GAUMEM: LEARNING WHAT TO KEEP IN A FIXED-BUDGET GAUSSIAN MEMORY FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

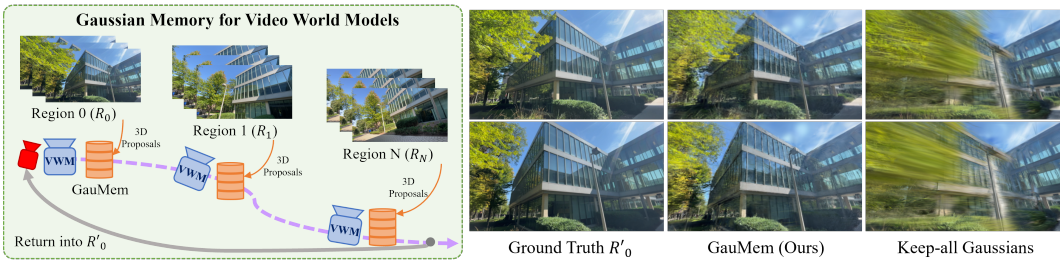

Figure 1: **Persistent 3D scene memory for video world models.** To ensure physically consistent generation, video world models require persistent spatial representations of visited locations. During sequential generative rollouts, **GauMem** consolidates streaming 3D Gaussian proposals emitted by a feed-forward frontend into a strictly bounded capacity. When a rollout loops back to the initial space after extended exploration ($\mathcal{R}_0 \rightarrow \mathcal{R}_1 \cdots \rightarrow \mathcal{R}_N \rightarrow \mathcal{R}'_0$), naive accumulation ("Keep-all Gaussians") suffers from severe visual clutter and unbounded memory expansion, whereas our fixed-budget GauMem reliably preserves 3D geometry to guide downstream video generation.

ABSTRACT

Video world models require long-horizon spatial consistency across extended rollouts. Existing designs either grow unboundedly or rely on hand-crafted pruning heuristics that fail to anticipate retention needs upon delayed revisits. To resolve this, we propose **GauMem**, a capacity-bounded 3D Gaussian scene memory for video world models, governed by an end-to-end learned write policy. GauMem centers on a *Read-Before-Write (RBW)* mechanism: rendering the memory into incoming viewpoints diagnoses discrepancies to isolate novel content from existing structures prior to updates. Guided by this write-time evidence, *Evidence-Gated Merge (EGM)* consolidates consistent incoming proposals into stored slots, after which *Joint Insertion–Eviction (JIE)* ranking prioritizes stored slots and residual proposals within a unified pool up to the hard budget. The entire write policy is trained end-to-end on delayed revisits to optimize long-horizon spatial recall. On DL3DV, GauMem substantially outperforms heuristic write and fusion rules, as well as an unbounded store accumulating 7–15$\times$ more primitives under severe scene interference. When decoded through a frozen video diffusion model, GauMem consistently preserves superior visual fidelity and tight 3D camera trajectory consistency.

1 INTRODUCTION

Recent advances in video generation, propelled by large-scale video diffusion models (Ho et al., 2022; Blattmann et al., 2023; Kong et al., 2024; Yang et al., 2025; Wan et al., 2025), have given rise to video world models, which predict what an agent would observe given its next actions (Yang et al., 2023; Hu et al., 2023; Bruce et al., 2024; Wang et al., 2026b). Because actions arrive sequentially, such models synthesize future rollouts conditioned on historical context. Maintaining spatial persistence therefore requires carrying observed content forward: an agent that leaves a space and

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.

later returns should find the environment as it was left. However, because the generator's internal context is limited to the recent past and suffers from compounding drift beyond it (Po et al., 2025), continuous exploration necessitates an external memory operating under a strictly bounded capacity. Under such a hard budget, simply expanding storage merely defers the fundamental retention decision; the memory is inevitably forced to decide which content to preserve and which to discard.

Existing video world model memories prove inadequate when constrained to such a strict budget. While retrieval-based buffers (Xiao et al., 2025; Yu et al., 2025; Li et al., 2025a) scale unboundedly and condition generation on disjoint 2D views lacking projective consistency, bounded implicit designs (Wei et al., 2026; Wang et al., 2025b) compress history into latent tokens which lack explicit geometric grounding, leaving rollouts prone to trajectory drift. Conversely, explicit 3D stores (Ren et al., 2025; Wu et al., 2025) rely on unprojected point clouds that continually accumulate projection errors without slot-level retention control. Compared to point clouds, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) provides continuous, hole-free rendering whose learnable opacities and spatial extents naturally support primitive consolidation. However, managing Gaussians under a bounded budget with standard heuristics, such as thresholding opacity, recency, or instantaneous rendering contribution, leads to severe visual clutter and catastrophic forgetting upon return. Since hand-crafted write rules cannot foresee which structures remain essential upon delayed revisits, the memory write policy should be learned end-to-end directly for long-horizon spatial recall.

To address these limitations, we propose **GauMem**, a fixed-budget 3D scene memory built upon 3DGS and governed by a learned write policy. Our key insight is that the memory should *read itself before it writes*. Under this *Read-Before-Write (RBW)* mechanism, the memory is first rendered into the incoming camera viewpoint to evaluate its current scene representation; the computed disagreement against the observation provides direct write-time evidence, where uncovered pixels indicate novel content, photometrically consistent regions yield redundancy, and depth discrepancies expose geometric inconsistency. Guided by this evidence, an *Evidence-Gated Merge (EGM)* first refines existing memory by fusing consistent incoming proposals into their nearest stored slots via a learned, evidence-weighted gate. Conditioned on the updated evidence, *Joint Insertion–Eviction (JIE) ranking* then prioritizes stored slots and residual incoming proposals jointly up to the hard budget, unifying insertion and eviction into a single decision while suppressing primitives that have already transferred their evidence. The entire write policy, comprising only $0.67\,\mathrm{M}$ parameters across all modules, is trained end-to-end by backpropagating rendering losses on views revisited later in the trajectory, directly optimizing for long-horizon spatial recall.

We evaluate GauMem by simulating interactive video world model rollouts: an initial space is observed and encoded into memory, the agent traverses extended intervening scenes, and eventually returns to the starting environment without re-accessing the written frames of R_0 during intervening writes. On DL3DV (Ling et al., 2024), our learned write policy renders the revisited space with substantially higher visual fidelity than heuristic write policies for 3D Gaussians, including an unbounded store accumulating 7–15$\times$ more primitives. This spatial persistence directly benefits downstream video generation: when a frozen video diffusion model (Wan et al., 2025; Jiang et al., 2025) synthesizes return rollouts conditioned on the renderings from GauMem, it maintains superior visual consistency while keeping re-estimated camera trajectories markedly closer to the commanded path. Our contributions are summarized as follows:

- We introduce **GauMem**, a capacity-bounded 3D Gaussian scene memory for video world models, governed by an end-to-end learned write policy that directly optimizes long-horizon spatial recall across delayed revisits.
- We design a learned write policy comprising: (i) the **Read-Before-Write (RBW)** mechanism to diagnose geometric and photometric discrepancies prior to updates, (ii) an **Evidence-Gated Merge (EGM)** to fuse consistent incoming proposals into stored slots, and (iii) a **Joint Insertion–Eviction (JIE)** ranking that unifies insertion and eviction under a strict capacity budget.
- We establish an evaluation protocol tailored for spatial persistence in video world models, rigorously benchmarking delayed revisits under intervening scene interference across both raw 3D renderings and downstream video generation rollouts.
- We demonstrate that GauMem substantially outperforms heuristic write rules and unbounded stores under a fixed budget, while enforcing 3D geometric consistency in downstream video generators on DL3DV.

2 RELATED WORK

Feed-forward 3D Gaussian splatting as proposal frontends. Enforcing long-horizon 3D consistency in video world models necessitates an explicit spatial representation (Ren et al., 2025) that can be rendered directly into camera viewpoints. Among explicit formulations, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) offers continuous, hole-free geometry and natural primitive consolidation over sparse point clouds; however, standard per-scene optimization is computationally intractable for online streaming. Feed-forward 3DGS frameworks circumvent this latency by lifting incoming observations into 3D Gaussian proposals in a single forward pass (Charatan et al., 2024; Chen et al., 2024; Xu et al., 2025; Bui et al., 2026; Zhao et al., 2026), often leveraging geometric priors from vision foundation backbones (Wang et al., 2025a; Lin et al., 2025). Crucially, as these predictions are stateless across chunks, naively accumulating their proposals quickly exceeds practical storage budgets while compounding geometric redundancy. Alternatively, large multi-view reconstruction models (Ziwen et al., 2025; Kang et al., 2026) process video sequences simultaneously; yet, they are non-causal, require full re-computation as new frames arrive, and generate unconstrained Gaussian sets that scale with sequence length. Addressing these limitations necessitates a dedicated write policy to manage 3D Gaussian proposals as a persistent world-model memory.

