HLA-WM: HYBRID LINEAR ATTENTION FOR LONG-HORIZON VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon video world models require persistent memory to preserve scene consistency over extended rollouts. Softmax attention retains the full generation history through a growing KV cache, whereas recurrent linear attention compresses history into fixed-size states with substantially lower memory cost. However, we identify severe long-range forgetting in Gated DeltaNet (GDN), where information from distant but relevant scenes is progressively attenuated by subsequent state updates. To address this limitation, we propose HLA-WM, a training-free *hybrid linear-attention framework* that combines coarse-grained geometry-guided retrieval with fine-grained recurrent linear-state computation. HLA-WM exploits the affine structure of GDN to cache compact chunk-wise transition summaries, retrieve scene-relevant historical chunks using camera geometry, and recompose them into query-specific recurrent states. On the 60-second SANA-WM-Bench, HLA-WM improves all six aggregate revisit-consistency and camera-control metrics of the base autoregressive generator without additional training, including a 0.74 dB PSNR gain and a 28.5% reduction in rotation error. The improvements persist after downstream refinement and generalize to MBench-A, where HLA-WM consistently improves all three revisit-consistency metrics across all four subsets and all evaluated inference modes over 547 samples. At a 60-second context, HLA-WM reduces historical-state memory by $12\times$ relative to full KV caching while incurring at most a 1.6% reduction in inference throughput. These results demonstrate that selectively addressable recurrent memory can improve long-range scene recall while preserving the efficiency advantages of GDN.

1 INTRODUCTION

World models predict environment dynamics, enabling agents to anticipate future observations and reason about the consequences of their actions (Ha & Schmidhuber, 2018; Hafner et al., 2025). Recent video world models realize such simulation through autoregressive generation conditioned on agent actions or camera trajectories (Zhang et al., 2025; He et al., 2025; Mao et al., 2025; Zhu et al., 2026b;a). Beyond short-term temporal continuity, long-horizon simulation requires previously observed scenes to remain recoverable after they leave the recent context (Wu et al., 2025; Hong et al., 2025; Wang et al., 2026b). When the camera later returns, the model should recover the earlier scene structure rather than generate a semantically similar but spatially inconsistent view. Full-history key-value (KV) caching provides direct access to past observations, but its storage cost grows continuously with the rollout length (Yang et al., 2026; Yesiltepe et al., 2026). Recurrent linear attention offers a compact alternative by compressing history into fixed-size states (Katharopoulos et al., 2020; Li et al., 2026; Zhu et al., 2026a). This raises a central question for long-horizon world modeling: *how can recurrent models retain efficient memory while still recovering distant scene information when it becomes relevant again?*

Among modern recurrent linear-attention architectures, Gated DeltaNet (GDN) combines data-dependent decay with delta-rule updates to manage information within a compact recurrent state (Yang et al., 2025). Nevertheless, we observe a clear **long-range forgetting** problem in GDN-based video world models. As illustrated in Fig. 1, SANA-WM exhibits a pronounced decline in revisit consistency as the temporal gap increases, while the qualitative example shows substantial changes in scene structure after the camera leaves and later returns. Such revisits are particularly challenging because information that is no longer locally relevant may become essential again after a

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.

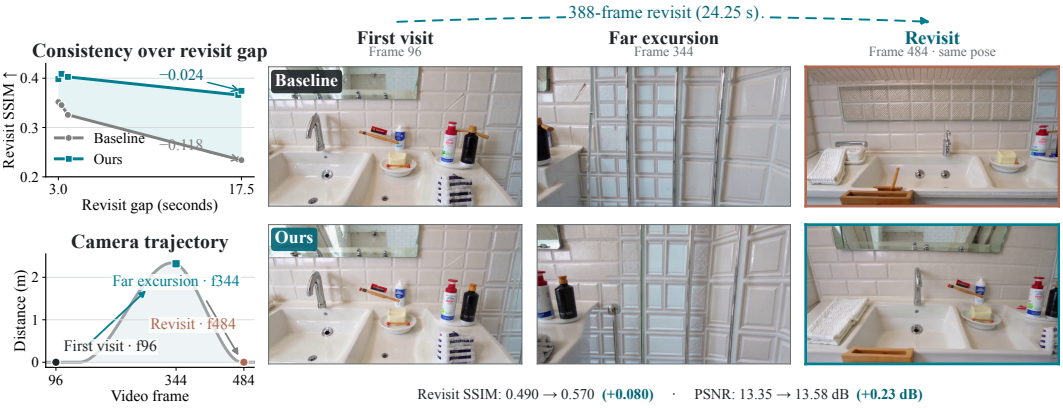

Figure 1: **Long-range forgetting in GDN-based world-model rollouts. Upper left:** Revisit SSIM over increasing temporal gaps. **Lower left:** Camera displacement along the qualitative trajectory. **Right:** Qualitative comparison of the first visit, far excursion, and revisit at the similar camera pose. After the long excursion, the baseline SANA-WM shows noticeable changes in scene layout and object arrangement, whereas our HLA-WM better preserves the first-visit scene content.

long sequence of intermediate observations. This behavior suggests that compact recurrent memory alone is insufficient to reliably preserve distant scene information over extended rollouts.

To understand this limitation, we analyze the recurrent state transition in GDN. Since the full history is compressed into a single recurrent state, earlier information must survive all intervening updates to remain recoverable, motivating a memory organization in which compact historical states remain independently addressable.

We therefore ask whether a pretrained recurrent video world model can recover distant scene history without retaining a full token-level KV cache or requiring additional training. To this end, we propose **HLA-WM**, a training-free *hybrid linear-attention framework* that combines coarse-grained geometry-guided retrieval with fine-grained recurrent linear-attention access. HLA-WM exploits the affine structure of GDN to cache compact chunk-wise summaries of historical state transitions, selectively retrieves relevant summaries using camera geometry, and recomposes them into a query-specific recurrent state. This establishes an intermediate memory regime between a single recurrent state and token-level full-history caching, enabling selective access to distant history while retaining compact linear-state storage.

For camera-controlled world-model rollouts, camera metadata provides a natural addressing signal without requiring a learned router (Yu et al., 2025; Xiao et al., 2025). HLA-WM uses a lightweight camera-frustum overlap proxy derived from poses, intrinsics, and a scene-level depth estimate to identify relevant historical chunks. The selected summaries, together with the first chunk as a persistent sink and recent context for local continuity, are recomposed as retrieved context in chronological order without replaying historical visual tokens. Selective recomposition is applied only to the main GDN memory, while camera-control and local convolutional states retain their original recurrent updates. All pretrained parameters remain unchanged. Thus, HLA-WM changes how historical memory is stored and accessed rather than retraining the underlying generator.

Experiments on SANA-WM-Bench (Zhu et al., 2026a) and MBench-A (Zhang et al., 2026) demonstrate consistent improvements in long-range revisit fidelity. At Stage 1, where HLA-WM directly modifies recurrent memory, it also improves camera controllability by reducing accumulated structural drift and maintaining a more stable spatial reference during generation. Downstream refinement introduces mode-dependent trade-offs, while the long-range consistency gains remain evident across both benchmarks. On MBench-A, all three revisit-consistency metrics improve across all four subsets and all evaluated inference modes. At a 60-second context, HLA-WM requires $12\times$ less historical-state storage than full KV caching while incurring at most a 1.6% reduction in inference throughput across the evaluated pipelines. These results show that selectively addressable recurrent memory can substantially improve long-range scene recall.

Our main contributions are summarized as follows:

- We empirically characterize long-range forgetting in GDN-based video world models and show how accumulated recurrent transitions progressively weaken access to distant scene information.
- We propose HLA-WM, a training-free hybrid linear-attention framework that equips recurrent video world models with selectively addressable historical memory through compact affine summaries, camera-guided addressing, and chronological state recomposition.
- Extensive experiments on SANA-WM-Bench and MBench-A demonstrate improved long-range scene consistency. On the 60-second SANA-WM-Bench, HLA-WM improves the base autoregressive generator by 0.74 dB in PSNR and reduces rotation error by 28.5% without training.

2 Related Work

Video World Models. Recent advances in video generation have enabled world models to predict future visual observations conditioned on actions, camera trajectories, or other control signals, supporting interactive and long-horizon environment simulation (Zhang et al., 2025; He et al., 2025; Mao et al., 2025; Zhu et al., 2026b;a). To maintain scene consistency over extended rollouts, existing approaches preserve historical information in different forms. Frame-based methods retain selected historical observations as references for future generation (Wu et al., 2025; Wang et al., 2026b). Token- or KV-based methods preserve historical visual tokens or attention caches to provide direct access to past context (Hong et al., 2025; Samuel et al., 2026; Wu et al., 2026). Feature-memory methods summarize past observations into compact latent representations or dedicated memory tokens (Wei et al., 2026). Spatial-memory approaches instead maintain explicit geometric or 3D representations that can be queried according to the current viewpoint (Wu et al., 2025; Wang et al., 2026b; Team et al., 2026). Although these mechanisms improve long-range consistency, they introduce additional storage and retrieval overhead that can grow with the rollout. In particular, frame-based memories may require retrieved observations to be re-encoded and their attention caches recomputed, while token, feature, and spatial memories require maintaining and querying additional historical representations, limiting their efficiency for long-horizon generation.

Linear Attention for World Models. Linear attention reformulates attention through associative computation, allowing historical information to be summarized into fixed-size recurrent states instead of maintaining a KV cache that grows with sequence length (Katharopoulos et al., 2020; Schlag et al., 2021). Building on this formulation, subsequent works improve the expressiveness and memory management of linear states through more sophisticated recurrent updates. Gated Linear Attention introduces data-dependent decay to selectively forget historical information (Yang et al., 2024a), while DeltaNet employs the delta rule to enable targeted updates to the recurrent memory (Yang et al., 2024b). Gated DeltaNet (GDN) further combines these two mechanisms, providing both adaptive forgetting and selective memory updates within a fixed-size state (Yang et al., 2025). More recently, recurrent linear attention has been extended to autoregressive video generation, where linear states serve as compact cross-frame memory for long-form generation (Li et al., 2026; Zhu et al., 2026a). Despite their favorable memory efficiency, these methods continuously compress historical observations into a limited recurrent state, making distant information increasingly susceptible to attenuation and interference from subsequent updates during long-horizon generation.

