TOKENWEAVE: CONSOLIDATING TEMPORAL KV HISTORY INTO PERSISTENT SCENE MEMORY FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming autoregressive video world models naturally accumulate history as a temporal sequence of key–value (KV) states. However, long interactions repeatedly observe the same scene content across distant times, making such temporally organized history increasingly redundant and eventually forcing old observations to be discarded. We introduce TOKENWEAVE, a training-free framework that reorganizes temporal KV history into persistent scene memory by associating and consolidating repeated patch-level observations. Rather than managing history as a collection of separate observations, we reorganize repeated observations around recurring scene content. It operates entirely on the frozen model's internal KV states, grouping observations that likely correspond to the same scene content and integrating them into persistent memory entries. These shared entries are updated with position- and multiplicity-aware fusion, while genuinely novel content is retained as new memory. Memory therefore grows mainly with new or unmatched content rather than with time. Experiments across multiple domains and rollout lengths demonstrate substantial improvements in long-horizon revisit consistency without memory-specific training. Additional videos are available at https://tokenweave-iclr.github.io/.

1 INTRODUCTION

Video world models are increasingly used for interactive simulation, embodied agents, controllable content creation, and open-ended virtual exploration (Bruce et al., 2024; Hafner et al., 2025; He et al., 2025). During long interactions, however, the same scene content may be observed repeatedly at widely separated times. Standard streaming autoregressive (AR) models record these observations as a growing temporal sequence of key–value (KV) states (Huang et al., 2025; Liu et al., 2026), even when multiple entries correspond to the same persistent part of the world. As a result, long-term memory naturally accumulates substantial redundancy, while bounded-context models must eventually discard earlier observations altogether.

Existing methods mainly address this problem by managing the accumulated history. Sliding-window models simply evict old states. Retrieval-based methods such as WorldKV (Yi et al., 2026) and WorldCam (Nam et al., 2026) preserve historical observations and recover relevant ones using camera pose, action correspondence, or learned features (Li et al., 2026). Learned memory methods (Hong et al., 2025; Wu et al., 2026) instead compress or distill history into a compact representation. These strategies differ in how history is stored, retrieved, or compressed, but they largely retain the same underlying abstraction: past observations are first accumulated as historical evidence and only later selected or summarized.

We take a different perspective. In a persistent world, repeated observations of the same scene content should not necessarily remain separate memory entries. Instead, they can be treated as repeated evidence of a shared underlying entity or region. This suggests reframing long-term world-model memory from a problem of ***managing a temporal history*** into ***content-centric memory organization***: when a new historical observation arrives, the model should determine whether it corresponds to previously stored content and, if so, integrate it into the existing memory rather than storing another

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.

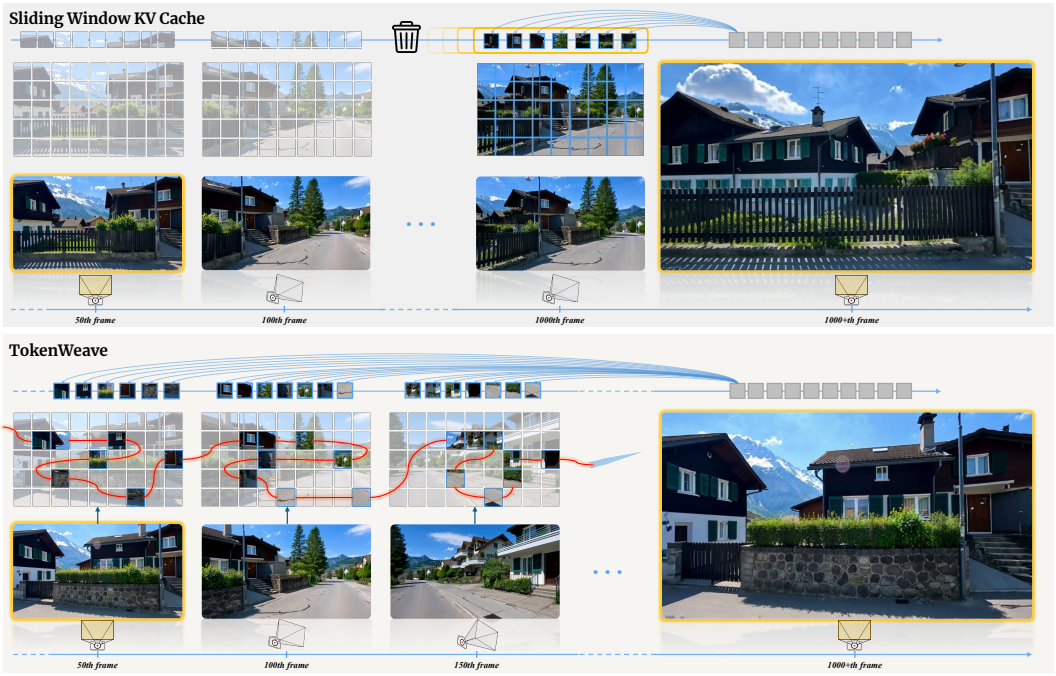

Figure 1: (Top) A standard sliding-window KV cache completely discards tokens once they leave the window, limiting the model to only a few seconds of recent context and causing previously observed content to be forgotten. (Bottom) In contrast, TOKENWEAVE weaves related historical tokens into a compact, persistent memory, preserving scene information across long trajectories and enabling consistent recall over minute-long time scales.

redundant copy. Under this view, memory expands with scene novelty rather than simply with time, aligning memory growth with the structure of the observed world.

In this work, we introduce TOKENWEAVE, a training-free method for transforming temporal KV history into a persistent scene-organized memory. As illustrated in Figure 1, instead of treating observations from different times as independent historical entries, TOKENWEAVE identifies patch-level observations that correspond to the same underlying scene content and consolidates them into shared persistent memory slots. By weaving related observations across time, TOKENWEAVE reorganizes history around persistent scene content rather than observation order. This turns long-term memory from temporal accumulation into scene-based consolidation, while providing a finer-grained memory representation than frame- or chunk-level retrieval.

