolution. Confidence intervals use 10,000 scene-clustered bootstrap resamples, retaining all five styles for every sampled scene. The comparison is action-matched but not pose-identical because the two systems expose different camera interfaces.

Table 7: **Complete matched comparison with PermaVid on 500 scene–style cases.** Outputs are resized to a common evaluation resolution before scoring. The paired difference is PermaVid minus MemWorld; intervals are 95% scene-clustered bootstrap confidence intervals. XEdit compares pre-edit views with their post-edit revisits, whereas Post compares views within the updated world. ↓ indicates lower is better.

Metric	PermaVid	MemWorld	Paired Δ [95% CI]
Pre-edit baseline			
CSD Pre ↑	0.4934	0.4610	$+0.0323\ [-0.0379,\ 0.1017]$
MegaStyle Pre ↑	0.3764	0.4422	$-0.0657\ [-0.1372,\ 0.0053]$
Gram Pre ↓	1.1217	1.5009	$-0.3792\ [-0.6382,\ -0.1188]$
SigLIP-T Pre ↑	-0.0033	-0.0181	$+0.0148\ [0.0103,\ 0.0197]$
Post-edit fidelity and edit response			
CSD Post ↑	0.7537	0.7355	$+0.0182\ [-0.0097,\ 0.0438]$
MegaStyle Post ↑	0.6401	0.6775	$-0.0374\ [-0.0917,\ 0.0179]$
Gram Post ↓	0.8069	1.1064	$-0.2995\ [-0.4800,\ -0.1194]$
SigLIP-T Post ↑	0.0687	0.0680	$+0.0007\ [-0.0093,\ 0.0127]$
CSD Δ ↑	0.2603	0.2744	$-0.0141\ [-0.0906,\ 0.0637]$
MegaStyle Δ ↑	0.2637	0.2354	$+0.0284\ [-0.0546,\ 0.1107]$
Gram Δ ↑	0.3148	0.3945	$-0.0797\ [-0.2084,\ 0.0555]$
SigLIP-T Δ ↑	0.0720	**0.0861**	$-0.0141\ [-0.0238,\ -0.0020]$
Persistence and structural consistency			
CSD Retention ↑	-0.1154	**0.0449**	$-0.1603\ [-0.1944,\ -0.1288]$
MegaStyle Retention ↑	-0.1505	-0.0076	$-0.1430\ [-0.1843,\ -0.1041]$
XEdit DepthCorr ↑	0.8299	**0.9472**	$-0.1173\ [-0.1764,\ -0.0648]$
XEdit DepthSSIM ↑	0.8049	**0.9106**	$-0.1057\ [-0.1266,\ -0.0842]$
Post DepthCorr ↑	0.8539	**0.9891**	$-0.1353\ [-0.2000,\ -0.0809]$
Post DepthSSIM ↑	0.8258	**0.9649**	$-0.1392\ [-0.1579,\ -0.1205]$
Summary quantities			
Edit Axis ↑ †	1.0000	1.0000	$0.0000\ [-0.0888,\ 0.0033]$
Consistency Axis ↑	0.4471	**0.9722**	$-0.5251\ [-0.7852,\ -0.2931]$
Composite ↑ †	0.6179	0.9859	$-0.3680\ [-0.6903,\ -0.1763]$

† The Edit Axis saturates for both systems; the Composite is therefore a deterministic transform of the Consistency Axis and is not independent ranking evidence.

The matched results separate edit adoption from persistence. Overall CSD and MegaStyle edit responses, as well as absolute post-edit style fidelity, show no significant difference. After continued generation, however, MemWorld has higher CSD and MegaStyle retention and stronger cross-edit and post-edit depth consistency. The paired confidence intervals exclude zero for CSD retention,

MegaStyle retention, and all four depth-consistency metrics. The retention and consistency gaps also have the same direction for all five styles.

The frozen Edit Axis saturates at 1.000 for both systems. Consequently, the reported Composite becomes the deterministic transform $2C/(1 + C)$ of the Consistency Axis and is not independent ranking evidence. Our conclusions therefore rest on raw metrics and paired differences. Gram distance favors PermaVid both before and after the intervention, while Gram Δ is not significantly different. Because the pre-edit gap is already present after the common evaluation resize, we treat it as a baseline shift rather than evidence of better editing.

Table 8: **Mechanism-level comparison with PermaVid.** Both systems use geometry-aware retrieval and edit-aware RGB–depth validity. Their main difference is how the selected evidence enters generation. This table records architectural facts and is not a same-backbone causal ablation.