3 Empirical Insights

We first investigate the long-range forgetting behavior of GDN in SANA-WM (Zhu et al., 2026a). At the chunk level, the recurrent update of chunk i can be written as

$$S_i = S_{i-1}A_i + B_i, \tag{1}$$

where A_i denotes the state transition applied to historical memory and B_i represents the information newly written by the current chunk. Expanding the recurrence over multiple chunks gives

$$S_j = \cdots + B_i A_{i+1} A_{i+2} \cdots A_j + \cdots + B_j. \tag{2}$$

Thus, information written by an early chunk is repeatedly transformed by all subsequent transition matrices before it can influence a distant future scene. We quantify its cumulative retention from chunk i to chunk j using the average norm of the intervening transition product:

$$W_{i\to j} = \mathrm{AvgNorm}\left(A_{i+1}A_{i+2}\cdots A_j\right), \tag{3}$$

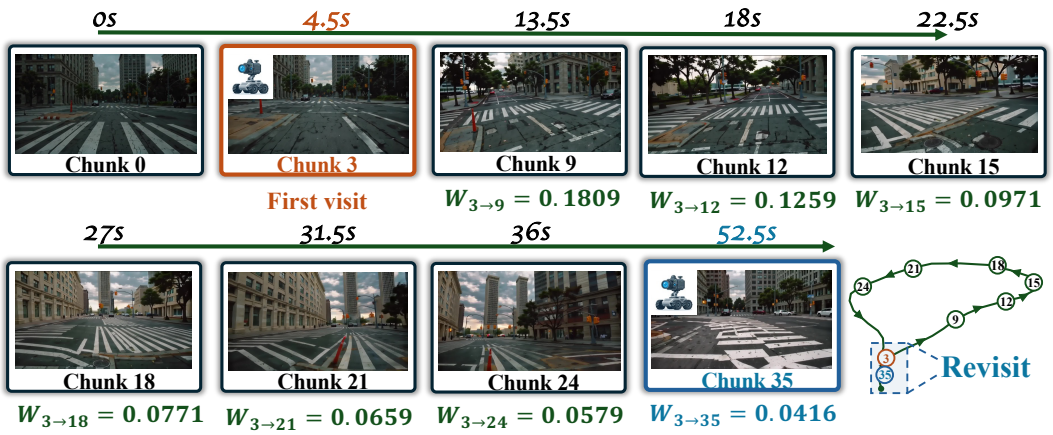

Figure 2: **Long-range forgetting in GDN.** The camera follows a long loop and eventually returns near the region observed at Chunk 3 around Chunk 35. We visualize representative generated chunks along the trajectory together with the cumulative retention $W_{3\to j}$ of Chunk 3. Although the early scene becomes relevant again near the end of the rollout, its retained influence continuously decreases as intermediate chunks are generated, reaching only 0.0416 at Chunk 35.

where a smaller $W_{i\to j}$ indicates stronger attenuation from chunk i to j.

Figure 2 illustrates this phenomenon on a representative trajectory from the SANA-WM Benchmark. The camera leaves the region observed at Chunk 3, traverses a large loop, and eventually returns to a nearby viewpoint at Chunk 35. Since the two chunks observe highly overlapping regions, the information stored at Chunk 3 should remain particularly relevant when generating Chunk 35. However, its retained influence steadily diminishes as the rollout progresses and drops to only 0.0416 by the time the camera revisits the scene. Correspondingly, the scene generated at Chunk 35 deviates substantially from the earlier observation, indicating that the model can no longer effectively recover the distant but relevant scene information.

More importantly, most intermediate chunks correspond to different viewpoints and are only **weakly related** to the scene revisited at Chunk 35. Nevertheless, their state-transition matrices are still sequentially applied to the memory written by Chunk 3, causing that memory to undergo repeated transformations even when the intermediate observations provide little useful information for the later revisit. As a result, scene information that becomes relevant again can be substantially attenuated before it is needed. This exposes a fundamental mismatch in conventional GDN memory: *historical information is updated according to temporal progression, whereas its relevance in world-model rollouts is often determined by spatial and geometric proximity*. This mismatch motivates selectively preserving and composing scene-relevant historical states instead of propagating all past information uniformly through every intermediate chunk.

4 Method

4.1 Overview

We propose HLA-WM, a hybrid linear-attention framework for long-horizon video world models. The key idea is to replace the single accumulated GDN memory with a collection of independently addressable chunk-wise memories. Specifically, we partition the generated sequence into temporal chunks and summarize the GDN transition induced by each chunk. For each query chunk, HLA-WM uses camera geometry to retrieve a sparse set of scene-relevant historical chunks and selectively composes their summaries into the recurrent state used for generation. As illustrated in Fig. 3, this design combines coarse-grained geometry-guided memory retrieval with recurrent linear attention, allowing distant but relevant information to bypass unrelated intermediate updates while preserving the compactness of linear-state memory.

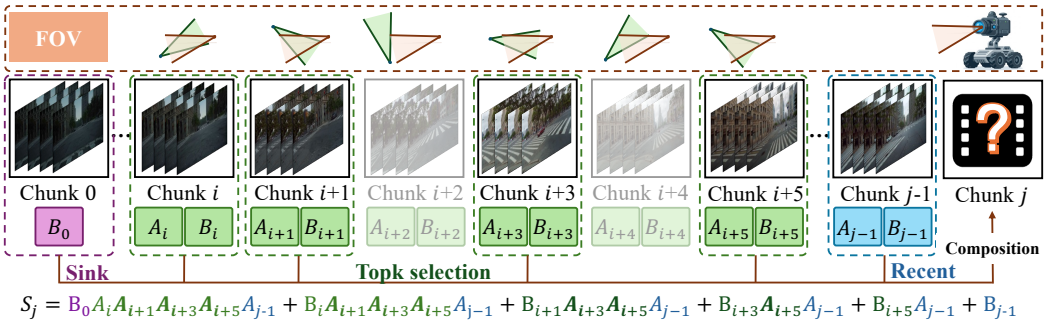

Figure 3: Overview of HLA-WM. Historical video chunks are represented by independent GDN summaries. Given the current query, geometry-guided retrieval selects scene-relevant historical chunks together with persistent sink and recent context. The selected summaries are then composed in chronological order to construct a query-specific recurrent state.

4.2 Chunk-wise GDN Memory

Gated DeltaNet maintains a recurrent linear state through token-wise updates (Yang et al., 2025). Let t index tokens in the recurrent scan order. For token t, the state update is

$$S_t = \gamma_t S_{t-1} + \beta_t \left(v_t - S_{t-1} k_t\right) k_t^\top, \tag{4}$$

where k_t and v_t are the key and value vectors projected from the current token representation, γ_t controls memory decay, and β_t controls the strength of the delta update. Rearranging Eq. 4 gives

$$S_t = S_{t-1} A_t + B_t, \tag{5}$$

where

$$A_t = \gamma_t I - \beta_t k_t k_t^\top, \qquad B_t = \beta_t v_t k_t^\top. \tag{6}$$

This affine form can be directly composed over consecutive tokens. Consider chunk i containing L tokens, and let (i, ℓ) denote its ℓ-th token. Unrolling the recurrence within the chunk gives

$$S_i = S_{i-1} A_{i,1} A_{i,2} \cdots A_{i,L} + B_{i,1} A_{i,2} \cdots A_{i,L} + \cdots + B_{i,L}.$$

Therefore, the entire chunk can be represented by a single affine transition

$$S_i = S_{i-1} A_i + B_i, \tag{7}$$

where

$$A_i = A_{i,1} A_{i,2} \cdots A_{i,L}, \qquad B_i = \sum_{\ell=1}^{L} B_{i,\ell} \prod_{u=\ell+1}^{L} A_{i,u}. \tag{8}$$

Here, the products follow the recurrent scan order, with an empty product defined as the identity. The pair (A_i, B_i) therefore provides a sufficient summary of chunk i: A_i describes how the chunk transforms existing memory, while B_i represents the information written by the chunk itself.

Equivalently, their summaries follow the composition rule

$$(A_i, B_i) \star (A_j, B_j) = (A_i A_j, B_i A_j + B_j). \tag{9}$$

The operation is associative, enabling an arbitrary subset of historical chunk summaries to be efficiently recomposed without replaying their original visual tokens through the model.

4.3 Geometry-Guided Memory Retrieval

Given the individually addressable chunk-wise memories, we retrieve historical chunks according to their geometric relevance to the current view. For each query chunk q and an eligible historical chunk i, we estimate their overlap using only camera poses, intrinsics, and a scene-level median-depth estimate. For a sampled image-plane location (u, v), we construct the camera-space ray

$$\mathbf{r}(u, v) = \left[\frac{u - c_x}{f_x}, \frac{v - c_y}{f_y}, 1\right]^\top, \tag{10}$$

and instantiate proxy 3D points at several depths around the scene-level median depth,

$$\mathcal{D} = \{0.5d_{\text{med}},\, d_{\text{med}},\, 1.5d_{\text{med}}\}. \tag{11}$$

These proxy points are transformed into world coordinates using the corresponding camera poses and projected into the camera views of the other chunk.

Let $V(i \to q)$ denote the fraction of proxy points constructed from chunk i that project with positive depth into at least one camera view of chunk q. We define the symmetric frustum-overlap score as

$$R(i, q) = \frac{1}{2}\left[V(i \to q) + V(q \to i)\right]. \tag{12}$$

The score provides a lightweight proxy for geometric overlap rather than exact 3D scene intersection. Similar geometry-based criteria have been used to retrieve relevant historical context for long-horizon video generation and world simulation (Yu et al., 2025; Xiao et al., 2025). Importantly, our retrieval requires neither generated RGB content nor learned semantic features. We rank the eligible historical chunks according to $R(i, q)$ and retrieve the top-K most relevant ones. Further details of the camera sampling and overlap computation are provided in Appendix A.1.1.

In addition to the retrieved historical chunks, we retain the first chunk as a persistent sink to preserve global context, together with recent chunks to maintain local temporal continuity. When the change in camera translation direction exceeds a threshold τ_{turn}, we temporarily expand the recent context to the three most recent chunks to improve robustness under abrupt camera motion.