TOKENWEAVE performs this reorganization directly in the frozen model's KV space. It leaves the recent context unchanged and processes only tokens that are about to leave the active window. For each outgoing patch token, we use camera pose as a soft spatial prior to identify plausible memory candidates, and rely on the model's native keys to verify content identity. Verified matches are fused through RoPE-aligned K/V updates, while a multiplicity-aware correction preserves the attention mass of consolidated observations. Tokens without a reliable match initialize new slots, so repeated content is progressively absorbed into persistent memory while novel content expands it. We evaluate our method on long-horizon scene revisitation across multiple domains and rollout lengths. TOKENWEAVE consistently improves revisit consistency on frozen backbones and matches or surpasses trained memory systems on alignment-aware scene-consistency metrics, despite requiring no memory-specific training. Our contributions are threefold:

- We reformulate long-term world-model memory as consolidating repeated scene observations, rather than accumulating historical entries.
- We introduce TOKENWEAVE, a training-free, patch-level method that reorganizes KV history into persistent scene memory through pose-guided key matching.
- We demonstrate strong long-horizon scene consistency across domains, with consistent gains on frozen backbones and competitive results against trained memory systems.

2 Related Work

Autoregressive video generation. Recent video diffusion models increasingly adopt causal or autoregressive generation to support long rollouts and real-time interaction (Chen et al., 2024; Yin et al., 2025; He et al., 2025). Self Forcing (Huang et al., 2025) reduces the train–test gap by training on self-generated histories, while Rolling Forcing (Liu et al., 2026) jointly denoises a moving frame window and retains attention sinks for stable streaming generation. These methods make interactive generation practical by reusing KV states within a bounded context. Their rolling caches, however, primarily preserve local temporal continuity: content that falls outside the active window is no longer directly accessible. Our work is complementary to these generation strategies and focuses on extending their memory beyond the recent context without modifying the generative backbone.

Long-term memory for video world models. Existing memory mechanisms can be broadly grouped by how they represent and recover the past. Explicit spatial memories reconstruct scene content in 3D and condition future generation through geometric reprojection (Wu et al., 2025); for example, EvoWorld (Wang et al., 2025) evolves an explicit 3D reconstruction during exploration. Such representations provide a strong spatial anchor, but their quality is coupled to depth, pose, and reconstruction accuracy, and static geometry is difficult to reconcile with dynamic scene changes. Retrieval-based approaches instead preserve historical observations and recall a subset when needed (Xiao et al., 2025; Yu et al., 2025). WorldPlay (Sun et al., 2026) reconstructs context from temporal and spatial memories, WorldCam (Nam et al., 2026) indexes observations by global camera pose, and WorldKV (Yi et al., 2026) retrieves evicted KV chunks using camera or action correspondence. I3DM (Li et al., 2026) improves view selection and injection with features from a pretrained novel-view synthesis model. These methods retain detailed observations, but the frame or chunk is a coarse memory unit and performance depends on selecting the right history. Learned compact memories take another route (Oshima et al., 2025; Yu et al., 2026; Peng et al., 2026): RELIC (Hong et al., 2025) encodes compressed camera-aware historical tokens, while Infinite-World (Wu et al., 2026) recursively distills history into a fixed-budget hierarchical representation. They demonstrate the value of learned memory, but rely on memory-aware training, distillation, or fine-tuning. TokenWeave pursues compact long-term memory from a complementary, training-free perspective: it consolidates the frozen model's historical KV states at patch granularity, without first reconstructing the scene or learning a separate memory representation.

KV-cache compression. Token and KV-cache compression have become effective tools for scaling long-context language models, where substantial context can be discarded or summarized while preserving task performance (Xiao et al., 2024; Li et al., 2024; Zhang et al., 2023; 2024). This perspective is particularly promising for video generation. Unlike text, consecutive video frames densely sample a continuously evolving scene: neighboring patches are spatially correlated, adjacent frames often differ only by small motion, and revisiting a location creates repeated observations over much longer intervals. Historical video tokens therefore contain pronounced spatial, temporal, and cross-visit redundancy. Recent video methods begin to exploit this structure. Forcing-KV (Ji et al., 2026) combines head-aware structured pruning with similarity-based token pruning. FlowCache (Ma et al., 2026) combines chunk-wise feature caching with importance–redundancy-aware KV compression for efficient autoregressive video generation. These approaches primarily adapt KV compression to the computational structure of video generation. TokenWeave further adapts it to the memory requirements of interactive world models: it uses camera pose or action as a spatial prior for deciding which historical tokens may describe the same scene content, and optimizes the compressed cache for long-term revisitation rather than generation efficiency alone. It thus consolidates related tokens across distant frames into a continuously updated memory organized around persistent scene content.

3 Method

An autoregressive world model records observations in temporal order, yet the same wall, building, or object can recur at widely separated times. Our insight is that *long-term memory should be organized around persistent scene content*. We introduce TokenWeave, a training-free framework that maps historical KV states to persistent slots, each accumulating observations associated with the same scene region. The unit of memory becomes scene content shared across time. Repeated observations update existing slots; compression follows naturally from this scene-centric organization.