Streaming 3D reconstruction. While online 3D Gaussian splatting incrementally updates scene representations by dynamically merging or pruning primitives (Li et al., 2025b; Huang et al., 2026), the lack of an explicit capacity bound causes storage to scale with sequence length, with retention guided strictly by immediate frame overlap rather than delayed revisits. Recurrent and autoregressive reconstruction methods instead maintain learned latent or hidden feature states (Wang et al., 2025b; Wang & Agapito, 2025; Wang et al., 2026a; LAN et al., 2026). However, under fixed capacities, these states suffer from progressive overwriting as new observations arrive and lack explicit, slot-level control to preserve geometric consistency. GauMem resolves this dilemma by maintaining an explicit, strictly capacity-bounded 3D Gaussian memory governed by a learned write policy that directly optimizes primitive retention for long-horizon revisits.

Memory for video world models. While early world models captured environmental history through recurrent latent dynamics (Ha & Schmidhuber, 2018), scaling to complex and explorable environments has motivated diverse external memory paradigms. Implicit memory designs retain past context within transformer attention windows or recurrent feature states (Po et al., 2025); however, representations outside the context window are forgotten, and compressing past frames into fixed tokens degrades high-frequency details while lacking explicit geometric grounding to enforce strict projective consistency across camera trajectories (Wei et al., 2026). Retrieval-based approaches instead maintain historical 2D keyframes or latents, fetching relevant context via viewpoint overlap, geometric indexing, or prompt relevance (Xiao et al., 2025; Yu et al., 2025; Li et al., 2025a; Huang et al., 2025a; Ji et al., 2025); yet, their memory footprint scales continuously over time, leaving overlapping observations unconsolidated into a coherent 3D scene. Explicit 3D caches unproject observations into accumulated point clouds queried through rendering (Ren et al., 2025; Huang et al., 2025b; Wu et al., 2025); nevertheless, they rely on heuristic appending or fixed-weight fusion, causing projection errors to compound over time without slot-level capacity control. A closely related work (Chen et al., 2026) anchors the generator's own diffusion latents to Gaussian primitives and consolidates them with a fixed 2×2 downsample and a fixed-cell voxel merge, which bounds redundancy but not capacity: the map grows with the explored volume, does not explicitly formulate insertion and eviction under a global primitive budget, and its primitives carry latent features rather than color, so memory is coupled to the generator's latent representation. In contrast, GauMem bridges the gap between explicit 3D spatial fidelity and bounded computation: by training an end-to-end write policy on delayed revisits, GauMem selectively consolidates and evicts 3D Gaussians under a strict hard budget, outperforming hand-crafted retention rules under the same budget.

3 METHOD

We present **GauMem**, an explicit 3D Gaussian memory framework for video world models (VWMs) that learns what to retain under a strict capacity budget. Fig. 2 illustrates an overview of our GauMem framework. Given sequential video chunks, GauMem diagnoses novelty and geometric inconsistency through a Read-Before-Write (RBW) disagreement signal, softly fuses incoming evidence into

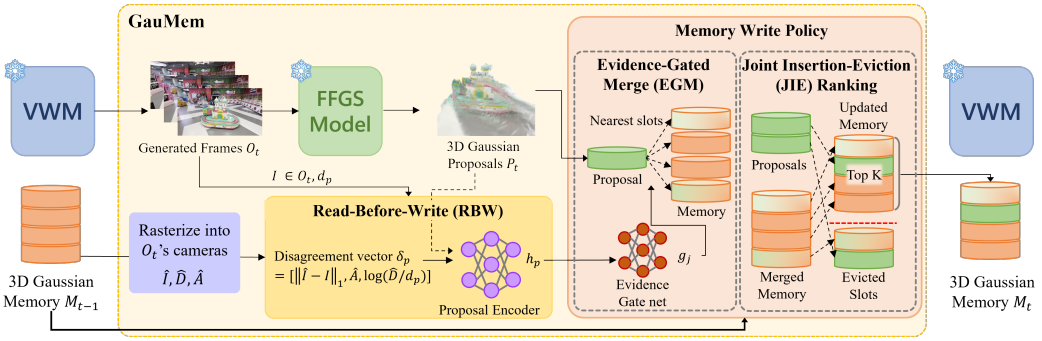

Figure 2: **Overview of the GauMem framework.** Given incoming frames O_t from a video world model (VWM), a frozen feed-forward Gaussian splatting (FFGS) frontend proposes candidate 3D Gaussians P_t. The Read-Before-Write (RBW) module renders the current memory M_{t-1} into the camera viewpoints of O_t to compute a per-primitive disagreement vector δ_p across photometric error, accumulated opacity, and depth mismatch. Then, the Evidence-Gated Merge (EGM) softly fuses each proposal into its four nearest stored memory slots and Joint Insertion–Eviction (JIE) ranking evaluates all stored slots and incoming proposals in a single unified pool, to select the top-K primitives under a strict budget to form M_t. The write policy is trained end-to-end via revisit supervision on ground-truth video streams to enforce long-term spatial consistency.

stored primitives via Evidence-Gated Merge (EGM), and enforces the capacity bound through Joint Insertion–Eviction (JIE) ranking in a unified candidate pool. This continuous write cycle maintains an explicit, capacity-bounded 3D representation that faithfully preserves visited scenes to provide coherent visual conditioning across extended camera trajectories.

3.1 Problem Formulation

Consider a video world model that sequentially predicts future visual observations along extended camera trajectories conditioned on historical context. To maintain long-term spatial consistency, the video world model interfaces with a persistent 3D memory through two operations: reading and writing. To *read*, the memory is rendered from queried camera viewpoints Π to provide visual conditioning for subsequent video rollouts. To *write*, incoming observations chunked at step t, denoted as $O_t = \{I\}$, are integrated into the memory state. A frozen feed-forward Gaussian splatting (FFGS) frontend first lifts O_t into a set of candidate Gaussian proposals $P_t = \{p\}$. Each proposal p carries 3DGS rendering parameters θ_p alongside its source pixel coordinates u_p and predicted depth d_p. The persistent memory M_t is strictly constrained to a capacity budget of at most K primitive slots ($|M_t| \leq K$). Each stored slot $j \in M_t$ occupies 42 B in total, comprising its rendering parameters θ_j, a compact latent feature f_j, a scalar priority logit π_j, an accumulated mass m_j, and an age counter a_j.

3.2 Read-Before-Write

To make grounded retention and merge decisions, the write policy must determine whether incoming observations introduce novel surfaces, redundant geometry, or conflicting viewpoints. GauMem achieves this through a Read-Before-Write (RBW) mechanism: before proposals P_t are absorbed, the existing memory M_{t-1} is rendered under stop-gradient into the camera viewpoints of the incoming chunk O_t. For each observed frame $I \in O_t$, this yields a rendered image $\hat{I}$, a depth map $\hat{D}$, and an accumulated opacity map $\hat{A} \in [0, 1]$. Sampling the rendered buffers at each proposal's source pixel u_p yields the disagreement vector:

$$\delta_p = \Big[\|\hat{I}(u_p) - I(u_p)\|_1,\ \hat{A}(u_p),\ \Delta d_p,\ \mathbb{I}_{\text{valid}} \Big], \tag{1}$$

where $\Delta d_p = \log\big(\hat{D}(u_p)/d_p\big)$ captures the relative depth mismatch, and $\mathbb{I}_{\text{valid}}$ is a binary flag indicating whether the memory is already populated. In addition, the policy receives a budget context b encoding current memory occupancy, incoming proposal load, and the normalized target capacity.

A lightweight MLP embeds each proposal from its rendering parameters θ_p, disagreement δ_p, and budget context b into a latent feature h_p. This design requires only that the frontend provides pixel-aligned proposals with depth estimates, keeping the write architecture modular.

3.3 Evidence-Gated Merge

When navigating continuous scenes, incoming chunks frequently re-observe surfaces already represented in memory. Directly appending these overlapping proposals causes severe primitive redundancy, whereas naive overwriting erases historical consistency. The Evidence-Gated Merge (EGM) softly consolidates incoming proposals into existing memory slots prior to discrete eviction, simultaneously deriving the structural signals required for downstream ranking.

Specifically, each proposal $p \in P_t$ identifies its four spatially nearest stored slots, predicting a soft association weight $w_{pj} \in (0, 1)$ for each candidate slot. Each slot $j \in M_{t-1}$ aggregates its total assigned evidence $e_j = \sum_p w_{pj}$, alongside the evidence-weighted proposal feature $\bar{h}_j$ and candidate parameters $\bar{\theta}_j$. We update the stored rendering parameters via a gated interpolation:

$$\theta_j \leftarrow (1 - g_j)\,\theta_j + g_j\,\bar{\theta}_j, \quad g_j = \sigma\Big(\log(e_j/m_j) + \phi_{\text{gate}}(f_j, \bar{h}_j, b)\Big), \tag{2}$$

followed by the mass update $m_j \leftarrow m_j + e_j$. Here, the gating network ϕ_{gate} is zero-initialized. Consequently, the gate initially defaults to a standard mass-weighted running average, $g_j = \sigma(\log(e_j/m_j)) = e_j/(e_j + m_j)$, providing a stable inductive bias from which the policy learns content-aware adjustments conditioned on the memory state f_j, incoming proposal feature $\bar{h}_j$, and budget context b. Finally, the slot feature f_j is updated with $\bar{h}_j$.