Let $\mathcal{H}_q$ denote the remaining historical chunks after excluding the sink and recent context. The selected history is

$$\mathcal{I}_q = \mathcal{I}_{\text{sink}} \cup \mathcal{I}_{\text{recent}} \cup \text{TopK}_{i \in \mathcal{H}_q} R(i, q). \tag{13}$$

This design combines sparse retrieval of geometrically relevant long-range memory with adaptive preservation of short-term context, reducing interference from unrelated intermediate scenes while maintaining local continuity during abrupt camera motion.

4.4 Selective State Recomposition

After retrieval, the selected chunks are restored to their original chronological order. Denote the ordered set as

$$\pi_q = (\pi_1, \pi_2, \ldots, \pi_M). \tag{14}$$

Using the associative rule in Eq. 9, their summaries can be directly composed as

$$A_{\pi_q} = \prod_{m=1}^{M} A_{\pi_m}, \tag{15}$$

and

$$B_{\pi_q} = \sum_{m=1}^{M} B_{\pi_m} \prod_{n=m+1}^{M} A_{\pi_n}. \tag{16}$$

The resulting historical recurrent state is

$$S_q^{\text{hist}} = S_0 A_{\pi_q} + B_{\pi_q}, \tag{17}$$

where S_0 is the initial GDN state.

Standard GDN composes every historical chunk in temporal order, causing information written by an early scene to be repeatedly transformed by all subsequent chunks. In contrast, HLA-WM shortens this propagation path by excluding geometrically irrelevant intermediate chunks. Distant but relevant memories are therefore exposed to substantially fewer unrelated transitions before being reused. Meanwhile, the original GDN dynamics within each selected chunk are fully preserved, rather than replacing them with direct averaging or summation of linear states.

The recomposition is performed directly from cached (A_i, B_i) summaries and does not require re-forwarding the selected historical chunks. When all previous chunks are selected, the composition reduces to the original chronological GDN recurrence, providing a direct consistency check for

Table 1: Main results on SANA-WM-Bench under different generation stages and refinement modes. We report revisit consistency, camera controllability, and inference throughput. Best results within each setting are shown in bold.

Pipeline	Method	Revisit Consistency			Camera Control			Efficiency
		PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	FPS ↑
Simple-Trajectory Split								
Stage-1	SANA-WM	9.18	0.1729	0.6327	19.7040	2.2625	2.3974	**22.403**
	HLA-WM	**9.91**	**0.1999**	**0.6049**	**13.5864**	**2.0928**	**2.1798**	22.053
Stage-1 + AR Refine	SANA-WM	14.44	0.2770	0.5738	20.9315	**2.1587**	**2.3121**	**8.280**
	HLA-WM	**14.82**	**0.2872**	**0.5682**	**20.2933**	2.1910	2.3240	8.211
Stage-1 + Bi. Refine	SANA-WM	13.04	0.2945	0.5786	11.3848	1.9893	2.0545	**7.701**
	HLA-WM	**13.56**	**0.3148**	**0.5614**	**8.8823**	**1.8637**	**1.9127**	7.700
Hard-Trajectory Split								
Stage-1	SANA-WM	9.37	0.1774	0.6044	20.3759	1.9952	2.1507	**22.403**
	HLA-WM	**10.11**	**0.1997**	**0.5848**	**15.0887**	**1.8388**	**1.9482**	22.053
Stage-1 + AR Refine	SANA-WM	14.12	0.2719	0.5621	25.6330	2.0099	2.2079	**8.280**
	HLA-WM	**14.47**	**0.2812**	**0.5557**	**22.0165**	**1.9351**	**2.1043**	8.211
Stage-1 + Bi. Refine	SANA-WM	13.05	0.2972	0.5468	13.1713	1.6649	1.7694	**7.701**
	HLA-WM	**13.51**	**0.3142**	**0.5384**	**10.5268**	**1.5449**	**1.6230**	7.700

the chunk-wise formulation. For the accompanying softmax-attention pathway, we apply the same selected chunk set to the corresponding historical KV cache.

Efficient State Recomposition Kernel. We implement direct state recomposition on top of the chunk-wise Triton GDN kernels. During the original forward pass, the affine summaries (A_i, B_i) of each chunk are computed once and cached independently of the incoming recurrent state. For each query chunk q, the summaries in $\mathcal{I}_q$ are gathered in chronological order and directly composed through an efficient forward scan to reconstruct S_q^{hist}.

Memory and Computational Overhead. HLA-WM introduces additional storage only for a fixed-size affine summary (A_i, B_i) and lightweight camera metadata for each temporal chunk. Since historical information is cached at the chunk level rather than the token level, the memory footprint scales with the number of chunks instead of the number of visual tokens. During generation, only the selected summaries are involved in state composition, while geometry-based retrieval operates directly on camera information with minimal overhead. Consequently, HLA-WM provides selective access to long-range memory while largely preserving the memory and computational efficiency of recurrent linear attention.

5 Experiments

5.1 Experimental Setup

Datasets and Baselines. We conduct experiments on SANA-WM-Bench (Zhu et al., 2026a) and MBench-A (Zhang et al., 2026). SANA-WM-Bench contains simple and hard splits, each with 80 camera-controlled trajectories spanning four scene categories: indoor, outdoor city, outdoor nature, and game-style environments. Each trajectory contains 961 frames at 16 fps, corresponding to approximately 60 seconds, with a resolution of 1280×704. MBench-A contains 547 samples across four official subsets: Causal, Human, Environment, and Object. We use the official 1.6B SANA-WM as our baseline. Its Stage 1 model performs 4-step distilled autoregressive streaming generation with recurrent GDN states, followed by optional causal AR or bidirectional refinement. HLA-WM modifies only the Stage 1 memory mechanism and requires no additional training or fine-tuning, while downstream refinement remains unchanged. Within each setting, the baseline and HLA-WM use identical inputs, camera trajectories, prompts, and generation configurations.

Implementation Details. All video-generation experiments are conducted on a single NVIDIA H200 GPU with 140 GB of memory. We follow the official SANA-WM autoregressive streaming pipeline (Zhu et al., 2026a), including the chunk-causal LTX-2 refiner and causal VAE. Following the default 4-step distilled configuration, we use the sampling schedule $[1000, 960, 889, 727, 0]$ and

Table 2: Overall results on MBench-A under different inference modes. Best results within each setting are shown in bold.

Pipeline	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Stage-1	SANA-WM	10.17	0.2545	0.6251
	HLA-WM	**11.01**	**0.2848**	**0.5823**
Stage-1 + AR Refine	SANA-WM	12.36	0.2996	0.6554
	HLA-WM	**12.73**	**0.3122**	**0.6447**
Stage-1 + Bi. Refine	SANA-WM	12.94	0.3310	0.5597
	HLA-WM	**13.52**	**0.3431**	**0.5470**

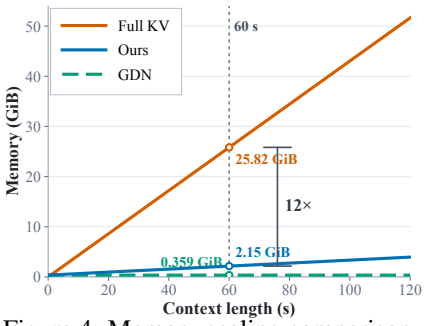

Figure 4: Memory scaling comparison.

three latent frames per autoregressive block. Unless otherwise specified, HLA-WM retains one sink chunk, the most recent chunk, and the top-1 historical chunk ranked by geometric overlap. When the within-chunk change in camera translation direction exceeds $\tau_{\text{turn}} = 25^\circ$, the recent context is temporarily expanded from one to three chunks.

Evaluation Metrics. We evaluate long-range scene consistency using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) between frames corresponding to repeated or near-repeated camera poses. Camera controllability is measured by rotation error (RotErr), relative translation error ($\text{TransErr}_{\text{rel}}$), and relative camera-motion consistency ($\text{CamMC}_{\text{rel}}$), following the SANA-WM evaluation protocol (Zhu et al., 2026a). Camera trajectories are estimated using Pi3X, an enhanced implementation of π^3 (Wang et al., 2026a).

5.2 Main Results

Performance on SANA-WM-Bench. Table 1 evaluates HLA-WM across different stages of the SANA-WM generation pipeline. At Stage 1, where HLA-WM directly modifies the recurrent memory, it improves all revisit-consistency and camera-control metrics on both trajectory splits. On Hard trajectories, PSNR and SSIM increase by 0.74 dB and 0.0223, LPIPS decreases by 0.0196, and RotErr is reduced by 26.0%. Similar improvements are observed on Simple trajectories, including a 0.73 dB PSNR gain and a 31.0% reduction in RotErr. These camera-control gains are consistent with improved scene stability, as preserving previously observed geometry reduces structural drift and provides a more reliable spatial reference for camera-conditioned generation. The improvements largely persist after downstream refinement. Under causal AR refinement, all three revisit-consistency metrics improve on both splits. On Hard trajectories, all three camera-control metrics also improve, while on Simple trajectories RotErr decreases but TransErr and CamMC increase slightly. Full-sequence refinement preserves the benefits more consistently and improves all six metrics on both splits, including PSNR gains of 0.52/0.46 dB and RotErr reductions of 22.0%/20.1% on the Simple/Hard splits. Overall, selective long-range memory consistently improves revisit fidelity while generally enhancing camera controllability over long rollouts. Detailed category-level results, together with extensive ablations on component contributions, GDN/KV alignment, retrieval budget, and historical selection strategies, are provided in the Appendix.

Performance on MBench. We further evaluate HLA-WM on the full MBench-A benchmark with 547 samples to assess its generalization beyond SANA-WM-Bench. As shown in Table 2, HLA-WM consistently improves PSNR, SSIM, and LPIPS across all three inference modes. The gains are most pronounced at Stage 1, where PSNR and SSIM increase by 0.84 dB and 0.0303, respectively, while LPIPS decreases by 0.0428. The improvements remain consistent after downstream refinement, with PSNR gains of 0.37 dB and 0.58 dB under causal AR and bidirectional refinement, respectively. These results demonstrate that the proposed memory mechanism generalizes effectively across different refinement modes and diverse scene and interaction patterns. Detailed subset-level results are provided in the Appendix.

