Addressable Memory for Video World Models

Xindi Wu[1,2] **Sven Elflein**[1,3,4] **James Lucas**[1] **Olga Russakovsky**[2] **Laura Leal-Taixé**[1] **Despoina Paschalidou**[1]
Jonathan Lorraine[1] **Aljosa Osep**[1]
[1]NVIDIA [2]Princeton University [3]University of Toronto [4]Vector Institute
https://research.nvidia.com/labs/sil/projects/WorldTrace/

Abstract

We study visual persistence in autoregressive video world models: the Key-Value (KV) cache accumulates a growing visual memory, but once rollouts extend beyond the training horizon, the model can no longer reliably address stored content. We identify out-of-distribution temporal positional encodings as the root cause, as generation exceeds the training context, relative RoPE offsets enter phase regimes the model was never trained on, silently corrupting attention-based retrieval regardless of what is stored in the cache. This reveals that long-horizon failure is fundamentally a problem of addressability: no memory compression scheme can improve visual persistence if past observations cannot be reliably retrieved once they fall outside the training context We propose WorldTrace, a training-free framework that keeps compressed memory addressable by assigning each slot a fixed, in-distribution position relative to the current frame. With addressable memory, we explore two retention approaches: WorldTrace-Field (coherence-oriented) aggregates history in a rotation-invariant space, improving TempSSIM by $+15.5\%$ while reducing scene drift; WorldTrace-Landmark (recall-oriented) stores verbatim scene traces at detected scene transitions. The recall-oriented variant sustains scene reconstruction over long rollouts, with stronger long-range recall and smooth temporal coherence.

1 Introduction

Autoregressive video world models [Ha and Schmidhuber, 2018, Team et al., 2026, He et al., 2025b, Decart et al., 2024] generate scenes chunk-by-chunk, with each chunk attending over a Key-Value (KV) cache of prior context. These models promise interactive applications such as next-generation game engines [Valevski et al., 2025, Decart et al., 2024, He et al., 2025b] and closed-loop robot simulators [NVIDIA, 2025, NVIDIA Spatial Intelligence Lab, 2026], in which users can move freely through a visual world and revisit prior locations. Such revisits demand *visual persistence*: when an agent returns, the model must re-create the scene's original appearance, not a plausible-looking alternative.

In practice, visual persistence degrades rapidly once generation exceeds the training context length. A natural way to address this issue is to compress the linearly growing KV cache into a fixed-size memory [Zhang et al., 2023, Li et al., 2024, Cai et al., 2025, Kim et al., 2026] equal to the size of the context seen during training. However, we find that the bottleneck is not simply whether past content is stored, but whether it remains addressable. As the generation horizon extends beyond the training context, positional encodings [Su et al., 2024] with a training-bounded index range become out-of-distribution (OOD), degrading attention-based memory retrieval. In other words, even if past memories are stored in the KV cache, the model cannot retrieve them. As a result, compressed memory becomes effectively inaccessible, regardless of its construction. *This reveals that long-horizon failure is fundamentally a problem of addressability:* no memory scheme can improve visual persistence if past observations cannot be reliably accessed once they fall outside the training context.

To address this problem, we propose WorldTrace, a training-free memory framework that makes compressed memory *addressable* at arbitrary horizons. We assign each memory entry a fixed, in-distribution temporal position relative to the current frame, keeping past observations within the temporal positional encoding range the model was trained on. Addressable memory exposes a design space for memory retention. We organize

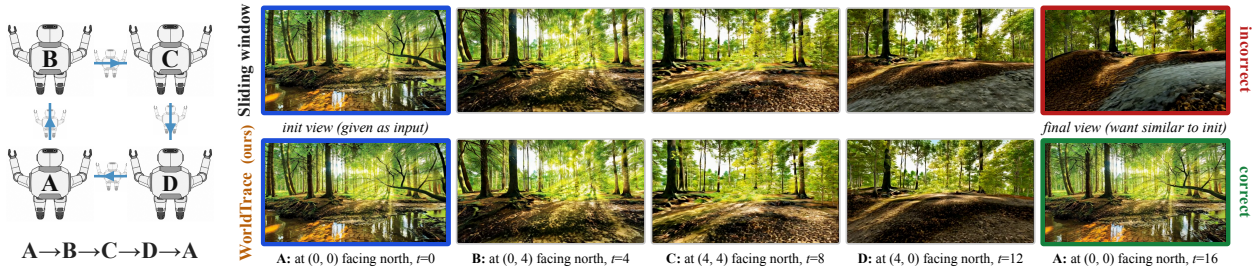


Figure 1: **Addressable memory restores long-horizon visual persistence.** *(Left)* Scripted topology A→B→C→D→A. *(Right)* Representative frames on Matrix-Game-2 (MG2) along the path: the initial scene A (blue border) and waypoints B–D; the last column is the return to A. *Sliding window* (top) mismatches the reference appearance at the return (red border). WorldTrace with frozen landmark keys (*ours*; green border) matches scene A, illustrating addressable long-horizon recall.

the cache into two tiers: a recent window for immediate context, plus summary slots for the distant past. Two complementary approaches summarize the distant horizon into summary slots: WorldTrace-Field aggregates past information in position-invariant space to maintain temporal coherence, and WorldTrace-Landmark detects and preserves high-fidelity scene traces, enabling long-horizon recall.

Our experiments show that addressable memory enables qualitatively new long-horizon behavior. In controlled return-to-origin navigation tasks, where the model must reconstruct a previously visited scene after a long detour, a standard sliding-window cache rapidly loses structural fidelity on the return frame, whereas WorldTrace recovers the original scene appearance (Fig. 1). This extends the effective memory horizon: while existing models begin to forget scene structure after only a few seconds, our approach maintains consistent scene reconstruction over minutes-long rollouts. At the same time, it preserves smooth frame-to-frame evolution, allowing us to control temporal coherence and long-horizon recall.

Our contributions are as follows: (i) We identify out-of-distribution positional encodings as the primary cause of long-horizon failure in autoregressive video world models, and show that under the evaluated architecture, this suppression is complete: any compressed summary with OOD positions produces outputs identical to a sliding-window cache, regardless of summary content; (ii) We propose WorldTrace, a training-free memory framework that keeps compressed memory in-distribution at any generation horizon, enabling reliable retrieval of past observations without modifying the underlying model. (iii) We propose a scripted evaluation benchmark (LoopMem) to demonstrate that our approach extends visually persistent generation from seconds to minutes-long rollouts, maintaining consistent scene reconstruction while existing methods degrade substantially.

2 Rethinking Memory in Video World Models

Autoregressive video world models retain prior context as cached keys and values during generation. Long-horizon recall, therefore, depends on both whether older, relevant context remains physically inside the cache and whether the current query can still address them. Two coupled bottlenecks arise. First, cached tokens must remain *addressable* (Sec. 2.1): attending to a distant token requires a temporal RoPE offset that may lie outside the training distribution, rendering the token unreadable even when stored. Second, compressed summaries must preserve *content* (Sec. 2.2): when old keys are compressed naively, RoPE phase differences can destroy the summary's intended signal. Both bottlenecks are cache-side: we only change the cache content and virtual positions; the generator and rollout state remain fixed.

2.1 Position Determines Whether Memory Is Addressable

During training, the model attends only within its local attention window:

$$|\delta_{q,t}| = |q - t| \leq \Delta t_{\text{train}}.$$

For temporal RoPE pair k at angular frequency θ_k, the attention-logit contribution is:

$$\ell_k(q,t) = \mathrm{Re}\big(A^{(k)}_{q,t} e^{i\theta_k \delta_{q,t}}\big)\,,$$

where $\mathrm{Re}(\cdot)$ is the real part, $\ell_k(q,t)$ the contribution of frequency k to the query-key logit, and $A^{(k)}_{q,t}$ the content-dependent inner product in canonical (unrotated) coordinates. During inference, absolute positions increase, causing relative offsets to be larger than during training. The affected RoPE components then enter phase regimes that the model was never trained to interpret. Attention may still assign non-negligible weight to such tokens via key content, but the logits impose rotations the model was not trained to invert (App. F.2, per-frequency severity).

Concurrent methods address related positional issues. Infinity-RoPE [Yesiltepe et al., 2026] proposes block-relativistic RoPE (Block-Rel), which expresses cached-token positions relative to the current AR block, thus bounding $\delta_{q,t}$. While this reduces RoPE extrapolation, Block-Rel caps offsets at Δt_{train}, so any summary slot representing context beyond L_{train} blocks is assigned the same temporal position. With $N_s{>}1$ summary slots spanning diverse history, Block-Rel makes them positionally indistinguishable once the horizon exceeds a few training windows (Rem. 1). MemRoPE [Kim et al., 2026] builds on this with dual EMA memory tokens, each assigned a distinct Block-Rel position. MemRoPE's fixed two-slot summary structure sidesteps this partially, but does not extend to N_s-slot caches at arbitrary horizons. Crucially, both methods are designed for video generation rollout stability; neither addresses the interactive, arbitrary-time memory access required by world models, which motivates our design. We provide a detailed discussion of related work in App. B.

2.2 Content Determines Whether Memory Is Useful

Even if a compressed summary occupies an in-distribution temporal position, what fills it matters. The most literal way to compress a block of cached keys is averaging [Bolya et al., 2023, Rae et al., 2020]. With the standard implementation of RoPE-rotated keys, this corresponds to:

$$\bar{K}^{(k)}_{\text{naive}} = \frac{1}{M}\sum\nolimits_{m=1}^{M} R(\theta_k t_m)\; K^{(k)}_{\text{cx},m},$$

where $K^{(k)}_{\text{cx},m}$ is the canonical (unrotated) key of source frame m at frequency k and $R(\theta_k t_m)$ is its RoPE rotation. The failure is visible directly: each source frame is rotated by a different angle $\theta_k t_m$ before averaging, so $\bar{K}^{(k)}_{\text{naive}}$ sums vectors pointing in different directions. When source frames span a small interval, their angles are similar, and the average preserves content. As the interval approaches or exceeds a RoPE period, contributions point in opposite directions and partially cancel, attenuating that frequency component regardless of what it encodes. Naive averaging in RoPE-rotated space thus erases the signal a summary should carry.

The two failures are coupled. Position and content are usually separated: prior work independently picks a position scheme (*e.g.* Block-Rel) and a compression rule (EMA [Kim et al., 2026], averaging, eviction). We argue and show in Sec. 3.5 that they are coupled: compressors tuned for one positional regime degrade when applied to another, which our proposed method in Sec. 3 addresses jointly.

3 WorldTrace

We propose WorldTrace, a training-free KV-cache framework for autoregressive video world models. WorldTrace pairs a two-tier memory system using temporal position assignment that makes every summary entry *addressable* (Sec. 3.2) with two content compression strategies that operate in canonical key space: WorldTrace-Field, described in Sec. 3.3, optimizes for temporal coherence; WorldTrace-Landmark, described in Sec. 3.4, focuses on visual persistence.

3.1 Cache Structure

As shown in Fig. 2, we structure the memory into two parts that together span the model's training window of L_{train} AR blocks (F latent frames each): a **recent window** $\mathcal{R}$ of W_r verbatim blocks, and a **summary cache** $\mathcal{S}$ of N_s slots storing older, compressed history with constraint $N_s + W_r = L_{\text{train}}$. This creates a trade-off: more

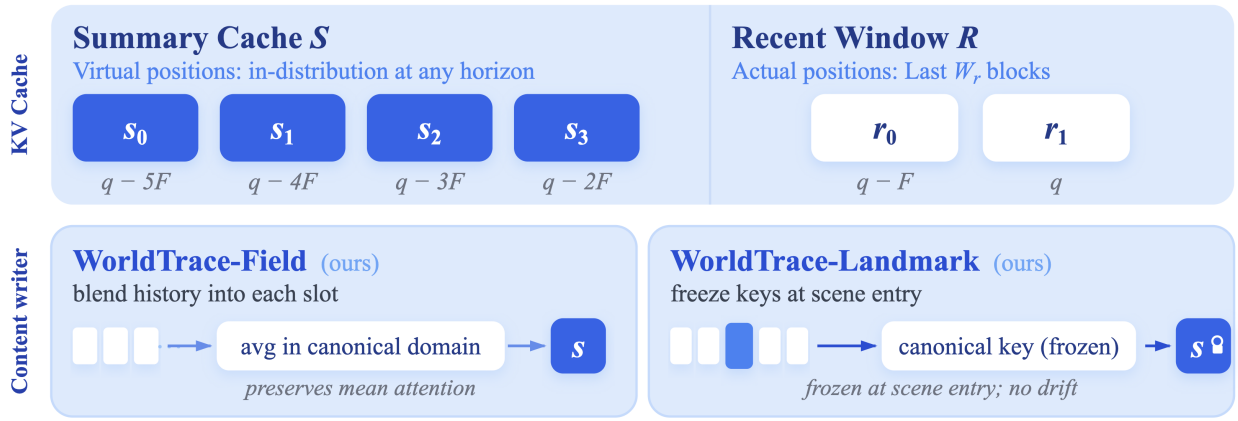


Figure 2: **WorldTrace overview.** The KV cache splits into a compressed *summary* cache $\mathcal{S}$ (left) and a verbatim *recent* window $\mathcal{R}$ (right), spanning the trained attention horizon (Sec. 3.1); each tile is one AR block of F frames at the labelled lead-frame position. **Bottom left (WorldTrace-Field):** keys unrotate, average canonically, then re-encode at each slot's virtual position (Defs. 1, 2). **Bottom right (WorldTrace-Landmark):** scene-entry landmarks freeze canonical keys and re-rotate to the current virtual position (Eq. (4)), without cumulative drift.

summary slots favor recall-dominant tasks, while more recent-window slots favor coherence-dominant ones.

Scope. WorldTrace assumes (i) temporal RoPE on keys before the attention dot product, and (ii) a fixed autoregressive KV cache with known training context length L_{train}. Single-shot video diffusion (no AR cache) and content-agnostic positional encodings (*e.g.* learned absolute embeddings) fall outside this scope.

3.2 Virtual Position Assignment: WorldTrace Slot Indexing

Any memory position scheme should satisfy four properties: (i) linearity in q and s for a simple, predictable mapping; (ii) in-distribution positions so the model can attend; (iii) horizon-stability so slot offsets stay fixed regardless of generation length; and (iv) injectivity so distinct slots are distinguishable. Using the same q as in Sec. 2.1, summary token positions must stay within the range seen during training:

$$t^v_{\min} = \max\big(0,\ q - (L_{\text{train}}{-}1)F\big), \quad t^v_{\max} = q - W_r F. \tag{1}$$

We propose to assign positions by *slot indexing*, formalized next:

Definition 1 (WorldTrace Slot Indexing)**.** Given current query position q, training context length L_{train}, latent frames per AR block F, and summary cache size N_s, the virtual position of summary slot s (0-indexed from oldest) is

$$t_v^{(s)} \ = \ q - (L_{\text{train}}{-}1{-}s)\,F, \quad s = 0, \ldots, N_s{-}1. \tag{2}$$

The assignment covers the in-distribution range with fixed slot-rank offsets relative to the current frame. By construction, $t_v^{(s)} \in [t^v_{\min}, t^v_{\max}]$ at any generation length: offsets depend only on slot rank and q, not the absolute horizon N.

An alternative, *centroid linear* (each slot's mean source timestamp linearly mapped into $[t^v_{\min}, t^v_{\max}]$), satisfies (i) and (ii) but violates horizon-stability and is less stable at long horizons.

