

Addressable Memory for Video World Models

Xindi Wu^{1 2} Sven Elflein^{1 3 4} James Lucas¹ Olga Russakovsky² Laura Leal-Taixé¹ Despoina Paschalidou¹

Jonathan Lorraine¹ Aljosa Osep¹

¹NVIDIA ²Princeton University ³University of Toronto ⁴Vector Institute

<https://research.nvidia.com/labs/sil/projects/WorldTrace/>

Abstract

We study visual persistence in interactive video world models. These models rely on a Key-Value (KV) cache as a growing visual memory to carry forward previously generated frames. However, we find that once rollouts extend beyond the training horizon, models can no longer reliably address stored content: once temporal Rotary Positional Embeddings (RoPE) offsets fall outside the range seen during training, the model struggles to retrieve the relevant visual information through attention. Moreover, naively compressing the cache in the RoPE-rotated space corrupt memory by averaging together incompatible positional phases. To address this, we propose **WORLDTRACE**, a training-free memory framework for long-horizon visual persistence. **WORLDTRACE** keeps compressed memory addressable by assigning each summary slot a distinct, in-distribution virtual position. Within this addressable cache, we study two memory compression approaches: **WORLDTRACE-FIELD** compresses history for temporal coherence, while **WORLDTRACE-LANDMARK** stores verbatim scene traces at detected transitions for episodic recall. To measure recall under controlled revisits, we introduce **LOOPMEM**, a benchmark evaluating whether a compressed cache can reconstruct a previously visited scene after a long detour. **WORLDTRACE-FIELD** improves temporal consistency by +15.5%, and **WORLDTRACE-LANDMARK** improves episodic recall by +19.5% on **LOOPMEM**, extending visually persistent generation without retraining.

1 Introduction

Memory is what makes a generated world persistent. A world model should not only predict what comes next, but also preserve what has already been seen. World models aim to simulate worlds we can explore [Ha and Schmidhuber, 2018, Hafner et al., 2025, Google DeepMind, 2025], and recent autoregressive video world models [Team et al., 2026, He et al., 2025b, Decart et al., 2024] pursue this vision by generating video 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 [Agarwal et al., 2026, Basant et al., 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 generated results should be consistent with 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 in the cache, but whether it remains addressable and how it is compressed. As the generation horizon extends beyond the training context length, the positional encoding these models rely on (e.g., RoPE [Su et al., 2024]) is queried at position indices outside its training-bounded range, so the model needs to attend over relative distances that are 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

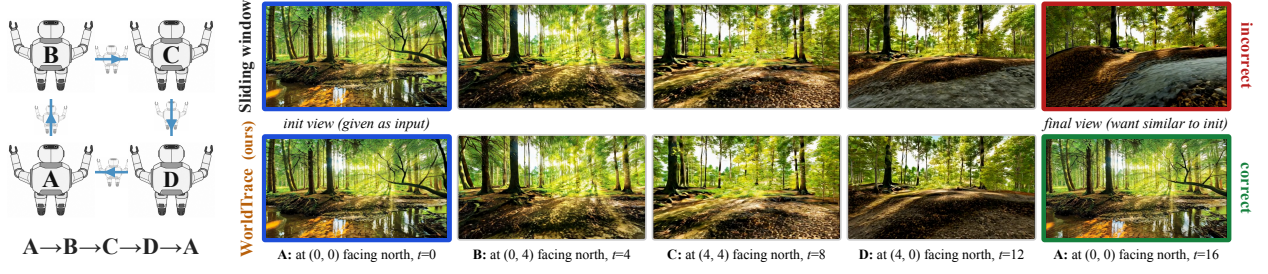


Figure 1: **Addressable memory maintains long-horizon visual persistence.** (Left) Loop topology $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$. (Right) Representative frames on Matrix-Game-2 along the path: the initial scene A (blue border), waypoints B, C, and D, and the return to A. *Sliding window* (top) mismatches the reference appearance at the return (red border). Our **WORLDTRACE** with frozen landmark keys (green border) matches scene A, confirming addressable long-horizon recall.

problem of *addressability*: no memory scheme can improve visual persistence if past observations cannot be reliably accessed once they fall outside the context window.