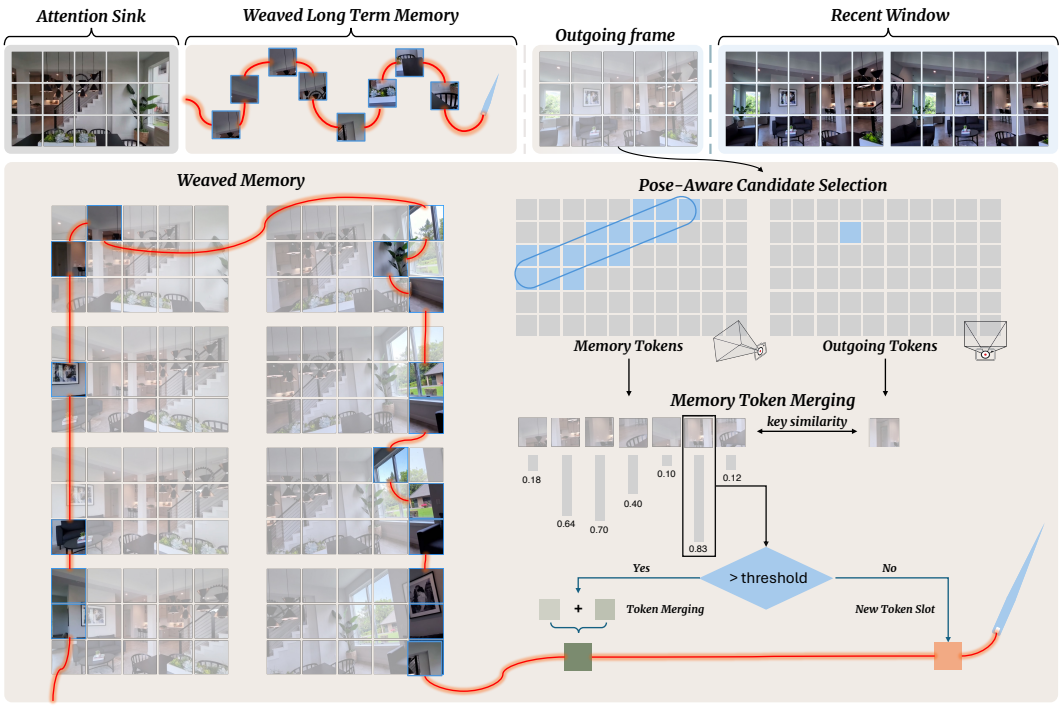

Figure 2: **Scene-centric memory organization in TOKENWEAVE.** (Top) Observations leaving the recent window enter persistent scene memory. (Bottom left) Related observations across time share a memory slot. (Bottom right) Pose-guided candidates and native-key similarity associate each outgoing token with an existing slot or register new scene content.

3.1 FROM TEMPORAL HISTORY TO SCENE-CENTRIC MEMORY

We consider a frozen streaming video transformer with L layers and $P = H_p W_p$ patch tokens per latent frame. At generation step t, its cache is

$$\mathcal{C}_t = [\mathcal{S},\, \mathcal{M}_t,\, \mathcal{R}_t], \tag{1}$$

where $\mathcal{S}$ contains attention sinks, $\mathcal{R}_t$ is the native recent window, and $\mathcal{M}_t$ is persistent scene memory. Temporal order remains useful for short-term motion, so sinks and recent states retain their original form. When a frame leaves the window, its tokens are associated with scene slots in $\mathcal{M}_t$: matched observations refresh existing slots, and unmatched observations register new ones (Figure 2). Memory allocation thus follows newly encountered or unmatched content, while repeated observations can share storage across long temporal gaps.

Each slot j stores layer-wise states $\{m_j^{K,\ell}, m_j^{V,\ell}\}_{\ell=1}^{L}$ and metadata $(p_j, r_j, a_j, w_j, n_j)$: its registration patch and frame, latest observation time, averaging weight, and number of represented tokens. These are patch-level scene associations expressed directly in the model's feature space, requiring no explicit 3D reconstruction or learned memory encoder. The same slot assignment is used across all layers, and future generation reads the slots through the backbone's existing attention mechanism.

3.2 CROSS-TIME SCENE ASSOCIATION

For an outgoing token at patch p in frame f, the key decision is whether it observes content already represented in memory. A false association mixes distinct scene regions; a missed association stores another observation of the same region. We combine camera geometry, which identifies plausible correspondences, with native keys, which verify their visual compatibility.

Pose-guided candidate selection (PGCS). Let $(R, \mathbf{t})$ map camera f to the registration camera r_j, K be the intrinsics expressed in patch-grid coordinates, and x the current-frame ray normalized to $x_z = 1$. At inverse depth $\zeta = 1/z$, its projection is $\pi(A + \zeta B)$, where $A = KRx$, $B = K\mathbf{t}$, and

$\pi(v) = v_{xy}/v_z$. Starting at $x_\infty = \pi(A)$, depth uncertainty induces displacement along the unit epipolar direction $u \propto B_{xy}A_z - A_{xy}B_z$. Using a scale-free near-depth prior $\rho = Z_{\min}/\|\mathbf{t}\|$ for nonzero translation, we estimate the parallax bound over depths $z \geq Z_{\min}$ as

$$d_{\max}(x) \approx \frac{\|B_{xy}A_z - A_{xy}B_z\|}{Z_{\min}A_z\,(A_z + B_z/Z_{\min})}. \tag{2}$$

For valid projections, we admit slots inside a capsule around this segment, together with a local fallback for imperfect camera control:

$$\begin{aligned}\mathcal{G}(p) = \big\{j : \operatorname{dist}\big(p_j, \operatorname{seg}[x_\infty, x_\infty + d_{\max}u]\big) \leq b(f - r_j)\big\} \\ \cup \big\{j : \|p_j - p\|_\infty \leq r\big\}, \qquad b(\Delta) = b_0 + \gamma \min(\Delta, \Delta_{\max}).\end{aligned} \tag{3}$$

Here r is the local fallback radius, b_0 the initial capsule width, γ its growth rate, and $\Delta_{\max}$ the maximum age used to widen it. The bound assumes $A_z > 0$ and $A_z + B_z/Z_{\min} > 0$, so the depth interval projects in front of the registration camera. Invalid projections are excluded from the geometric gate. With zero parallax, the segment reduces to x_∞ without evaluating u. Pose only restricts candidate associations and never overwrites the stored features.