Remark 1 (Block-Rel saturation)**.** *Per Sec. 2.1, Block-Rel [Yesiltepe et al., 2026] caps the relative look-back at $(L_{\text{train}}{-}1)F$, so any summary slot whose centroid falls past this cap is clamped to the same virtual position $t^v_{\min}$. Under uniform-bucket averaging, the source pool feeding the N_s summary slots grows by F frames per chunk while the cap is fixed, so for any choice of N_s all slot centroids eventually drift past the cap and become indistinguishable to attention. Def. 1 avoids this because the per-slot offset $(L_{\text{train}}{-}1{-}s)F$ stays distinct across slots and is independent of N.*

With addressable summary slots, the remaining design choice is the *content writer*: the rule that decides what to store in each slot. Both WorldTrace variants use the same canonical (un-rotated) key do-

main: WorldTrace-Field compresses all history into each slot via canonical key averaging (Sec. 3.3); WorldTrace-Landmark selects a small number of verbatim frames at detected scene boundaries and freezes their canonical keys (Sec. 3.4).

3.3 WorldTrace-Field: Canonical Key Averaging

WorldTrace-Field targets temporal coherence, not episodic recall; it averages all history into each summary slot and cannot recover a specific past scene. To avoid the phase-cancellation failure of Sec. 2.2, we fix the compression domain by aligning all source keys to a shared phase before averaging, i.e., the canonical (unrotated) representation. The operator is:

Definition 2 (Canonical Key Averaging (WorldTrace-Field operator))**.** For each temporal head-dimension pair k, the compressed key at virtual position t_v is:

$$K^{(k)}_{\text{field}}(t_v) \;:=\; R(\theta_k t_v)\;\frac{1}{M}\sum\nolimits_{m=1}^{M} R(-\theta_k t_m)\;K^{(k)}_{t_m}, \tag{3}$$

i.e. we unrotate each key to its canonical content $K^{(k)}_{\text{cx},m} = R(-\theta_k t_m)\;K^{(k)}_{t_m}$, average in that space, and then re-encode at virtual position t_v. Values are not rotated as only keys carry RoPE; slot values are computed as the mean of the source-block values at the same intra-block frame index. The count M is not a separate hyperparameter: all $T_{\text{old}} = (N - W_r)F$ frames that have left the recent window are split into N_s contiguous temporal groups (oldest to newest, slot s receives group s), giving $M \approx T_{\text{old}}/N_s$ per slot, a ratio that grows linearly with generation length N.

Remark 2 (Mean attention preservation, informal)**.** *The compressed key from Def. 2 preserves the mean attention logit that the source keys $\{K^{(k)}_{t_m}\}_{m=1}^{M}$ would receive after being reassigned the shared virtual position t_v (logits only; softmax non-linearity precludes the same statement for attention weights). A formal statement is in App. C (Prop. 1). We identify phase cancellation as the failure mode and show that this preservation property holds.*

3.4 WorldTrace-Landmark: Landmark Traces with Frozen Keys

To optimize for episodic recall, WorldTrace-Landmark fills the N_s slots with verbatim high-value past frames rather than averaged summaries (the name evokes verbatim-anchor recall in the spirit of Landmark Attention [Mohtashami and Jaggi, 2023]; our scene-entry detection, frozen canonical keys, and slot-rank virtual positions differ from token-level landmark insertion in trained transformers).

Landmark selection. We identify recall-relevant frames using the canonical-K representation from Eq. (3): for each incoming frame, we compute the cosine distance between consecutive canonical-K signatures and mark frames with a spike above threshold τ as scene-entry events. WorldTrace-Landmark fills the N_s summary slots with the most recent scene-entry frames verbatim: when fewer than N_s landmarks have fired, the remaining slots repeat the oldest available landmark; once more than N_s have fired over a long rollout, the oldest is evicted to make room for the newest (FIFO over scene entries). Virtual positions (Eq. (2)) apply unchanged; each landmark frame inherits the slot-rank position of its summary slot, not its absolute timestamp.

Frozen landmark keys. Slots are ordered from oldest ($s{=}0$) to newest, so each new chunk shifts every cached frame one slot toward the older end: a landmark frame at slot s at chunk n occupies slot $s{-}1$ at chunk $n{+}1$. Under standard summary updates, each shift unrotates the cached Key to its canonical form, then re-rotates it to its new virtual position. Over many shifts, these unrotate-rerotate cycles accumulate floating-point errors. This is exacerbated in practice by bfloat16 precision, as shown by Wang et al. [2025]. WorldTrace-Landmark removes this drift by freezing each selected key in canonical form: at landmark time, we store the canonical key once and apply a single fresh rotation to the current virtual position at every subsequent shift:

$$K^{(k)}_{\text{land}}\left(t_v^{(s)}\right) \;:=\; R(\theta_k t_v^{(s)})\;R(-\theta_k t_{\ell^*})\;K^{(k)}_{t_{\ell^*}}, \tag{4}$$

where t_{ℓ^*} is the original timestamp of the selected landmark frame. Compared to Eq. (3), the form is identical (unrotate to canonical, re-rotate to virtual position), with $M{=}1$ and no averaging; the canonical key $R(-\theta_k t_{\ell^*})\;K^{(k)}_{t_{\ell^*}}$ is computed once at landmark time and reused across all subsequent shifts. The canonical-

caching mechanism is shared with concurrent work that stores RoPE keys canonically and rotates at attention time (*e.g.* MemRoPE [Kim et al., 2026]); WorldTrace-Landmark differs in (i) applying it selectively at detected scene-entry events rather than to every cached key, and (ii) combining it with the slot-rank position assignment of Sec. 3.2, so each frozen landmark occupies a distinct in-distribution virtual position at any horizon.

3.5 Position-Content Coupling and Cache Update

Prior KV-cache work treats position assignment and content compression as separable, independently choosing a position scheme (Block-Rel, Infinity-RoPE) and a content writer (EMA, averaging, eviction) [Yesiltepe et al., 2026, Kim et al., 2026, Zhou et al., 2025b, He et al., 2025a]. This is unsafe for autoregressive video memory: a writer accumulates statistics at write-time offsets but is queried at read-time offsets, so when the two regimes disagree, the stored statistic no longer matches the query. Two failures follow: writing in the rotated key space under saturating positions reduces compression to sliding-window eviction once addressability collapses (Rem. 1), and writing a position-conditioned average under one offset distribution and reading at another shifts the effective phase of every cached token, breaking the mean-attention preservation that motivates Def. 2. WorldTrace addresses both by construction: Def. 1 fixes the positional regime, and Def. 2 writes content in the rotation-invariant canonical domain that absorbs the source-to-virtual shift; slot content (averaged frames or frozen landmarks) is then orthogonal and content-only within this coupled framework.

4 Experiments

We organize our experiments around our three claims. **(Q1, Sec. 4.2)** Is position, not content, the binding constraint? **(Q2, Sec. 4.2)** Does WorldTrace-Field improve *coherence* over naive averaging and a sliding-window cache? **(Q3, Sec. 4.3)** Can WorldTrace-Landmark improve *episodic recall* at extended horizons?

4.1 Setup

Models. Our evaluation uses Matrix-Game-2 (MG2-1.3B) [He et al., 2025b], a distilled 1.3B-parameter autoregressive game world model based on Wan 1.3B T2V [Wan Team, 2025] with 3D-RoPE, training-time KV-cache extent $L_{\text{train}} = 6$ AR blocks (F=3 latent frames per block), and a local attention window of 2 blocks, *i.e.* Δt_{train}+1=6 latent frames. We provide additional experiments for LingBot-Fast [Team et al., 2026] in App. D. We use N_s=2/W_r=4 for coherence (Sec. 4.2) and N_s=4/W_r=2 for episodic recall (Sec. 4.3).

Baselines. We compare against the sliding-window approach used in MG2 by default. Next, we consider a baseline that performs frame averaging in rotated key space without RoPE disentanglement (Naive+Block-Rel), and canonical averaging+Block-Rel (canonical-domain key compression with Block-Rel positions; isolates the position contribution of WorldTrace). *Canonical-K anchoring*: Latent re-anchor (re-injects scene-A KV into the most recent buffer slot). *Verbatim recall*: Landmark+Block-Rel isolates the slot-rank scheme on top of verbatim landmarks. *Concurrent training-free*: MemRoPE [Kim et al., 2026] (Block-Rel + dual-rate EMA) and YaRN [Peng et al., 2024] (NTK-aware temporal RoPE rescaling on sliding-window KV retention).

Benchmark. We introduce **LoopMem**, a scripted-navigation benchmark for episodic recall in AR world models. The model executes a scripted navigation path that returns to a previously visited location; the regenerated return frame is scored against the original scene appearance at geometrically matched positions, requiring no external reference. LoopMem organizes difficulty along four axes: waypoint count (K, Fig. 3), edge length (L), camera-orientation closure, and multi-revisit depth (R); the full benchmark gallery with all evaluated configurations is in Fig. 6 (App. E).

Metrics. Our evaluation focuses on two key aspects relevant to world models: the temporal *coherence* of generated outputs and the ability to correctly recall revisited locations. We evaluate *coherence* with $\mathrm{TempSSIM}$ (↑, SSIM [Wang et al., 2004] between consecutive decoded frames) and Local Scene Drift (↓, mean per-chunk CLIP feature distance to the preceding chunk); *recall* with PAC (↑, CLIP-ViT-H/14 cosine similarity between geometrically paired return- and forward-leg frames in loops); and LatentDiff (↓, MSE between consecutive

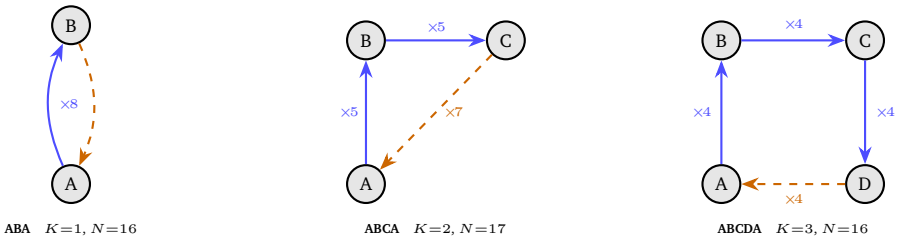


Figure 3: **Three LoopMem geometries.** Blue solid: outward; dashed orange: return. ABA is a straight reversal, same-heading. ABCA is an approximate L-triangle (hypotenuse $5\sqrt{2} \approx 7$), rotated-heading. ABCDA is a square, same-heading. The model must regenerate scene A after each path. Full gallery (edge length, orientation, multi-revisit): App. E, Fig. 6.

latent frames) as a latent-space sanity check (easily confounded with low motion). Full definitions are in Tab. 16 (App. F.5).

4.2 Coherence: WorldTrace-Field

Position is the binding constraint. Holding the content operator fixed (canonical key averaging, Eq. (3)) and varying only the position assignment reveals that Block-Rel offsets beyond the training window receive zero softmax weight under MG2's local-attention mask, collapsing canonical averaging to the same sliding-window eviction; compressing keys without addressable positions provides no benefit. Centroid-linear avoids saturation but uses an N-dependent formula that destabilizes past the training window. Only WorldTrace assigns positions by slot rank alone, keeping every summary slot in-distribution at every horizon: WorldTrace leads Block-Rel by $+5.9\%$ TempSSIM at $N{=}8$ (0.413 vs. 0.390) and $+2.8\%$ at $N{=}16$ (0.545 vs. 0.530).

Position ablation. Canonical averaging fixed; only positions vary.

Position	$N{=}8$	$N{=}16$
Block-Rel	0.390	0.530
Centroid linear	0.377	0.479
WorldTrace (ours)	**0.413**	**0.545**

WorldTrace-Field combines canonical-domain key averaging with WorldTrace slot-rank positions to maintain scene coherence under compression. We compare it against the sliding-window baseline and two canonical-K averaging variants that share WorldTrace-Field's content operator (Eq. (3)) but differ in positions: *Block-Rel* [Yesiltepe et al., 2026], which saturates summary-slot offsets to a single capped position; *centroid*, which maps each slot's mean source-frame timestamp into the in-distribution range (slots stay distinguishable but the formula is N-dependent). Only WorldTrace-Field uses positions that depend on slot rank alone (Eq. (2)), horizon-independent and addressable.

Fig. 4 shows the effect directly: by $N{=}18$, the sliding-window baseline has lost scene context and generates incoherent geometry, while WorldTrace-Field stays coherent through $N{=}48$. Tab. 1 quantifies it at two horizons ($N{=}32, 48$). At $N{=}32$, WorldTrace-Field leads TempSSIM (0.613 vs. 0.571); centroid variants reach the lowest Drift but at 0.04 lower TempSSIM. At $N{=}48$ ($24\times$ training horizon), WorldTrace-Field leads on *both* dimensions ($+15.5\%$ relative lift on TempSSIM, lowest Drift), and the cluster of canonical-averaging variants spreads non-monotonically as horizon grows (centroid 0.479, fullcombo 0.449 at $N{=}48$), consistent with the horizon-instability of N-dependent position formulas. This shows that stacking content-domain heuristics cannot recover what an unstable position scheme loses; the position assignment is the binding constraint as the horizon grows past the training window.

4.3 Episodic Recall: WorldTrace-Landmark

WorldTrace-Landmark results on LoopMem. Tab. 2 reports results at $N_s{=}4$, $W_r{=}2$. Across all the topologies, WorldTrace-Landmark improves PAC by $+19.3\%$ on ABA, $+17.7\%$ on ABCA, and $+20.0\%$ on ABCDA. The edge-length sweep (Tier 2) reveals the compression trend clearly: the gap widens from $+6.3\%$ on ABA-short ($N{=}8$) to $+19.8\%$ on ABA-long ($N{=}32$), confirming verbatim landmark injection matters most when KV-context distance to scene A is largest. Camera-orientation pans (Tier 3) show gains at $90°$ ($+3.2\%$) and $180°$ ($+11.0\%$) but near-zero gain at $360°$ ($+1.8\%$), consistent with the known limitation of the approximate yaw-scaling protocol. In multi-revisit (Tier 4), WorldTrace-Landmark reaches PAC$=0.941$ on ABABA vs.

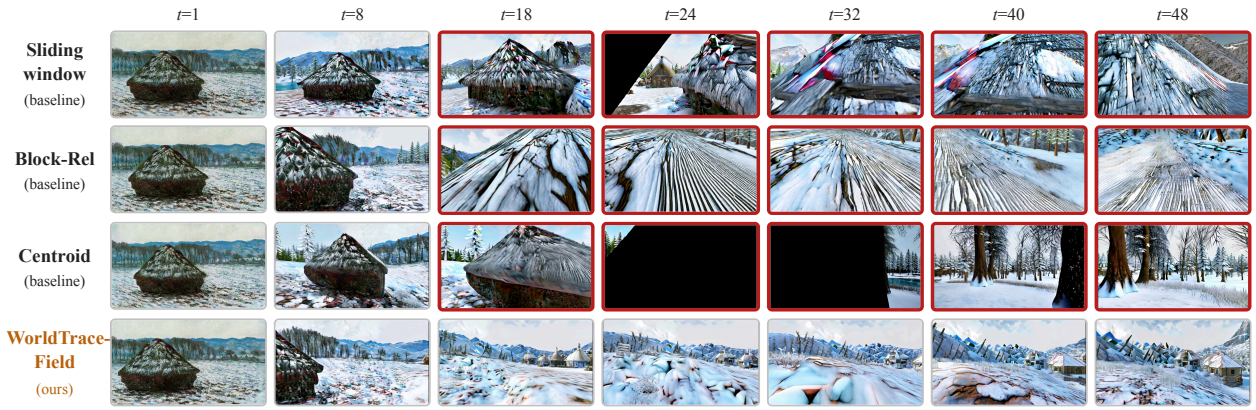


Figure 4: **WorldTrace-Field at N=48.** Four position-encoding methods from the same t=1 conditioning image: three baselines (sliding window, Block-Rel, Centroid) and WorldTrace-Field (ours). The sliding window loses initial-scene context by t=18, with Block-Rel and Centroid following at t=24; once a baseline diverges, it does not recover, and subsequent frames no longer match the trajectory's earlier appearance. Diverged frames are bordered in red. WorldTrace-Field retains memory of earlier chunks and remains coherent across the horizon. The camera path across the seven columns is plotted in Fig. 9 (App. F).