Component	PermaVid	MemWorld
Retrieval	Field-of-view-overlap retrieval from a dual RGB–depth bank	Camera-frustum TopK retrieval with recent-context exclusion
Validity scope	Version-aware filtering of retrieved RGB and depth memory	One world-version policy for retrieved memory and the five-latent recent-history window
Generation interface	Selected images enter as VACE `reference-image` conditions; the latest frame supplies the recent video condition	Retrieved memory, recent history, and current noisy latents form $[\mathbf{M}_t \mid \mathbf{H}_t \mid \mathbf{X}_t^\sigma]$ inside native self-attention
Adaptation	Training-free memory routing around the published VACE-based generator	Two-stage LoRA adaptation: RGB recall followed by RGB–depth mixing under self-generated context

The main architectural distinction is the integration interface. PermaVid filters a dual RGB–depth memory bank and provides selected images to the generator as VACE reference-image conditions. MemWorld instead places geometrically retrieved memory, recent history, and current noisy latents in a single camera-aware self-attention sequence. This gives nonlocal and recent evidence a shared, trained route into every denoising block; the validity policy also covers the five-latent recent-history window, not only the retrieved bank. The two-stage training procedure first establishes RGB recall and then adapts the same pathway to RGB–depth mixtures under self-generated context. The matched performance gap is consistent with this integration advantage, while the within-MemWorld ablation in Sec. B.1 provides diagnostic evidence consistent with the importance of routing recent history as well as memory. Because PermaVid uses a Wan2.1-VACE-14B backbone whereas MemWorld uses a Wan2.2-TI2V-5B backbone, the cross-system comparison alone does not causally isolate VACE conditioning from joint self-attention.

C.3 Editing Persistence over Time

Figure 6 resolves the post-edit rollout over normalized time, complementing the aggregate results in Table 6. The first two panels distinguish a persistent edit from either a high average dominated by early frames or an isolated initial response that later disappears: they report CSD similarity to each method's edited anchor and the corresponding change relative to the pre-edit window. The third panel reports DepthCorr only at pose-matched pre-/post-edit views. The WSLRRLL trajectory provides such matches at six of the 12 normalized post-edit bins; the intervening bins are deliberately left unconnected rather than interpolated. MemWorld maintains the highest CSD-to-anchor and edit-response curves throughout the rollout. Its geometry values should be interpreted only at the displayed matched viewpoints, rather than as a claim of continuous geometric measurement during the off-view excursion.

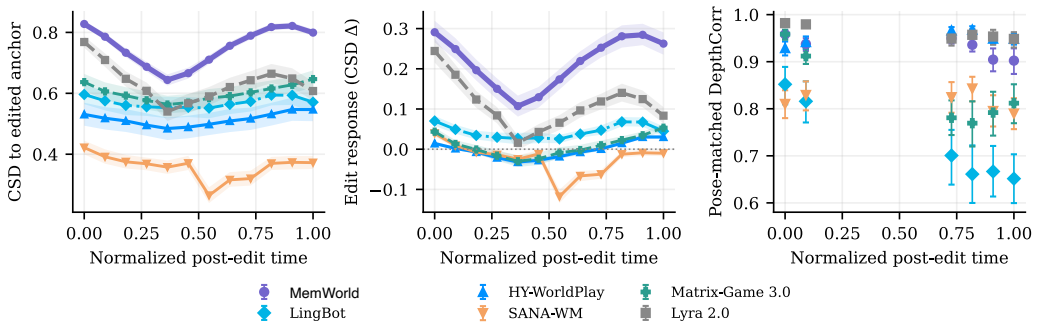

Figure 6: **Editing persistence across normalized post-edit time.** The first two panels show CSD similarity to each method's edited anchor and the CSD change relative to the pre-edit window; shaded bands are 95% confidence intervals from image-clustered bootstrap resampling. The third panel shows pose-matched DepthCorr only at the six trajectory bins with valid pre-/post-edit camera matches; error bars give the same confidence interval and no line is drawn across unmatched bins. All six methods are evaluated on the complete 500-case benchmark.

C.4 Full R2M-Bench Results

Table 9 gives the complete per-metric MemoryGain and NMR results. Benchmark construction and validation are reported in the original R2M-Bench paper (Gu et al., 2026b).