Native-key association. To compare observations across times and image locations, we remove both spatial and temporal rotary position embedding (RoPE) phases (Su et al., 2021): $\bar{k}^{\ell,h}_{f,p} = \mathrm{RoPE}^{-1}_{f,p}(k^{\ell,h}_{f,p})$ and $\bar{m}^{\ell,h}_j = \mathrm{RoPE}^{-1}_{a_j,p_j}(m^{K,\ell,h}_j)$, where $\mathrm{RoPE}_{t,p}$ denotes the full spatiotemporal rotary encoding at frame t and patch p, and h indexes heads. Matching is performed on these unrotated keys; positional phases are reapplied only after association, when preparing the keys for consolidation. Aggregating cosine similarity over selected layers and heads gives

$$s_{p,j} = \operatorname*{Agg}_{\ell,h} \frac{\langle \bar{k}^{\ell,h}_{f,p}, \bar{m}^{\ell,h}_j \rangle}{\|\bar{k}^{\ell,h}_{f,p}\|\,\|\bar{m}^{\ell,h}_j\|}, \qquad j^* = \arg\max_{j \in \mathcal{G}(p)} s_{p,j}. \tag{4}$$

We associate the token with j^* only if $s_{p,j^*} > \tau$; an empty candidate set or a failed match creates a new slot. Thus, temporally distant tokens can share memory when both spatial plausibility and native features support the association.

3.3 Maintaining Persistent Scene Memory

Stable anchors, evolving observations. A persistent slot must retain a consistent spatial reference as its appearance is refreshed. We separate an immutable *geometric clock* r_j, paired with registration patch p_j, from a mutable *appearance clock* a_j used for temporal RoPE. Updating a_j leaves (p_j, r_j) intact: a later observation can arrive at another image location, so pairing the original patch with the new camera would invalidate future geometric associations. These two clocks let scene memory evolve while preserving its registration anchor.

Observation consolidation. For a matched token, we reapply RoPE to both unrotated keys at spatial anchor p_j and temporal position f. Denote the aligned outgoing key by $\widetilde{k}^{\ell}_{f,p\to j}$ and the temporally restamped slot key by $\widetilde{m}^{K,\ell}_{j\to f}$. We update every layer as

$$\begin{aligned} m^{K,\ell}_j &\leftarrow \frac{w_j \widetilde{m}^{K,\ell}_{j\to f} + \widetilde{k}^{\ell}_{f,p\to j}}{w_j + 1}, \\ m^{V,\ell}_j &\leftarrow \frac{w_j m^{V,\ell}_j + v^{\ell}_{f,p}}{w_j + 1}. \end{aligned} \tag{5}$$

We then set $a_j \leftarrow f$, $w_j \leftarrow \min(w_j + 1, w_{\max})$, and $n_j \leftarrow n_j + 1$. Capping w_j allows accumulated features to adapt to subsequent observations. An unmatched token is stored verbatim with $(p_j, r_j, a_j, w_j, n_j) = (p, f, f, 1, 1)$.

Evidence-aware memory reading (EMR). A slot can represent many observations, but ordinary softmax attention counts it only once. Following proportional attention (Bolya et al., 2023), we retain its uncapped multiplicity n_j and use

$$\ell_{q,j} = \frac{\langle q, m^K_j \rangle}{\sqrt{d}} + \alpha \log n_j, \tag{6}$$

Table 1: **Quantitative comparison on the long-horizon revisit benchmark.** wPSNR and wSSIM denote warped PSNR and SSIM, respectively. Best and second-best results in each column are shown in **bold** and underline, respectively; Δ is the change from the corresponding unmodified backbone. TOKENWEAVE improves all four metrics on both backbones without additional training.

Method	25s / 400 frames				45s / 720 frames				60s / 960 frames			
	PSNR ↑	SSIM ↑	wPSNR ↑	wSSIM ↑	PSNR ↑	SSIM ↑	wPSNR ↑	wSSIM ↑	PSNR ↑	SSIM ↑	wPSNR ↑	wSSIM ↑
Without Memory Training												
LingBot-World-v2	10.5271	0.3347	11.2603	0.4025	9.2543	0.2293	8.6500	0.3140	9.3678	0.2436	9.2848	0.3398
+TOKENWEAVE	12.8546	0.3420	**17.3594**	**0.5473**	12.2026	0.3270	**16.4742**	**0.5398**	12.0081	0.3226	15.1799	**0.4880**
Δ vs. base	+2.3275	+0.0073	+6.0991	+0.1448	+2.9483	+0.0977	+7.8242	+0.2258	+2.6403	+0.0790	+5.8951	+0.1482
LingBot-World-v1	10.9880	0.2907	11.8777	0.3574	9.5291	0.2502	8.9157	0.2796	9.3549	0.2467	9.3307	0.2803
+TOKENWEAVE	12.9176	0.3484	16.5737	0.5014	12.0522	0.3150	13.8774	0.4176	11.3166	0.2753	13.4967	0.3702
Δ vs. base	+1.9296	+0.0577	+4.6960	+0.1440	+2.5231	+0.0648	+4.9617	+0.1380	+1.9617	+0.0286	+4.1660	+0.0899
WorldKV	10.9347	0.2467	11.7949	0.3286	9.9749	0.2091	10.2098	0.2676	9.4458	0.1854	9.3125	0.2586
With Memory Training												
Infinite-World	13.9309	0.3115	17.3213	0.4825	**14.5645**	**0.3895**	16.1538	0.4232	**14.3310**	0.3748	15.9188	0.4562
HY-WorldPlay	**14.6285**	**0.4075**	16.4048	0.4653	13.5140	0.3840	15.4515	0.4372	13.9593	**0.3962**	**15.9295**	0.4567

where q is a query, d the key dimension, and α the correction strength. At $\alpha = 1$, the correction exactly preserves the attention mass of n_j identical keys; for consolidated observations, it approximates their aggregate contribution. Thus, w_j controls feature adaptation, while n_j preserves the amount of evidence represented by a slot.

Online inference and cost. After each generated chunk, outgoing frames are processed once in temporal order, making updated slots available to subsequent frames. The backbone remains frozen; without camera information, association uses the local gate. For M_t slots and L_m matching layers, our dense implementation costs $O(L_m P M_t d)$ per outgoing frame, and generation attends to $|\mathcal{S}| + M_t + |\mathcal{R}_t|$ tokens. Repeated observations reduce cache growth through slot reuse. Memory can still grow with unmatched content: the implementation allocates a capacity and requires the rollout to fit within it, without an additional eviction policy.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Long-horizon revisit benchmark. We construct a cross-domain benchmark from 165 initial frames drawn from MBench (Zhang et al., 2026), ABot-World-0 (Jiang et al., 2026), and SceneFly (Peng et al., 2026), covering real photographs, game imagery, and Unreal Engine renderings across indoor and outdoor environments. Each image is paired with six candidate revisit sequences across three durations (25, 45, and 60 seconds), giving 990 candidates. To account for differences in camera control, we calibrate commands for each model and select one candidate per duration with sufficient overlap between visits. This yields $165 \times 3 = 495$ scene–sequence pairs per model; each backbone and its TOKENWEAVE variant use identical selected pairs.