Table 1: **WorldTrace-Field leads on TempSSIM at both horizons and on Local Drift.** At N=32, WorldTrace-Field has the highest TempSSIM (0.613) though centroid variants have lower Drift; at N=48, WorldTrace-Field leads on both metrics. *Centroid* replaces Block-Rel's saturating offset by linearly mapping each slot's mean source-frame timestamp into $[t^v_{\min}, t^v_{\max}]$; *norm* rescales the canonical mean to preserve per-frame norm; *geom* uses geometric group sizes; *knorm* weights frames by their canonical-key L2 norm.

	N=32		N=48	
Method	TempSSIM ↑	Local Drift ↓	TempSSIM ↑	Local Drift ↓
Sliding window (baseline)	0.571	0.0229	0.472	0.0305
Canonical averaging + Block-Rel (uniform)	0.585	0.0215	0.530	0.0339
Canonical averaging + Centroid (uniform)	0.573	**0.0211**	0.479	0.0297
WorldTrace-Field (ours)	**0.613**	0.0250	**0.545**	**0.0295**

0.892 for sliding-window; the advantage is smaller on palindromes (ABCBA, ABCDBA) where B-side revisits partially restore scene context. The PAC horizon sweep (Tab. 9, App. D.2) confirms only verbatim recall sustains high PAC at N=256 (0.989 vs. 0.610 for canonical-K anchoring). Concurrent training-free baselines (MemRoPE, YaRN), per-topology breakdowns, history-selection, slot-sensitivity: App. D.1, App. D.2, App. D.4.

4.4 Ablation

We isolate two design axes using the ABA recall protocol with a fixed cache budget (N_s+W_r=6 slots). **(1) Position assignment** (WorldTrace vs. Block-Rel): holding the content operator fixed and varying only positions isolates whether position, not content, is the binding constraint for WorldTrace-Field. **(2) Canonical vs. naive key compression**: canonical-domain averaging avoids phase cancellation (Sec. 2.2); its benefit surfaces on coherence (TempSSIM, LatentDiff) rather than recall.

Table 4: **Phase cancellation: Naive vs. canonical key averaging.** LatentDiff (↓) at N=16 for varying N_s. Canonical averaging avoids the phase cancellation that corrupts low frequencies under naive RoPE-space averaging.

N_s	Naive LatentDiff ↓	Canonical LatentDiff ↓
1	0.257	**0.224**
2	0.261	**0.257**
4	0.312	**0.233**

(1) Position assignment is the binding constraint for WorldTrace-Field (Tab. 3). Canonical averaging

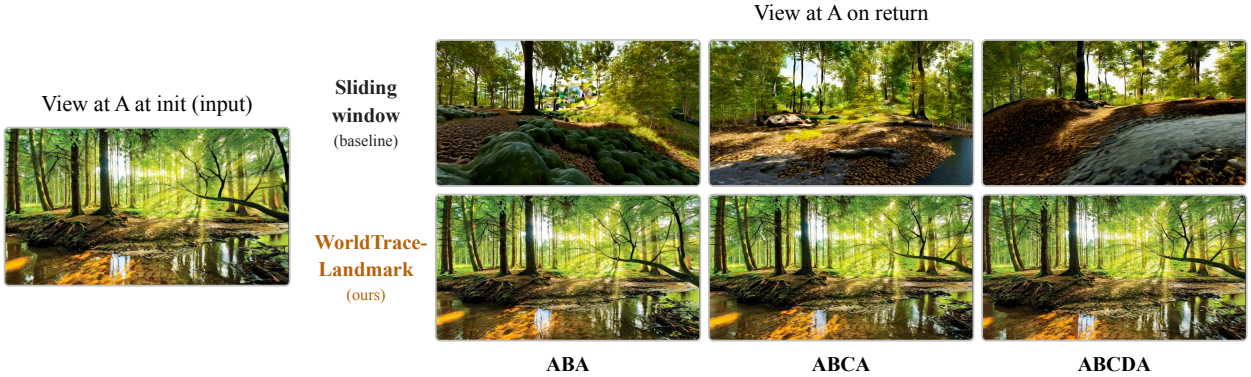


Figure 5: **LoopMem qualitative results** on ABA, ABCA, and ABCDA loops. The sliding-window baseline mismatches the reference on return; WorldTrace-Landmark matches scene A in all three, keeping the scene-origin trace addressable across waypoints.

Table 2: **WorldTrace-Landmark improves episodic recall across topology, edge length, camera orientation, and multi-revisit depth.** Each tier groups methods (rows) by loop configuration (columns). Primary metric (left of /) is PAC for ABA-topology and multi-revisit. Secondary metric (right of /) is TempSSIM over the return leg.

Tier 1: varying topology (K)			
	ABA (N=16)	ABCA (N=17)	ABCDA (N=16)
Sliding window	$0.723_{\pm 0.013}$ / 0.761	$0.673_{\pm 0.014}$ / 0.740	$0.666_{\pm 0.013}$ / 0.748
WorldTrace-Landmark	$\mathbf{0.864}_{\pm 0.009}$ / **0.786**	$\mathbf{0.792}_{\pm 0.011}$ / **0.758**	$\mathbf{0.799}_{\pm 0.011}$ / **0.771**
Tier 2: varying edge length (L, ABA)			
	ABA short (L=4, N=8)	ABA (L=8, N=16)	ABA long (L=16, N=32)
Sliding window	$0.859_{\pm 0.007}$ / 0.780	$0.723_{\pm 0.013}$ / 0.761	$0.627_{\pm 0.013}$ / 0.771
WorldTrace-Landmark	$\mathbf{0.922}_{\pm 0.004}$ / **0.789**	$\mathbf{0.864}_{\pm 0.009}$ / **0.786**	$\mathbf{0.825}_{\pm 0.009}$ / **0.787**
Tier 3: camera-orientation (agent fixed at A)			
	Pan 90° (N=4)	Pan 180° (N=8)	Pan 360° (N=8)
Sliding window	$0.829_{\pm 0.008}$ / 0.480	$0.671_{\pm 0.014}$ / 0.458	$0.559_{\pm 0.015}$ / 0.455
WorldTrace-Landmark	$\mathbf{0.861}_{\pm 0.007}$ / **0.493**	$\mathbf{0.781}_{\pm 0.010}$ / **0.486**	$\mathbf{0.577}_{\pm 0.015}$ / **0.467**
Tier 4: multi-revisit (R>1)			
	ABABA (R=2, N=32)	ABCBA (R_B=2, N=20)	ABCDBA (R_B=2, N=20)
Sliding window	$0.892_{\pm 0.004}$ / 0.765	$0.825_{\pm 0.010}$ / 0.751	$0.842_{\pm 0.010}$ / 0.758
WorldTrace-Landmark	$\mathbf{0.941}_{\pm 0.005}$ / **0.789**	$\mathbf{0.863}_{\pm 0.009}$ / **0.771**	$\mathbf{0.876}_{\pm 0.009}$ / **0.775**

Table 3: **Position assignment & canonical compression ablation (Axes 1 & 2).** PAC_{near} (↑) and TempSSIM (↑) on ABA recall with N_s+W_r=6 slots. Field averaging with Block-Rel matches sliding-window eviction byte-for-byte (OOD offsets suppress summary-slot attention).

Method	PAC_{near} ↑				TempSSIM (N=16) ↑
	N=16	N=64	N=128	N=256	
Axes 1 & 2: position assignment & canonical compression (compression tier)					
Sliding window (baseline)	0.540	0.412	0.504	0.631	0.9945
+ Naive averaging (Block-Rel)	0.570	0.443	0.486	0.598	0.9944
+ Field averaging (Block-Rel)	0.540	0.412	0.504	0.631	0.9945
+ WorldTrace **(WorldTrace-Field, ours)**	0.555	0.442	0.495	0.602	**0.9948**

with Block-Rel positions collapses to the sliding-window baseline byte-for-byte (OOD offsets receive zero softmax weight), while naive averaging in RoPE-rotated space scores marginally higher on PAC_{near} at N=16 (0.570) due to incidental OOD suppression rather than genuine recall. Only WorldTrace keeps every summary slot in-distribution: WorldTrace-Field recovers PAC_{near}=0.555 vs. 0.540 for the sliding-window

baseline ($+0.015$), growing to $+0.030$ at $N{=}64$, and achieves the best TempSSIM (0.9948 vs. 0.9945).

(2) Canonical key averaging avoids phase cancellation and improves coherence (Tab. 4). Holding positions fixed at Block-Rel, switching from naive to canonical averaging reduces LatentDiff from 0.312 to 0.233 at $N_s{=}4$ (-25.3%). Under naive averaging, LatentDiff grows with N_s (from 0.257 at $N_s{=}1$ to 0.312 at $N_s{=}4$) as wider temporal groups amplify phase cancellation, while canonical averaging is unaffected (0.224–0.257). Averaging in canonical (unrotated) space is therefore necessary to preserve the low-frequency content of compressed summary Keys.

5 Conclusion

Autoregressive video world models generate scenes by attending over a growing KV cache of prior context. Long-horizon recall depends on whether old cached keys remain both physically stored and addressable by the current query. Two coupled bottlenecks emerge: first, temporal RoPE offsets must stay in-distribution (Sec. 2.1); attending to distant tokens requires positional encodings outside the training range, rendering them unreadable even when cached. Under soft attention, logit rotations stay miscalibrated even when weights are non-zero. Second, compressed summaries must preserve content (Sec. 2.2); naive averaging in RoPE-rotated space sums vectors pointing in different directions, causing phase cancellation when source frames span intervals approaching a RoPE period. Prior work independently picks a position scheme and a compression rule. We show in Sec. 3.5 that they are coupled: compressors tuned for one positional regime degrade when applied to another. WorldTrace addresses both jointly: each summary slot receives a fixed slot-rank offset that keeps the compressed cache addressable at any horizon, paired with two writers for non-overlapping failure modes: WorldTrace-Field (canonical key averaging) for coherence, and WorldTrace-Landmark (frozen canonical landmarks) for episodic recall, all training-free.

Acknowledgements

We thank our collaborators at NVIDIA's Spatial Intelligence Lab and Princeton University for early feedback and discussions.

References

Dmitry Akulov, Mohamed Sana, Antonio De Domenico, Tareq Si Salem, Nicola Piovesan, and Fadhel Ayed. Kv-compose: Efficient structured kv cache compression with composite tokens. *arXiv preprint arXiv:2509.05165*, 2025. 22

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024. 24

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.06205. 23, 31

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 22

Jesse Bettencourt, Xindi Wu, Matan Atzmon, James Lucas, and Jonathan Lorraine. Variance reduction for expectations with diffusion teachers. *arXiv preprint arXiv:2605.21489*, 2026. 37

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 (ICLR)*, 2023. 3, 22

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video gen-

eration models as world simulators. OpenAI Technical Report, https://openai.com/research/video-generation-models-as-world-simulators, 2024. 24

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024. 24

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 20

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.21058, OpenReview: https://openreview.net/forum?id=y6XJZlEC2x. 25

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. In *Conference on Language Modeling (COLM)*, 2025. 1, 22

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 (NeurIPS)*, 2024. 24, 31

Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past- and future-informed kv cache policy with salience estimation in autoregressive video diffusion. *arXiv preprint arXiv:2601.21896*, 2026a. 25

Jintao Chen, Chengyu Bai, Junjun Hu, Xinda Xue, and Mu Xu. Grounded forcing: Bridging time-independent semantics and proximal dynamics in autoregressive video synthesis. *arXiv preprint arXiv:2604.06939*, 2026b. 25

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026c. 25

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 23

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026d. 25

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025a. Project page https://sites.google.com/view/vrag. 25

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025b. arXiv:2410.21216. 23

Giulio Corallo and Paolo Papotti. FINCH: Prompt-guided key-value cache compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1517–1532, 2024. 23

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914*, 2026a. 25

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-Forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.02283. 24

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2978–2988, 2019. arXiv:1901.02860. 20

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. https://oasis-model.github.io, 2024. 1, 24

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning (ICML)*, 2024. 23

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *International Conference on Computer Vision (ICCV)*, 2025. 26, 37

Xinhang Gao, Junlin Guan, Shuhan Luo, Wenzhuo Li, Guanghuan Tan, and Jiacheng Wang. Memcam: Memory-augmented camera control for consistent video generation. *arXiv preprint arXiv:2603.26193*, 2026. 26

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024. 23

Google DeepMind. Genie 2: A large-scale foundation world model. DeepMind Blog, https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model, 2024. 24

Google DeepMind. Genie 3: A new frontier for world models. DeepMind Blog, https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models, 2025. 24, 37

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. 22

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 24, 26

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. MineWorld: A real-time and open-source interactive world model on Minecraft. *arXiv preprint arXiv:2504.08388*, 2025. 24

Yanjun Guo, Zhengqiang Zhang, Pengfei Wang, Xinyue Liang, Zhiyuan Ma, and Lei Zhang. Memorize when needed: Decoupled memory control for spatially consistent long-horizon video generation. *arXiv preprint arXiv:2604.18215*, 2026. 25

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 24

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. 24

Junhui He, Junna Xing, Nan Wang, Rui Xu, Shangyu Wu, Peng Zhou, Qiang Liu, Chun Jason Xue, and Qingan Li. A^2ATS: Retrieval-based KV cache reduction via windowed rotary position embedding and query-aware vector quantization. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025a. arXiv:2502.12665. 6, 23

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*, 2025b. 1, 6, 24

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable

long video generation from text. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2403.14773. 24

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2403.13298. 23

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. 25

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.06654. 23

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2412.17739. 23

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on Minecraft. *arXiv preprint arXiv:2510.03198*, 2025a. 25

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. LIVE: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026a. 24

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2505.14357. 24

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 (NeurIPS)*, 2025b. Spotlight; arXiv:2506.08009. 24, 31

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 35

DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. Block-recurrent transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.07852. 20

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. MemFlow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025. 25

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020. 21

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.19466. 23

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. *arXiv preprint arXiv:2603.12513*, 2026. 1, 3, 6, 24, 26, 27, 28, 36

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video

generation models as world models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025a. arXiv:2502.20694. 26

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026a. 25

Jia Li, Xiaomeng Fu, Xurui Peng, Weifeng Chen, Youwei Zheng, Tianyu Zhao, Jiexi Wang, Fangmin Chen, Xing Wang, and Hayden Kwok-Hay So. Train short, inference long: Training-free horizon extension for autoregressive video generation. *arXiv preprint arXiv:2602.14027*, 2026b. 25

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025b. 24

Kunyang Li, Mubarak Shah, and Yuzhang Shang. PackCache: A training-free acceleration method for unified autoregressive video generation via compact KV-cache. *arXiv preprint arXiv:2601.04359*, 2026c. 24

Ruibin Li, Tao Yang, Fangzhou Ai, Tianhe Wu, Shilei Wen, Bingyue Peng, and Lei Zhang. Long-horizon streaming video generation via hybrid attention with decoupled distillation. *arXiv preprint arXiv:2604.10103*, 2026d. 25

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *International Conference on Computer Vision (ICCV)*, 2025c. arXiv:2506.18903. 25

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025d. 25

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 (NeurIPS)*, 2024. arXiv:2404.14469. 1, 22, 35, 36

Kewei Lian, Shaofei Cai, Yilun Du, and Yitao Liang. Toward memory-aided world models: Benchmarking via spatial consistency. *arXiv preprint arXiv:2505.22976*, 2025. Loop-based Minecraft navigation benchmark for spatial consistency in world models. 26

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 (ICLR)*, 2026. arXiv:2509.25161. 25

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2-bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.02750. 22

Jonathan Lorraine. *Scalable nested optimization for deep learning*. PhD thesis, University of Toronto, 2024. Ph.D. thesis; arXiv:2407.01526. 37

Jonathan Lorraine and Safwan Hossain. JacNet: Learning functions with structured Jacobians. In *ICML Workshop on Invertible Neural Networks and Normalizing Flows (INNF)*, 2019. 27