Table 9: Template-balanced MemoryGain (MG; top) and Normalized Memory Ratio (NMR; bottom), each averaged equally over Out-and-back, Translation–rotation, and Closed-loop. Memory types: RC = raw context, FoV-R = field-of-view retrieval, G-R = geometric retrieval, and SSM = state-space recurrence. WonderWorld (gray) is a separate 3D rendering reference. For video models, **bold** and underlined values mark the best and second-best results. Overall NMR first averages metrics within each family and then the five families; MG is not aggregated because its units are metric-specific.

MemoryGain (MG; higher is better)

			Appearance Fidelity			Scene Identity			Object Identity		Local Geometry		Persistent State
Method	**Params**	**Mem. Type**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↑**	**DINO-sim ↑**	**BoQ-sim ↑**	**MVPR-sim ↑**	**App. Persist ↑**	**Sem. Persist ↑**	**Feat. Corr. ↑**	**Geo. Inlier ↑**	**State Cons. ↑**
WonderWorld	–	Explicit 3D	12.828	0.398	0.548	0.369	0.633	0.327	0.254	0.067	0.535	0.147	2.676
LingBot-World-Fast	14B	RC	0.283	0.018	0.007	0.029	0.041	0.027	0.022	0.012	0.040	0.008	0.186
LingBot-World-2-Fast	14B	RC	0.271	0.009	0.011	0.083	0.013	0.057	0.031	0.014	0.015	0.023	0.262
Matrix-Game 3.0	5B	FoV-R	1.039	0.024	0.060	0.145	0.176	**0.104**	**0.098**	**0.045**	0.145	0.037	0.476
HY-WorldPlay	8B	FoV-R	2.357	0.055	0.155	0.146	0.208	0.069	0.087	0.029	0.180	0.018	0.510
SANA-WM	2.6B	SSM	0.393	0.008	0.009	-0.004	0.003	-0.019	-0.001	-0.003	-0.001	0.009	-0.008
Lyra-2	14B	G-R	1.318	0.041	0.090	0.129	0.134	0.086	0.058	0.022	0.096	0.031	0.422
MemWorld	5B	FoV-R	**4.647**	**0.120**	**0.247**	**0.185**	**0.262**	0.103	0.095	0.027	**0.184**	**0.046**	**0.912**

Normalized Memory Ratio (NMR; higher is better)

			Appearance Fidelity			Scene Identity			Object Identity		Local Geometry		Persistent State	
Method	**Params**	**Mem. Type**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↑**	**DINO-sim ↑**	**BoQ-sim ↑**	**MVPR-sim ↑**	**App. Persist ↑**	**Sem. Persist ↑**	**Feat. Corr. ↑**	**Geo. Inlier ↑**	**State Cons. ↑**	**Overall ↑**
WonderWorld	–	Explicit 3D	3.253	3.076	1.868	1.110	1.249	1.096	1.221	1.319	1.166	1.132	1.247	1.510
LingBot-World-Fast	14B	RC	0.047	0.083	0.030	0.073	0.080	0.103	0.126	0.245	0.098	0.043	0.103	0.100
LingBot-World-2-Fast	14B	RC	0.198	0.144	0.045	0.097	0.040	0.115	0.112	0.311	0.040	0.092	0.216	0.141
Matrix-Game 3.0	5B	FoV-R	0.257	0.192	0.251	0.449	0.400	0.470	0.572	**0.896**	0.385	0.226	0.300	0.403
HY-WorldPlay	8B	FoV-R	0.446	0.308	0.488	**0.734**	0.589	0.626	0.602	0.752	0.488	0.243	0.318	0.485
SANA-WM	2.6B	SSM	0.094	0.071	0.051	0.002	0.012	-0.049	0.016	-0.007	0.002	0.020	0.006	0.016
Lyra-2	14B	G-R	0.284	0.286	0.306	0.457	0.330	0.435	0.337	0.442	0.236	0.180	0.255	0.310
MemWorld	5B	FoV-R	**1.034**	**1.093**	**0.981**	0.700	**0.692**	**0.662**	**0.641**	0.782	**0.533**	**0.355**	**0.654**	**0.706**

C.5 Qualitative Limitation: Fine-Grained Local Consistency

The aggregate R2M-Bench results measure scene-level revisit consistency, but our learned generation-and-memory pipeline does not guarantee exact persistence of every local detail. Figure 7 shows two selected MemWorld rollouts from the standard R2M-Bench evaluation protocol. These examples are qualitative diagnostics rather than estimates of failure frequency.