Revisit pairs and metrics. Each pair $(I^{(1)}, I^{(2)})$ contains two visits to an overlapping view, separated long enough for the first observation to leave the recent context. Neither is the sink-protected input frame, ensuring that the comparison tests memory beyond the active window. We report PSNR and SSIM (Wang et al., 2004), together with alignment-aware variants that account for residual view misalignment. Let $\mathcal{W}(I^{(2)}; I^{(1)})$ be the image-space warp from the second visit to the first and Ω its valid overlap mask. For a metric D,

$$D_{\text{warp}} = D\Big(I^{(1)}|_{\Omega}, \mathcal{W}(I^{(2)}; I^{(1)})|_{\Omega}\Big). \tag{7}$$

Direct scores measure end-to-end revisit behavior, whereas warped scores focus on appearance and structural consistency after compensating for small view offsets. We apply the same overlap criterion to all methods and aggregate first within each trajectory and then across scenes. Revisit-pair construction and warping details are provided in Appendix B.3–B.4.

Compared systems. We compare five base systems that differ in how they retain and access past observations. LingBot-World-v1 (Robbyant Team et al., 2026) and LingBot-World-v2 (Gao et al.,



Figure 3: **Qualitative comparison over a 45-second revisit trajectory.** Rows show TOKENWEAVE, LingBot-World-v2, HY-WorldPlay, Infinite-World, and WorldKV. Red boxes mark the first-visit and revisit frames.

2026) are both 14B-parameter autoregressive models that use a sliding context window. WorldKV (Yi et al., 2026) retrieves historical keyframes, and Infinite-World (Wu et al., 2026) encodes the past in an implicit memory. HY-WorldPlay (Sun et al., 2026) reconstructs context from temporal and spatial memories. We integrate TOKENWEAVE into both LingBot-World backbones and compare each augmented model with its unmodified counterpart to assess whether the same training-free memory mechanism improves revisit consistency across backbones. Each baseline uses its public checkpoint and the default inference settings provided by its official repository.

4.2 QUANTITATIVE COMPARISON

Table 1 groups systems by whether their memory mechanism requires training. The within-backbone comparisons isolate the effect of adding TOKENWEAVE, while comparisons across systems also reflect differences in camera control and generation capability.

Consistent training-free memory gains across backbones. Adding TOKENWEAVE improves all four metrics on both LingBot-World-v1 and LingBot-World-v2 at 25, 45, and 60 seconds: all 24 within-backbone comparisons yield positive gains. Warped PSNR gains are 4.17–4.96 dB on v1 and 5.90–7.82 dB on v2. At 60 seconds (960 frames), warped SSIM still improves by 0.0899 and 0.1482, respectively.

Residual camera error still affects warped scores. For v2 + TOKENWEAVE at 45 seconds, alignment raises PSNR from 12.2026 to 16.4742 and SSIM from 0.3270 to 0.5398. The smaller changes for Infinite-World and HY-WorldPlay suggest that direct scores understate TOKENWEAVE's scene consistency when camera-control error remains. However, warping only partially compensates for these errors: imperfect registration and viewpoint-dependent effects can leave residual misalignment even within the valid overlap. These residuals still penalize PSNR and SSIM, so even the warped scores can underestimate TOKENWEAVE's scene consistency. They should therefore be interpreted alongside the video results, rather than as measurements of memory quality alone.

Comparison with trained memory systems. In our evaluated rollouts, LingBot-World revisit frames can retain pose offsets, whereas Infinite-World and HY-WorldPlay exhibit smoother camera control. Direct scores reflect both viewpoint mismatch and scene-content changes; warping reduces the contribution of view misalignment. After alignment, v2 + TOKENWEAVE achieves the highest warped SSIM at all three durations and the highest warped PSNR at 25 and 45 seconds (Table 1). At 60 seconds, trained memory systems retain higher warped PSNR. On v1, TOKENWEAVE ranks second in warped SSIM at 25 seconds, ahead of both trained memory systems.



Figure 4: **Qualitative comparison over a 25-second revisit trajectory.** Row order and red-box annotations follow Figure 3.

4.3 Qualitative Comparison

We encourage reviewers to visit our *anonymous project website* for video results, which complement the frame comparisons below. Figures 3 and 4 show street and snowy residential revisits at 45 and 25 seconds, respectively. Because camera control differs across methods, we compare first-visit and revisit frames within each row; these examples assess end-to-end revisit behavior rather than memory quality alone. TokenWeave preserves the street layout and building appearance in Figure 3, and the house's dark wooden facade, roof shape, and window arrangement in Figure 4.

For the 45-second coastal trajectory with parallax, occlusion, and newly revealed regions (Figure 5), TokenWeave retains the wooden shelter, foreground trees, and ocean background after exploring the lighthouse and coastline. Baseline revisit frames show substantial changes relative to their own first visits.

4.4 Analysis and Ablation

Full-history KV cache. Table 2 compares TokenWeave with sliding-window and full-history caches; the latter retains all generated K/V tokens and attends over the entire history. Compared with full history, TokenWeave improves direct scores at 25 seconds with slightly lower warped scores; at 45 seconds, PSNR improves, SSIM decreases slightly, and warped scores are nearly unchanged. At 60 seconds, all four metrics improve, including 1.39 dB in PSNR and 0.0231 in warped SSIM. We hypothesize that consolidation mitigates *attention dispersion*, in which weakly relevant tokens dilute attention to useful evidence as history grows (Vasylenko et al., 2026). Across the three durations, KV-cache memory is reduced by 60.5–68.2%, with 1.16–1.40× throughput.