WORLDTRACE. To tackle this problem, we propose **WORLDTRACE**, a training-free approach for addressable compressed memory over long 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 so that distant tokens remain retrievable. The cache is organized into a recent window and summary slots for the distant horizon. Within this addressable cache, the remaining challenge is what the world model should choose to remember. We view this as a structured projection of a growing visual history into a fixed set of retrievable traces. We propose two complementary memory writers: **WORLDTRACE-FIELD** which preserves the coarse field of past information in rotation-invariant space for temporal coherence, and **WORLDTRACE-LANDMARK** which preserves sparse scene-entry traces for long-horizon recall.

Our experiments show that addressable memory extends the effective memory horizon. To evaluate episodic recall under controlled scene revisits, we propose **LOOPMEM**, a *memory* benchmark that tests whether the compressed KV cache can retrieve a previously visited scene after a long detour. While the sliding-window baseline begins to forget scene structure after only a few seconds and loses structural fidelity when revisiting the same location, our **WORLDTRACE** maintains consistent scene reconstruction over long rollouts and recovers the original scene appearance (Fig. 1).

Our contributions are:

1. We identify two coupled challenges for long-horizon generation in autoregressive video world models, *where* memory is placed and *what* it stores: out-of-distribution temporal offsets make cached tokens unaddressable, and RoPE phase cancellation makes naively compressed summaries uninformative (Sec. 2).
2. We propose **WORLDTRACE**, a training-free approach that keeps compressed memory both addressable and useful without retraining. It assigns each summary an in-distribution virtual position (Sec. 3.1) and organizes the growing visual history into a fixed set of retrievable traces, realized by two complementary operators: **WORLDTRACE-FIELD** for temporal coherence (Sec. 3.3) and **WORLDTRACE-LANDMARK** for episodic recall (Sec. 3.4).
3. We introduce **LOOPMEM**, a *memory* benchmark of controlled scene revisits that measures whether a compressed cache can recover a previously visited scene (Sec. 4). We show that **WORLDTRACE** extends the effective memory horizon over the default sliding-window cache: **WORLDTRACE-FIELD** improves temporal consistency by +15.5% on long rollouts, and **WORLDTRACE-LANDMARK** improves episodic recall by +19.5% on **LOOPMEM** ABA revisits, both without retraining.

2 Rethinking Memory in Video World Models

Autoregressive video world models retain prior context as cached keys and values during generation. Long-horizon memory therefore depends both on whether relevant past context remains inside the cache and

on whether the model can effectively attend to it during future generation. Two coupled bottlenecks arise. First, cached tokens need to remain *addressable* (Sec. 2.1): attending to a distant token requires a temporal RoPE offset that may lie outside the training distribution, leaving the token unreadable even when stored. Second, compressed summaries need to remain *informative* (Sec. 2.2): when old keys are compressed naively, mismatched RoPE phases can cancel out the signal the summary is supposed to carry.

2.1 Position Determines Whether Memory Is Addressable

The first bottleneck is *addressability*: a cached token helps only if the current query can still attend to it. The further back a token lies, the larger its temporal RoPE offset from the current query; once this offset grows beyond the range seen during training, the model can no longer read the entry even though it still sits in the cache. We summarize the notations in App. A.

Concretely, during training the query only attends over a bounded range of relative offsets $\delta_{q,t}$ between the current frame and each cached frame:

$$0 \leq \delta_{q,t} = q - t \leq \Delta t_{\text{train}},$$

where q is the absolute position of the current query frame, t that of a cached key frame, with $t \leq q$ under causal autoregressive attention, and Δt_{train} the largest offset seen during training.

For temporal RoPE pair k at angular frequency θ_k , its contribution $a_k(q, t)$ to the query-key attention score is:

$$a_k(q, t) = \text{Re}(A_{q,t}^{(k)} e^{i\theta_k \delta_{q,t}}),$$

where $\text{Re}(\cdot)$ is the real part and $A_{q,t}^{(k)}$ is the content-dependent inner product in canonical coordinates between the query at q and the key at t . During long-horizon inference, these offsets grow ever larger, and $\delta_{q,t}$ can exceed Δt_{train} . The affected RoPE components then take on rotation angles the model never saw during training. This effect is frequency-dependent: the fastest temporal components degrade into positional noise, while slower ones stay close to their trained range and keep carrying usable signal [Peng et al., 2024, Barbero et al., 2025]. We quantify this per frequency in App. C.1.