Figure 7: **Local consistency limitations despite stable scene-level structure.** Orange boxes mark the relevant regions, and the second row additionally enlarges the sign. **(a)** During a close camera pass, the red sculpture, tower, and plaza layout remain recognizable, but the nearby glass shelter develops transient texture stretching and later changes shape. **(b)** Across a wider camera sweep, the pickup truck and salt-flat layout remain stable, while the small lettering on the sign varies across viewpoints. Frame indices and timestamps refer to the corresponding 16-FPS generated rollouts.

Both effects are local: the surrounding scene and principal landmarks remain consistent while near-field geometry or high-frequency symbolic details drift. They therefore expose a gap between scene-level memory and pixel-accurate local reconstruction rather than a loss of the remembered environment. Finer-grained correspondence and stronger near-view geometric constraints are promising directions for reducing these artifacts.

C.6 Full MBench Results

Table 10 reports all nine MBench sub-dimensions. MemWorld ranks second overall, with its strongest results in object texture and rendering light and its clearest remaining gaps in human identity and interaction action.

Table 10: Results on MBench. The nine sub-dimensions evaluate object, human, spatial, rendering, and interaction memory; Mean is the benchmark aggregate. All metrics are higher-is-better. **Bold** and underlined values denote the best and second-best result, respectively.

Method	Obj-Geom ↑	Obj-Tex ↑	Hum-Id ↑	Hum-App ↑	Spa-Epi ↑	Spa-Rep ↑	Ren-Light ↑	Ren-Style ↑	Int-Action ↑	Mean ↑
Matrix-Game 2.0	14.62	28.99	1.22	0.94	14.78	3.08	38.79	73.78	47.86	23.68
Matrix-Game 3.0	44.15	58.22	42.38	47.91	61.99	32.86	62.06	95.17	81.93	55.72
HY-WorldPlay	47.12	68.54	**52.46**	**66.58**	**83.86**	**68.17**	82.67	**98.23**	85.69	**69.57**
Yume-1.5	**60.96**	49.99	17.41	40.57	51.86	24.55	51.21	92.05	62.20	58.52
Lingbot-World	33.20	44.54	11.57	33.53	22.12	7.57	40.06	85.87	63.32	47.93
Infinite-World	35.70	61.88	23.08	46.85	74.04	61.51	62.63	96.87	**86.37**	61.39
MemWorld	52.27	**71.51**	31.37	57.98	65.40	66.19	**97.46**	97.95	76.20	68.48

C.7 R2M-Bench Metric Implementation Details

We follow the public R2M-Bench protocol (Gu et al., 2026b). Metrics are computed on individual frame pairs and averaged over the revisit, baseline, and short-range sets before computing MemoryGain and NMR. Unless stated otherwise, RGB frames are evaluated at their decoded resolution.

Appearance Fidelity. PSNR and SSIM use RGB values in $[0, 255]$, with a 7×7 SSIM window. LPIPS uses the AlexNet backbone and inputs scaled to $[-1, 1]$. PSNR and SSIM are higher-is-better; LPIPS is direction-flipped for MemoryGain and NMR.

Scene Identity Preservation. DINOv2 uses ViT-B/14 features from ImageNet-normalized 518×518 crops. BoQ uses its DINOv2-backed VPR model at 322×322, and MutualVPR uses 512-

dimensional descriptors at 518×518. All descriptors are ℓ_2-normalized and compared by cosine similarity.

Object Identity. Objects are re-detected in each pair rather than tracked through the rollout. GroundingDINO detections use box/text thresholds 0.25/0.20, NMS IoU 0.60, and a minimum SAM2 mask-area ratio of 0.002. We retain the top ten anchors by detection confidence times mask area and at most five candidates per concept in the paired frame. Masked DINOv2 crop similarity gives appearance compatibility; GroundingDINO confidence and CLIP image–text similarity give semantic compatibility. Their equal-weighted score accepts a candidate at 0.50. Appearance and semantic persistence average the accepted scores over all first-frame anchors, assigning zero to unmatched anchors; a pair with no anchor also receives zero.

Local Geometric Correspondence. SuperPoint detects at most 1024 keypoints per image with threshold 5×10^{-4}; LightGlue uses matching threshold 0.1. The feature score is the number of matches divided by the smaller keypoint count. OpenCV RANSAC estimates a fundamental matrix with a 3-pixel threshold and confidence 0.99; the second score is the inlier fraction. It is set to zero when fewer than eight matches are available or estimation fails. Both metrics are higher-is-better.