3.4 Joint Insertion–Eviction Ranking

Following EGM, the memory enforces the strict capacity constraint $|M_t| \le K$ via Joint Insertion–Eviction (JIE) ranking. Rather than decoupling primitive admission and pruning into disjoint heuristic stages, JIE casts memory update as a unified competition between stored slots and incoming proposals within a single candidate pool. This joint formulation enables salient novel geometry to immediately displace inconsistent primitives, while preventing unverified proposals from overwriting established structures.

For each stored slot $j \in M_{t-1}$, a scoring head predicts an increment Δ_j conditioned on its merged feature f_j, age a_j, mass m_j, assigned evidence e_j, and budget context b. Its priority updates via leaky accumulation, $\pi_j \leftarrow 0.9\,\pi_j + \Delta_j$, defining the slot retention score $s_j = \pi_j$. Concurrently, for each incoming proposal $p \in P_t$, an insertion score s_p is predicted from its embedding h_p and disagreement vector δ_p, and subsequently penalized by its total merged evidence ($s_p \leftarrow s_p - \sum_j w_{pj}$) to suppress the re-admission of content already consolidated into stored slots. To enable fair competition across both sets, scores are zero-centered by group means and bounded with a positive offset, preserving relative rankings while maintaining stable visibility gating (Eq. 5). The capacity-bounded memory is then formed by selecting the top-K primitives:

$$M_t = \text{TopK}_K\Big(\{s_j\}_{j \in M_{t-1}} \cup \{s_p\}_{p \in P_t}\Big). \tag{3}$$

Retained slots increment their age, while newly admitted proposals instantiate fresh slots with initial age $a_p = 0$ and mass $m_p = 1$.

Cold-start heuristic at evaluation. At the initial step, where the memory is not yet populated ($\mathbb{I}_{\text{valid}} = 0$), we initialize the first K slots via uniform random subsampling from the initial incoming proposals to ensure balanced spatial coverage before learned relative scoring takes over.

3.5 Revisit-Supervised Training

Training objective and curriculum. Let $R(M, \Pi)$ denote the differentiable rasterization of memory M into camera viewpoints Π. Every supervised rendering is evaluated with a composite photometric and perceptual metric, $\ell(I, \hat{I}) = \text{MSE}(I, \hat{I}) + 0.1\,\text{LPIPS}(I, \hat{I})$ (Zhang et al., 2018). To

balance long-term retention with immediate reconstruction fidelity, the write policy is optimized via:

$$\mathcal{L} = 2 \sum_{t \in \mathcal{T}_{\text{rev}}} \ell\big(R(M_t, \Pi_{\text{rev}}),\, I_{\text{rev}}\big) + \mathcal{L}_{\text{aux}}, \tag{4}$$

where the revisit term supervises held-out targets I_{rev} observed at earlier timestamps $\tau < t$ (excluded from intervening writes) to penalize catastrophic forgetting. Complementarily, the auxiliary term $\mathcal{L}_{\text{aux}}$ supervises the reconstruction of the latest chunk, including held-out frames, compelling the policy to absorb novel surfaces from current observations.

Learning priorities via visibility gating. The discrete retention decision in Eq. 3 applies a hard TopK_K selection, introducing a non-differentiable bottleneck in the write pipeline. To train the scoring network end-to-end without resorting to high-variance gradient estimators, the priority score serves a dual role: during forward rendering, each retained slot $j \in M_t$ is modulated by an effective opacity:

$$\alpha_j = \sigma(o_j)\, \sigma(\pi_j), \tag{5}$$

where o_j is the stored opacity logit in θ_j and π_j is the priority score established by the ranking in Eq. 3. Subsequent rendering losses treat $\sigma(\pi_j) \in (0, 1)$ as a continuous visibility gate, allowing supervisory gradients to backpropagate directly into the scoring heads and proposal embeddings. During inference, GauMem reuses these learned priority scores directly for competitive Top-K truncation. This formulation acts as a principled surrogate objective: while evicted primitives receive no rendering gradients, the policy inherently learns to assign higher priorities to primitives critical for future visual fidelity.

4 Experiments

4.1 Experimental Setup

Evaluation protocol. We design an evaluation protocol that directly tests the key capability of a persistent video world model: long-horizon retention upon revisit. Specifically, the initial region $\mathcal{R}_0$ (12 views) is written first into the memory. Next, a stream of N subsequent regions $\{\mathcal{R}_i\}_{i=1}^{N}$ from the same scene (each containing 12 views) is sequentially written into the memory under the fixed budget constraint. Finally, the camera loops back to the initial space along a return trajectory, denoted as $\mathcal{R}'_0$, where the model is evaluated on held-out intermediate frames between the written views of $\mathcal{R}_0$ that were never exposed to the memory during writing. Crucially, no frame of $\mathcal{R}_0$ is shown to the memory again before evaluation. Because the intervening regions belong to the same 3D scene, their visual frustums partially overlap with $\mathcal{R}_0$, actively competing for stored slots. We evaluate scenarios across $N \in \{4, 10\}$ intervening regions. Unlike protocols permitting target re-observation, our setting strictly evaluates long-term retention instead of recency caching.

Implementation details. We evaluate on 140 test scenes from the DL3DV benchmark (Ling et al., 2024) at a resolution of 252×448; the write policy is trained exclusively on the official training split. We employ StructSplat (Zhao et al., 2026) as the frozen FFGS frontend, which yields 913,752 proposals per 12-view chunk. Prior to writing, each chunk is locally refined for 150 steps using the Adam optimizer (Kingma & Ba, 2014) against its own incoming views, identically across all evaluated policies without accessing stored frames. The memory capacity is capped at $K = 655{,}360$ slots (27.5 MB), where the incoming stream delivers approximately $15\times$ more primitives than the memory budget over ten regions.

Baselines. Every baseline consumes identical 3D Gaussian proposals to isolate write policy effects. As an unconstrained baseline, *Keep-all* stores all proposals without enforcing a budget. Also, we benchmark five alternative retention strategies: (i) *Recency eviction (LRU)*, which strictly evicts the oldest primitives; (ii) *Uniform random eviction*, which removes primitives uniformly at random; (iii) *Contribution pruning*, which greedily retains primitives with the highest opacity; (iv) *Tuned surfel fusion*, merging proposals sharing spatial-normal bins with stored surfels; and (v) *Downsample + voxel fusion* (Chen et al., 2026), which downsamples proposals on a 2×2 grid before spatial voxel pooling.

Table 1: **Quantitative comparison of write rules on the DL3DV test set (140 scenes)**. Under bounded capacity ($K = 655$k), our GauMem outperforms all baselines. The slot column gives the mean number of Gaussians stored at the return (N=4 / N=10 where they differ). ↑ (↓) indicates that higher (lower) is better.

		N=4 intervening regions			N=10 intervening regions		
Write policy	Slots	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Keep-all (unbounded)	4.6M / 10.1M	15.11	0.501	0.481	13.85	0.421	0.590
Recency eviction (LRU)	655k	11.09	0.328	0.560	11.07	0.330	0.554
Uniform random eviction	655k	14.61	0.425	0.489	13.63	0.400	0.493
Contribution pruning	655k	15.03	0.462	0.435	13.87	0.374	0.497
Tuned surfel fusion	593k / 623k	14.86	0.358	0.540	14.05	0.305	0.582
Downsample + voxel fusion	1.1M / 2.2M	16.92	0.545	0.414	15.54	0.475	0.497
GauMem (Ours)	655k	**19.57**	**0.707**	**0.242**	**19.01**	**0.668**	**0.276**



Figure 3: **Qualitative comparison on the revisit trajectory $\mathcal{R}'_0$ after ten intervening regions of the same scenes** ($N = 10$). GauMem suppresses floater clutter and preserves fine architectural textures where other baselines collapse.

4.2 Comparison with Write Rules

Table 1 compares quantitative results on write rules for the DL3DV test set. As shown, our GauMem reliably preserves held-out revisit views of the initial region ($\mathcal{R}'_0$) across extended exploration, outperforming all training-free heuristics by a substantial margin under both moderate and long interference horizons. While tuned surfel fusion provides a competitive baseline, its rigid spatial-normal binning is acutely fragile to depth noise, causing retention quality to deteriorate. Furthermore, unconstrained *Keep-all* retains every proposal without budget limits—accumulating substantially more primitives than our budget—yet falls considerably behind GauMem in rendering fidelity. Without active consolidation, unbounded accumulation simply stacks noisy floaters across overlapping views, whereas GauMem selectively filters this clutter. Similarly, while downsampled voxel fusion achieves stronger results among baselines, it still lags far behind GauMem while consuming more storage (1.1M / 2.2M slots). Because its write rule operates without a read-time geometric discrepancy check, it fails to preserve fine structural details. Finally, recency eviction (LRU) suffers catastrophic forgetting upon return, collapsing almost entirely on extended sequences. Because LRU

Table 2: **Comparison of memory representations read through a video generator** on DL3DV test set (140 scenes, $N{=}4$). Memories condition a frozen Wan2.1-VACE-1.3B along the 17-frame return path of $\mathcal{R}_0'$. Geometric metrics (rot°, transl°, ATE %, and AbsRel) are re-estimated from the generated video alone via an independent reconstruction model (Lin et al., 2025) without memory access (real frames achieve $0.14°$ rotation error).