Efficiency Analysis. HLA-WM offers a practical trade-off between full-KV attention and recurrent linear attention. While full KV caching preserves direct access to the entire history at a memory cost that grows rapidly with context length, native GDN maintains constant memory but compresses the entire history into a single recurrent state. HLA-WM retains selectively addressable long-range

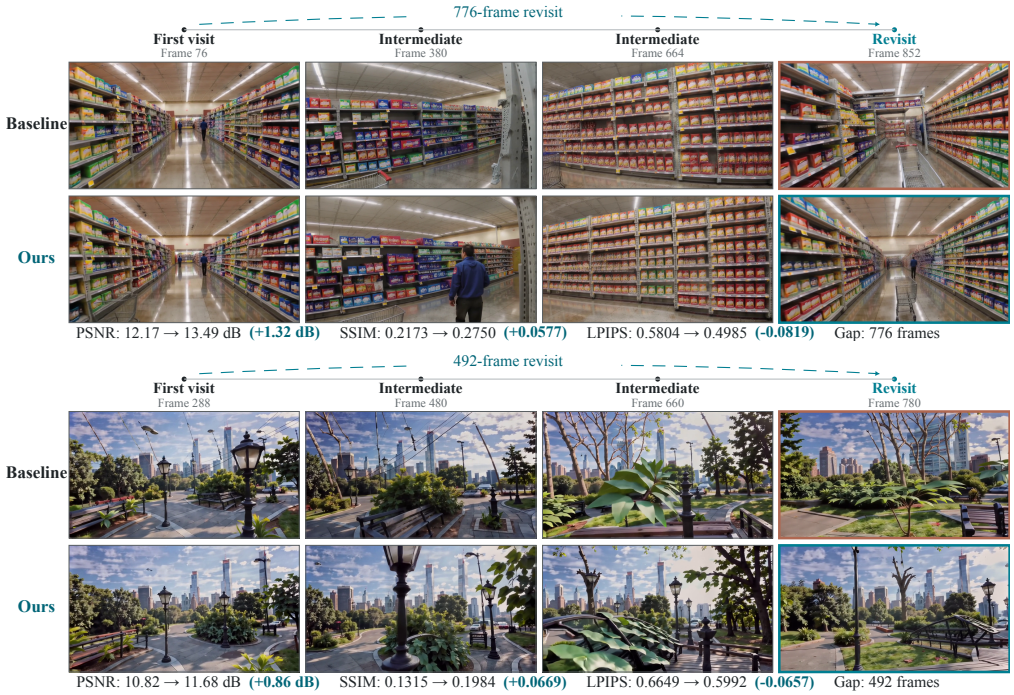

Figure 5: Qualitative comparison on long-range revisit scenarios under causal AR refinement (top) and bidirectional refinement (bottom). Each example shows the first visit, intermediate views, and a later revisit along the same camera trajectory. Revisit metrics are computed between the first-visit and revisit frames within each method.

memory through compact chunk-wise summaries, requiring only 2.15 GiB of historical-state memory at a 60-second context, 12× less than full KV caching, while incurring at most a 1.6% reduction in inference throughput across the evaluated settings.

Qualitative Results. Figure 5 further illustrates the long-range memory improvements of HLA-WM under different refinement modes. In the AR-refined indoor example, the baseline exhibits substantial structural drift after a long revisit: the aisle geometry changes, the shelf arrangement becomes inconsistent, and the alignment between the shelves and ceiling lights deviates noticeably from the first visit. In contrast, HLA-WM better preserves the two-sided shelf layout, central aisle structure, and overall scene perspective. In the bidirectionally refined outdoor example, the baseline shows larger changes in the foreground composition and spatial arrangement of the walkway, vegetation, and distant buildings. HLA-WM more consistently preserves these scene elements and their relative geometry upon revisit. Additional visualizations are provided in the Appendix.

6 Conclusion

We identify long-range forgetting as a key limitation of recurrent linear attention in video world models. Our chunk-level analysis of GDN shows that distant but relevant scene information is progressively attenuated by subsequent, often unrelated, state transitions. Based on this insight, we propose HLA-WM, which retrieves geometrically relevant chunk-wise memories and selectively recomposes them for long-range recall. Experiments on SANA-WM-Bench demonstrate consistent improvements in revisit fidelity and camera controllability, while retaining substantially lower memory cost than full KV caching and introducing only modest inference overhead.

Limitations and Future Work. HLA-WM is currently training-free, which limits the achievable gains. Future work could integrate selective state retrieval into training to better preserve and exploit long-range memory. In addition, our retrieval relies solely on camera geometry and FOV overlap, without explicitly modeling occlusion, visibility, or semantic relevance. Incorporating richer or learned retrieval signals may further improve robustness in complex scenes.

AI Assistance Statement

Generative AI tools were used to assist with manuscript drafting and language polishing, literature retrieval and organization, and research ideation and execution. In particular, they were used to help refine technical writing, identify potentially relevant literature, brainstorm experimental analyses, assist with implementation and debugging, and organize and present experimental results. All retrieved references were independently checked by the authors, and all AI-assisted technical content, implementations, analyses, and experimental conclusions were reviewed and verified by the authors. All reported experiments and numerical results were obtained from the authors' experimental pipelines, and the authors take full responsibility for the correctness and integrity of the manuscript.

Reproducibility Statement

We provide detailed descriptions of the proposed method, including the token- and chunk-level GDN formulation, geometry-guided historical retrieval, selective state recomposition, and the complete inference procedure. The Appendix further specifies the camera-overlap computation, turn-aware recent-context strategy, implementation details, retrieval budgets, and evaluation protocols. We report the model configuration, generation settings, benchmark splits, evaluation metrics, ablation settings and efficiency measurements. These details are provided to enable reproduction of the reported results.

References

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

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.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pp. 5156–5165. PMLR, 2020.

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.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *arXiv preprint arXiv:2602.01801*, 2026.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *ICML*, pp. 9355–9366. PMLR, 2021.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, et al. Alayaworld: Interactive long-horizon world modeling–full technical report. *arXiv preprint arXiv:2607.18367*, 2026.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, volume 2026, pp. 10481–10497, 2026a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *NeurIPS*, volume 38, Main Conference, pp. 49371–49393. Curran Associates, Inc., 2025. doi: 10.52202/085713-1651. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/467655d26fcc207bca08915dc91964c6-Paper-Conference.pdf.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable memory for video world models. *arXiv preprint arXiv:2608.07408*, 2026.

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

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *ICLR*, 2026. URL https://openreview.net/forum?id=nCAODkpsPJ.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523. PMLR, 21–27 Jul 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *NeurIPS*, 37:115491–115522, 2024b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *ICLR*, volume 2025, pp. 29687–29707, 2025.

Hidir Yesiltepe, Jiazhen Hu, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, Hoda Eldardiry, and Pinar Yanardag. Videomla: Low-rank latent kv cache for minute-scale autoregressive video diffusion. *arXiv preprint arXiv:2605.30351*, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. 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 *CVPR*, pp. 586–595. IEEE, 2018.

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

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *ICLR*, volume 2026, pp. 79167–79184, 2026b.

A Appendix

A.1 Additional Implementation Details

A.1.1 Geometry-Aware Historical Retrieval

For each autoregressive chunk, we estimate its geometric overlap with historical chunks from the prescribed camera trajectory. The retrieval uses only camera intrinsics, extrinsics, and a scene-level median-depth scalar from the trajectory metadata, without relying on generated RGB frames, optical flow, semantic features, or dense depth maps.

For a video with N output frames and C autoregressive chunks, let a_c and b_c denote the starting and ending frame indices assigned to chunk c, respectively. We compute them as

$$a_c = \mathrm{round}\left(\frac{cN}{C}\right), \qquad b_c = \max\left(a_c + 1, \mathrm{round}\left(\frac{(c+1)N}{C}\right)\right). \tag{18}$$

In our 60-second setting, $N = 961$ and $C = 40$, yielding approximately 24 camera poses per chunk.

For each camera frame, we uniformly sample a 7×7 image-plane grid over 5%–95% of the image extent. Each sampled ray is instantiated at three proxy depths,

$$\mathcal{D} = \{0.5d_{\mathrm{med}},\ d_{\mathrm{med}},\ 1.5d_{\mathrm{med}}\}, \tag{19}$$

where d_{med} denotes the scene-level median-depth proxy. For a pixel coordinate (u, v), the corresponding camera-space ray is

$$\mathbf{r}(u, v) = \left[\frac{u - c_x}{f_x}, \frac{v - c_y}{f_y}, 1\right]^\top. \tag{20}$$

Each proxy point at depth $d \in \mathcal{D}$ is then transformed into world coordinates using the camera-to-world transformation.

Let $\mathcal{P}_i$ denote the set of proxy points constructed from chunk i. We define the directed visibility score $V(i \to j)$ as the fraction of points in $\mathcal{P}_i$ that project with positive depth into the image plane of at least one camera frame in chunk j. The symmetric overlap score is then

$$R(i, j) = \frac{1}{2}\left[V(i \to j) + V(j \to i)\right]. \tag{21}$$

This score provides a lightweight proxy for frustum overlap rather than exact 3D scene intersection, as it does not explicitly model occlusion or scene geometry.

For each query chunk c, we retain the sink and recent chunks and retrieve the top-K remaining historical chunks according to $R(i, c)$. The selected chunks are reordered chronologically before GDN state recomposition.

A.1.2 Turn-Aware Recent Context

To preserve local continuity under abrupt trajectory changes, we adapt the recent-history budget based on the motion within the current query chunk. For a regular 24-frame camera chunk with translation vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_{23}\}$, we define

$$\mathbf{d}_{\mathrm{early}} = \mathbf{p}_7 - \mathbf{p}_0, \qquad \mathbf{d}_{\mathrm{late}} = \mathbf{p}_{23} - \mathbf{p}_{16}, \tag{22}$$

and compute their angular difference as

$$\theta_c = \arccos \frac{\mathbf{d}_{\mathrm{early}}^\top \mathbf{d}_{\mathrm{late}}}{\|\mathbf{d}_{\mathrm{early}}\|_2 \|\mathbf{d}_{\mathrm{late}}\|_2}. \tag{23}$$

The recent-history budget is then set to

$$r_c = \begin{cases} 3, & \theta_c \geq \tau_{\mathrm{turn}}, \\ 1, & \text{otherwise}, \end{cases} \tag{24}$$

where $\tau_{\mathrm{turn}} = 25^\circ$. The expanded recent context is applied only to the current chunk; subsequent chunks revert to the default budget unless the criterion is triggered again.