Jonathan Lorraine, Nihesh Anderson, Chansoo Lee, Quentin De Laroussilhe, and Mehadi Hassen. Task selection for AutoML system evaluation. *arXiv preprint arXiv:2208.12754*, 2022a. 29

Jonathan Lorraine, Paul Vicol, Jack Parker-Holder, Tal Kachman, Luke Metz, and Jakob Foerster. Lyapunov exponents for diversity in differentiable games. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2022b. 31

Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. ATT3D: Amortized text-to-3D object synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 37

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. 25

Weian Mao, Xi Lin, Wei Huang, Yuxin Xie, Tianfu Fu, Bohan Zhuang, Song Han, and Yukang Chen. TriAttention: Efficient long reasoning with trigonometric KV compression. *arXiv preprint arXiv:2604.04921*, 2026a. 22

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. PackForcing: Short video training suffices for long video sampling and long context inference. *arXiv preprint arXiv:2603.25730*, 2026b. 25

Nikhil Mehta, Jonathan Lorraine, Steve Masson, Ramanathan Arunachalam, Zaid Pervaiz Bhat, James Lucas, and Arun George Zachariah. Improving hyperparameter optimization with checkpointed model weights. In *ECCV Workshop on Efficient Deep Learning Foundation Models (EFM)*, 2024. 36

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023. Notable Top 5%; arXiv:2209.00588. 24

Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16300. 5, 20

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024. 20

NVIDIA. KVPress: A compression library for transformer KV caches. `https://github.com/NVIDIA/kvpress`, 2024. 22, 23, 36

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 1, 24

NVIDIA Spatial Intelligence Lab. OmniDreams: Real-time generative world model for closed-loop autonomous vehicle simulation. `https://research.nvidia.com/labs/sil/projects/omnidreams-blog/`, 2026. 1, 24

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025. 25

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.00071. 6, 23, 27, 28, 31

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022. 23

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020. 3, 20

Suraj Ranganath, Vaishak Menon, and Anish Patnaik. KV cache quantization for self-forcing video generation: A 33-method empirical study. *arXiv preprint arXiv:2603.27469*, 2026. 22

Jessie Richter-Powell, Antonio Torralba, and Jonathan Lorraine. Score distillation sampling for audio: Source separation, synthesis, and beyond. In *ICML Workshop on AI Heard That!*, 2025. arXiv:2505.04621. 37

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. 25

Sand AI. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 24

Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. LongRoPE2: Near-lossless LLM context window scaling. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.20082. 23

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning (ICML)*, 2025a. arXiv:2502.06764. 24

Yanke Song, Jonathan Lorraine, Weili Nie, Karsten Kreis, and James Lucas. Multi-student diffusion distillation for better one-step generators. In *ICML Workshop on Efficient Systems for Foundation Models (ES-FoMo)*, 2025b. arXiv:2410.23274. 31, 36

Sebastian Stapf, Pablo Acuaviva, Aram Davtyan, and Paolo Favaro. Composition of memory experts for diffusion world models. In *International Conference on Learning Representations (ICLR)*, 2026. OpenReview: https://openreview.net/forum?id=sUEdpZCHdp. 25

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1, 23

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025a. 25

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *International Conference on Machine Learning (ICML)*, 2025b. 22

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2212.10554. 23

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context LLM inference. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2406.10774. 22

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. 1, 6, 28, 37

Yuxuan Tian, Zihan Wang, Yebo Peng, Aomufei Yuan, Zhiming Wang, Bairen Yi, Xin Liu, Yong Cui, and Tong Yang. KeepKV: Achieving periodic lossless KV cache compression for efficient LLM inference. In *AAAI Conference on Artificial Intelligence*, 2026. arXiv:2504.09936. 22

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 24

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6, 24, 31

Haonan Wang, Qian Liu, Chao Du, Tongyao Zhu, Cunxiao Du, Kenji Kawaguchi, and Tianyu Pang. When

precision meets position: BFloat16 breaks down RoPE in long-context training. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2411.13476. 5, 23

Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. LLaMA-Mesh: Unifying 3D mesh generation with language models. *arXiv preprint arXiv:2411.09595*, 2024. 37

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. 6, 33, 35

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026a. 24, 37

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. AnchorWeave: World-consistent video generation with retrieved local spatial memories. *arXiv preprint arXiv:2602.14941*, 2026b. 25

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoRoPE: What makes for good video rotary position embedding? In *International Conference on Machine Learning (ICML)*, 2025. Oral; arXiv:2502.05173. 23

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*, 2026a. 25

Shunlong Wu, Hai Lin, Shaoshen Chen, Tingwei Lu, Yongqin Zeng, Shaoxiong Zhan, Hai-Tao Zheng, and Hong-Gee Kim. Semanticache: Efficient kv cache compression via semantic chunking and clustered merging. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19562–19566. IEEE, 2026b. 22

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 (NeurIPS)*, 2025. arXiv:2506.05284, project page https://spmem.github.io. 25

Xindi Wu, Despoina Paschalidou, Jun Gao, Antonio Torralba, Laura Leal-Taixé, Olga Russakovsky, Sanja Fidler, and Jonathan Lorraine. Motion attribution for video generation. In *International Conference on Machine Learning (ICML)*, 2026c. arXiv:2601.08828. 37

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 20

Haocheng Xi, Shuo Yang, Yilong Zhao, Muyang Li, Han Cai, Xingyang Li, Yujun Lin, Zhuoyang Zhang, Jintao Zhang, Xiuyu Li, Zhiying Xu, Jun Wu, Chenfeng Xu, Ion Stoica, Song Han, and Kurt Keutzer. Quant VideoGen: Auto-regressive long video generation via 2-bit KV-cache quantization. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.02958. 22

Chendong Xiang, Jiajun Liu, Jintao Zhang, Xiao Yang, Zhengwei Fang, Shizun Wang, Zijun Wang, Yingtian Zou, Hang Su, and Jun Zhu. Geometry-aware rotary position embedding for consistent video world model. *arXiv preprint arXiv:2602.07854*, 2026. 23

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2402.04617. 23

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024b. 22

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.10819. 23

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 (NeurIPS)*, 2025b. 24

Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. LATTE3D: Large-scale amortized text-to-enhanced 3D synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 37

Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, Siqi Yan, Rajasi Saha, Albert Pumarola, Wenchen Wang, and Peng Li. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation. *arXiv preprint arXiv:2604.21221*, 2026a. 25

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. UCM: Unifying camera control and memory with time-aware positional encoding warping for world models. *arXiv preprint arXiv:2602.22960*, 2026b. 23, 25, 36

Yan Team. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025. 24

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 *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2509.22622. 24

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region RoPE for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026b. 25

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026. 26, 37

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. ∞-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2511.20649. 3, 4, 6, 7, 24, 26, 33

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025. 24, 26

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 24, 31

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*, 2025a. 25

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth P. Chakaravarthy, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. MosaicMem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026. 23, 25, 36

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yang-Tian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. VideoSSM: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025b. 25

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight; arXiv:2504.12626. 24

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 36

Xuan Zhang, Cunxiao Du, Chao Du, Tianyu Pang, Wei Gao, and Min Lin. SimLayerKV: A simple framework for layer-level KV cache reduction. *arXiv preprint arXiv:2410.13846v1*, 2024. 22

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 (NeurIPS)*, 2023. 1, 22, 36

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.15894. 23

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed KV-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026. 25

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 26, 37

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI, 2025a. 35

Yuhao Zhou, Sirui Song, Boyang Liu, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Zhihao Zhang, Wei Li, and Xuanjing Huang. EliteKV: Scalable KV cache compression via RoPE frequency selection and joint low-rank projection. *arXiv preprint arXiv:2503.01586*, 2025b. 6, 23

Appendices

A Notation

Symbols used throughout the paper. Tab. 5 lists model and cache notation, indices, and virtual positions; Tab. 6 lists RoPE and key representations. Evaluation metrics: Tab. 16 (App. F.5); LoopMem parameters: App. E.

A.1 Core Notation

A.2 RoPE and Key Representations

B Related Work

B.1 Memory-Augmented Transformers

WorldTrace inherits the two-tier pattern (a recent verbatim window plus a compressed older history) from a lineage of memory-augmented transformers. Transformer-XL [Dai et al., 2019] pairs segment-level recurrence with a relative position encoding so recurrence does not corrupt absolute indices; Compressive Transformers [Rae et al., 2020] add a compressed memory of older tokens on top, and the two-tier structure we adopt. Memorizing Transformers [Wu et al., 2022] attach an external kNN-retrieved memory; Recurrent Memory Transformer [Bulatov et al., 2022] and Block-Recurrent Transformers [Hutchins et al., 2022] pass learned memory tokens between segments; Infini-attention [Munkhdalai et al., 2024] fuses a sliding window with a compressive long-term memory in one block. Landmark Attention [Mohtashami and Jaggi, 2023] introduces *landmark* tokens that gate cross-block retrieval; this antecedent shares vocabulary with our

Table 5: **Core notation.** Methods, cache structure, indices, virtual positions, scene-boundary threshold, conventions.

Symbol	Description
Method Names	
WorldTrace	Training-free KV-cache framework (Sec. 3); assigns each summary slot a fixed slot-rank virtual position
WorldTrace-Field	WorldTrace variant that compresses history by canonical key averaging (coherence; Sec. 3.3, Def. 2)
WorldTrace-Landmark	WorldTrace variant keeping verbatim scene-entry frames with frozen canonical keys (recall; Sec. 3.4, Eq. (4))
Cache Structure	
N	Total number of AR chunks in a rollout (generation length)
N_s	Number of summary slots in the KV cache
W_r	Number of recent-window slots (verbatim, newest W_r chunks)
L_{train}	Training context length in AR blocks; $N_s + W_r = L_{\text{train}}$
F	Latent frames per AR block (chunk size)
M	Source frames compressed into one summary slot under WorldTrace-Field
T_{old}	Frames outside the recent window: $T_{\text{old}} = (N - W_r)F$ for $N > L_{\text{train}}$ (linear in N)
s	Summary slot index, $s = 0, \ldots, N_s - 1$ (0 = oldest)
$\mathcal{R}$	Recent-window cache: ordered set of verbatim KVs for the newest W_r blocks
$\mathcal{S}$	Summary cache: N_s compressed slots indexed by s
K_*	Block just popped from $\mathcal{R}$ when a new chunk is appended; $K^{(k)}_{\text{cx},K_*,f}$ denotes its canonical key at intra-block frame f
SE	Scene-entry indicator: $\text{SE} = 1$ if cosine-distance spike on canonical-K exceeds τ
J	Number of detected scene-entry landmarks at the current horizon ($J \leq N_s$)
Indices (dummy / iteration)	
k	RoPE frequency-pair index
m	Source-frame iteration index (inside $\sum_{m=1}^{M}$)
n	AR-chunk iteration index
f	Intra-block frame index, $f \in \{0, \ldots, F-1\}$ (per-frame slot of an AR block; Alg. 1)
μ	Algorithm mode selector ($\mu \in \{$WorldTrace-Field, WorldTrace-Landmark$\}$, Alg. 1)
Virtual Position Assignment (Slot Indexing, Def. 1)	
q	Absolute timestamp of the current query frame
t	Absolute timestamp of a cached key frame
$t^v_{\min}$	In-distribution lower bound: $\max(0,\ q - (L_{\text{train}} - 1)F)$
$t^v_{\max}$	In-distribution upper bound for summary slots: $q - W_r F$
t_v	Generic virtual position (no slot index)
$t_v^{(s)}$	Virtual position of slot s: $q - (L_{\text{train}} - 1 - s)F$
Scene-Boundary Detection	
τ	Scene-boundary detection threshold (τ=0.15; used by WorldTrace-Landmark)
Experimental Conventions	
n	Number of distinct initial scenes (test instances) per evaluation condition
seed	Global random seed controlling initial scene selection and denoising noise
$N_s + W_r = L_{\text{train}}$	Capacity constraint: total cache always fills the training window

WorldTrace-Landmark variant but differs on three axes: their landmarks are *trained* representatives that gate attention to off-cache blocks at 1D position, while WorldTrace-Landmark is training-free, stores verbatim canonical-frame keys inside the $O(1)$ cache, and freezes them against unrotate–rerotate drift in 3D RoPE. The broader full-attention-to-compressed-state spectrum spans linear attention [Katharopoulos et al.,

Table 6: **RoPE and key representations.** Parameters, rotations, and key/query variants used across the equations.

Symbol	Description
RoPE Parameters	
θ	RoPE base frequency (θ=10000 for MG2)
θ_k	RoPE angular frequency for temporal head-dimension pair k; $\theta_k = \theta^{-k/c_t}$
c_t	Number of temporal RoPE complex pairs per head
c_h	Number of height-spatial RoPE complex pairs per head
c_w	Number of width-spatial RoPE complex pairs per head
n_ℓ	Number of transformer layers
$R(\alpha)$	Rotation matrix by angle α (applied per frequency pair)
Key / Query Variants	
$K^{(k)}_{t_m}$	RoPE-rotated key at absolute position t_m, frequency pair k
$K^{(k)}_{\mathrm{cx},m}$	Canonical (unrotated) key content: $R(-\theta_k t_m)\ K^{(k)}_{t_m}$
$\bar{K}^{(k)}_{\mathrm{cx}}$	Canonical mean across M source frames (implicit operator output of WorldTrace-Field)
$\bar{K}^{(k)}_{\mathrm{naive}}$	Naive RoPE-space average of M rotated keys
$K^{(k)}_{\mathrm{field}}(t_v)$	WorldTrace-Field compressed key at virtual position t_v (Def. 2)
$K^{(k)}_{\mathrm{land}}\left(t_v^{(s)}\right)$	WorldTrace-Landmark frozen canonical key at slot s (Eq. (4))
t_{ℓ^*}	Original timestamp of a selected landmark frame
$K^{(k)}_{t_{\ell^*}}$	Landmark source key at t_{ℓ^*}, frequency pair k
$Q^{(k)}_q$	RoPE-rotated query at q, frequency pair k
Attention Quantities	
$\delta_{q,t}$	Query-key temporal offset: $q - t$
$\Delta t_{\mathrm{train}}$	Largest temporal offset within the local attention window during training (latent frames; *e.g.* $\Delta t_{\mathrm{train}}$=5 for MG2). Distinct from the training cache extent $(L_{\mathrm{train}}-1)F$
$\ell_k(q,t)$	Attention-logit contribution from frequency pair k at offset $\delta_{q,t}$
$A^{(k)}_{q,t}$	Query-key content term in canonical coordinates for $\ell_k(q,t)$
Math Operators	
$\mathrm{Re}(\cdot)$	Real part of a complex expression (used in $\ell_k(q,t)$)
e, i	Euler's number and imaginary unit (italic; complex exponentials in Eq. (3))
sp	Spatial-axis label superscript ($e^{i\theta^{\mathrm{sp}}\cdot(h,w)}$, App. C)

2020], Mamba [Gu and Dao, 2024], and Test-Time Training [Sun et al., 2025b]; we target the fixed-budget, training-free regime in AR video diffusion.

B.2 KV Cache Compression for Large Language Models

Window and eviction methods. Window methods [Beltagy et al., 2020] retain the most recent w tokens; StreamingLLM [Xiao et al., 2024b] augments windows with initial "sink" tokens. Token eviction methods select high-importance tokens via accumulated attention mass (H_2O [Zhang et al., 2023]), key L2-norm (KnormPress [NVIDIA, 2024]), pooled attention (SnapKV [Li et al., 2024]), per-layer budget (PyramidKV [Cai et al., 2025]), lazy-layer drop (SimLayerKV [Zhang et al., 2024]), or query-aware page criticality (Quest [Tang et al., 2024]).