Table 2: **KV-cache quality–efficiency trade-off.**

Cache	PSNR↑	SSIM↑	wPSNR↑	wSSIM↑	KV (GiB)↓	FPS↑
25s / 400 frames						
Full history	12.42	0.3289	**17.60**	**0.5563**	118.16	1.93
Original	10.53	0.3347	11.26	0.4025	**30.72**	2.14
TokenWeave	**12.85**	**0.3420**	17.36	0.5473	46.72	**2.23**
45s / 720 frames						
Full history	11.93	**0.3283**	16.43	0.5396	210.24	1.46
Original	9.25	0.2293	8.65	0.3140	**30.72**	**2.12**
TokenWeave	**12.20**	0.3270	**16.47**	**0.5398**	66.88	1.85
60s / 960 frames						
Full history	10.62	0.2808	15.16	0.4649	233.20	1.25
Original	9.37	0.2436	9.28	0.3398	**30.72**	**2.12**
TokenWeave	**12.01**	**0.3226**	**15.18**	**0.4880**	77.76	1.75

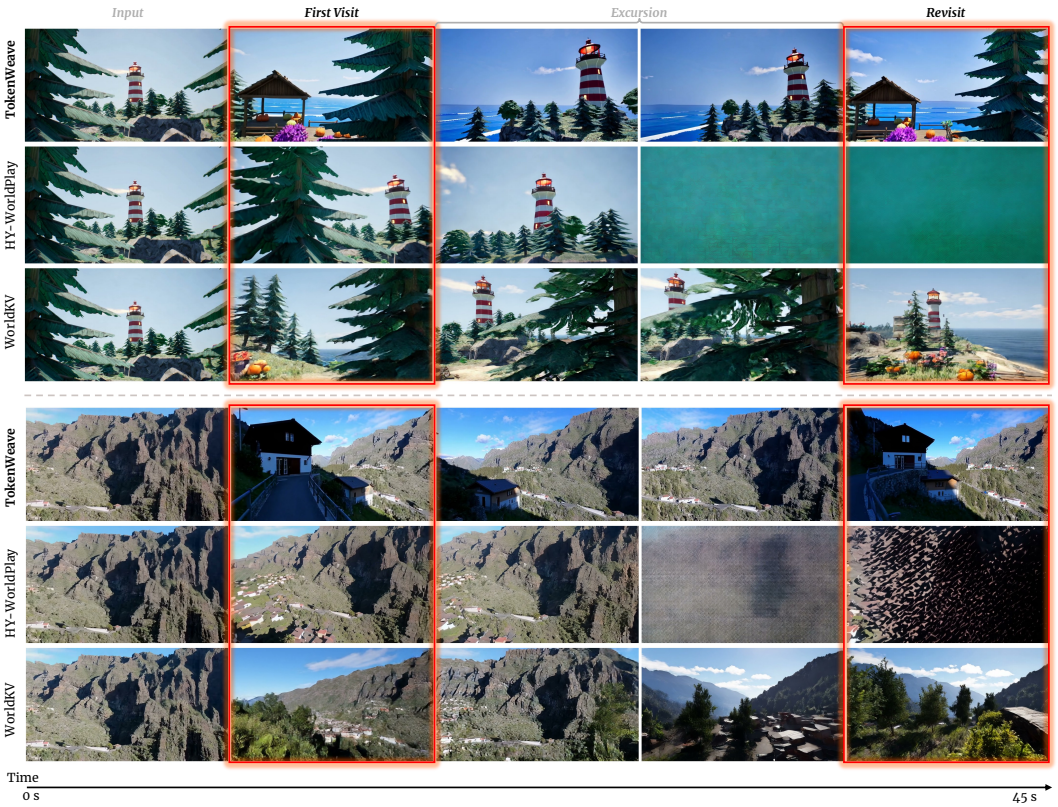

Figure 5: **Revisiting under more complex camera motion.**

Table 3: **Ablation on components.**

Method	PSNR ↑	SSIM ↑	wPSNR ↑	wSSIM ↑
Full	11.4237	0.2882	**15.3932**	**0.4999**
w/o EMR	**11.5930**	**0.2960**	14.4095	0.4572
w/o PGCS	10.8681	0.2341	14.3632	0.4097

Table 4: **Sensitivity to the merging threshold τ.**

τ	PSNR ↑	SSIM ↑	wPSNR ↑	wSSIM ↑	Comp. ratio ↑
0.90	**11.6344**	**0.3017**	**15.8453**	**0.5121**	2.65×
0.85	11.4237	0.2882	15.3932	0.4999	3.63×
0.80	10.9190	0.2832	14.6496	0.4636	7.01×
0.70	10.2115	0.2778	12.4478	0.4040	42.0×

Ablation on components. Removing evidence-aware memory reading (EMR; Equation 6) lowers wPSNR by 0.98 dB and wSSIM by 0.0427, despite slight increases in direct PSNR and SSIM. The lower warped scores indicate weaker scene-content consistency after alignment, supporting the benefit of accounting for each slot's accumulated evidence. Removing pose-guided candidate selection degrades all four metrics, including decreases of 1.03 dB in wPSNR and 0.0903 in wSSIM, supporting spatial constraints for reliable token association. The full method achieves the best warped scores in Table 3, showing that both components contribute to preserving previously observed content.

Sensitivity to the merging threshold. TOKENWEAVE maintains good revisit quality across the tested thresholds $\tau \in \{0.80, 0.85, 0.90\}$, indicating limited sensitivity within this range (Table 4). Warped PSNR remains 14.65–15.85 dB while compression spans 2.65–7.01×. The default $\tau = 0.85$ balances quality and compactness. More aggressive merging at $\tau = 0.70$ yields 42× compression but lowers wPSNR to 12.45 dB.