Algorithm 1 Overall Inference Procedure of HLA-WM

Require: Prescribed camera trajectory, retrieval budget K, and turn threshold τ_{turn}
Compute pairwise geometric overlap scores $R(i, j)$ using Eq. 21
Initialize the chunk-wise GDN summary bank and streaming states
for query chunk $q = 0, \ldots, C-1$ **do**
  Compute the recent-history budget r_q using Eqs. 23–24
  if insufficient history for selective retrieval **then**
    $\mathcal{I}_q \leftarrow \{0, \ldots, q-1\}$
  else
    Select historical chunks $\mathcal{I}_q$ according to Eq. 13
    Sort $\mathcal{I}_q$ in chronological order
    Gather the cached affine summaries of chunks in $\mathcal{I}_q$
    Recompose the query-specific historical state S_q^{hist}
    Replace the main-GDN recurrent state with S_q^{hist}
  end if
  Generate chunk q using the resulting streaming cache
  Perform the $t = 0$ cache-writing forward pass
  Cache the affine summary of chunk q
  Update camera-GDN and local convolutional states chronologically
end for

A.2 Implementation of State Recomposition

The affine state composition described in the main text is implemented directly on top of the original GDN kernels. After each chunk is completed, we cache its frame-wise Phase-A sufficient statistics, which characterize its affine effect on an arbitrary incoming recurrent state. For each query chunk, the statistics of the selected historical chunks are gathered in chronological order and passed to the original Phase-B forward scan to reconstruct the query-specific GDN state. This avoids re-evaluating historical chunks through the Transformer.

Selective recomposition is applied only to the main GDN matrix and normalization states. The camera-GDN recurrent state, GDN short-convolution state, and FFN temporal-convolution state retain their original chronological updates. We additionally batch compatible GDN blocks during recomposition to reduce kernel-launch and data-movement overhead without changing the underlying recurrence.

A.2.1 Overall Inference Procedure

Algorithm 1 summarizes the complete inference procedure of HLA-WM. The pairwise geometric overlap scores are first computed from the prescribed camera trajectory using Eq. 21. At each autoregressive step, we determine the recent-history budget according to the within-chunk trajectory change, select the historical chunks using Eq. 13, and recompose the corresponding main-GDN state. The current chunk is then generated using the resulting query-specific memory. After generation, an additional $t = 0$ forward pass caches the affine summary of the completed chunk for subsequent retrieval. Throughout this process, camera-control and local convolutional states retain their original chronological updates.

The retrieval procedure accesses only historical chunks, and the selected chunks are always recomposed in their original temporal order. Importantly, HLA-WM recomposes only the main-GDN recurrent memory; the camera-GDN recurrent state and local convolutional states follow the original streaming recurrence. Historical chunks are therefore never re-evaluated through the Transformer.

A.3 Component Ablation

Table 3 evaluates the major components of HLA-WM with $K = 1$ on the Hard-Trajectory split. Here, "w/o Select A" retains the transition terms A_i of all historical chunks while selectively composing their write terms, whereas "w/o Select B" retains all historical write terms B_i while selectively composing the transitions. Removing the sink degrades both revisit consistency and camera

controllability, highlighting the importance of persistent global context. Retaining the transition terms of unselected chunks substantially reduces PSNR and SSIM and increases camera errors, supporting our observation that irrelevant intermediate transitions contribute to long-range forgetting. Similarly, retaining all historical write terms consistently degrades both revisit fidelity and camera controllability, indicating that selective composition is beneficial for both transition and write components of the recurrent memory. Removing the turn-aware recent context also reduces revisit consistency and worsens camera control, confirming its role in preserving local continuity under abrupt camera motion. Overall, the full HLA-WM achieves the best performance across all six metrics, demonstrating the complementary contributions of selective state composition, persistent sink memory, and adaptive recent context.

Table 3: Component ablation on the Hard-Trajectory split using Stage 1 outputs. Best results are shown in bold.

Variant	Revisit Consistency			Camera Control		
	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
HLA-WM	**10.11**	**0.1997**	**0.5848**	**15.0887**	**1.8388**	**1.9482**
w/o Sink	9.71	0.1867	0.6019	15.9351	1.9326	2.0531
w/o Select A	9.61	0.1887	0.5901	15.2426	1.8552	2.0562
w/o Select B	9.64	0.1892	0.5921	15.2326	1.9830	2.0564
w/o Turn-aware	9.96	0.1917	0.5876	15.1784	2.1096	1.9899

A.4 GDN-State Recomposition and KV-Cache Alignment

We further isolate the roles of GDN-state recomposition and softmax KV-cache alignment under the Top-1 retrieval setting. *GDN-only* applies selective recomposition to the main GDN memory while retaining the native softmax KV history. *KV-only* applies the selected history only to the softmax KV cache while keeping the native GDN states. The full HLA-WM applies the same retrieved history to both memory pathways. Tables 4 and 5 report the results on SANA-WM-Bench and MBench-A, respectively.

The two memory pathways play complementary but asymmetric roles. At Stage 1, GDN-only already yields substantial gains in revisit consistency on both SANA-WM-Bench splits, whereas KV-only provides comparatively smaller improvements. Their combination further improves long-range recall, with the full HLA-WM achieving the best PSNR, SSIM, and LPIPS on both splits. KV alignment additionally improves camera stability, although its effect varies across trajectory types. Under causal AR refinement, the full method achieves the strongest revisit performance on both splits and improves all six metrics on Hard trajectories. These results indicate that GDN-state recomposition is the primary source of long-range memory improvement, while KV-cache alignment provides complementary stabilization of the attention-based context.

The bidirectional setting further illustrates the interaction between the two pathways. On Hard trajectories, GDN-only substantially improves SSIM but increases RotErr from 13.17° to 33.04°. This suggests that modifying only the recurrent history can create a mismatch with the accompanying softmax context. Applying the same historical selection to the KV cache restores consistency between the two memory pathways, and the full method achieves the best PSNR, LPIPS, and all three camera-control metrics. Thus, KV alignment is particularly important for maintaining stable generation when the recurrent history is selectively recomposed.

MBench-A exhibits an even clearer pattern. Across Stage 1, causal AR refinement, and bidirectional refinement, GDN-only consistently provides larger gains than KV-only, while the full HLA-WM achieves the best PSNR, SSIM, and LPIPS in every setting. This confirms that directly modifying the recurrent GDN memory is the main driver of improved long-range recall, whereas KV-cache alignment alone provides limited gains. Combining the two yields the strongest and most consistent overall performance.

Table 4: Ablation of GDN-state recomposition and KV-cache alignment on SANA-WM-Bench under Top-1 retrieval. Best results within each split and inference setting are shown in bold.

Pipeline	Method	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
Simple-Trajectory Split							
Stage 1	Baseline	9.18	0.1729	0.6327	19.7040	2.2625	2.3974
	GDN-only	9.45	0.1812	0.6219	17.9934	2.2832	2.4083
	KV-only	9.50	0.1866	0.6149	14.4283	**1.9569**	**2.0556**
	HLA-WM	**9.91**	**0.1999**	**0.6049**	**13.5864**	2.0928	2.1798
Stage 1 + AR Refine	Baseline	14.44	0.2770	0.5738	20.9315	**2.1587**	**2.3121**
	GDN-only	14.65	0.2784	0.5687	21.2344	2.3596	2.5147
	KV-only	13.88	0.2703	0.5812	20.9084	2.2087	2.3662
	HLA-WM	**14.82**	**0.2872**	**0.5682**	**20.2933**	2.1910	2.3240
Stage 1 + Bi. Refine	Baseline	13.04	0.2945	0.5786	11.3848	1.9893	2.0545
	GDN-only	13.26	0.2998	0.5746	10.7564	**1.8627**	1.9310
	KV-only	13.17	0.2999	0.5762	9.6939	1.8866	1.9417
	HLA-WM	**13.56**	**0.3148**	**0.5614**	**8.8823**	1.8637	**1.9127**
Hard-Trajectory Split							
Stage 1	Baseline	9.37	0.1774	0.6044	20.3759	1.9952	2.1507
	GDN-only	9.78	0.1890	0.5929	20.0838	1.9502	2.1075
	KV-only	9.52	0.1805	0.6000	**13.3992**	**1.7635**	**1.8616**
	HLA-WM	**10.11**	**0.1997**	**0.5848**	15.0887	1.8388	1.9482
Stage 1 + AR Refine	Baseline	14.12	0.2719	0.5621	25.6330	2.0099	2.2079
	GDN-only	14.35	0.2797	0.5576	23.1860	1.9519	2.1296
	KV-only	13.54	0.2599	0.5721	24.1145	2.0425	2.2302
	HLA-WM	**14.47**	**0.2812**	**0.5557**	**22.0165**	**1.9351**	**2.1043**
Stage 1 + Bi. Refine	Baseline	13.05	0.2972	0.5468	13.1713	1.6649	1.7694
	GDN-only	13.49	**0.3304**	0.5601	33.0406	2.1843	2.4584
	KV-only	13.26	0.3047	0.5469	11.4688	1.5860	1.6696
	HLA-WM	**13.51**	0.3142	**0.5384**	**10.5268**	**1.5449**	**1.6230**

Table 5: Ablation of GDN-state recomposition and KV-cache alignment on MBench-A under Top-1 retrieval. Best results within each inference setting are shown in bold.