Several concurrent methods investigate related positional issues. Infinity-RoPE [Yesiltepe et al., 2026] proposes Block-relative RoPE which reindexes each cached frame relative to the current AR block so that $\delta_{q,t}$ always stays within the trained range, keeping generation in-distribution without a fixed memory budget. It targets infinite video generation rather than memory compression. Applied to a *compressed* cache, however, this reindexing caps every offset at Δt_{train} , so distinct summary slots collapse onto the same position and become indistinguishable (Rem. 1). MemRoPE [Kim et al., 2026] compresses history into two EMA memory tokens, a long-term and a short-term stream, which avoids this collapse simply because only two summaries need distinct positions. The design is fixed to two streams and does not extend to an N -slot cache, where capped Block-relative positions would again collide. Unlike these, our WORLDTRACE targets an N -slot compressed cache directly: it assigns every slot a distinct, in-distribution slot-rank position, so an arbitrary number of summaries stay individually addressable at any horizon. We provide a detailed discussion of related work in App. D.

2.2 Content Compression Determines Whether Memory Is Informative

Even if a compressed summary is within an in-distribution temporal position, the second bottleneck is *informativeness*: the slot must still carry an informative summary after compression. A common way to compress cached keys is averaging [Bolya et al., 2023, Rae et al., 2020]. With RoPE-rotated keys, at frequency k this corresponds to:

$$\bar{K}_{\text{naive}}^{(k)} = \frac{1}{M} \sum_{m=1}^M R(\theta_k t_m) K_m^{(k)},$$

where $K_m^{(k)}$ is the canonical (unrotated) key of frame m at frequency k and $R(\theta_k t_m)$ is its RoPE rotation. Each frame is rotated by a different angle $\theta_k t_m$ before averaging, so $\bar{K}_{\text{naive}}^{(k)}$ sums vectors that point in different

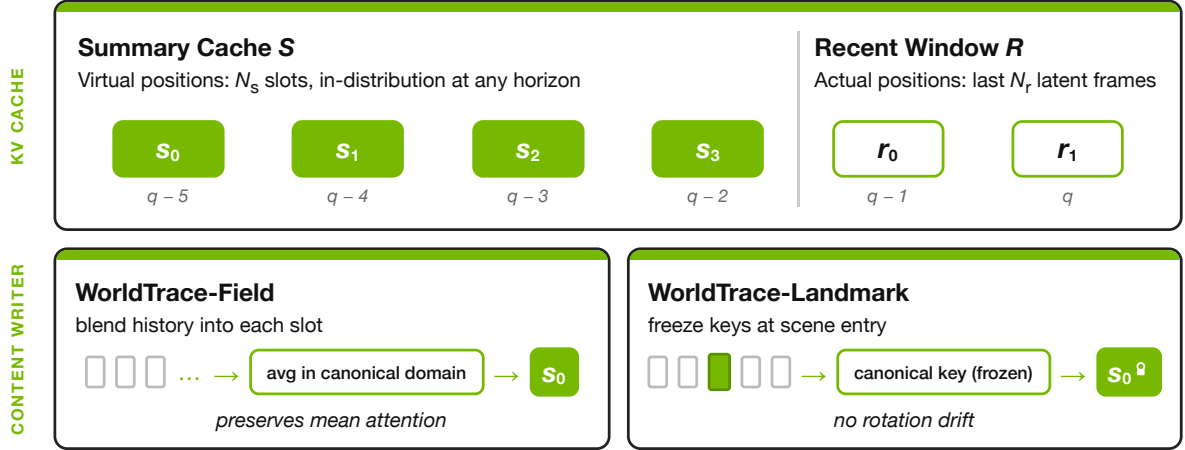


Figure 2: **WORLDTRACE overview.** **WORLDTRACE** partitions the context window into a compressed *summary* cache $\mathcal{S}$ for the distant past and a verbatim *recent* window $\mathcal{R}$ (top). Each slot occupies one latent-frame position in the cache, with recent slots holding one verbatim frame and summary slots holding compressed history. **WORLDTRACE-FIELD** (bottom left): Keys are unrotated and averaged in canonical space, then re-rotated at each slot’s virtual position. **WORLDTRACE-LANDMARK** (bottom right): Scene-entry landmarks store frozen canonical keys and are re-rotated to the current virtual position.

directions. If the frames come from nearby timestamps, their rotation angles are close, so averaging largely preserves the signal. If they come from distant timestamps, their angles can point in opposite directions and partially cancel, weakening that frequency’s contribution to the summary regardless of what it encoded. Naive averaging in RoPE-rotated space thus corrupts the compressed memory.