Memory	PSNR↑	SSIM↑	LPIPS↓	rot° ↓	transl° ↓	ATE %↓	AbsRel↓
Point cloud (position and color, no orientation)							
Keep-all (Ren et al., 2025)	15.66	0.361	0.590	0.90	2.24	1.44	0.312
Fixed-cell fusion (Wu et al., 2025)	14.68	0.281	0.683	4.20	13.21	8.78	0.374
Budget-matched fusion	13.78	0.273	0.717	5.18	18.34	11.25	0.408
Frames							
Top-4 frame retrieval (Xiao et al., 2025; Li et al., 2025a)	17.24	0.573	0.341	1.44	4.32	3.04	0.157
In-context frames (Yu et al., 2025)	15.52	0.453	**0.308**	2.73	9.58	8.34	0.132
Gaussian primitives							
Keep-all (Ren et al., 2025)	15.01	0.493	0.523	1.01	2.32	1.58	0.305
Tuned surfel fusion (Wu et al., 2025)	15.02	0.374	0.576	1.30	3.04	2.06	0.314
Downsample + voxel fusion (Chen et al., 2026)	16.80	0.530	0.466	0.31	0.94	0.53	0.226
GauMem	**19.36**	**0.656**	0.311	**0.21**	**0.77**	**0.40**	**0.121**

strictly favors the most recently observed content, the long stream of intervening regions evicts $\mathcal{R}_0$ completely before the camera revisits it.

The qualitative comparisons in Fig. 3 highlight the underlying geometric fidelity of each representation when rendering the memory upon revisit. Unconstrained accumulation (Keep-all) suffers from severe floater accumulation, creating a hazy, ghosting veil that obscures background structures. Tuned surfel fusion mitigates primitive explosion, but its rigid voxel binning struggles with camera pose noise and depth discrepancies, resulting in noisy, broken surfaces and missing architectural elements. In contrast, GauMem cleanly resolves geometric conflicts and consolidates redundant proposals into sharp, persistent surfaces.

4.3 Comparison of Memory Representations in Video World Models

A video world model decodes stored memory through a generative backbone, which can either absorb or amplify discrepancies between memory representations. To isolate memory representation from generative capacity, all formats are evaluated through an identical, frozen video reader (Wan2.1-VACE-1.3B (Wan et al., 2025; Jiang et al., 2025)) in a zero-shot setting. We benchmark three distinct paradigms: (1) *point clouds* (position and color without orientation or scale); (2) *frames*, via retrieval (Xiao et al., 2025; Li et al., 2025a) or context-window conditioning (Yu et al., 2025); and (3) *3D Gaussians*, comparing unconstrained accumulation (Ren et al., 2025), surfel fusion (Wu et al., 2025), downsampled voxel fusion (Chen et al., 2026), and GauMem. Each memory is rendered along the 17-frame return trajectory of $\mathcal{R}_0$ (where odd-indexed frames were never written, testing pure spatial generalization) to provide visual conditioning, with uncovered pixels (accumulated opacity below 0.5) inpainted by the reader.

To assess camera-path consistency, we re-estimate camera poses from each generated 17-frame RGB clip using Depth Anything 3 (Lin et al., 2025) and evaluate them against the commanded trajectory. We report relative rotation error (rot°) over all $\binom{17}{2} = 136$ unordered frame pairs, translation-direction error (transl°) over the same pairs excluding degenerate baselines, and camera-center RMSE after a single global $\mathrm{Sim}(3)$ alignment, normalized by the path extent and reported as ATE (%). Depth consistency is measured via median-scaled AbsRel against real-frame estimates on the eight never-written views.

As shown in Table 2, GauMem attains the best PSNR and SSIM across every representation and is level with in-context frames on LPIPS (0.311 against 0.308), despite operating under a strictly bounded 27.5 MB footprint. Examining geometry, trajectory drift concentrates heavily in fused point clouds and in-context frames, whereas unfused point clouds (*Keep-all*) and frame retrieval incur moderate drift. Among Gaussian baselines, downsampled voxel fusion achieves competitive camera stability (0.53% ATE and $0.31°$ rotation error); however, its photometric fidelity (16.80 dB PSNR, 0.466 LPIPS) and depth accuracy (AbsRel $= 0.226$) drop markedly due to aggressive 2×2

Table 3: **Ablation study of GauMem components and training objective.** Each variant is trained from scratch under the fixed budget and evaluated on held-out revisit views ($\mathcal{R}'_0$) across 140 DL3DV scenes at $N{=}4$ and $N{=}10$ horizons. This isolates the contribution of each core component—RBW, EGM, JIE, and revisit supervision.

	$N{=}4$ intervening regions			$N{=}10$ intervening regions		
Methods	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
without RBW	17.10	0.602	0.330	16.62	0.590	0.345
without EGM	19.20	0.679	0.262	17.74	0.603	0.321
without JIE	14.30	0.422	0.497	13.16	0.391	0.506
without Revisit supervision	11.28	0.341	0.850	11.44	0.355	0.852
GauMem	**19.57**	**0.707**	**0.242**	**19.01**	**0.668**	**0.276**

proposal discarding and unguided spatial pooling. In contrast, GauMem reliably suppresses camera drift across the entire clip, achieving a rotation error of 0.21° (closely tracking the 0.14° real-frame baseline) and an ATE of 0.40%, while leading in visual fidelity. Furthermore, GauMem achieves the lowest depth error (AbsRel $= 0.121$), outperforming even in-context frames (0.132) despite the latter benefiting from having real frames embedded directly inside its rollout clip. These results support the effectiveness of the proposed Gaussian-memory write policy in preserving view-aligned conditioning for downstream generation.

4.4 Ablation Study

To validate each design choice, Table 3 ablates individual modules of GauMem alongside the revisit training objective across $N{=}4$ and $N{=}10$ intervening regions on DL3DV test split. First, omitting **RBW** incurs a consistent penalty of ≈2.4–2.5 dB across both horizons, confirming that write-time self-rendering provides essential feedback to diagnose novelty and geometric conflicts. Next, excluding **EGM** exhibits a widening gap as interference accumulates (falling by 1.27 dB at $N{=}10$): without soft evidence-gated merging, redundant proposals continuously compete for bounded capacity, cluttering memory and prematurely evicting vital geometry. Replacing **JIE** with uniform random eviction causes catastrophic degradation, dropping PSNR by 5.27–5.85 dB (down to 13.16 dB) and demonstrating the necessity of coordinating stored slots and incoming proposals in a unified priority pool. Most crucially, training without **Revisit supervision** (optimizing solely for immediate reconstruction of current chunks) leads to complete memory collapse, falling to ≈11.3 dB with severe perceptual degradation (0.85 LPIPS). This highlights that without explicit revisit feedback to signal future utility, the network defaults to short-sighted updates that purge previously visited geometry. These results verify that maintaining robust long-horizon retention requires the full integration of all core components.

5 Conclusion

We presented **GauMem**, an explicit and capacity-bounded 3D Gaussian memory framework for video world models with a write policy learned end-to-end. As interactive agents explore unbounded environments, persistent world models must operate under finite storage, necessitating principled policies to decide what content to retain, consolidate, or discard. GauMem addresses this challenge through an a closed-loop write cycle: it renders existing memory prior to writing to diagnose novelty and physical conflict via Read-Before-Write disagreement, softly fuses re-observed evidence into stored slots through Evidence-Gated Merge, and coordinates candidate selection within a single unified pool via Joint Insertion–Eviction ranking. Trained end-to-end through revisit supervision, GauMem learns to resist catastrophic forgetting without relying on fragile heuristic thresholds. Evaluations on DL3DV demonstrate that GauMem outperforms both spatial fusion heuristics and unbounded accumulation across moderate and extended horizons. When integrated with a frozen video reader, GauMem acts as a reliable geometric anchor, eliminating severe trajectory drift and enforcing 3D geometric consistency in generated videos.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing. The authors reviewed and verified all AI-assisted material and take full responsibility for the final content and results.

ETHICS STATEMENT

This work introduces an explicit 3D Gaussian memory consolidation framework for video world models. All empirical evaluations are conducted using publicly available benchmark datasets (DL3DV, 3RScan, RealEstate10K), strictly adhering to their intended research licenses and usage policies; no human subjects, sensitive personal data, or crowdsourced annotations were involved.