Pipeline	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Stage 1	Baseline	10.17	0.2545	0.6251
	GDN-only	10.61	0.2626	0.6052
	KV-only	10.43	0.2606	0.6091
	HLA-WM	**11.01**	**0.2848**	**0.5823**
Causal AR Refine	Baseline	12.36	0.2996	0.6554
	GDN-only	12.57	0.3064	0.6492
	KV-only	12.35	0.3016	0.6547
	HLA-WM	**12.73**	**0.3122**	**0.6447**
Bidirectional Refine	Baseline	12.94	0.3310	0.5597
	GDN-only	13.28	0.3389	0.5477
	KV-only	13.04	0.3325	0.5563
	HLA-WM	**13.52**	**0.3431**	**0.5470**

A.5 Number of Retrieved Historical Chunks

We study the effect of the retrieval budget K on the Hard-Trajectory split using Stage 1 outputs. All variants use the same sink memory and turn-aware recent-context strategy, differing only in the number of historical chunks retrieved by geometric overlap. As shown in Table 6, performance remains broadly comparable across different retrieval budgets, and increasing K does not consistently improve either revisit fidelity or camera controllability. Notably, $K = 1$ achieves the best PSNR,

SSIM, and LPIPS while using the smallest retrieval budget. Larger values provide modest gains in some camera-control metrics, with $K = 5$ achieving the lowest RotErr and $K = 4$ the lowest TransErr and CamMC. These results indicate that a single highly relevant historical chunk is often sufficient for effective long-range recall. We therefore adopt $K = 1$ as the default setting for its favorable trade-off between retrieval quality and efficiency.

Table 6: Ablation on the retrieval budget K on the Hard-Trajectory split using Stage 1 outputs. Best results are shown in bold.

K	Revisit Consistency			Camera Control		
	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
1	**10.11**	**0.1997**	**0.5848**	15.0887	1.8388	1.9482
3	10.06	0.1973	0.5865	14.1537	1.7658	1.8661
4	10.00	0.1936	0.5885	12.2795	**1.6985**	**1.7832**
5	9.89	0.1886	0.5915	**11.5435**	1.7200	1.7970

A.6 Effectiveness of Geometry-Guided Historical Selection

We evaluate whether geometry provides an effective signal for selecting historical states by comparing HLA-WM with three geometry-agnostic selection strategies on Hard80 under Stage 1 inference. To more clearly expose the differences among selection strategies, we conduct this comparison with a larger retrieval budget of $K = 4$, allowing each method to operate over a broader historical context. *Random* deterministically samples up to six states from the available history, *Recent* retains the six most recent states, and *Uniform* selects six states approximately uniformly over the entire history. In contrast, HLA-WM retains a sink state and recent context while selecting distant historical chunks according to camera-frustum overlap. All variants use the same model, generation setting, seed, and evaluation protocol, covering 80 scenes and 400 revisit pairs. The three heuristic baselines use a fixed budget of at most six historical states, whereas HLA-WM normally uses six states but temporarily expands the recent context when the turn-aware criterion is triggered. We therefore regard this experiment as a comparison of historical selection strategies rather than a strictly equal-compute ablation.

As shown in Table 7, uniformly covering the history already provides a substantially stronger baseline than random or purely recent selection, highlighting the importance of maintaining access to distant context. Uniform selection remains competitive, likely because the one-minute evaluation horizon allows uniformly sampled states to cover much of the observed scene history. Nevertheless, geometry-guided selection further improves all three revisit-consistency metrics, increasing PSNR from 9.91 to 10.00 and SSIM from 0.1910 to 0.1936, while reducing LPIPS from 0.5957 to 0.5885. It also achieves slightly lower RotErr and CamMC, while Uniform obtains a marginally lower TransErr. These results indicate that explicitly selecting view-relevant historical states provides additional benefit beyond broad temporal coverage, supporting camera geometry as an effective criterion for long-range memory retrieval.

A.7 Effect of Revisit Interval

Figure 6 analyzes the effect of revisit interval on the improvement of HLA-WM. We group revisit pairs into short (< 10 s), medium (10–40 s), and long (≥ 40 s) intervals. Across Stage 1, AR refinement, and bidirectional refinement, the gains are generally larger for medium and long revisits. In particular, PSNR improvement increases consistently with the revisit interval across all three generation modes. LPIPS exhibits a similar trend, shifting from minor degradation at short intervals to clear improvement for long revisits. SSIM also shows substantially larger gains at medium or long intervals for Stage 1 and bidirectional refinement, although the trend under AR refinement is not monotonic. Overall, these results indicate that the benefit of selective long-range memory becomes more pronounced as the temporal distance between repeated observations increases, consistent with its effectiveness in mitigating long-horizon forgetting.

Table 7: Comparison of historical state selection strategies on Hard80 under Stage 1 inference with $K = 4$. Higher PSNR and SSIM are better, while lower LPIPS and camera errors are better.

Selection Policy	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
Random	9.70	0.1864	0.6049	14.1883	1.7822	1.8924
Recent	9.29	0.1696	0.6179	27.6876	2.0142	2.2408
Uniform	9.91	0.1910	0.5957	12.3324	**1.6881**	1.7838
HLA-WM	**10.00**	**0.1936**	**0.5885**	**12.2795**	1.6985	**1.7832**

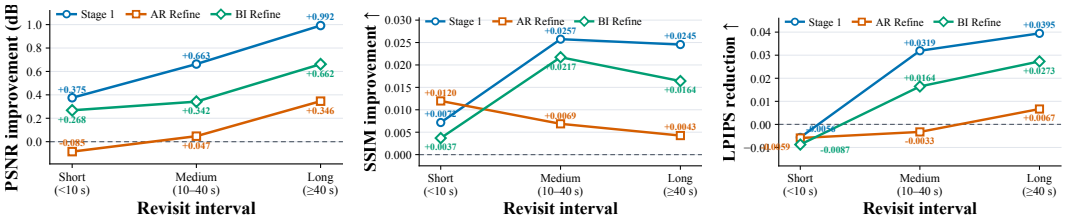

Figure 6: Performance improvement over SANA-WM across different revisit intervals on the Hard-Trajectory split. We report PSNR improvement (left), SSIM improvement (middle), and LPIPS reduction (right) for Stage 1, AR refinement, and bidirectional refinement.

A.8 Detailed Benchmark Results

We provide detailed category- and subset-level results to complement the aggregate results in the main paper. We first report the four scene categories of SANA-WM-Bench, followed by the four official subsets of MBench-A.

A.8.1 SANA-WM-Bench Category Results

SANA-WM-Bench contains four scene categories—Game Style, Indoor, Outdoor City, and Outdoor Nature—with 20 scenes per category in each trajectory split. Tables 8–10 report category-level results under Stage 1, causal AR refinement, and bidirectional refinement. The gains are broadly distributed across scene categories, with Stage 1 showing the most consistent improvements and downstream refinement introducing several category-dependent trade-offs.

Overall, the category-level results show that the improvements are not concentrated in a single scene type. Stage 1 provides the most consistent gains, while downstream refinement introduces several category-specific trade-offs. At the aggregate level, HLA-WM improves all three revisit-consistency metrics under each evaluated inference mode.

A.8.2 MBench-A Subset Results

MBench-A contains 547 samples across four official subsets: Causal, Human, Environment, and Object. For each sample, we first average its matched revisit pairs and then compute the macro average across samples. Table 11 reports the complete subset-level results under the three evaluated inference modes. Across all 12 mode–subset combinations, HLA-WM simultaneously improves PSNR, SSIM, and LPIPS, demonstrating consistent generalization across diverse scene and interaction types.

The subset-level results confirm that the aggregate improvements are not driven by a particular subset. Although the gain magnitude varies across subsets and refinement modes, HLA-WM improves all three revisit-consistency metrics in every evaluated MBench-A setting.

Table 8: Category-level results on SANA-WM-Bench using Stage 1 outputs. Best results within each category are shown in bold.

Category	Method	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
Simple-Trajectory Split							
Game Style	SANA-WM	9.85	0.2070	0.6197	16.5064	**1.9242**	**2.0482**
	HLA-WM	**11.06**	**0.2639**	**0.5866**	**12.8640**	2.0047	2.0939
Indoor	SANA-WM	9.38	0.2105	0.6576	38.6940	1.6993	2.0227
	HLA-WM	**10.05**	**0.2338**	**0.6306**	**29.0323**	**1.6196**	**1.8481**
Outdoor City	SANA-WM	8.96	0.1323	0.6390	12.2424	2.0782	2.1372
	HLA-WM	**9.25**	**0.1561**	**0.6240**	**5.9851**	**1.5976**	**1.6170**
Outdoor Nature	SANA-WM	8.52	0.1419	0.6145	11.3733	3.3483	3.3814
	HLA-WM	**9.29**	**0.1459**	**0.5785**	**6.4643**	**3.1493**	**3.1603**
Hard-Trajectory Split							
Game Style	SANA-WM	9.84	0.2023	0.5823	**14.2161**	1.8745	1.9719
	HLA-WM	**11.41**	**0.2550**	**0.5505**	14.3713	**1.8250**	**1.9349**
Indoor	SANA-WM	9.22	0.2085	0.6390	36.5361	1.2446	1.6297
	HLA-WM	**9.63**	**0.2305**	**0.6330**	**26.6061**	**1.0675**	**1.3277**
Outdoor City	SANA-WM	9.17	0.1463	0.6206	16.0176	1.8409	1.9254
	HLA-WM	**9.56**	**0.1611**	**0.6122**	**9.9119**	**1.6669**	**1.7070**
Outdoor Nature	SANA-WM	9.23	**0.1526**	0.5756	14.7337	3.0208	3.0760
	HLA-WM	**9.84**	0.1521	**0.5437**	**9.4654**	**2.7958**	**2.8233**

Table 9: Category-level results on SANA-WM-Bench after causal AR refinement. Best results within each category are shown in bold.