Token merging methods. Merging methods combine similar tokens rather than evicting them: ToMe [Bolya et al., 2023], KVCompose [Akulov et al., 2025] (composite tokens), SemantiCache [Wu et al., 2026b] (semantic clusters), and KeepKV [Tian et al., 2026] (lossless via attention-score adjustment). Pre-RoPE Q and K vectors concentrate around stable directional centers in LLMs; TriAttention [Mao et al., 2026a] scores each token by angular alignment plus magnitude and retains the top-scoring ones verbatim.

Quantization, architectural, and retrieval alternatives. KIVI [Liu et al., 2024] quantizes the cache (orthogonal to and composable with our position-content axis); on the video side, Quant VideoGen [Xi et al., 2026] achieves 2-bit KV quantization with semantic-aware smoothing for AR video diffusion, and Ranganath et al. [2026] systematically benchmark 33 quantization variants under self-forcing rollouts; quantization is

empirically decoupled from the position-OOD failure mode WorldTrace addresses. DuoAttention [Xiao et al., 2025a] partitions heads into retrieval (full-cache) and streaming (constant-size + sinks) families, parallel to our verbatim-recent vs. compressed-summary split; InfLLM [Xiao et al., 2024a] combines sliding-window attention with memory-unit retrieval for training-free long-context extrapolation. RULER [Hsieh et al., 2024] is the standard synthetic retrieval/aggregation benchmark on the LLM side; our LoopMem (App. E) extends this paradigm to closed-loop video generation.

Position handling in KV compression. Most LLM-side methods operate only on content and do not address what virtual position to assign a summary representing multiple merged tokens. EliteKV [Zhou et al., 2025b] identifies per-head frequency preferences for low-rank compression but compresses individual tokens, so position reassignment does not arise; A^2ATS [He et al., 2025a] decouples positional dependency for retrieval rather than reassigning positions for summaries. The unrotate–rerotate primitive appears in FINCH [Corallo and Papotti, 2024] and the KeyRerotationPress module of KVPress [NVIDIA, 2024], both applied per-retained-token after compression; our contribution is its use as an *averaging* operator over multiple canonical keys (WorldTrace-Field), coupled with the slot-rank positions of WorldTrace.

Why KV compression differs in our setting. The KV context spans separately denoised AR blocks rather than a single token stream, and the RoPE-OOD failure mode is benign for LLM deployments operating within the trained window (the orthogonal long-context-extension problem is in the next subsection). The angular-concentration prerequisite of TriAttention does not hold under denoising-timestep-conditioned Q: on MG2, TriAttention collapses to a norm-based eviction that matches canonical key averaging or underperforms both it and sliding-window eviction. Layer-allocation methods (PyramidKV, SimLayerKV) reallocate budget across layers but do not reassign positions.

B.3 Position Extrapolation for RoPE

LLM-side analysis. A long line of work addresses the same mechanism we exploit: RoPE [Su et al., 2024] is not robust to relative offsets beyond the training distribution. Position Interpolation [Chen et al., 2023] down-scales position indices; YaRN [Peng et al., 2024] refines this with NTK-by-parts scaling; LongRoPE [Ding et al., 2024] reaches 2M-token context via non-uniform per-dimension interpolation; LongRoPE2 [Shang et al., 2025] attributes residual OOD to undertrained higher RoPE dimensions, aligned with the per-frequency severity we measure in App. F.2. ALiBi [Press et al., 2022] replaces RoPE with a fixed linear distance bias; xPos [Sun et al., 2023] adds an exponential-decay term; Kazemnejad et al. [2023] (NoPE) show position-free transformers can outperform RoPE/ALiBi/APE on length generalization (clarifying that our RoPE-OOD argument concerns models *already* trained with RoPE). CoPE [Golovneva et al., 2024] makes positions content-conditional; Barbero et al. [2025] dissect which RoPE frequencies carry position vs. semantics. HoPE [Chen et al., 2025b] and FoPE [Hua et al., 2025] perform per-frequency RoPE-OOD analysis at the LLM scale via cascade-failure and Non-Uniform Discrete Fourier Transform theory, respectively, anchoring our per-frequency view (App. F.2).

Vision and video. RoPE-ViT [Heo et al., 2024] adapts RoPE to vision transformers via a 2D RoPE-Mixed split; VideoRoPE [Wei et al., 2025] introduces a 3D RoPE structure with low-frequency temporal allocation and the V-NIAH-D long-context retrieval test for video LLMs; ViewRope [Xiang et al., 2026] replaces screen-space coordinates with camera-ray geometry to maintain loop-closure consistency in pose-conditioned video world models; RIFLEx [Zhao et al., 2025] achieves training-free length extrapolation in video diffusion via per-frequency intrinsic-frequency reduction.

Position-content coupling. Wang et al. [2025] show position couples to numerical precision (bf16 deviates RoPE from its intended relative encoding over long contexts), consistent with our position-content coupling thesis (Sec. 3.5). On the video side, UCM [Xu et al., 2026b] reassigns 3D positional encodings via time-aware warping for camera-controlled world models, and MosaicMem [Yu et al., 2026] introduces "Warped RoPE" that reprojects positional encodings of memory patches into the target view. Both are virtual-position-assignment operators in our sense, but their warping is camera/geometry-driven and paired with trained content writers; ours is slot-rank-driven and training-free.

Why RoPE extrapolation differs in our setting. These works adapt position embeddings for LLMs that read a single growing sequence (or for video LLMs operating on a fixed-length input); our setting differs in two structural ways: (i) inference proceeds across many independently denoised AR video blocks rather than a single autoregressive token stream, so the cache must be *compressed* rather than just *re-mapped*, and (ii) compressing M source frames into one slot raises the question of *which* virtual position to assign that summary, a question that does not arise when each cached token retains its own identity. WorldTrace addresses (ii) by assigning each slot a fixed slot-rank offset.

B.4 Autoregressive Video World Models

Lineage and training paradigms. Recent autoregressive video diffusion models [Brooks et al., 2024, Bruce et al., 2024, Google DeepMind, 2024, 2025, He et al., 2025b, Decart et al., 2024, Valevski et al., 2025, Guo et al., 2025, Li et al., 2025b, Yan Team, 2025] generate interactively via AR chunk prediction. The dominant training paradigm Self-Forcing [Huang et al., 2025b] (built on Diffusion Forcing [Chen et al., 2024], DFoT [Song et al., 2025a], and CausVid [Yin et al., 2025]) rolls out on the student's own KV cache, inheriting a local attention window not designed for arbitrarily long inference. Self-Forcing++ [Cui et al., 2026b] extends this to multi-minute video via long-rollout self-distillation; FAR [Gu et al., 2025] reformulates AR video as next-frame prediction over an unbounded cache; Vid2World [Huang et al., 2026b] causalizes bidirectional video diffusion into interactive world models; LIVE [Huang et al., 2026a] stabilizes long horizons via a cycle-consistency objective. Our work is complementary, operating purely at inference time on top of an already-trained model. Open foundation backbones [Wan Team, 2025, NVIDIA, 2025, Sand AI, 2025] provide the capacity that makes long-horizon AR rollout viable.

World-model and RL lineage. The world-model framing traces to Ha and Schmidhuber [2018] through the Dreamer line [Hafner et al., 2025]; transformer-based world models like IRIS [Micheli et al., 2023] and DIAMOND [Alonso et al., 2024] establish the autoregressive transformer as a sample-efficient world-model backbone, and open foundation platforms [NVIDIA, 2025, Wan Team, 2025] provide the capacity our inference-time intervention runs on. OmniDreams [NVIDIA Spatial Intelligence Lab, 2026] is a real-time generative world model for closed-loop autonomous vehicle simulation.

B.5 Memory and KV Cache for Video World Models

Existing memory approaches. Existing systems use one of three cache approaches. *Sliding-window KV caches* (MG2 [He et al., 2025b], LongLive [Yang et al., 2026a]) bound memory at $O(1)$ but discard history and leave cached positions OOD. *Memory re-encoding* (MG3 [Wang et al., 2026a]) selects past frames via field-of-view similarity and re-encodes at each AR step, sidestepping both problems at the cost of roughly doubled per-step latency; FramePack [Zhang et al., 2025] compresses input frame contexts by frame-wise importance, and WorldMem [Xiao et al., 2025b] attaches a state-aware memory of pose-tagged frames re-injected via auxiliary attention. *Architectural memory modules* include StreamingT2V [Henschel et al., 2025], which pairs a short conditional-attention window with a long-term appearance-preservation module anchored on the first chunk.

Concurrent training-free cache work. Closest to ours, four concurrent training-free works share our diagnosis (temporal RoPE OOD as the long-horizon bottleneck) but pair it with a different content writer: Infinity-RoPE [Yesiltepe et al., 2026] uses Block-Rel offsets with KV Flush (sink + most recent block only); FAR [Gu et al., 2025] introduces FlexRoPE for temporal-decay extrapolation on an unbounded $O(T)$ cache; Deep Forcing [Yi et al., 2025] dedicates $\sim$50% of the window to "Deep Sink" tokens with re-aligned RoPE phases; MemRoPE [Kim et al., 2026] pairs Block-Rel positions with a dual-rate EMA summary cache. The first three either operate on existing tokens (verbatim or sink-aligned) or scale to unbounded caches; WorldTrace, instead, assigns positions to *compressed summaries* that represent multiple merged frames within a fixed budget. Deep Forcing's importance-pruning is content-side and composes with slot-rank. Tab. 8 shows that transplanting WorldTrace positions into MemRoPE without retuning its EMA content regime regresses performance, confirming position and content must be jointly designed. PackCache [Li et al., 2026c] is also training-free and uses a "spatially preserving position embedding" for cache removal; it applies virtual position assignment per token rather than per summary slot, paired with cross-frame decay rather than canonical-key

averaging or verbatim landmarks. FlowCache [Ma et al., 2026] introduces chunk-wise denoising-step caching and an importance–redundancy KV compression scheme for autoregressive video; its goal is per-step compute reduction rather than long-horizon recall, but the chunk-aware cache structure is similar to the recent–summary split we adopt. TempCache [Samuel et al., 2026] merges near-duplicate cached keys across AR frames via approximate-nearest-neighbor temporal correspondence, paired with sparse self- and cross-attention; merging operates at per-token similarity level rather than reassigning summary-slot virtual positions in canonical space.

Closest concurrent training-time analogs. Two contemporaneous works share WorldTrace's diagnosis and propose related fixes through training-time mechanisms. Grounded Forcing [Chen et al., 2026b] introduces a Dual Memory KV Cache (Local Temporal Memory plus Global Consistency Memory) coupled with Dual-Reference RoPE Injection that stores raw pre-RoPE keys and injects fixed $t{=}0$ for global anchors and relative offsets for local frames; their dual-reference indexing is the training-time analog of our slot-rank virtual positions. Anchor Forcing [Yang et al., 2026b] partitions the cache into sink, junction, and local regions, each with its own RoPE reference origin capped at the pretrained limit, and learns the assignment via RoPE re-alignment distillation. WorldTrace differs in being inference-time-only with no distillation or retraining, in decomposing retention into orthogonal coherence (WorldTrace-Field) and verbatim recall (WorldTrace-Landmark) operators, selectable per task, and in pinning summary-slot position rather than content as the binding constraint. Composition of Memory Experts [Stapf et al., 2026], the long-term spatial-memory framework of Wu et al. [2025], and Mixture of Contexts [Cai et al., 2026] provide training-time decompositions: a contrastive product-of-experts over short-term, long-term episodic, and spatial memory; point-cloud spatial memory plus sparse episodic keyframes; and sparse top-k routing with mandatory text/intra-shot anchors. Context Forcing [Chen et al., 2026d] addresses the same student-teacher mismatch by training a long-context student under a long-context teacher with a Slow-Fast Memory architecture that bounds attention cost while preserving 20 s+ effective context. Stable Video Infinity [Li et al., 2025d] bridges the train-test hypothesis gap via Error-Recycling Fine-Tuning that injects historical errors as supervisory prompts so the DiT learns to correct its own drift. Hybrid Forcing [Li et al., 2026d] combines a linear-attention summary state for evicted tokens with block-sparse local attention through a decoupled-distillation pipeline. All three are training-time analogs that complement WorldTrace's inference-time-only operation; WorldTrace can, in principle, compose with any of them at the cache level.

Concurrent trained video-memory architectures. A second cluster of concurrent works addresses the same long-horizon memory problem but with trained components. RELIC [Hong et al., 2025] stores compressed historical latents in the KV cache with both relative-action and absolute-camera-pose annotations (camera-aware memory); MosaicMem [Yu et al., 2026] pairs Warped RoPE (geometry-driven virtual position assignment) with Warped Latent injection; UCM [Xu et al., 2026b] reassigns 3D positional encodings via time-aware warping for camera-controlled world models; PackForcing [Mao et al., 2026b] partitions the cache into sink/mid-compressed/recent and applies a Continuous Temporal RoPE Adjustment to re-align position gaps left by dropped tokens (zero-shot or 5s-clip-trained). Compared to these, WorldTrace pursues the strict zero-fine-tune regime, depends on slot-rank rather than camera/geometry signals, and decomposes the cache update into orthogonal coherence (WorldTrace-Field) and recall (WorldTrace-Landmark) operators that can be selected per task without retraining.

Other concurrent video-cache and memory-bank work. Other concurrent training-free work spans position-side fixes (LoL [Cui et al., 2026a], FLEX [Li et al., 2026b]), content-side selection (PaFu-KV [Chen et al., 2026a], Rolling Sink [Li et al., 2026a], Relax Forcing [Zhao et al., 2026], Rolling Forcing [Liu et al., 2026]), world-model architectures (WorldPack [Oshima et al., 2025], WorldPlay [Sun et al., 2025a], VideoSSM [Yu et al., 2025b]), and memory-bank designs (Memory Forcing [Huang et al., 2025a], VMem [Li et al., 2025c], Context as Memory [Yu et al., 2025a], AnchorWeave [Wang et al., 2026b], Infinite-World [Wu et al., 2026a], HyDRA [Chen et al., 2026c], MemFlow [Ji et al., 2025], Memorize-When-Needed [Guo et al., 2026], VRAG [Chen et al., 2025a]). Sparse Forcing [Xu et al., 2026a] learns native block-sparse attention plus persistent spatiotemporal anchors via a Persistent Block-Sparse Attention kernel, observing the same "implicit spatiotemporal memory" on

Table 7: **KV cache comparison.** WorldTrace makes compressed memory addressable; WorldTrace-Field uses a field vs. WorldTrace-Landmark retaining selected landmarks.

Method	Fixes pos. OOD?	Bounded memory?	Preserves history?	Training-free?
Full KV	No	No ($O(N)$)	Yes (exact)	Yes
Sliding window	No	Yes ($O(1)$)	Recent only	Yes
KV Flush + Block-Rel [Yesiltepe et al., 2026] (KV-Flush variant)	Yes	Yes ($O(1)$)	No (amnesic)	Yes
Naive + Block-Rel	Partial	Yes ($O(1)$)	Corrupted (Sec. 2.2)	Yes
MemRoPE [Kim et al., 2026]	Partial	Yes ($O(1)$)	EMA-diluted	Yes
FAR (FlexRoPE) [Gu et al., 2025]	Yes	No ($O(T)$)	Yes (exact)	Yes
Deep Forcing [Yi et al., 2025]	Partial	Yes ($O(1)$)	Sink + importance-pruned recent	Yes
KnormEvict + WorldTrace	Yes	Yes ($O(1)$)	Partial (1 token/slot)	Yes
WorldTrace-Field (ours)	**Yes**	**Yes ($O(1)$)**	**Yes (compressed field)**	**Yes**
WorldTrace-Landmark (ours)	**Yes**	**Yes ($O(1)$)**	**Yes (verbatim landmarks)**	**Yes**

persistent KV slots that motivates WorldTrace-Landmark's scene-entry detection but training the sparsity end-to-end. MemCam [Gao et al., 2026] pairs a context-compression module with co-visibility-based historical-frame selection for camera-controlled long video generation; the selection is geometry-driven rather than canonical-K-spike-driven, and the framework is trained. MemFlow retrieves the most relevant historical frames per chunk and activates only the top-k tokens in attention; Memorize-When-Needed adds a decoupled memory branch with camera-aware gating that conditions generation on memory only when meaningful historical references are present; VRAG augments interactive video generation with explicit global-state retrieval to reduce compounding errors. The memory-bank designs store geometry-anchored or retrieval-indexed frames *outside* the standard KV cache and re-inject them via auxiliary attention; WorldTrace instead compresses history into a constant-size summary *within* the standard KV cache, supporting both coherence (WorldTrace-Field) and verbatim recall (WorldTrace-Landmark) under a single $O(1)$ budget without re-encoding or retrieval.