While advances in high-fidelity video generation carry inherent dual-use risks regarding synthetic media and misinformation, our research focuses strictly on bounded memory management, geometric consistency, and slot consolidation rather than unconstrained video synthesis. By enabling scalable, long-horizon 3D scene persistence under tight memory constraints, we believe this work primarily contributes to beneficial downstream applications, including embodied AI, robotics, and persistent spatial navigation.

REPRODUCIBILITY STATEMENT

To facilitate reproduction, our core methodology and theoretical formulations are detailed in Sec. 3 and Appendix. Comprehensive training hyperparameters—including optimizer schedules, loss formulations, and batch specifications—as well as hardware runtime configurations are documented. Furthermore, the evaluation protocols are elaborated in Sec. 4. Our source code, along with trained model checkpoints, will be made publicly available upon acceptance.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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.

Minh-Quan Viet Bui, Jaeho Moon, and Munchurl Kim. Airsplat: Alignment and rating for robust feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2603.25129*, 2026.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467. IEEE, 2024.

Hanlin Chen, Jiaxin Wei, Xibin Song, Yifu Wang, Steve Wang, Hongdong Li, Pan Ji, and Gim Hee Lee. Robust dreamer: Deviation-aware latent gaussian memory for action-controlled ar video generation. *arXiv preprint arXiv:2605.30855*, 2026.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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

Guichen Huang, Ruoyu Wang, Xiangjun Gao, Che Sun, Yuwei Wu, Shenghua Gao, and Yunde Jia. Longsplat: Online generalizable 3d gaussian splatting from long sequence images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4994–5002, 2026.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025b.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Gyeongjin Kang, Seungkwon Yang, Seungtae Nam, Younggeun Lee, Jungwoo Kim, and Eunbyung Park. Multi-view pyramid transformer: Look coarser to see broader. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37380–37390, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yushi LAN, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STream3r: Scalable sequential 3d reconstruction with causal transformer. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RTTYGeC2Io.

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, 2025a.

Yang Li, Jinglu Wang, Lei Chu, Xiao Li, Shiu-hong Kao, Ying-Cong Chen, and Yan Lu. Streamgs: Online generalizable gaussian splatting reconstruction for unposed image streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25841–25850. IEEE, 2025b.

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.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8733–8744. IEEE, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7657–7666. IEEE, 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chen Wang, Hao Tan, Wang Yifan, Zhiqin Chen, Yuheng Liu, Kalyan Sunkavalli, Sai Bi, Lingjie Liu, and Yiwei Hu. tttlrm: Test-time training for long context and autoregressive 3d reconstruction. *arXiv preprint arXiv:2602.20160*, 2026a.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

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*, 2026b.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, et al. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=HbTxc6U1fO.

Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=c6CAVKlKmU.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463. IEEE, 2025.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

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. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Jia-Chen Zhao, Beiqi Chen, Xinyang Chen, Guangcong Wang, and Liqiang Nie. Structsplat: Generalizable 3d gaussian splatting from uncalibrated sparse views. *arXiv preprint arXiv:2606.28321*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4349–4359. IEEE, 2025.

APPENDIX

This appendix provides supplementary specifications and extended evaluations. Specifically, we detail our implementation setup and training protocols (Sec. A), report additional quantitative analyses across alternative memory budgets and traversal regions (Sec. B and C), and evaluate zero-shot generalization across diverse frontends and datasets (Sec. D and E). Finally, we discuss model limitations (Sec. F) alongside further quantitative results and qualitative visual comparisons (Sec. G). In addition, we provide accompanying videos in the supplementary material featuring side-by-side comparisons of memory renderings.

A IMPLEMENTATION DETAILS

Memory primitive representation and byte budget. To enforce strict capacity constraints, each primitive slot in GauMem is compactly serialized to exactly 42 B, compared to 28 B for standard heuristic baselines. The 28 B baseline representation stores core 3DGS rendering parameters θ in 16-bit half precision (`fp16`): 3D position (3×2 B = 6 B), rotation quaternion (4×2 B = 8 B), 3D scale (3×2 B = 6 B), opacity logit (1×2 B = 2 B), and diffuse RGB color (3×2 B = 6 B). The additional 14 B in each learned GauMem slot comprises its consolidation metadata: an 8-dimensional compact latent feature f_j (8×1 B `fp8` = 8 B), a scalar priority logit π_j (1×2 B `fp16` = 2 B), an accumulated evidence mass m_j (1×2 B `fp16` = 2 B), and an integer age counter a_j (1×2 B `uint16` = 2 B). At the evaluation budget of $K = 655{,}360$ slots, GauMem occupies strictly $655{,}360 \times 42$ B ≈ 27.5 MB.

Architecture and feature encoding. The learned write policy consists of lightweight multi-layer perceptrons (MLPs) with GELU activations, totaling 0.67 M parameters across all sub-modules. In the Read-Before-Write (RBW) vector δ_p, semi-transparent background primitives often elevate rendered opacity $\hat{A}$ even across unvisited free space; the learned MLP inherently learns to place primary reliance on the signed depth discrepancy $\Delta d_p = \log(\hat{D}/d_p)$ and photometric error $\|\hat{I} - I\|_1$ to discriminate true surface occlusions from redundant coverage. The evidence-gating network ϕ_{gate} uses 2-layer MLPs, while the scoring heads for JIE predict Δ_j and s_p via linear projections over pooled latent states.

Training protocol, curriculum, and hardware. We optimize the write policy with AdamW ($\beta = (0.9, 0.98)$, $\epsilon = 10^{-6}$, weight decay 10^{-4}), gradient clipping at 1.0, and an exponential moving average (EMA) with a decay factor of 0.999. Training runs for 48,000 steps with a learning rate of 5×10^{-5} and a batch size of 4 streams across two NVIDIA B200 GPUs (183 GB each, using $\approx$140 GB per GPU) with two accumulation steps, taking $\approx$85 h. Each training stream is sampled from continuous trajectories in the DL3DV training set at 252×448, consisting of consecutive 12-view chunks with natural camera loop closures and spatial overlaps, where revisit losses supervise held-out views from previously visited chunks ($\tau < t$) across varying temporal horizons (spanning 4 to 10 intervening chunks). Gradients propagate through unrolled multi-chunk writes via truncated backpropagation through time (TBPTT).

Video world model decoding and downstream evaluation. Downstream video rollouts are synthesized by feeding rendered memory guidance to a frozen Wan2.1-VACE-1.3B video diffusion reader (Wan et al., 2025; Jiang et al., 2025). For each query view along the return trajectory, the memory is rasterized to produce conditioning RGB frames and opacity maps. Pixels with accumulated opacity below 0.5 are marked as missing in a binary conditioning mask, instructing the reader to generatively inpaint occluded or unobserved regions. Generation uses 15 denoising steps with shared noise seeds and prompts across all baselines to ensure strictly paired comparisons.

Inference latency. At inference, processing and writing a single 12-view chunk takes $\approx$20 s end to end. The learned write policy (RBW rendering, EGM merging, and JIE ranking) accounts for $\approx$8 s, while the remaining $\approx$12 s is consumed by the feed-forward frontend forward pass, pose alignment, and a 150-step local primitive refinement shared across all baselines. Memory evaluation requires $\approx$80 GB of GPU memory, while the video reader synthesizes a 17-frame clip at 480×832 in $\approx$10 s using $\approx$22 GB of VRAM.

Table 4: **Comparison across memory budgets and write policies** on held-out return views of $\mathcal{R}'_0$ across the DL3DV test set (140 scenes). Storage footprints reflect return-time memory (42 B/slot for GauMem vs. 28 B for heuristics).