The two failures are coupled. Most prior methods fix only one of the two: one line of work adjusts *where* memories are placed, reparameterizing or capping offsets (e.g. Block-relative [Yesiltepe et al., 2026]), while another decides *what* each summary stores, whether by smoothing past keys with dual-rate EMA [Kim et al., 2026], merging them by temporal averaging [Bolya et al., 2023, Rae et al., 2020], or evicting entries [Zhang et al., 2023, Li et al., 2024, Cai et al., 2025]. Addressable positions cannot help if compression cancels the signal a summary should carry, and an informative summary cannot help if it sits at an out-of-distribution position. Thus we need to keep compressed memory both addressable and informative once generation exceeds the training window. For a video world model, this is precisely what visual persistence requires: as an agent explores and later revisits a location, the cache must keep those distant scenes readable across long detours (addressability) and retain enough of their content to regenerate them faithfully (informativeness). Our **WORLDTRACE** is built to satisfy both, which is what lets a generated world stay consistent over long, interactive rollouts.

3 WORLDTRACE

We propose **WORLDTRACE**, a training-free approach for addressable memory in autoregressive video world models. **WORLDTRACE** maintains a bounded trace of distant history in fixed summary slots via canonical-space key assignment, while keeping each slot *addressable* (Sec. 3.1). We formulate this cached memory compression for visual persistence as structured sparse attention (Sec. 3.2, App. B). Both **WORLDTRACE-FIELD** (Sec. 3.3) and **WORLDTRACE-LANDMARK** (Sec. 3.4) are its two variants, for temporal coherence and episodic recall, respectively.

3.1 Virtual Position Assignment: **WORLDTRACE** Slot Indexing

Cache structure. As shown in Fig. 2, we partition the local attention window L_{attn} into a **recent window** $\mathcal{R}$ of N_r verbatim latent frames and a **summary cache** $\mathcal{S}$ of N_s slots storing compressed history, so that $N_s + N_r = L_{\text{attn}}$.

Given this cache structure, the remaining question is how to position each summary slot relative to the query. A

memory position scheme assigns a virtual position to summary slot s for a query at timestamp q ; to be usable by the attention mechanism, it 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 virtual positions stay in-distribution at any generation length, and (iv) distinct positions for each slot so attention can tell them apart. Summary slot positions must stay within the range seen during training:

$$t_{\min}^v = \max(0, q - \Delta t_{\text{train}}), \quad t_{\max}^v = q - N_r. \quad (1)$$

The position of each summary slot is assigned as follows:

Definition 1 (WORLDTRACE Slot Indexing). Given current query position q , local attention window L_{attn} ($= N_s + N_r$), and summary cache size N_s , the virtual position of summary slot s is:

$$t_s^v = q - (L_{\text{attn}} - 1 - s), \quad s = 0, \dots, N_s - 1. \quad (2)$$

The assignment covers the in-distribution range with fixed slot-rank offsets relative to the current frame. By construction, $t_s^v \in [t_{\min}^v, t_{\max}^v]$ at any generation length, offsets depend only on slot rank and q , and, importantly, *not* the absolute horizon N .

Remark 1 (Block-relative collapse). *Block-relative positions* [Yesiltepe et al., 2026] cap all KV-cache offsets at Δt_{train} (Sec. 2.1). As generation grows, summary slots spanning history beyond Δt_{train} steps back all share the same minimum virtual position $t_{\min}^v$, making the compressed slots positionally indistinguishable. Our WORLDTRACE slot indexing (Def. 1) avoids this by anchoring each virtual position to slot rank: $t_s^v = q - (L_{\text{attn}} - 1 - s)$, where the offset $(L_{\text{attn}} - 1 - s)$ is fixed by slot rank alone, so all N_s slots remain distinct at any generation horizon N .

Why canonical keys. Each slot’s virtual position t_s^v is set by slot rank (Def. 1), independently of the original timestamps $\{t_m\}$ of its source frames, so the slot-rank assignment can place t_s^v far from any t_m . Placing compressed content at t_s^v therefore requires a two-step operation: first unrotate each key to remove its original RoPE rotation, then re-rotate to the target phase at t_s^v . Writing directly in the RoPE-rotated domain fails for two reasons. First, keys from different timestamps each carry a distinct rotation $R(\theta_k t_m)$; combining them without first removing those rotations causes contributions at each frequency to point in different directions, partially canceling the signal the compression should preserve. Second, even if one avoids phase cancellation, the compressed key would still carry the phase of the source timestamps rather than the required phase of t_s^v , making it misaligned with the virtual position the attention mechanism expects. Operating in the canonical (unrotated) domain resolves both issues: unrotating each key removes its timestamp-specific rotation, compression then acts on content-only vectors without phase misalignment, and re-rotation at t_s^v encodes the position needed for the assigned slot rank.