5 CONCLUSION

We presented TOKENWEAVE, a training-free framework that consolidates temporal KV history into persistent scene memory for video world models. By associating and fusing repeated patch-level observations, it organizes memory around recurring scene content within a frozen backbone. Experiments demonstrate its effectiveness in preserving long-horizon scene consistency with compact memory across different scenarios with multiple backbones. More broadly, our work suggests that better memory organization can complement advances in model architecture and training, offering a practical path toward more persistent and coherent interactive worlds.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper. The authors take full responsibility for the accuracy and integrity of the final content.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.09461.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.01392.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026. URL https://arxiv.org/abs/2607.07534.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training Agents Inside of Scalable World Models. *arXiv preprint arXiv:2509.24527*, 2025. URL https://arxiv.org/abs/2509.24527.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. URL https://arxiv.org/abs/2508.13009.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025. URL https://arxiv.org/abs/2512.04040.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.08009.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, Xitai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid kv cache compression for efficient autoregressive video diffusion models. *arXiv preprint arXiv:2605.09681*, 2026. URL https://arxiv.org/abs/2605.09681.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, Junfeng Ni, Hongyu Pan, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026. URL https://arxiv.org/abs/2607.19191.

Jia Li, Han Yan, Yihang Chen, Siqi Li, Xibin Song, Yifu Wang, Jianfei Cai, Tien-Tsin Wong, and Pan Ji. I3dm: Implicit 3d-aware memory retrieval and injection for consistent video scene generation. *arXiv preprint arXiv:2603.23413*, 2026. URL https://arxiv.org/abs/2603.23413.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.14469.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25161.

Yuexiao Ma, Xuzhe Zheng, Jing Xu, Xiwei Xu, Feng Ling, Xiawu Zheng, Huafeng Kuang, Huixia Li, Xing Wang, Xuefeng Xiao, Fei Chao, and Rongrong Ji. Flow caching for autoregressive video generation. *arXiv preprint arXiv:2602.10825*, 2026. URL https://arxiv.org/abs/2602.10825.

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, Seungryong Kim, and Yang Zhou. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation. *arXiv preprint arXiv:2603.16871*, 2026. URL https://arxiv.org/abs/2603.16871.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Dynamic Frame Compression for Long-context Video World Modeling. *arXiv preprint arXiv:2512.02473*, 2025. URL https://arxiv.org/abs/2512.02473.

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

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2512.14614.

Pavlo Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos Treviso. Long-Context Generalization with Sparse Attention. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.16640.

Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, and Jieneng Chen. Evoworld: Evolving panoramic world generation with explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025. URL https://arxiv.org/abs/2510.01183.

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.

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

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video World Models with Long-term Spatial Memory. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.05284.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.12369.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026. URL https://arxiv.org/abs/2605.22718.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.07772.

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. *arXiv preprint arXiv:2506.03141*, 2025. URL https://arxiv.org/abs/2506.03141.

Wangbo Yu, Kunhao Liu, Wenbo Hu, Shenghai Yuan, Chaoran Feng, Haiyang Zhou, Yukun Huang, Yiran Wang, Wang Zhao, Yingmin Luo, and Ying Shan. WorldCrafter: Consistent Video World Model with Implicit 3D-aware Memory. *arXiv preprint arXiv:2609.24984*, 2026. URL https://arxiv.org/abs/2609.24984.

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

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. CaM: Cache merging for memory-efficient LLMs inference. In *International Conference on Machine Learning*, pp. 58840–58850, 2024. URL https://proceedings.mlr.press/v235/zhang24n.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.14048.

A More Qualitative Results

We encourage readers to visit our anonymous project website to view video results:

https://tokenweave-iclr.github.io/

B Implementation Details and Evaluation Protocol

B.1 Hardware, Backbones, and Inference Settings

We integrate TokenWeave into LingBot-World-v1 (Robbyant Team et al., 2026) and LingBot-World-v2 (Gao et al., 2026), both 14B-parameter autoregressive video world models. The pretrained weights remain frozen throughout inference, and TokenWeave introduces no memory-specific training or parameter updates. All methods are evaluated under the same hardware environment and inference precision when reporting memory usage and throughput.

Baseline configurations. We run LingBot-World-v1, LingBot-World-v2, WorldKV, Infinite-World, and HY-WorldPlay using their public checkpoints and the default inference settings provided by their respective official repositories. Model-specific generation settings follow the corresponding official configuration, while rollout lengths and camera commands are specified by our revisit benchmark. For each LingBot-World backbone, the original model and its TokenWeave variant use the same checkpoint, input image, selected camera trajectory, resolution, frame rate, sampling schedule, and random seed. This paired setup isolates the effect of the inference-time memory mechanism.

Table B.1: **Experimental configuration.** Backbone-specific inference settings follow the corresponding official repository. The merging threshold is fixed at $\tau = 0.85$ except in the threshold sensitivity study.

Setting	Configuration
Backbones	LingBot-World-v1 and LingBot-World-v2
Backbone size and type	14B parameters each; autoregressive
Baseline inference	Official repository defaults
TokenWeave training	None; both backbones remain frozen
Default merging threshold	$\tau = 0.85$
Threshold sweep	$\tau \in \{0.70, 0.80, 0.85, 0.90\}$
Rollout durations	25, 45, and 60 seconds
Reported frame counts	400, 720, and 960, respectively
Recent context and sinks	Retained in their native form
Memory update	Outgoing frames processed in temporal order

B.2 Memory Configuration and Positional Alignment

Association and storage. We use a default merging threshold of $\tau = 0.85$ on both backbones. An outgoing token is associated with its highest-scoring eligible memory slot only when the cosine similarity exceeds this threshold. Candidate slots are selected by the union of the pose-guided geometric gate and the local fallback in Equation 3. A failed match or an empty candidate set creates a new slot. The resulting assignment is shared across layers, and only tokens leaving the native recent window are consolidated. Attention sinks and recent states retain their original form.