		N=4 intervening regions			N=10 intervening regions		
Write policy	memory (MB)	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Tuned surfel fusion	16.5 / 17.5	14.86	0.358	0.540	14.05	0.305	0.582
Tuned surfel at byte parity	24.6 / 25.6	14.95	0.382	0.536	13.54	0.320	0.599
Keep-all	128 / 281	15.11	0.501	0.481	13.85	0.421	0.590
Fixed-cell fusion	14.9 / 25.5	13.94	0.428	0.789	13.42	0.422	0.792
Opacity-density pruning	18.4	14.95	0.459	0.441	13.76	0.380	0.497
Redundancy culling	99 / 183	15.42	0.529	0.464	13.42	0.441	0.590
Downsample + voxel fusion	29.7 / 62.1	16.92	0.545	0.414	15.54	0.475	0.497
GauMem	27.5	**19.57**	**0.707**	**0.242**	**19.01**	**0.668**	**0.276**

B Evaluation Across Memory Budgets and Alternative Write Policies

Since each learned slot in GauMem occupies 42 B against 28 B for a heuristic primitive, we test whether our gains come from the write policy rather than raw storage capacity (Table 4). Alongside *Keep-all* and *Tuned surfel fusion*, we benchmark five alternative retention paradigms adapted to our streaming setup. Each is re-implemented on our primitives and consumes identical proposals; citations indicate the foundational methodology each row adapts rather than the end-to-end original framework:

- *Tuned surfel at byte parity:* Allocates 983,040 slots ($1.5\times$ our primitive budget) to match GauMem's maximum storage budget of 27.5 MB.
- *Fixed-cell fusion:* Inspired by voxel-wise fusion in Wu et al. (2025), this baseline pools Gaussian attributes within a growing voxel grid using fixed-weight running averages, rather than integrating TSDF values as in the original method. We sweep cell sizes from 0.04 to 0.10 times the mean inter-camera distance.
- *Opacity-density pruning:* Motivated by reducing redundant history (Wei et al., 2026), we construct a heuristic that ranks primitives by opacity scaled by inverse local spatial density. This score is specific to our Gaussian setting and does not implement the original GP-based mutual-information criterion.
- *Redundancy culling:* Inspired by the ID-based removal and overlap-derived supervision in LongSplat (Huang et al., 2026), this baseline removes stored primitives identified as redundant through rendered Gaussian IDs and spatial overlap, then appends incoming proposals. Our implementation uses world-axis-aligned bounding boxes and directly applies overlap-based decisions across swept thresholds, rather than using the original learned mask predictor. No hard capacity limit is imposed.
- *Downsample + voxel fusion:* Inspired by the reduced-resolution Gaussian prediction and voxel fusion in Robust Dreamer (Chen et al., 2026), this baseline subsamples the supplied proposals on a 2×2 pixel grid before pooling primitives within spatial voxels. We replace the original learned confidence weights with opacity-weighted pooling or distance-penalized softmax weights. We sweep voxel sizes to obtain storage footprints both above and below GauMem's budget; no hard capacity limit is imposed.

For baselines without a hard capacity limit, we report the measured storage footprint alongside reconstruction quality.

As reported in Table 4, GauMem consistently outperforms all alternative paradigms across both moderate and extended exploration horizons. Granting 50% more primitive slots to surfel fusion under byte parity yields marginal improvement ($+0.09$ dB at N=4), confirming that raw slot count is not the limiting bottleneck. Among unbounded baselines, the two that never merge what they keep (*Keep-all* and *Redundancy culling*) incur massive memory expansion—retaining substantially

Table 5: **Quantitative evaluation across held-out revisit views of all regions** ($\mathcal{R}'_0 \dots \mathcal{R}'_N$) on the DL3DV test set (140 scenes). Mirroring the delayed revisit protocol of $\mathcal{R}'_0$, each traversed region $\mathcal{R}_i$ is queried on held-out revisit views $\mathcal{R}'_i$ that were excluded during memory writes.

		N=4 intervening regions			N=10 intervening regions		
Write policy	Memory (MB)	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Keep-all	128 / 281	14.90	0.497	0.504	13.76	0.418	0.611
Tuned surfel fusion	16.5 / 17.5	14.58	0.354	0.550	13.94	0.305	0.584
Downsample + voxel fusion	29.7 / 62.1	16.85	0.544	0.422	15.56	0.476	0.506
GauMem	27.5	**19.51**	**0.702**	**0.256**	**19.12**	**0.689**	**0.288**

more primitives—while stacking severe floating clutter across overlapping viewpoints. Most notably, against a concurrent Gaussian-memory baseline adapted to our explicit 3DGS setting—the downsampled voxel fusion adapted from Robust Dreamer (Chen et al., 2026)—GauMem maintains a substantial lead, while enforcing a strictly smaller and bounded storage budget (27.5 MB vs. up to 62.1 MB). This performance gap reflects distinct architectural trade-offs: its write rule discards three quarters of the proposals on a fixed 2×2 grid before any consolidation, so content absent from the retained quarter cannot be recovered, and it operates without a write-time signal to diagnose geometric discrepancy or surface conflict before merging into spatial bins. In contrast, GauMem couples write-time discrepancy diagnosis with evidence-weighted consolidation and joint ranking, preserving persistent 3D geometry under a strictly bounded capacity.

C Global Scene Retention Across All Visited Regions

Table 5 extends this revisit protocol across *all* traversed regions. Mirroring the design of $\mathcal{R}'_0$, each region $\mathcal{R}_i$ ($i \in \{0, \dots, N\}$) is evaluated on held-out revisit trajectories $\mathcal{R}'_i$ consisting of intermediate views that were never shown to the memory during write steps. Evaluating the final consolidated memory M_N across $\{\mathcal{R}'_i\}_{i=0}^{N}$ verifies whether the memory genuinely preserves a globally consistent representation of the entire explored environment.

As shown in Table 5, GauMem reliably maintains high visual fidelity across all traversed regions under both moderate (N=4) and extended (N=10) horizons. Notably, its multi-region performance closely mirrors its recall on $\mathcal{R}'_0$ alone, exhibiting minimal degradation even when distributing its strict 27.5 MB budget across 11 sequential regions. In contrast, unbounded accumulation (Keep-all) suffers from compounding floater clutter despite substantial storage growth, while tuned surfel fusion deteriorates due to rigid voxel overwriting. Although downsampled voxel fusion (Chen et al., 2026) bounds primitive explosion, its unguided pooling yields noticeably lower visual fidelity while consuming more than twice our storage budget at $N = 10$. These comparisons confirm that GauMem does not adopt an initial-freezing shortcut, but dynamically balances primitive retention to preserve a globally consistent representation of the explored scene.

D Generalization Across Diverse Frontends and Resolutions

Because the write policy interfaces strictly through proposal rendering attributes θ_p, the Read-Before-Write disagreement vector δ_p, and the budget context b, it remains decoupled from internal frontend feature representations. Consequently, GauMem functions as a modular memory framework that can readily interface with distinct feed-forward proposal generators. In Table 6, we evaluate this flexibility across two representative frontends, Depth Anything 3 (Lin et al., 2025) and DepthSplat (Xu et al., 2025), spanning varied proposal formulations, image resolutions, and operating capacity budgets.

Evaluation on Depth Anything 3. When paired with proposals emitted by Depth Anything 3 GIANT model (Lin et al., 2025) at 252×448, GauMem consistently outperforms other baselines. This closely parallels the advantages achieved with StructSplat under the identical 164k-slot budget, verifying that our revisit-supervised write policy reliably optimizes long-term retention when driven by vision foundation model backbones.

Table 6: **Generalization across diverse frontends and resolutions** evaluated on held-out revisit views of $\mathcal{R}'_0$ across the DL3DV test set (140 scenes). We evaluate GauMem on StructSplat (Zhao et al., 2026), Depth Anything 3 (Lin et al., 2025), and DepthSplat (Xu et al., 2025) across different resolutions (252×448 and 512×960) and operating slot budgets.

				N=4 intervening regions			N=10 intervening regions		
Frontend	Write rule	Resolution	Slots	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
StructSplat	Keep-all	252×448	1.2M/2.5M	14.37	0.460	0.635	12.82	0.411	0.739
	Tuned surfel fusion	252×448	156k/161k	14.64	0.383	0.606	13.43	0.356	0.645
	GauMem	252×448	164k	**17.64**	**0.606**	**0.394**	**17.12**	**0.564**	**0.442**
Depth Anything 3	Keep-all	252×448	423k/931k	16.79	0.513	0.603	14.73	0.460	0.693
	Tuned surfel fusion	252×448	147k/162k	16.17	0.464	0.658	14.45	0.422	0.728
	GauMem	252×448	164k	**18.75**	**0.659**	**0.357**	**18.15**	**0.654**	**0.385**
DepthSplat	Keep-all	512×960	1.8M/4.1M	15.35	0.576	0.478	13.68	0.504	0.593
	Tuned surfel fusion	512×960	240k/246k	13.64	0.376	0.598	12.93	0.360	0.634
	GauMem	512×960	262k	**19.23**	**0.709**	**0.316**	**18.77**	**0.678**	**0.356**

Table 7: **Generalization across diverse scene datasets** evaluated on held-out revisit views of $\mathcal{R}'_0$. The write policy, trained solely on DL3DV, is evaluated zero-shot on 3RScan (62 scenes) (Wald et al., 2019) and RealEstate10K (140 scenes) (Zhou et al., 2018).

		N=4 intervening regions			N=10 intervening regions		
Dataset	Write rule	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3RScan	Keep-all	16.68	0.722	0.511	14.63	0.690	0.579
	Tuned surfel fusion	15.76	0.649	0.633	14.00	0.635	0.667
	GauMem	**18.26**	**0.765**	**0.425**	**17.48**	**0.701**	**0.457**
RealEstate10K	Keep-all	17.57	0.621	0.371	16.26	0.578	0.424
	Tuned surfel fusion	16.43	0.489	0.513	15.34	0.444	0.557
	GauMem	**18.58**	**0.658**	**0.354**	**17.91**	**0.627**	**0.384**

Evaluation on DepthSplat. To examine scalability across higher resolutions and differing multi-view proposal distributions, we evaluate GauMem on DepthSplat (Xu et al., 2025) at 512×960. Here, we directly deploy the write policy checkpoint trained on StructSplat without per-architecture fine-tuning. Even under this increased resolution and domain shift, GauMem maintains a solid lead over tuned surfel fusion and substantially suppresses floater accumulation compared to unbounded Keep-all. This demonstrates that the learned spatial consolidation dynamics generalize effectively out of the box to higher-resolution proposal streams.