Evaluation benchmarks for world generation. WorldScore [Duan et al., 2025], MIND [Ye et al., 2026], and VBench-2.0 [Zheng et al., 2025] provide unified long-horizon evaluation; WorldModelBench [Li et al., 2025a] explicitly positions video generation as world modeling. Closer to our setting, Lian et al. [2025] constructs a Minecraft loop-navigation benchmark that assesses spatial consistency during revisits to previously seen locations. Our LoopMem benchmark (App. E) targets the orthogonal axis of *closed-loop* episodic recall (return-to-origin trajectories) on top of pretrained autoregressive video world models, complementing rather than replacing these broader quality benchmarks.

Differentiation summary. Tab. 7 summarizes four axes: whether positional OOD is fixed, whether memory is bounded, whether intermediate history is preserved, and whether retraining is required.

C WorldTrace-Field: Additional Properties

This section expands Eq. (3), summarised at the end of Sec. 3.3.

Mean attention preservation. The WorldTrace-Field compression preserves the mean attention logit: a single summary token $K^{(k)}_{\text{field}}(t_v)$ reproduces exactly the average of the logits that the M individual source keys would produce if they were all relocated to the same virtual position t_v. This holds for every query $Q^{(k)}_q$ and every temporal RoPE frequency k, so no query can distinguish a slot that was built from one frame versus many.

Proposition 1 (Mean attention preservation)**.** *Fix a temporal RoPE pair k and a virtual position t_v. For any query $Q^{(k)}_q$,*

$$\langle Q^{(k)}_q, K^{(k)}_{\text{field}}(t_v)\rangle = \frac{1}{M}\sum_{m=1}^{M}\langle Q^{(k)}_q, R(\theta_k t_v)\, K^{(k)}_{\text{cx},m}\rangle.$$

By Eq. (3), $K^{(k)}_{\text{field}}(t_v) = R(\theta_k t_v)\,\bar{K}^{(k)}_{\text{cx}}$ with $\bar{K}^{(k)}_{\text{cx}} = \frac{1}{M}\sum_m K^{(k)}_{\text{cx},m}$; linearity of the inner product in its second argument then gives $\langle Q^{(k)}_q, K^{(k)}_{\text{field}}(t_v)\rangle = \langle Q^{(k)}_q, R(\theta_k t_v)\,\bar{K}^{(k)}_{\text{cx}}\rangle = \frac{1}{M}\sum_m \langle Q^{(k)}_q, R(\theta_k t_v)\,K^{(k)}_{\text{cx},m}\rangle$. The summary token contributes the mean attention logit that the source keys would produce if their content were first moved to the shared virtual position t_v.

Concretely, $K^{(k)}_{\text{field}}(t_v)$ acts as if all M source frames were simultaneously repositioned to t_v: no individual timestamp t_m appears in the compressed representation, only the canonical content average $\bar{K}^{(k)}_{\text{cx}}$. This fully

decouples the summary content from its temporal placement, enabling WorldTrace to freely reassign virtual positions (Def. 1) without distorting the stored signal.

Characterization. Eq. (3) is the linear compression operator that (a) maps M RoPE-encoded keys to a single token at virtual position t_v and (b) satisfies the mean attention preservation property of Prop. 1: unrotate each key to canonical space, average, and re-encode at t_v. The content of the compressed token is fully determined by property (b); any linear operator satisfying it must produce the same canonical average $\bar{K}_{\text{cx}}^{(k)}$, since the property pins the output attention logit for all queries. Only t_v remains free, and WorldTrace chooses it (Sec. 3.2). Pinning a learned operator by a structural property of its derivative (here, the block-diagonal phase rotation of RoPE that Eq. (3) inverts) has antecedents in structured-Jacobian parameterizations [Lorraine and Hossain, 2019]; WorldTrace-Field is the closed-form, training-free analog for the temporal RoPE block. Non-linear alternatives (*e.g.* selecting the highest-norm token verbatim) satisfy neither property and lose phase coherence across compressed frames. We confirm this empirically: KnormEvict+WorldTrace (single highest-norm canonical key per slot) achieves PAC 0.553 vs. 0.574 for WorldTrace-Field (-3.7%, p=0.074, n=10, paired at N=16), with TempSSIM also significantly lower (p=0.007), consistent with information loss from discarding all history except the single highest-norm frame.

Spatial RoPE is unaffected. Spatial RoPE (encoding height h and width w) is fixed across all frames at the same spatial position. The spatial phase $e^{i\theta^{\text{sp}}h}$ is identical for tokens (f_1, h, w) and (f_2, h, w); it factors out of the sum in Eq. (3) and is preserved exactly. The temporal unrotation and re-rotation in Eq. (3) touch only the first $2c_t$ head dimensions, at 2 FLOP per key per channel pair.

Norm collapse. Averaging M uncorrelated canonical keys reduces the norm by $\sim 1/\sqrt{M}$, suppressing the attention weight of summary tokens in the softmax. A norm-preserving rescale can counteract this; we evaluate it as canonical key averaging+Norm in App. D. The rescale does not improve our method at long horizons, so we omit it.

D Additional Experimental Results

D.1 Concurrent Training-Free Baselines

Tab. 8 compares MemRoPE [Kim et al., 2026] and YaRN [Peng et al., 2024] on the Sec. 4.3 ABA loops, factoring the position scheme (Block-Rel vs. WorldTrace) from the content writer at two horizons.

MemRoPE [Kim et al., 2026] (Block-Rel + dual-rate EMA with α_{long}=0.01 and α_{short}=0.1, matching the original release) survives OOD positions via EMA decay but is still beaten by Landmark+Block-Rel: verbatim canonical keys retain stronger query-key similarity than smoothed averages, and WorldTrace-Landmark adds in-distribution positions on top. MemRoPE+WorldTrace (Block-Rel swapped for WorldTrace, EMA kept) degrades (p<0.001): write/read offsets disagree.

YaRN [Peng et al., 2024] rescales temporal RoPE frequencies to bring offsets back in-distribution, gaining $+22\%$ over the sliding-window baseline at N=32 (p<0.001) and confirming position is the primary bottleneck; but it needs $O(N)$ memory (OOM by $\sim N$=100) and the rescale diverges with horizon. WorldTrace-Landmark exceeds it by a wide margin at N=32 (0.964 vs. 0.490) in $O(1)$ memory and stays near-constant through N=512.

D.2 PAC Episodic Recall Sweep

Tab. 9 gives the full PAC breakdown across $N \in \{16, 32, 48, 64, 128, 256\}$ on ABA loops (N_s=4, W_r=2) and partitions methods by retained past content into three tiers: *compression-only* (sliding window, Naive, WorldTrace-Field), *canonical-K anchoring* (Latent re-anchor, WorldTrace-Field with scene anchoring), and *verbatim recall* (WorldTrace-Landmark). Within compression, WorldTrace-Field beats the sliding-window baseline at N=32–64 (p<0.001, paired t-test). Anchoring lifts recall by 40–50 points at moderate horizons; only verbatim recall sustains it at N=256. At N=256 the canonical-K tier's PAC falls to 0.610 while WorldTrace-Landmark holds at 0.989.

Table 8: **Concurrent training-free baselines.** PAC (↑) at two horizons (ABA). YaRN [Peng et al., 2024] uses $O(N)$ memory; only at N=32. MemRoPE+WorldTrace swaps Block-Rel for WorldTrace; dual-rate EMA unchanged.

Method	Position	Content	N=32 (16×)	N=48 (24×)
$O(N)$ cache:				
Sliding window (baseline)	actual	Sliding window	0.401	0.388
YaRN [Peng et al., 2024]	NTK	Sliding window	0.490	0.412
$O(1)$ cache, Block-Rel positions:				
MemRoPE [Kim et al., 2026]	Block-Rel	dual EMA	0.651	0.706
Landmark + Block-Rel	Block-Rel	verbatim	0.929	0.934
$O(1)$ cache, WorldTrace positions:				
MemRoPE + WorldTrace positions	WorldTrace	dual EMA	0.592	0.662
WorldTrace-Landmark **(ours)**	WorldTrace	verbatim	**0.964**	**0.972**

Table 9: **Episodic recall splits into three tiers; only verbatim recall scales.** ABA loops, PAC (↑); bold = best per column.

Method	N=16	N=32	N=48	N=64	N=128	N=256
Compression only:						
Sliding window (baseline)	0.540	0.401	0.388	0.412	0.504	0.631
Naive + Block-Rel	0.570	0.434	0.433	0.443	0.486	0.598
WorldTrace-Field (ours)	0.555	0.434	0.436	0.442	0.495	0.602
Canonical-K anchoring:						
Latent re-anchor	0.955	0.952	0.938	0.913	0.782	0.610
WorldTrace-Field with scene anchoring	0.955	0.951	0.935	0.911	0.777	0.624
Verbatim recall:						
WorldTrace-Landmark (ours)	**0.959**	**0.964**	**0.972**	**0.976**	**0.986**	**0.989**

PAC statistics (s.e.m.). Companion to Tab. 9. Sliding window: ±0.032 (N=16), ±0.029 (N=32), ±0.015 (N=128), ±0.013 (N=256). WorldTrace-Field: ±0.011 (N=256). Latent re-anchor: ±0.005 (N=16), ±0.004 (N=32), ±0.016 (N=128), ±0.022 (N=256). WorldTrace-Field with scene anchoring: ±0.006–±0.020 across N=16–64; ±0.022 (N=256).

D.3 Cross-Architecture Experiments

Tab. 10 reports PAC at 2×–8× training horizon on LingBot-Fast [Team et al., 2026] with WorldTrace applied. WorldTrace-Field matches or exceeds the sliding-window baseline at all horizons: Plücker camera conditioning already supplies the recall signal that KV-cache content provides on MG2, leaving no room for canonical averaging to add. WorldTrace-Landmark improves over sliding-window retention from 4× onward (+8.9%, +14.1%, +7.3% at 4×/6×/8×), with no significant gain at 2× (−0.006 absolute, p>0.1): verbatim key injection provides a complementary recall signal that strengthens as camera-pose priors alone become insufficient at longer horizons. A camera-ablation corroborates this: zeroing Plücker embeddings raises sliding-window PAC by +7.0% while depressing WorldTrace-Field by −23.2%, confirming that canonical averaging dilutes scene-origin content when the pose-based path is removed.

Table 10: **LingBot-Fast [Team et al., 2026] PAC at extended horizons.** WorldTrace-Landmark improves over the sliding-window baseline from 4× onward; WorldTrace-Field matches the sliding window at all horizons, consistent with Plücker conditioning supplying the primary recall signal.

N	Horizon	Sliding window	WorldTrace-Field	WorldTrace-Landmark
14	2×	0.657	**0.668**	0.651
28	4×	0.624	0.619	**0.680**
42	6×	0.591	0.620	**0.674**
56	8×	0.632	0.648	**0.678**

D.4 Slot Allocation

Summary/recent split sensitivity. The main experiments use $N_s{=}4$ summary slots and $W_r{=}2$ recent-window slots (total $N_s{+}W_r{=}6$, matching the training context size). To test sensitivity to this split, we swept five allocations ($N_s{+}W_r{=}6$, ABA loops at $N{=}32$ and $N{=}64$); results are in Tab. 11. $N_s{\leq}2$ collapses toward the sliding-window baseline: with one slot the B→A detector overwrites the scene-A landmark; with two slots the B-side traversal evicts it before the return. $N_s{=}3$ is intermediate ($0.652/0.693$ at $N{=}32/64$) but still evicts the landmark. $N_s{=}4$ retains the landmark through the B-side return; $N_s{=}5$ adds only $+0.009/{+}0.005$, confirming a four-slot plateau.

Table 11: **Slot Allocation.** PAC at $N{=}32$ and 64 as a function of N_s (total budget fixed at $N_s{+}W_r{=}6$). $N_s{\leq}2$ collapses toward the sliding-window baseline; four slots is the critical boundary for retaining the scene-A landmark through the B-side traversal.

N_s	W_r	PAC $N{=}32$	PAC $N{=}64$
Sliding window (reference)		0.401	0.412
1	5	0.413	0.437
2	4	0.419	0.426
3	3	0.652	0.693
4	2	0.964	0.976
5	1	0.973	0.981

Per-slot depth ablation. To isolate which slot carries the recall signal, we run WorldTrace-Landmark with only one summary slot active at a time (inactive summary slots revert to sliding-window eviction); results are in Tab. 12. Slot 3 alone recovers most of the four-slot PAC at $N{=}32$ and $N{=}48$ (Tab. 12: 0.821 vs. 0.964 at $N{=}32$; the shortfall vs. all four slots reaches 0.184 at $N{=}48$). Slot 1 only sits on a shallow plateau (0.630, 0.589); Slot 2 only improves moderately (0.628, 0.554) but remains far below Slot 3, with the gap widening over horizon ($+0.193$ at $N{=}32$ and $+0.234$ at $N{=}48$, Slot 3 only minus Slot 2 only).

Table 12: **Per-slot depth ablation.** PAC when exactly one summary slot runs WorldTrace-Landmark and the remaining summary slots revert to sliding-window eviction. Slot 1 only still collapses recall; Slot 2 only improves slightly yet remains far below full recall at longer horizons; **Slot 3 only** recovers most of the gain, and the full four-slot WorldTrace-Landmark achieves the best recall at every horizon.

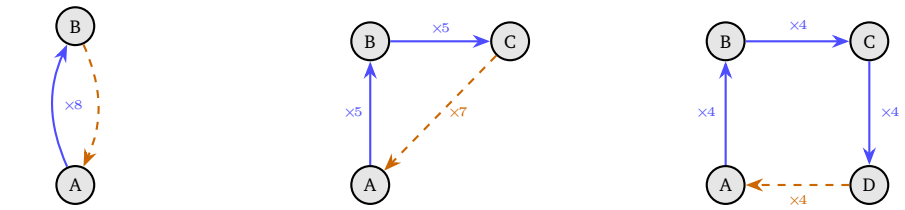
Method	$N{=}32$	$N{=}48$
Slot 1 only	0.630	0.589
Slot 2 only	0.628	0.554
Slot 3 only	**0.821**	**0.788**
WorldTrace-Landmark (all 4 slots)	0.964	0.972

E LoopMem: Scripted Navigation Benchmark

LoopMem evaluates episodic spatial recall in autoregressive world models: a model executes a scripted navigation path that returns to a previously visited location, and the generated return frame is scored against the original scene appearance at geometrically matched positions. Unlike global video-quality metrics, LoopMem requires no external reference; the forward path generates the target, and the return path is scored against it.