Removing and restoring RoPE. Key similarity is computed after removing both spatial and temporal RoPE phases from the outgoing and stored keys. With the notation of Section 3.2, the position of an outgoing key is (f, p), whereas a stored slot key is encoded at (a_j, p_j). Their unrotated representations are

$$\begin{aligned} \bar{k}_{f,p}^{\ell,h} &= \mathrm{RoPE}_{f,p}^{-1}(k_{f,p}^{\ell,h}), \\ \bar{m}_{j}^{\ell,h} &= \mathrm{RoPE}_{a_j,p_j}^{-1}(m_{j}^{K,\ell,h}). \end{aligned} \tag{B.1}$$

These unrotated keys are used in Equation 4. Once a match has been accepted, both keys are encoded at the same spatial anchor p_j and temporal position f before fusion:

$$\begin{aligned} \widetilde{k}^{\ell,h}_{f,p\to j} &= \mathrm{RoPE}_{f,p_j}(\bar{k}^{\ell,h}_{f,p}), \\ \widetilde{m}^{K,\ell,h}_{j\to f} &= \mathrm{RoPE}_{f,p_j}(\bar{m}^{\ell,h}_{j}). \end{aligned} \tag{B.2}$$

Equation 5 then fuses the aligned keys and their corresponding values. The registration anchor (r_j, p_j) remains fixed, while the appearance timestamp is updated to $a_j = f$. Thus, matching uses unrotated features, and stored keys retain the positional encoding needed by the backbone's attention mechanism.

Fusion and memory reading. The averaging weight w_j is capped at $w_{\max}$ to permit feature adaptation, whereas the represented-token count n_j remains uncapped. Memory reading adds $\alpha \log n_j$ to each slot's attention logit as in Equation 6. New slots are initialized with $w_j = n_j = 1$. Outgoing frames are processed once in temporal order, allowing subsequent frames to access the updated memory. Memory capacity is allocated to accommodate the evaluated rollout; the method does not apply an additional eviction policy to persistent slots.

Ablation settings. The threshold sensitivity study evaluates $\tau \in \{0.70, 0.80, 0.85, 0.90\}$, as reported in Table 4. The component study separately removes evidence-aware memory reading and pose-guided candidate selection (Table 3). The full-history comparison retains all historical K/V tokens instead of discarding or consolidating outgoing states, providing a reference for the quality and resource cost of uncompressed history (Table 2).

B.3 Trajectory and Revisit-Pair Construction

The benchmark contains 165 initial images from MBench, ABot-World-0, and SceneFly, spanning real photographs, game imagery, and Unreal Engine renderings. Each image is paired with six candidate revisit sequences across three rollout durations, yielding 990 candidates. A revisit sequence first observes a scene region, moves away during an excursion, and subsequently returns to an overlapping view.

Different world models respond differently to camera commands. We therefore calibrate the commands for each model and select one candidate per image and duration with sufficient overlap between the first visit and the revisit. This produces 165 selected trajectories at each duration, or 495 scene–sequence pairs per model. Within each LingBot-World comparison, the original model and its TokenWeave variant use identical selected trajectories. Across systems, the commands express semantically equivalent revisit behavior while accommodating differences in camera control.

Geometric revisit-pair construction. Let θ_t be the commanded yaw and $\mathcal{L}_1$ and $\mathcal{L}_3$ the initial and return sweeps. We select an anchor on $\mathcal{L}_1$ near a prescribed yaw $\theta_\star$ (default $|\theta_\star| = 65°$), beyond the sink-protected prefix. Since generated camera motion can deviate from the commands, we register each candidate on the return sweep directly to the anchor using SIFT matches and RANSAC. From this registration, we estimate the relative yaw $\widehat{\delta\theta}_t$ between each candidate and the anchor. We choose

$$t_a = \arg\min_{t\in\mathcal{L}_1} |\theta_t - \theta_\star|, \qquad t_r = \arg\min_{t\in\mathcal{C}_3} |\widehat{\delta\theta}_t|, \tag{B.3}$$

where $\mathcal{C}_3$ contains successfully registered candidates sampled every two frames over the full return sweep. Pair selection thus uses geometric agreement, without maximizing PSNR or SSIM. A valid pair requires a sign change in the estimated yaw curve, residual yaw at most 3°, at least 30 RANSAC inliers, and median inlier reprojection error at most 3 pixels. Failed pairs are marked invalid and excluded from the valid-pair metric averages. The two visits are separated beyond the recent context; the reported durations describe complete rollouts, not the interval between t_a and t_r.

B.4 Direct and Warped Revisit Metrics

For a first-visit image $I^{(1)}$ and its revisit $I^{(2)}$, direct PSNR and SSIM compare the two generated views without image-space alignment. These scores reflect both changes in scene content and residual viewpoint differences. To account for view misalignment, we additionally report warped PSNR (wPSNR) and warped SSIM (wSSIM).

Homography warp and valid overlap. For the selected pair, SIFT correspondences and RANSAC give a homography $H_{2\to1}$ mapping the revisit to the anchor. The reported warped scores use this fitted homography (`warp_raw` in the evaluator). With $\widetilde{\mathbf{x}}$ denoting homogeneous pixel coordinates and π perspective division,

$$\begin{aligned} \widehat{I}^{(2)}(\mathbf{x}) &= I^{(2)}\big(\pi\big(H_{2\to1}^{-1}\widetilde{\mathbf{x}}\big)\big)\,, \\ \Omega &= \mathrm{supp}(\mathcal{W}(\mathbf{1}; H_{2\to1}))\,. \end{aligned} \tag{B.4}$$

We resample the image bilinearly and warp an all-ones mask with nearest-neighbor interpolation to identify valid overlap. Scores are computed only within this overlap:

$$\begin{aligned} \mathrm{wPSNR} &= \mathrm{PSNR}\Big(I^{(1)}|_\Omega, \widehat{I}^{(2)}|_\Omega\Big)\,, \\ \mathrm{wSSIM} &= \mathrm{SSIM}\Big(I^{(1)}|_\Omega, \widehat{I}^{(2)}|_\Omega\Big)\,. \end{aligned} \tag{B.5}$$

PSNR averages squared error over color channels and valid pixels. SSIM is computed on grayscale images, with the overlap mask eroded by its 11×11 window to exclude neighborhoods touching warp boundaries. We use the same validity criteria for all methods and average valid-pair scores across scenes for each duration.

Warping reduces the influence of camera error but cannot eliminate it: registration error and viewpoint-dependent changes can still depress the aligned scores. Conversely, a fitted homography can absorb some generated geometric distortion. Warped scores are therefore not strict bounds on memory quality; we interpret them together with direct scores and the video results.