E Generalization Across Datasets

To determine whether the learned write policy overfits to DL3DV scene statistics or generalizes across distinct spatial distributions, we evaluate the fixed DL3DV checkpoint zero-shot on two challenging external benchmarks: 3RScan (Wald et al., 2019) and RealEstate10K (Zhou et al., 2018) (Table 7). Across both indoor RGB-D scans exhibiting dense household clutter (3RScan) and unconstrained real-estate camera trajectories (RealEstate10K), GauMem consistently outperforms all baseline strategies across both short ($N = 4$) and extended ($N = 10$) interference horizons. While tuned surfel fusion deteriorates due to domain-specific depth noise and unconstrained accumulation suffers from severe floater stacking across rapid camera pans, GauMem dynamically filters transient redundancy and consolidates persistent geometry. These results confirm that GauMem internalizes general, environment-agnostic consolidation rules rather than dataset-specific visual priors.

Table 8: **Comparison of memory representations read through a video generator** on DL3DV test set (140 scenes, N=10). Memories condition a frozen Wan2.1-VACE-1.3B along the 17-frame return path of $\mathcal{R}'_0$. Geometric metrics (rot°, transl°, ATE %, and AbsRel) are re-estimated from generated clips alone via an independent reconstruction model (Lin et al., 2025) without memory access (real frames achieve $0.14°$ rotation error).

Memory	PSNR↑	SSIM↑	LPIPS↓	rot° ↓	transl° ↓	ATE %↓	AbsRel↓
Point cloud (position and color, no orientation)							
Keep-all (Ren et al., 2025)	14.45	0.284	0.652	3.25	6.04	4.34	0.356
Fixed-cell fusion (Wu et al., 2025)	14.49	0.270	0.697	5.25	17.57	11.46	0.411
Budget-matched fusion	13.26	0.248	0.740	5.69	23.57	14.99	0.434
Frames							
Top-4 frame retrieval (Xiao et al., 2025; Li et al., 2025a)	16.51	0.535	0.370	1.13	5.02	3.75	0.177
In-context frames (Yu et al., 2025)	15.52	0.451	**0.305**	2.64	10.06	9.20	0.146
Gaussian primitives							
Keep-all (Ren et al., 2025)	13.79	0.422	0.625	3.46	8.35	5.85	0.366
Tuned surfel fusion (Wu et al., 2025)	14.26	0.324	0.615	2.66	6.88	5.14	0.361
Downsample + voxel fusion (Chen et al., 2026)	15.48	0.472	0.542	0.92	2.20	1.39	0.287
GauMem	**18.83**	**0.622**	0.313	**0.23**	**0.81**	**0.43**	**0.138**

F LIMITATIONS

While GauMem reliably preserves long-horizon spatial consistency under strict storage budgets, it has several limitations. First, our consolidation policy operates on primitives predicted by an upstream, frozen feed-forward 3DGS frontend; consequently, the reconstruction fidelity of the memory is inherently bounded by the frontend's geometric accuracy, and upstream depth artifacts or scale ambiguities can propagate into subsequent readouts. Second, GauMem assumes a predominantly static environment where visual disagreement stems from camera viewpoint exploration rather than non-rigid object motion. Extending learned 3D memory architectures to untangle transient object dynamics from permanent scene geometry remains an essential direction for future embodied world model research. Third, while GauMem demonstrates robust zero-shot generalization across distinct indoor and outdoor spatial distributions (Appendix E), the policy is trained exclusively on DL3DV camera and trajectory statistics. Deploying the memory in domains with substantially different camera dynamics (e.g., highly agile drone trajectories or autonomous driving scenarios) or distinct sensor modalities may require fine-tuning or domain-specific adaptation.

G FURTHER EXPERIMENTAL RESULTS

Downstream Video Generation under Extended Exploration (N=10). Table 8 extends the downstream video generation benchmark to an extended exploration horizon with N=10 intervening regions. As the stream lengthens, GauMem maintains a clear lead in photometric quality.

The contrast in geometric stability is particularly pronounced across horizons. While heuristic baselines exhibit compounding geometric degradation as interference accumulates—trajectory drift for unfused primitives (Keep-all) and surfel fusion increases to 4.34–5.85% ATE, and fused point clouds degrade beyond 11–15%—GauMem remains remarkably stable. Its trajectory errors barely shift between horizons, maintaining a rotation error of $0.23°$ (vs. $0.21°$ at N=4) and an ATE of 0.43% (vs. 0.40% at N=4), closely tracking the real-frame rotation baseline ($0.14°$). While in-context frame conditioning maintains marginally lower LPIPS (0.305)—benefiting from uncompressed real frames embedded directly in its context window—its camera trajectory exhibits severe drift (9.20% ATE and $2.64°$ rotation error). These results indicate that explicit, oriented 3D Gaussian memory provides persistent spatial grounding that prevents progressive camera drift over prolonged exploration.

Geometry of Raw Memory Renderings. In the downstream evaluation of Sec. 4.3, the video generator could potentially rectify geometric discrepancies or hallucinate missing surfaces. To isolate what the memory representation inherently preserves from generative post-processing, Table 9 evaluates the geometry of raw memory renderings directly. Specifically, we render each unconditioned memory state along the return trajectory of $\mathcal{R}'_0$ and feed the resulting views into the independent

Table 9: **Geometric fidelity of raw memory renderings** across $N{=}4$ and $N{=}10$ intervening regions on DL3DV test set (140 scenes). Metrics are re-estimated directly from unconditioned renderings along the return trajectory of $\mathcal{R}'_0$ via an independent reconstruction model (Lin et al., 2025). Real frames establish the estimator's empirical error floor.

	$N{=}4$ intervening regions				$N{=}10$ intervening regions			
Memory	rot° ↓	transl° ↓	ATE %↓	AbsRel↓	rot° ↓	transl° ↓	ATE %↓	AbsRel↓
Point cloud (position and color, no orientation)								
Keep-all	0.89	1.96	1.23	0.280	2.92	6.14	3.91	0.342
Fixed-cell fusion	2.76	9.35	5.60	0.390	3.59	11.75	6.95	0.409
Budget-matched fusion	4.28	18.39	10.22	0.424	4.81	23.44	11.61	0.445
Frames								
Top-4 frame retrieval	0.72	2.05	1.49	0.172	0.91	4.07	2.96	0.202
Gaussian primitives								
Keep-all	0.91	2.20	1.40	0.308	3.47	8.20	5.31	0.367
Tuned surfel fusion	1.06	2.46	1.66	0.313	2.25	5.70	4.05	0.365
Downsample + voxel fusion	0.25	0.78	0.41	0.225	0.79	2.29	1.63	0.291
GauMem	**0.17**	**0.64**	**0.33**	**0.117**	**0.20**	**0.67**	**0.35**	**0.132**
Real frames	0.14	0.58	0.28	—	0.14	0.58	0.28	—

reconstruction model (Lin et al., 2025). In-context frame conditioning baseline does not produce intermediate 3D renderings, thus omitted; evaluating real frames establishes the empirical measurement floor.

Comparing these raw renderings with downstream video generations yields three key insights: First, GauMem's raw memory closely tracks the real-frame error floor across all metrics and exhibits negligible drift as interference expands from $N{=}4$ to $N{=}10$. This confirms that the high camera stability and structural fidelity observed in generated rollouts are directly inherited from GauMem's explicit 3D memory, rather than synthesized post-hoc by the video generator. Second, while downsampled voxel fusion provides competitive camera stability at short horizons, its trajectory alignment deteriorates under extended interference ($N{=}10$), and its depth error remains noticeably higher due to unguided spatial pooling and primitive decimation. Finally, for point-cloud and heuristic fusion baselines, severe rotational drift and trajectory distortion are already entrenched within the raw memory buffers, which downstream video generation predictably fails to undo. These comparisons confirm that downstream trajectory drift is primarily driven by accumulated structural corruption within the memory itself, underscoring the necessity of GauMem's closed-loop diagnostic writes.

Qualitative analysis of generated return rollouts. In Figs. 4 and 5, we visualize frames generated by the frozen Wan2.1-VACE-1.3B generator when the camera returns along $\mathcal{R}'_0$ after traversing $N{=}4$ and $N{=}10$ intervening regions, evaluated at four never-written viewpoints against the ground truth. Each memory paradigm interfaces with the reader under its dedicated representation: point clouds lack orientation and spatial extents, rendered as isotropic splats; frame retrieval fetches the four most visible historical keyframes via an oracle buffer; and in-context conditioning directly injects the original recorded frames of $\mathcal{R}_0$ into the context window. Among explicit Gaussian representations, Keep-all, tuned surfel fusion, and GauMem ingest identical proposal streams, isolating the effect of the write policy.