Category	Method	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
Simple-Trajectory Split							
Game Style	SANA-WM	15.19	0.3225	0.5582	19.8894	2.1729	2.3106
	HLA-WM	**15.79**	**0.3505**	**0.5492**	**16.5280**	**2.0506**	**2.1496**
Indoor	SANA-WM	13.64	0.3127	0.6013	35.3855	**1.5308**	**1.8652**
	HLA-WM	**14.08**	**0.3172**	**0.5956**	**34.3931**	1.6263	1.9040
Outdoor City	SANA-WM	14.21	0.2353	0.5803	15.1111	1.9558	2.0376
	HLA-WM	**14.26**	**0.2510**	**0.5777**	**14.4444**	**1.9104**	**1.9801**
Outdoor Nature	SANA-WM	14.70	**0.2375**	0.5555	**13.3400**	**2.9753**	**3.0350**
	HLA-WM	**15.13**	0.2302	**0.5503**	15.8077	3.1768	3.2625
Hard-Trajectory Split							
Game Style	SANA-WM	14.95	0.3220	0.5269	23.5136	**1.8901**	2.0707
	HLA-WM	**15.73**	**0.3398**	**0.5140**	**18.0932**	1.8923	**2.0223**
Indoor	SANA-WM	**13.81**	0.3003	0.5923	35.8487	1.2361	1.6201
	HLA-WM	13.71	**0.3139**	**0.5920**	**30.6444**	**1.1457**	**1.4865**
Outdoor City	SANA-WM	13.56	0.2507	**0.5812**	20.0231	1.8331	1.9407
	HLA-WM	**13.74**	**0.2570**	0.5877	**19.7262**	**1.8103**	**1.9228**
Outdoor Nature	SANA-WM	14.18	**0.2147**	0.5478	23.1465	3.0804	3.2001
	HLA-WM	**14.71**	0.2143	**0.5290**	**19.6023**	**2.8921**	**2.9858**

A.9 Additional Qualitative Results

A.9.1 Stage 1

Figure 7 provides additional qualitative comparisons for the Stage 1 outputs before downstream refinement. In the first example, the baseline exhibits clear drift in the revisit view, with noticeable changes in the mountain contours, foreground ridge shape, and cloud distribution. In contrast, HLA-WM better preserves the overall mountain silhouette, the foreground terrain layout, and the spatial relationship between the ridge and the cloud layer. In the second example, the baseline alters the street geometry upon revisit, leading to inconsistent crosswalk patterns, road markings, and building alignment. HLA-WM more faithfully maintains the width and orientation of the crosswalk, the road layout, and the arrangement of nearby buildings and trees. In the third example, the baseline

Table 10: Category-level results on SANA-WM-Bench after bidirectional refinement. Best results within each category are shown in bold.

Category	Method	PSNR ↑	SSIM ↑	LPIPS ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓
Simple-Trajectory Split							
Game Style	SANA-WM	13.74	0.3351	0.5485	9.7648	**1.7510**	**1.8012**
	HLA-WM	**14.20**	**0.3553**	**0.5377**	**6.8672**	1.8842	1.9156
Indoor	SANA-WM	12.67	0.3174	0.6045	21.6798	1.4978	1.6675
	HLA-WM	**13.03**	**0.3401**	**0.5980**	**16.1626**	**1.0450**	**1.1800**
Outdoor City	SANA-WM	12.79	0.2479	0.5767	7.1726	1.7286	1.7495
	HLA-WM	**13.05**	**0.2676**	**0.5644**	**6.6284**	**1.6204**	**1.6402**
Outdoor Nature	SANA-WM	12.97	0.2776	0.5849	6.9223	2.9796	2.9997
	HLA-WM	**13.94**	**0.2960**	**0.5456**	**5.8709**	**2.9053**	**2.9150**
Hard-Trajectory Split							
Game Style	SANA-WM	13.75	0.3443	0.5140	**8.1256**	**1.4874**	**1.5398**
	HLA-WM	**14.55**	**0.3797**	**0.4956**	9.1194	1.5393	1.5961
Indoor	SANA-WM	12.96	0.3148	0.5846	22.9948	0.9594	1.2115
	HLA-WM	**12.98**	**0.3248**	**0.5774**	**14.8387**	**0.7795**	**0.9455**
Outdoor City	SANA-WM	12.46	0.2631	0.5644	13.5222	1.6371	1.7186
	HLA-WM	**12.89**	**0.2822**	**0.5585**	**11.4256**	**1.5799**	**1.6459**
Outdoor Nature	SANA-WM	13.03	0.2667	0.5242	8.0424	2.5758	2.6078
	HLA-WM	**13.61**	**0.2702**	**0.5220**	**6.7236**	**2.2810**	**2.3046**

largely loses the original library-like structure and replaces it with a substantially different interior configuration. HLA-WM better preserves the bookshelf arrangement, corridor layout, desk region, and window placement, resulting in a revisit view that remains much closer to the first visit. In the fourth example, the baseline shows substantial changes in the classroom structure, including distortions in the wall geometry, seating layout, and front display region. By comparison, HLA-WM better maintains the overall room geometry, the arrangement of desks and chairs, and the long horizontal display panels along the wall. Overall, these examples show that HLA-WM improves long-range scene consistency already at Stage 1, before any subsequent refinement.

A.9.2 Stage 1 + AR Refine

Figure 8 provides additional qualitative comparisons under causal AR refinement. Across all examples, HLA-WM better preserves previously observed scene content when the camera revisits the same region after an extended rollout. In the first example, the baseline substantially changes the distant skyline and foreground layout upon revisit, whereas HLA-WM better preserves the overall scene structure and characteristic background content. In the second example, the baseline noticeably alters the foreground geometry and loses much of the original structural layout, while HLA-WM better maintains the major scene elements and their spatial arrangement. In the third example, the baseline largely loses the distinctive foreground structure and changes the arrangement of surrounding objects, whereas HLA-WM more faithfully preserves these characteristic elements and their relative geometry. In the fourth example, HLA-WM better maintains the spatial arrangement of tables, people, and illuminated structures, while the baseline exhibits substantial changes in the revisited scene. Overall, these examples demonstrate improved preservation of both global scene structure and distinctive visual content over long-range revisits.

A.9.3 Stage 1 + Bidirectional Refine

Figure 9 provides additional qualitative comparisons under bidirectional refinement. In the first example, the baseline exhibits noticeable drift in the corridor geometry upon revisit, including changes in the floor pattern, wall modules, and spatial arrangement of the side structures. In contrast, HLA-WM better preserves the corridor layout, perspective, and characteristic interior elements. In the second example, the baseline substantially alters the lake surface, shoreline, and surrounding landscape, with pronounced distortions appearing in the revisited view. HLA-WM more faithfully maintains the mountain silhouette, tree line, shoreline geometry, and their reflections on the water. In the third example, the baseline introduces a large wooden structure that is absent from the first visit and

Table 11: Detailed Top-1 results on the four official MBench-A subsets under different inference modes. Best results within each subset and mode are shown in bold.

Subset	n	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Stage 1					
Causal	100	SANA-WM	10.4068	0.28982	0.65461
		HLA-WM	**11.0022**	**0.31830**	**0.61659**
Human	120	SANA-WM	9.5147	0.26845	0.67977
		HLA-WM	**10.2161**	**0.30275**	**0.63759**
Environment	227	SANA-WM	10.4140	0.24056	0.59436
		HLA-WM	**11.4626**	**0.27144**	**0.54857**
Object	100	SANA-WM	10.1755	0.23386	0.59953
		HLA-WM	**10.9412**	**0.25989**	**0.55808**
Causal AR Refine					
Causal	100	SANA-WM	12.4358	0.33422	0.67586
		HLA-WM	**12.6230**	**0.34587**	**0.66550**
Human	120	SANA-WM	11.8551	0.30453	0.70110
		HLA-WM	**12.2041**	**0.31959**	**0.68707**
Environment	227	SANA-WM	12.6355	0.29205	0.62765
		HLA-WM	**13.0913**	**0.30450**	**0.61762**
Object	100	SANA-WM	12.2440	0.27644	0.64322
		HLA-WM	**12.6283**	**0.28724**	**0.63458**
Bidirectional Refine					
Causal	100	SANA-WM	12.8197	0.36602	0.58537
		HLA-WM	**13.2118**	**0.37457**	**0.57395**
Human	120	SANA-WM	12.0657	0.33279	0.62407
		HLA-WM	**12.5213**	**0.34830**	**0.60919**
Environment	227	SANA-WM	13.3562	0.32284	0.52598
		HLA-WM	**14.1215**	**0.33725**	**0.51069**
Object	100	SANA-WM	13.1585	0.31224	0.53320
		HLA-WM	**13.6448**	**0.31892**	**0.52774**

significantly changes the foreground composition. HLA-WM avoids this spurious structural change and better preserves the open-water layout, ice distribution, and distant mountain ridge. In the fourth example, the baseline shows considerable changes in the rooftop geometry, foreground structures, and distant skyline, whereas HLA-WM better retains their relative arrangement and the overall scene identity. Together, these examples further demonstrate that selective long-range memory retrieval improves the preservation of both global geometry and distinctive scene content during bidirectional refinement.