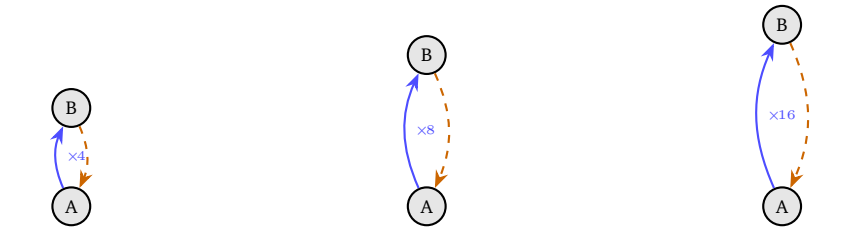
The benchmark organizes difficulty along four axes: (1) **waypoint count** (K): number of intermediate locations between departure and return to A (ABA: $K{=}1$; ABCDA: $K{=}3$); (2) **edge length** (L): AR chunks per directed edge, controlling KV-context distance; (3) **camera orientation**: heading on arrival at A relative to departure, testing viewpoint-invariant recall; and (4) **multi-revisit depth** (R): how many times a waypoint is re-entered within a single rollout. The four-axis decomposition follows the broader principle that benchmark suites should isolate task properties capable of producing distinguishable rankings between candidate systems [Lorraine et al., 2022a], here projected onto the position–content axes that the cache mechanism is hypothesized to probe. Fig. 6 shows representative configurations; those evaluated here are marked *.

Row 1: varying topology (K intermediate waypoints)



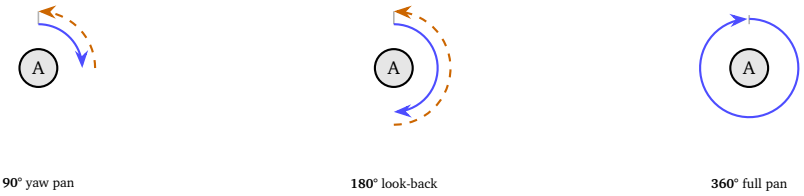
ABA* $K{=}1$, $N{=}16$ **ABCA*** $K{=}2$, $N{=}17$ **ABCDA*** $K{=}3$, $N{=}16$

Row 2: varying edge length (L chunks per leg, ABA topology)



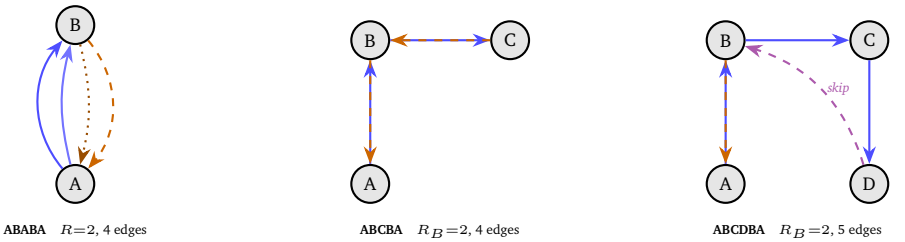
short $L{=}4$, $N{=}8$ *medium** $L{=}8$, $N{=}16$ *long** $L{=}16$, $N{=}32$

Row 3: camera-orientation pans (blue = pan away, orange dashed = return)



90° yaw pan **180°** look-back **360°** full pan

Row 4: multi-revisit patterns ($R > 1$; solid/lighter arcs = 1st/2nd traversal)



ABABA $R{=}2$, 4 edges **ABCBA** $R_B{=}2$, 4 edges **ABCDBA** $R_B{=}2$, 5 edges

Figure 6: **LoopMem benchmark gallery. Blue solid**: outward; **orange dashed**: return or palindrome; **violet dashed**: shortcut omitting intermediate waypoints. **Row 1**: topology (ABA/ABCA/ABCDA). **Row 2**: ABA edge length (longer edges push RoPE further OOD). **Row 3**: camera orientation (agent at A; blue = pan away, orange = return). **Row 4**: multi-revisit ($R > 1$): palindromes (ABCBA), shortcuts (ABCDBA).

Figure 7: **ABA qualitative results.** Each sample group of three frames shows trajectory keyframes at chunks 0 (A) / 7 (B) / 15 (A return) for one initial frame; the two rows compare the sliding-window baseline (top) vs. WorldTrace-Landmark (bottom). The sliding window drifts away from the scene-A appearance on the return leg in all four samples; WorldTrace-Landmark anchors the return to the original scene.

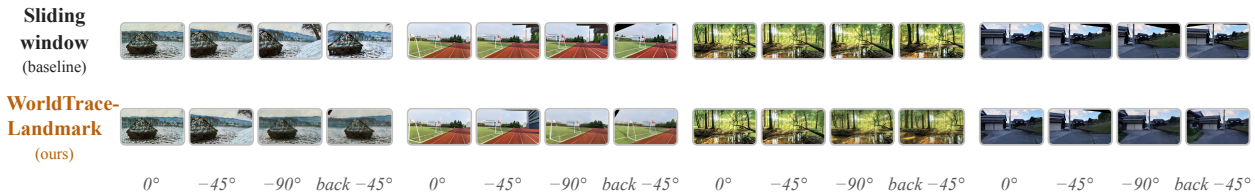


Figure 8: **Pan $90°$ qualitative results.** Camera-orientation Tier 3 (N=4): the agent is fixed at A while the camera pans right by $\sim 90°$ then returns. Each sample group of four frames shows keyframes at chunks 0 ($0°$) / 1 ($\sim 45°$) / 2 ($\sim 90°$) / 3 (back $\sim 45°$); the two rows compare the sliding-window baseline (top) vs. WorldTrace-Landmark (bottom). WorldTrace-Landmark restores the initial-view appearance on the return half-pan; the sliding window does not.

F Implementation Details

F.1 Model Architecture and Self-Forcing Background

We evaluate on MG2-1.3B, a 1.3B-parameter distilled autoregressive video world model based on Wan 1.3B T2V [Wan Team, 2025], with 30 transformer layers, 12 attention heads, and a head dimension of 128. Video is generated autoregressively in AR blocks of 3 latent frames each at spatial resolution 44×80 (352×640 pixels), yielding 880 tokens per frame $\times$ 3 frames = 2640 tokens per AR block per layer. Each new block is denoised via a multi-step flow-matching process conditioned on the KV cache of all prior blocks. During training with Self-Forcing [Huang et al., 2025b] (rollout on the student's own KV cache rather than teacher-forced context; *c.f.* Diffusion Forcing [Chen et al., 2024], CausVid [Yin et al., 2025], and one-step diffusion distillation [Song et al., 2025b]), attention is restricted to a local window of `local_attn_size`$= 6$ frames (2 AR blocks), so cross-frame temporal offsets stay $\leq \Delta t_{\text{train}}$=5. At inference, the rolling KV cache grows without bound; with a constant $O(1)$ budget, our method caps it at $L_{\text{train}} \times 2640 \times 12 \times 256 \times 30 \times 2 \approx 2.79$ GB per batch element (L_{train}=6 chunks, 2640 tokens/chunk, fp16; *c.f.* Tab. 14), independent of generation length. At long horizons, the OOD mismatch is not the presence of cached keys at large offsets, but the temporal RoPE rotation through which they are read: keys are stored verbatim, but the model was never trained to invert the phase rotations they accumulate beyond Δt_{train}.

F.2 Rotary Position Embeddings in 3D Video Attention

We inherit the 3D RoPE split of the Wan 2.1 backbone [Wan Team, 2025]: the 128-dimensional per-head embedding is split across three position axes, $2c_t$=44 dimensions for temporal position, and $2c_h$=$2c_w$=42 dimensions each for spatial height and width, with shared base frequency $\theta = 10000$. The k-th temporal frequency component follows $\theta_k = \theta^{-k/c_t}$ for $k = 0, \ldots, c_t - 1$, giving $\theta_0 = 1.0$ (fastest rotating) down to $\theta_{c_t-1} \approx 1.5 \times 10^{-4}$ (slowest). The attention score contribution from frequency k between the query at temporal position q and the key at t is proportional to $\cos(\theta_k(q-t))$. When $\theta_k|\delta_{q,t}| \ll \pi$, the cosine is near 1, and the component is position-invariant, carrying semantic content; when $\theta_k|\delta_{q,t}| \gg \pi$, it oscillates incoherently and constitutes positional noise. This wavelength-vs-context view follows YaRN's NTK-by-parts framing [Peng et al., 2024] and is consistent with the mechanistic finding that high-frequency RoPE components carry positional structure while low-frequency components carry semantics [Barbero et al., 2025]. Viewed as a long-horizon discrete dynamical system, the rapid-oscillation regime is the position-side analog of the local-divergence behavior studied via Lyapunov exponents in iterated maps [Lorraine et al., 2022b]: once $\theta_k|\delta_{q,t}| \gg \pi$ the cosine

kernel ceases to be Lipschitz in $\delta_{q,t}$ at training-scale resolution, so neighboring offsets receive uncorrelated attention weights and the recall signal decoheres across the rollout. At training max offset $\Delta t_{\text{train}} = 5$, components $k \geq 10$ satisfy $\theta_k \times 5 < 0.08$ rad and act as near-semantic carriers; at inference max offset $|\delta_{q,t}| = 30$, the three fastest components ($k \leq 2$) all exceed 3π rad, with k=3 at 8.5 rad and k=5 at 3.7 rad.

Table 13: **RoPE OOD severity by frequency.** Per-frequency RoPE phase at inference $|\delta_{q,t}|$=30 vs. training max $|\delta_{q,t}|\leq 5$ (MG2-1.3B). Low frequencies ($k\leq 2$) hit noise-level phases ($> 3\pi$ rad); high frequencies ($k\geq 10$) stay near in-distribution.

Freq k	θ_k	Train max ($\lvert\delta_{q,t}\rvert$=5)	Inference ($\lvert\delta_{q,t}\rvert$=30)	Status
k=0	1.000	5.0 rad	30.0 rad (9.5π)	Random noise
k=5	0.123	0.62 rad	3.70 rad (1.2π)	OOD
k=10	0.0152	0.076 rad	0.46 rad	Mild OOD
k=15	1.9×10^{-3}	0.0094 rad	0.056 rad	Near in-dist
k=18	5.4×10^{-4}	0.0027 rad	0.016 rad	Fully in-dist
k=21	1.5×10^{-4}	7.6×10^{-4} rad	0.0046 rad	Fully in-dist

Table 14: **MG2-1.3B model and generation statistics.** OOM horizon: shortest length where full-KV inference OOMs on one A100 80 GB; safe horizon: longest degradation-study length without OOM.

Property	**Value**
Architecture	
Transformer layers	30
Attention heads	12
Head dimension	128
Temporal RoPE complex pairs (c_t)	22
Video & Latent Space	
Decoded output FPS	16
VAE temporal compression	4×
VAE spatial compression	8× (each axis)
Latent resolution	44×80
Spatial tokens per latent frame	$22 \times 40 = 880$
Latent frames per chunk (F)	3
Decoded frames per chunk	12
Chunk duration	0.75 s
Training Context	
Training context length (L_{train})	6 chunks
Training context in latent frames	18
Training window duration	≈ 4.5 s
Tokens per chunk	$3 \times 880 = 2{,}640$
KV memory per chunk (fp16, all layers)	≈ 465 MB
Inference Horizon	
Safe inference horizon (no compression)	≈ 31 chunks (≈ 23 s)
OOM horizon (full KV cache, A100 80 GB)	≈ 150 chunks (≈ 112 s)
N=256 horizon	≈ 192 s (≈ 3 min)
N=512 horizon	≈ 384 s (≈ 6 min)

F.3 WorldTrace Algorithm and Architectural Baselines

WorldTrace cache update. Each autoregressive chunk appends fresh keys and values to the recent window; when that window overflows, the evicted block is merged into the summary tier using either uniform-bucket averaging (WorldTrace-Field) or scene-entry landmark insertion (WorldTrace-Landmark),

Algorithm 1 WorldTrace cache update (per AR chunk). $\mathcal{S}$ stores canonical (unrotated) keys; $R(\theta_k t_v^{(s)})$ is applied per slot at attention time via Def. 1, so positions recompute automatically as q advances (shared by both variants). Index $f \in \{0, \ldots, F-1\}$ ranges over the intra-block frames of an AR chunk; prev denotes the previously-popped block.

Require: recent cache $\mathcal{R}$, summary cache $\mathcal{S}$ (canonical), new chunk's KVs, mode $\mu \in \{\text{WorldTrace-Field}, \text{WorldTrace-Landmark}\}$
Ensure: updated $(\mathcal{R}, \mathcal{S})$

Append new KVs to $\mathcal{R}$ ▷ recent window is verbatim
if $|\mathcal{R}| > W_r$ **then**
  $K_* \leftarrow \text{PopOldest}(\mathcal{R})$
  if $\mu = \text{WorldTrace-Field}$ **then**
    $s^* \leftarrow$ summary slot whose source bucket now contains K_* ▷ uniform temporal grouping
    $\mathcal{S}[s^*] \leftarrow$ canonical mean of s^*'s source frames ▷ Def. 2
  else if $\mu = \text{WorldTrace-Landmark}$ **then**
    $\text{SE} \leftarrow \bigvee_{f=0}^{F-1} \left[\text{cosdist}(K^{(k)}_{\text{cx},K_*,f},\, K^{(k)}_{\text{cx,prev},f}) > \tau \right]$ ▷ per-frame check inside popped block
    if SE **then**
      $\mathcal{S} \leftarrow \text{ShiftLeft}(\mathcal{S})$ ▷ drop $\mathcal{S}[0]$
      $\mathcal{S}[N_s{-}1] \leftarrow K^{(k)}_{\text{cx},K_*}$ ▷ canonical landmark; *c.f.* Eq. (4)
    end if
    (Init.) If fewer than N_s landmarks are stored, fill empty slots with the oldest.
  end if
end if
Attention time: for $s{=}0, \ldots, N_s{-}1$, apply $R(\theta_k t_v^{(s)})$ to $\mathcal{S}[s]$ with $t_v^{(s)}$ from Def. 1.

as defined in the main text. Alg. 1 presents the full control flow in pseudocode, including the shared attention-time step that applies per-slot rotations to canonical summary keys using the virtual positions from Def. 1.

Default MG2 KV cache behavior. The stock MG2 inference pipeline implements a fixed-size sliding-window KV cache: each transformer layer maintains a pre-allocated buffer of size $L_{\text{train}} \times 880$ tokens, and when new tokens overflow the buffer, the oldest tokens are evicted without position correction or content compression (first-in-first-out over time). This sliding-window baseline is what we compare against throughout the paper.

F.4 Hyperparameters, Compute, and Evaluation Protocol

Inference hyperparameters. MG2-1.3B uses 3 distilled denoising steps, CFG scale 5.0, a single conditioning frame, and $F{=}3$ latent frames per AR block. The first chunk uses a clean KV context; later chunks receive the accumulated compressed cache.

WorldTrace-Field hyperparameters. Short-horizon experiments ($N{=}8$) use $N_s{=}2$ summary slots and $W_r{=}4$ recent-window slots ($N_s + W_r = 6 = L_{\text{train}}$). Long-horizon PAC and ablation experiments (Secs. 4.3–4.4) use $N_s{=}4$, $W_r{=}2$ (same total capacity) for both WorldTrace-Field and WorldTrace-Landmark, as described in the Setup section. The scripted loop evaluation (Sec. 4.3) uses $N_s{=}5$, $W_r{=}1$ for WorldTrace-Landmark (one additional landmark slot) and $N_s{=}4$, $W_r{=}2$ for WorldTrace-Field. Compression uses uniform temporal grouping: the T_{old} oldest frames are split into N_s equal groups; each group's keys are unrotated to canonical space (fp64 precision), averaged, and re-rotated at the group's virtual position in the original dtype. **Position assignment:** Tab. 4 reports canonical key averaging with Block-Rel virtual positions [Yesiltepe et al., 2026], where each summary slot is assigned position $\max(0, q - (L_{\text{train}} - 1)F)$, for a fair comparison against the Naive+Block-Rel baseline. Tab. 1 reports the long-horizon conditions (Sliding Window, Block-Rel, centroid linear, fullcombo) at $N{=}16$. The recent-window slot j keeps its absolute position across schemes.