C.8 Vision-Language Evaluation Prompt

We apply the same fixed prompt to every frame pair in the revisit, baseline, and short-range sets. The vision-language model is not told which set a pair comes from; it assigns a pairwise persistent-state consistency score using only the two images. The rubric first asks whether the images contain evidence of the same local scene, then accounts for viewpoint and occlusion when judging shared content. This distinction is important because a pair with no recognizable shared scene should receive a low consistency score rather than a perfect score merely because no local contradiction can be identified. The full prompt is shown below.

```
You are an expert evaluator of Video World Models.

Task: Assign a pairwise persistent-state consistency score to two images
sampled from the same generated video.

* Image A is earlier and Image B is later.
* Do NOT assume that the images depict the same place.
* This is not raw pixel similarity: viewpoint, occlusion, and mild rendering
  changes may differ without implying a state inconsistency.
* Score whether the images provide evidence of the same local scene and, when
  they do, whether the shared scene content remains persistent.

Procedure:

1. Classify the scene relation as same, partial, different, or uncertain.
2. If shared scene content exists, estimate the viewpoint change and identify
   the co-visible region.
3. Judge identity, geometry, and persistence only where content can be
   compared, discounting differences explained by viewpoint or occlusion.
4. If there is no recognizable shared scene content, assign low consistency;
   absence of comparable content is not evidence of perfect consistency.

Viewpoint exemptions for same or partially overlapping scenes: do NOT penalize:
* Any difference plausibly caused by camera motion, perspective, scale, or
  occlusion.
* Objects entering or leaving the frame due to the viewpoint shift.
* Minor lighting, exposure, or rendering-noise differences.
* When uncertain whether a difference is from viewpoint or a true change,
  assume viewpoint.

Inconsistency types (report only confident findings):
* scene: images depict different local scenes or share no recognizable place cues.
* identity: same object changed color, shape, material, texture, or category.
* geometry: impossible spatial layout or structural contradictions.
* persistence: object disappeared or appeared well inside the shared region,
  or count of repeated elements changed without occlusion explanation.

Scoring anchors:
* 5.0 = strong evidence of the same local scene; shared state is consistent.
* 4.0 = same local scene with only minor confident inconsistencies.
* 3.0 = partial or uncertain scene overlap, or moderate inconsistencies.
* 2.0 = weak shared-scene evidence or major state inconsistencies.
* 1.0 = clearly different local scenes or fundamentally incompatible states.
```

```
Intermediate values are allowed. Do not give a high score solely because
the images are individually plausible or because no content is comparable.

Output constraints: keep SHORT:
* scene_relation is one of: same, partial, different, uncertain
* viewpoint_difference <= 15 words
* at most 3 inconsistencies, each description <= 12 words
* reason <= 15 words
* Each inconsistency has ONLY keys: type, severity, description.

Output ONLY this JSON (no markdown):
{
  "scene_relation": "same|partial|different|uncertain",
  "viewpoint_difference": "...",
  "inconsistencies": [
    {
      "type": "scene|identity|geometry|persistence",
      "severity": "minor|moderate|major|catastrophic",
      "description": "..."
    }
  ],
  "score": <number from 1.0 to 5.0>,
  "reason": "..."
}
```

D R2M-BENCH SCORE DEFINITION

Following R2M-Bench (Gu et al., 2026b), let $\bar{s}_q^{\text{rev}}$, $\bar{s}_q^{\text{base}}$, and $\bar{s}_q^{\text{short}}$ denote the oriented mean score for metric q over pose-matched revisit pairs, same-video gap-matched non-revisit pairs, and short-range pairs. The benchmark defines

$$\text{MG}_q = \bar{s}_q^{\text{rev}} - \bar{s}_q^{\text{base}}, \tag{13}$$

$$\text{DR}_q = \bar{s}_q^{\text{short}} - \bar{s}_q^{\text{base}}, \tag{14}$$

$$\text{NMR}_q = \frac{\text{MG}_q}{\text{DR}_q + \epsilon}. \tag{15}$$

MemoryGain is the revisit advantage in each metric's native units, whereas NMR normalizes that advantage by the metric's short-range dynamic range. Overall NMR first averages metrics within each family and then weights the five families equally. An NMR near zero indicates no advantage over the long-gap baseline; a value near one reaches the corresponding short-range level. As in the original benchmark, NMR is interpreted together with MG because a small dynamic range can amplify the normalized value.