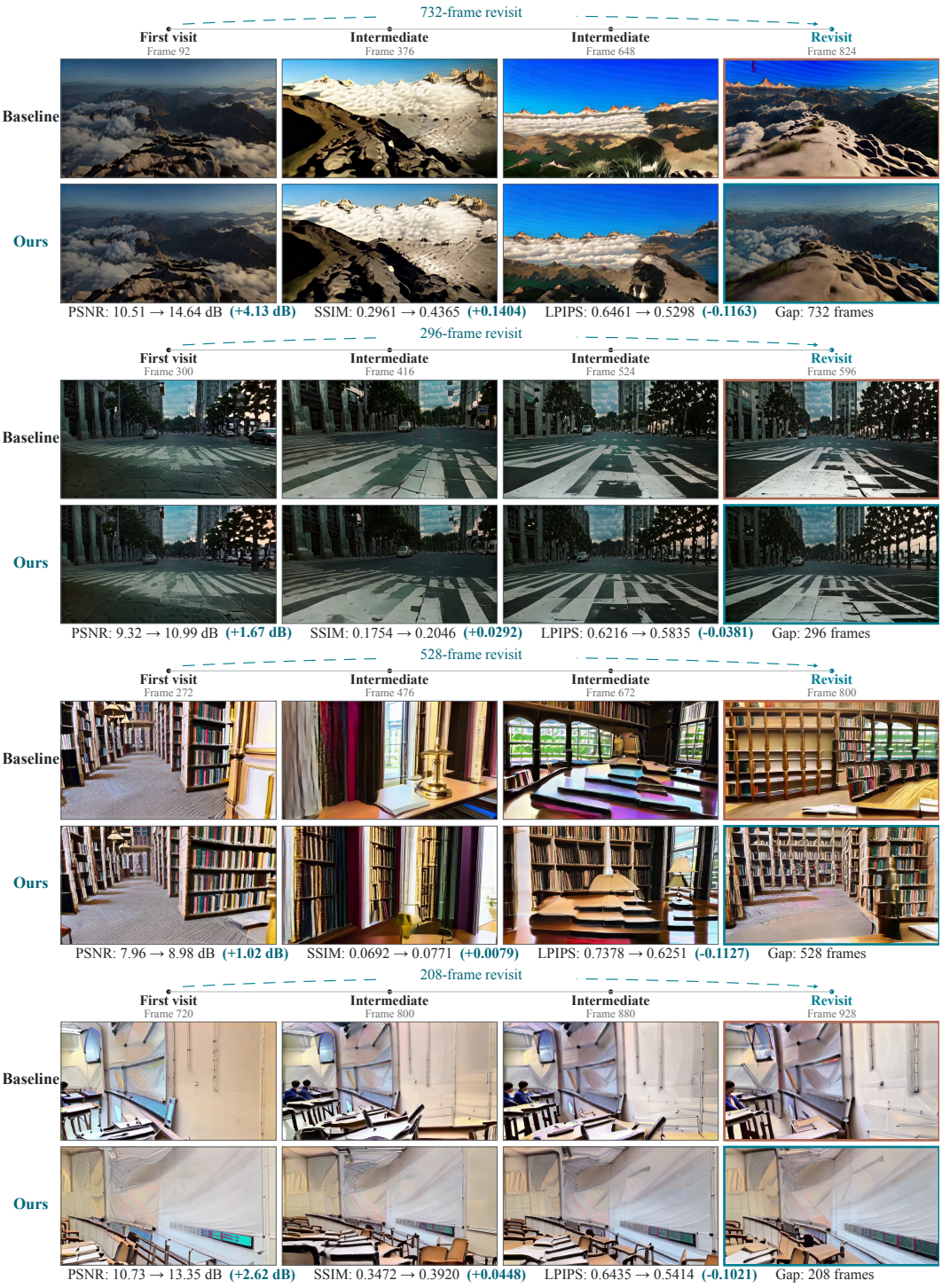

Figure 7: Additional qualitative comparisons for Stage 1 outputs. Each example shows the first visit, two intermediate views, and a later revisit along the same camera trajectory. Revisit consistency is evaluated between the first-visit and revisit frames within each method.

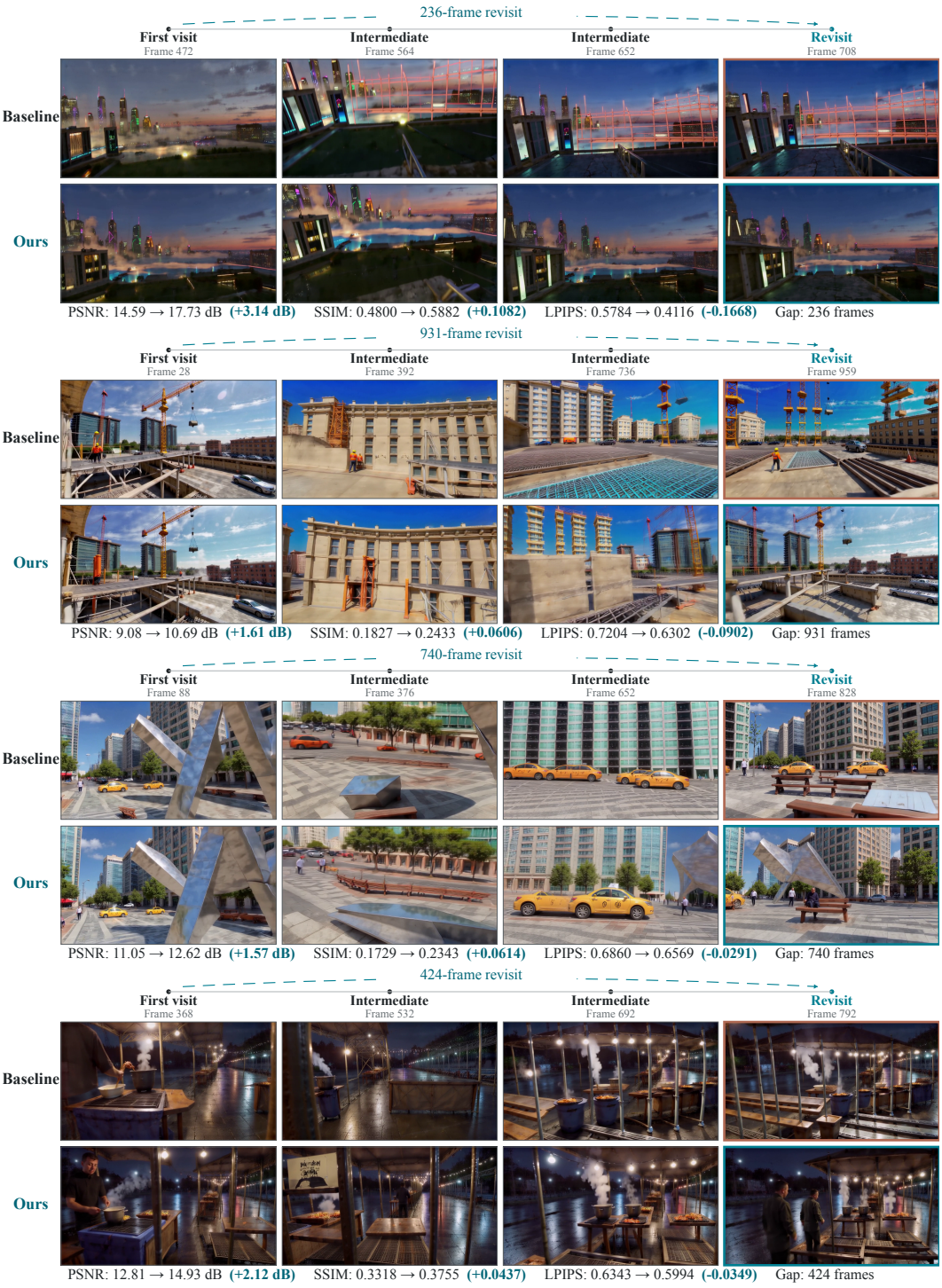

Figure 8: Additional qualitative comparisons under causal AR refinement. Each example shows the first visit, two intermediate views, and a later revisit along the same camera trajectory. Revisit consistency is evaluated between the first-visit and revisit frames within each method.

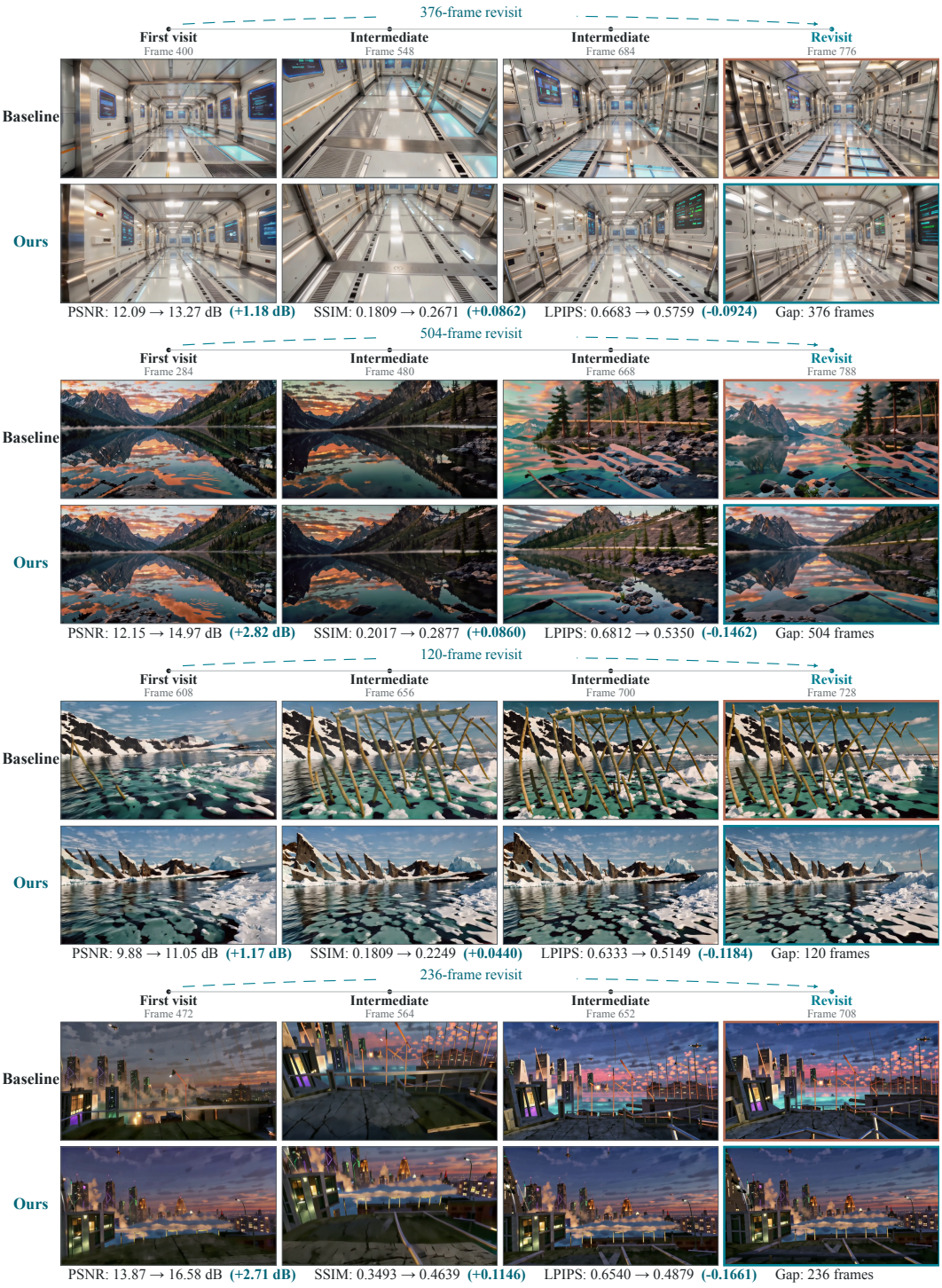

Figure 9: Additional qualitative comparisons under bidirectional refinement. Each example shows the first visit, two intermediate views, and a later revisit along the same camera trajectory. Revisit consistency is evaluated between the first-visit and revisit frames within each method.