The visual artifacts vividly reflect the limitations of heuristic retention. Because point clouds cannot model continuous surfaces, they either leave severe projective holes that force the generator to hallucinate geometry or blur high-frequency textures into diffuse splats. Unconstrained Gaussian accumulation (Keep-all) and tuned surfel fusion suffer from dense floating clutter and conflicting surface layers: fine structural elements are either duplicated, smeared, or overwritten across successive frames. In contrast, GauMem reliably consolidates redundant incoming proposals into coherent, persistent surfaces, anchoring the generator to render clean visual landmarks positioned identically to the ground truth.

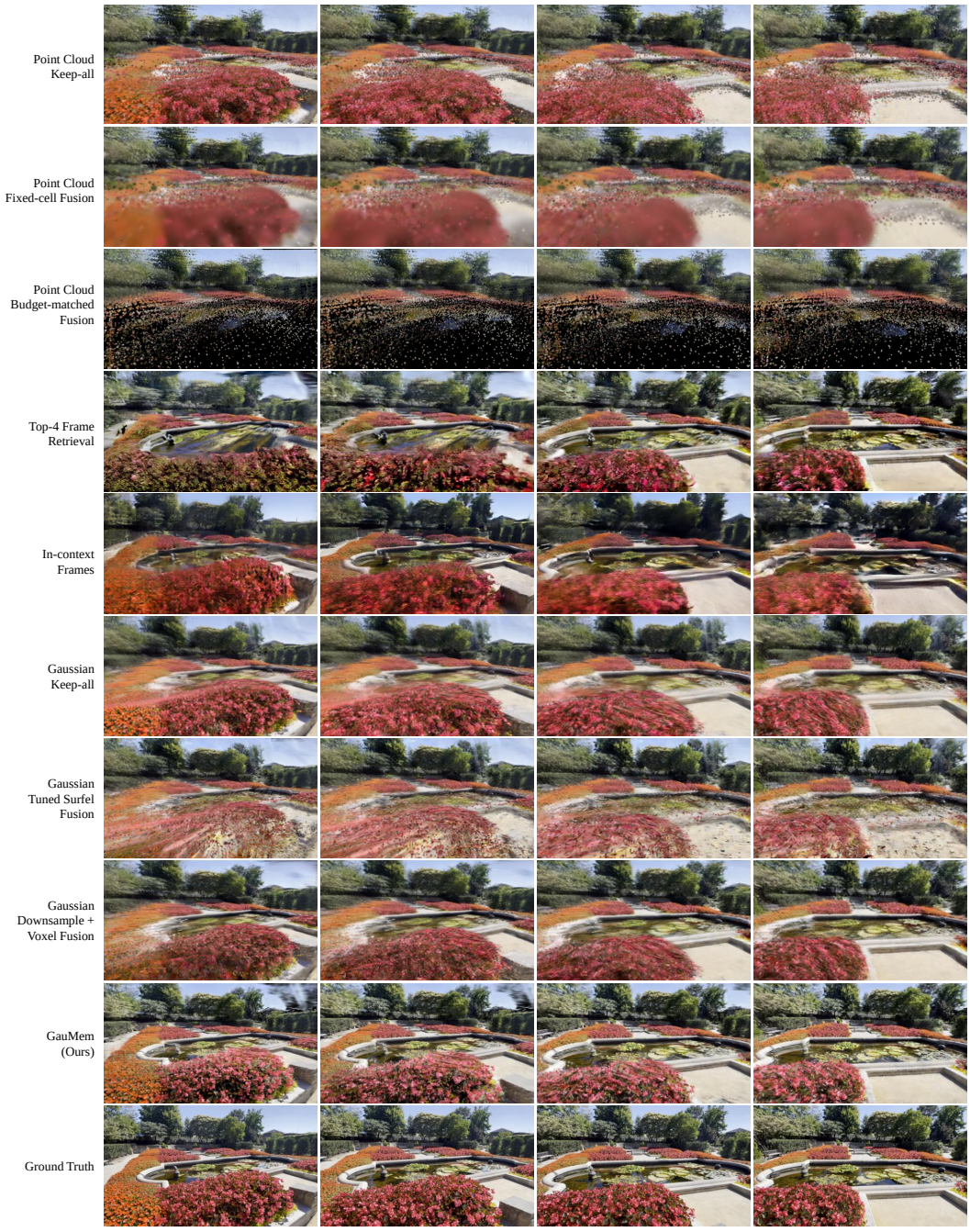

Figure 4: **Qualitative comparison of downstream video rollouts upon return** ($N{=}4$). Frames generated by the frozen Wan2.1-VACE-1.3B conditioned on different memory representations along $\mathcal{R}'_0$; the bottom row shows the ground truth.

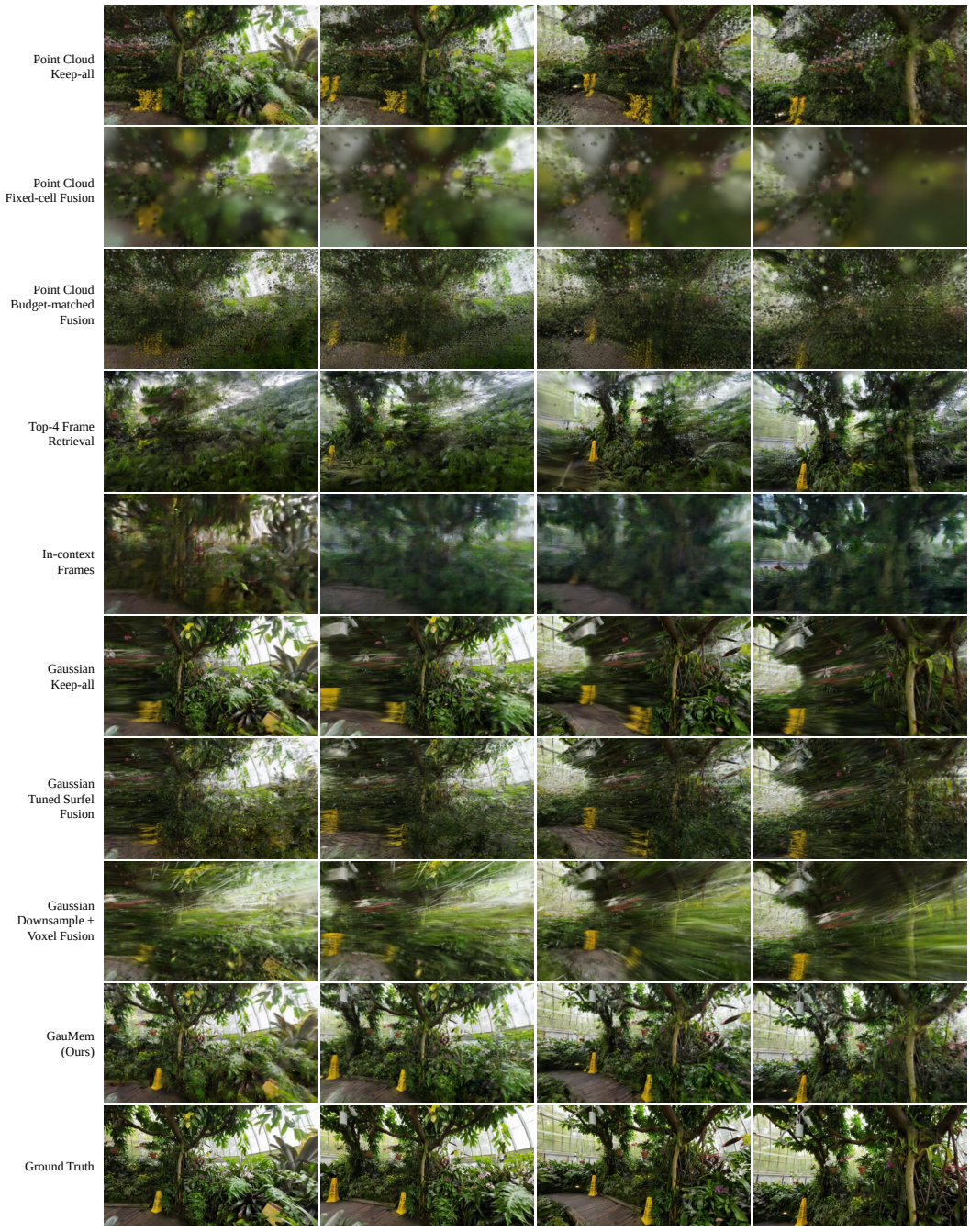

Figure 5: **Qualitative comparison of downstream video rollouts upon return** (N=10). Frames generated by the frozen Wan2.1-VACE-1.3B conditioned on different memory representations along $\mathcal{R}'_0$; the bottom row shows the ground truth.