Evaluation protocol. Each method is evaluated on 100 videos generated from initial frames. TempSSIM is computed on decoded RGB frames using SSIM [Wang et al., 2004] between consecutive frames. LatentDiff is the mean squared difference between consecutive latent frames (pre-VAE-decode), used in long-horizon ablations where VAE decoding is expensive. Multi-seed experiments use seeds $\{0, 42, 123, 456, 789\}$. All experiments run

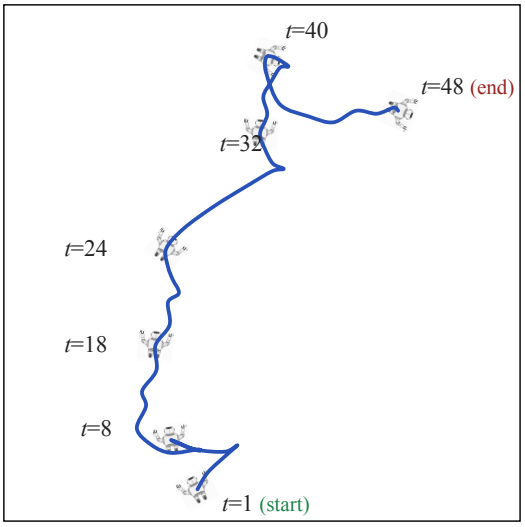


Figure 9: **Camera trajectory used for Fig. 4.** Top-down plan of the camera (x, z) path executed by the AR rollout for the WorldTrace-Field qualitative panel at $N{=}48$ ($\sim$36 s). Green: start ($t{=}1$); dark red: end ($t{=}48$). Orange ticks mark the chunk indices shown as columns in Fig. 4. The path leaves the initial scene, traverses novel territory, and lands at a different pose, so every column $t{>}1$ is out-of-window for the sliding-window baseline.

on a single NVIDIA A100 80 GB GPU.

Camera trajectory for the coherence qualitative panel. The qualitative comparison in Fig. 4 (Sec. 4.2) is generated along a single fixed camera path on Matrix-Game 2, played out over $N{=}48$ AR chunks ($\sim$36 s of decoded video). Fig. 9 shows that path as a top-down plan: the camera leaves an initial scene (start, $t{=}1$), explores a roughly counter-clockwise loop through novel territory across $t \in \{8, 18, 24, 32, 40\}$, and arrives at a distinct end pose ($t{=}48$). Because the world model is run autoregressively on the same control inputs across baselines, every method in Fig. 4 is evaluated at exactly the same intended camera pose at each timestep; differences between columns therefore reflect the cache mechanism, not the trajectory.

Compute. Generating one 8-chunk video (24 latent frames) with WorldTrace-Field takes approximately 7.3 s at batch size 1; VAE decode adds $\sim$2.5 s per chunk. The full ablation set required approximately 100 GPU-hours on single A100 80 GB GPUs. At $\sim$250 W typical A100 draw and $\sim$0.4 kg CO_2-eq/kWh, this is approximately 10 kg CO_2-eq.

Memory and compute. Cache storage is L_{train} blocks of KVs ($\sim$ 2.79 GB in fp16 for MG2-1.3B; constant in N, identical to the sliding-window baseline), while full-KV exhausts an 80 GB A100 at $N{\approx}150$ chunks ($\sim$112 s $\approx$1.9 min of decoded video; *c.f.* Tab. 14). The asymptotic per-token update cost is $O(T_{\text{old}} \cdot c_t \cdot n_\ell)$ ($\approx 1.7 \times 10^8$ FLOP at $N{=}100$); the FLOP cost itself is $O(1)$ in N. Wall-clock overhead per chunk grows modestly with horizon as bookmark accumulation and per-step kernel overhead take a larger share of each step: $+3.8\%$ ($N{=}8$) to $+9.4\%$ ($N{=}102$) for WorldTrace-Field, and $+5.2\%$ to $+27.2\%$ for WorldTrace-Landmark (Tab. 15). WorldTrace-Landmark replaces the averaging loop with a single unrotate-then-store per scene-entry frame; its growing per-chunk wall-clock at long horizons reflects scene-entry-bookmark accumulation rather than added FLOP.

Per-chunk runtime overhead. Tab. 15 reports wall-clock time per AR chunk on one A100 80 GB at batch size 1 (3 distilled denoising steps, 3 latent frames per chunk, $352{\times}640$ resolution). At short horizons ($N{=}8$), all methods are within 6% of the sliding-window baseline because the forward pass dominates and the cache is small. At $N{=}102$ ($\sim$77 s of decoded video), WorldTrace-Field adds $+9.4\%$ per chunk as the canonical unrotate/rerotate of the growing source set takes a larger share of each step, consistent with the $O(T_{\text{old}} \cdot c_t \cdot n_\ell)$ per-token form above. WorldTrace-Landmark's $+27.2\%$ at $N{=}102$ reflects per-step kernel overhead from bookmark accumulation rather than additional FLOP. The long-horizon column uses the `skip_decode` setting (no VAE decode during generation), so absolute times are lower than the $N{=}8$ column with VAE decode.

Table 15: **Runtime per chunk.** Wall-clock time on one A100 80 GB (batch 1, 3 distilled denoising steps, 352×640). Short: $N=8$ with VAE decode. Long: $N=102$, skip-decode. Δ vs. sliding window. WorldTrace-Landmark stores per-layer canonical-K bookmarks at detected scene-entry events; the per-chunk overhead grows with horizon as more bookmarks accumulate.

Method	$N=8$ (short)		$N=102$ (long)	
	s/chunk	Δ	s/chunk	Δ
Sliding window	0.95	–	0.57	–
Naive	0.97	+2.3%	0.60	+4.7%
WorldTrace-Field	0.99	+3.8%	0.63	+9.4%
WorldTrace-Landmark	1.00	+5.2%	0.73	+27.2%

F.5 Evaluation Metrics

Open-ended video world models have no unique ground truth: a generated rollout legitimately diverges from any reference within a few chunks, so reference-based metrics (PSNR, SSIM vs. GT, LPIPS) and single-window benchmarks (VBench [Huang et al., 2024], PAI-Bench [Zhou et al., 2025a]) penalize valid creative divergence and do not test long-horizon recall. We therefore evaluate with a suite of reference-free metrics organized into two groups, *quality* (frame-to-frame coherence) and *consistency* (long-range episodic recall), with one diagnostic (LatentDiff) as a cross-check (Tab. 16).

Table 16: **Metric definitions.** ↑: higher is better; ↓: lower is better.

Metric	**Role**	**Dir.**	**Definition**
TempSSIM	Coherence	↑	SSIM [Wang et al., 2004] between consecutive decoded frames, averaged over the rollout.
Local Scene Drift (SceneDrift)	Coherence	↓	Mean per-chunk CLIP feature distance to the preceding chunk.
PAC	Recall	↑	CLIP-ViT-H/14 cosine similarity between geometrically paired return- and forward-leg frames in ABA loops; PAC averages the final $N/8$ return chunks closest to scene A (Sec. 4.3).
LatentDiff	Diagnostic	↓	MSE between consecutive latent frames (pre-decode). Confounded: favors slowly-varying output. Used alongside pixel-domain metrics as a fast sanity check only.

The suite is designed so that no single metric can be trivially gamed: TempSSIM alone rewards frozen output (a model that repeats the same frame scores perfectly), LatentDiff alone rewards degenerate slowly-varying output (sliding-window eviction achieves the lowest LatentDiff despite the worst TempSSIM), and SceneDrift alone could miss longer-range consistency failures. Together, the metrics triangulate actual quality: high TempSSIM (local coherence), low SceneDrift (the model generates dynamic content without scene wandering), and high PAC (long-range episodic recall). All metrics are defined from first principles.

G Discussion

G.1 Limitations

Ego-motion and screen-coordinate aliasing. WorldTrace-Field decouples temporal from spatial RoPE: the unrotate/rerotate in Eq. (3) touches only the first $2c_t$ head dimensions and preserves the spatial-RoPE factors of Eq. (3) exactly (App. C). The canonical mean is exact *at fixed screen coordinates*, but under complex ego-motion (panning, strafing, rapid yaw), visually distinct objects can pass through the same (h, w) at different timestamps and be blended, losing object identity along motion paths. WorldTrace-Landmark sidesteps this by storing verbatim canonical keys, and pose-conditioned generators (e.g., LingBot-Fast's Plücker, App. F) supply ego-motion as side information.

Architecture and orthogonal cache mechanisms. WorldTrace assumes temporal RoPE on keys, and a fixed AR KV cache; single-shot generators and combinations with LLM-side eviction methods such as SnapKV [Li

et al., 2024] or H_2O [Zhang et al., 2023] are out of scope here. The null PAC result for WorldTrace-Field on LingBot-Fast is consistent with our recall/coherence split (Sec. 3.5): Plücker conditioning already supplies the recall signal, but a matched return-SSIM comparison under Plücker would test whether pose conditioning also saturates MG2's coherence gains.

Metrics and headline numbers. Episodic recall is measured with paired CLIP cosine similarity on scripted return legs (Sec. 4); this rewards semantic alignment rather than pixel-faithful identity, and headline PAC sweeps inherit the metric's geometric window scaling with horizon. Coherence is measured with frame-level TempSSIM, a local pixel-statistics measure that does not penalize a coherent rollout that has settled into the wrong scene; perceptually-aligned distances grow under such drift [Zhang et al., 2018]. Human studies and pixel-level visitation metrics could further provide more insights.

Training paradigm and comparative scope. MG2 inherits a Self-Forcing-style autoregressive training distribution with bounded local attention at write time (App. F.1); inference-time KV stitching does not alter that mismatch if future models train with different cache semantics. "Training-free" here excludes fine-tuning the generator (Sec. 3.1); methods that distill long-rollout consistency, enlarge context by training, or change attention masks lie outside this protocol. Multi-student distillation [Song et al., 2025b] and other one-step generator regimes change the per-block denoising budget and therefore the effective length over which an AR rollout accumulates RoPE OOD; how cache-side interventions like WorldTrace compose with such distillation pipelines is an interesting open direction.

Deployment overhead. WorldTrace adds canonical-domain key transforms and bookkeeping relative to sliding-window eviction; although peak cache memory scales as $O(1)$ in horizon (Sec. 3), wall-clock latency and bandwidth to move updated keys through fused attention kernels are not modeled here (Tab. 15 reports per-configuration timings under our reference stack).

Domain generalization. Evaluations emphasize game-engine and navigation-conditioned rollouts with strong layout and lighting structure; how WorldTrace-Field and WorldTrace-Landmark behave on natural video with thin scene boundaries, film cuts, or rapid appearance changes is not established, and the scene-entry heuristics may need different thresholds.

G.2 Future Directions

The position/content factorization lets several research threads extend WorldTrace without retraining the underlying video model.

Geometry-aware canonical keys. The canonical mean of Eq. (3) aggregates by screen coordinate. Coupling the unrotate/rerotate primitive with camera-pose warping (*e.g.* the Plücker-conditioned and Warped-RoPE writers of MosaicMem [Yu et al., 2026] and UCM [Xu et al., 2026b]) would let the same operator average over scene coordinates instead, removing the ego-motion aliasing above and unifying the strict zero-fine-tune regime with the trained camera-aware family.

Learned scene-entry policies. Replacing the canonical-key spike or gradient-onset rule with a small policy trained on action discontinuities, agent-pose deltas, or scene-segmentation logits would enable WorldTrace-Landmark to commit landmarks during continuous motion. Active recall, in which the policy decides *when* to commit and *which* of multiple stored landmarks to re-rotate at a given query, is a natural extension. The cache layout itself (N_s, W_r, scene-entry threshold) is fixed in our experiments; treating it as an autotunable schedule under a checkpoint-conditioned surrogate [Mehta et al., 2024] could let downstream practitioners adapt it per backbone or per horizon without re-running the full slot-sensitivity sweep of App. D.4.

Composition with content-side eviction. Because slot-rank positions are independent of which canonical content fills the slot, eviction heuristics from the LLM literature (H_2O [Zhang et al., 2023], SnapKV [Li et al., 2024], KnormPress [NVIDIA, 2024]) can be layered on top of WorldTrace-Field's canonical averages and WorldTrace-Landmark's frozen keys without re-deriving the position scheme. The MemRoPE [Kim et al., 2026] comparison in App. D.1 suggests the two interact, so principled co-design is open. A complementary direction for WorldTrace-Landmark is to spend an extra summary slot on a residual WorldTrace-

FIELD-style canonical mean of evicted landmarks, retaining a coarse coherence trace of the discarded scene-entry frames at no additional position-side cost.

Downstream consumers of pretrained generative teachers. WORLDTRACE targets the inference-time cache of an autoregressive video generator, but pretrained diffusion and AR generators feed a broader pipeline ecosystem whose budgets and biases interact with the cache design. Score-distillation pipelines for amortized 3D synthesis (ATT3D [Lorraine et al., 2023], LATTE3D [Xie et al., 2024]) and LLM-conditioned mesh generation (LLaMA-Mesh [Wang et al., 2024]) consume teacher gradients whose Monte Carlo variance, rather than long-horizon recall, dominates compute; compute-aware estimators for those gradients (CARV [Bettencourt et al., 2026]) and the corresponding analyses for non-vision teachers [Richter-Powell et al., 2025] are orthogonal axes to the position–content factorization studied here. On the data side, motion-attribution methods [Wu et al., 2026c] ask which training clips improved the temporal dynamics of a generator, complementing the cache-side question of which past slots a generator can still address at inference. Establishing whether the slot-rank virtual-position primitive transfers to these adjacent regimes is an open direction.

Fine-tuning extensions. WORLDTRACE stays within the training context length L_{train} (Sec. 3.1) precisely because it is training-free: the constraint $N_s + W_r = L_{\text{train}}$ keeps every summary slot at an in-distribution offset the generator was trained on. Two light fine-tuning paths would relax this budget without retraining the generator from scratch. (i) *Context-extension fine-tuning* on synthetic long rollouts would let WORLDTRACE allocate either more recent slots for coherence or more summary slots for longer recall horizons at the same in-distribution attention cost. (ii) *Position-aware fine-tuning* that exposes the model to the slot-rank offsets of Eq. (2) would tighten the canonical-mean approximation underlying WORLDTRACE-FIELD (Rem. 2), narrowing the residual long-horizon PAC gap. Both paths fit the nested-optimization template [Lorraine, 2024] of a frozen large inner model paired with a light outer adapter for the cache schedule, so the position/content factorization is preserved through the outer loop. Both are compatible with the WORLDTRACE cache layout and leave the position/content factorization intact.

Multi-tier and cross-architecture transfer. WORLDTRACE uses a two-tier split (recent verbatim, summary canonical). Adding intermediate tiers, with progressively larger source buckets and progressively deeper slot-rank offsets, could trade sharper recall for longer effective horizons. Porting the same factorization to KV-cache-bearing variants of MG3 [Wang et al., 2026a], Genie3 [Google DeepMind, 2025], or LingBot-Fast [Team et al., 2026] would test whether the position/content split is architecture-specific or generic.

Scaling LoopMem. LoopMem (App. E) exercises all four difficulty tiers on MG2-1.3B; camera-orientation Tier 3 uses MG2's mouse-yaw control, where chunk counts are nominal magnitudes rather than calibrated angles. Pose-conditioned generators with explicit pitch/roll would extend Tier 3, and broader cross-architecture leaderboards on Tiers 1, 2, and 4 would let the community compare position/content trade-offs at controlled compression ratios, complementing the broad video-generation benchmarks of WorldScore [Duan et al., 2025], MIND [Ye et al., 2026], and VBench-2.0 [Zheng et al., 2025]. The Pan 360° failure of WORLDTRACE-LANDMARK on visually continuous trajectories (Tab. 2) suggests learned or rule-based scene-entry policies as a concrete next step.

G.3 Broader Impact

This is an inference-time KV-cache modification: it does not alter training data, modify weights, or expand the base model's capabilities, so it introduces no new dual-use risks beyond those inherent to the underlying video model. The capability shift is minute-scale long-horizon generation at $O(1)$ peak cache memory; longer coherent clips raise the importance of provenance metadata and watermarking, and the lower memory footprint broadens both research access and the surface for misuse, so releases should follow the base model's content-policy guidance.