With addressable summary slots, the remaining design choice is the *content writer*. Both WORLDTRACE variants use the same canonical (un-rotated) key domain. Sec. 3.2 formalizes the writer as structured sparse attention, and we instantiate the two row structures in Sec. 3.3 and Sec. 3.4.

3.2 Memory Compression as Structured Sparse Attention

With slot positions fixed, the remaining choice is *what* to store in each slot. We view this memory compression as structured sparse attention: it replaces full attention over all T_{old} past frames with a fixed- L structured sparse pattern, where each of the L_{attn} cache slots summarizes a segment of the full history. The slot’s compressed key is $P_i K_m^{(k)}$, rotated to the slot’s virtual position, where $P \in \mathbb{R}^{L \times T_{\text{old}}}$ is the projection matrix. The recent window uses verbatim identity rows for the last N_r frames, and the N_s summary rows carry the compression policy (Fig. 6, App. B).

Projection view. Attending over all T_{old} past frames would produce a softmax-weighted combination of their values, with weights $\alpha_q \in \Delta_{T_{\text{old}}}$ determined by query-key similarity. WORLDTRACE replaces this with projected attention over L compressed keys: row i of P defines which past frames contribute to compressed key i , given by $\tilde{K}_i^{cx} = P_i K_m^{(k)}$. Re-encoding each compressed key at t_s^v keeps the approximation in-distribution

at any horizon. The approximation error decomposes into *distribution mismatch* between the true attention weights α_q and the weights induced by the compressed cache (Prop. 1, App. B). The choice of P therefore determines which parts of history are preserved under a fixed slot budget.

NMF connection. Treating the projection weights as free variables turns aggregate distribution mismatch into a nonnegative matrix factorization (NMF) problem [Lee and Seung, 1999]: uniform-partition rows correspond to cluster-mean compressed keys (WORLDTRACE-FIELD), while 0/1 rows correspond to medoid-style token selection (WORLDTRACE-LANDMARK). At inference, softmax coupling re-ties the projection weights to the compressed keys. The relaxed NMF view nonetheless clarifies why the two row structures arise and when each is close to optimal. We provide full derivations and optimality statements in App. B.

3.3 WORLDTRACE-FIELD: Canonical Key Averaging

WORLDTRACE-FIELD targets temporal coherence, not episodic recall. It compresses distant history into rotation-invariant slot traces by averaging all frames in 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:

$$K_{\text{field}}^{(k)}(t^v) = R(\theta_k t^v) \frac{1}{M} \sum_{m=1}^M R(-\theta_k t_m) K_{t_m}^{(k)}, \quad (3)$$

i.e. we unrotate each key to its canonical content $K_m^{(k)} = R(-\theta_k t_m) K_{t_m}^{(k)}$, 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-frame values assigned to the slot. The count M is not a separate hyperparameter: all T_{old} 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 score that the source keys $\{K_{t_m}^{(k)}\}_{m=1}^M$ would receive if each were reassigned the shared virtual position t^v . This guarantee is score-level only, as softmax prevents it from extending to attention weights. A formal statement is in App. C (Prop. 2). 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. This is akin to Landmark Attention [Mohtashami and Jaggi, 2023]. However, unlike token-level landmark insertion in trained transformers, our method uses scene-entry detection, frozen canonical keys, and slot-rank virtual positions.

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 scene-entry events have been detected, the remaining slots repeat the oldest available landmark. Once more than N_s scene-entry events have been detected 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-rotate 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:

$$K_{\text{land}}^{(k)}(t_v^{(s)}) = R(\theta_k t_s^v) R(-\theta_k t_{\ell^*}) K_{t_{\ell^*}}^{(k)}, \quad (4)$$

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 $R(-\theta_k t_{\ell^*}) K_{t_{\ell^*}}^{(k)}$ 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, so each frozen landmark occupies a distinct in-distribution virtual position at any horizon.

Comparing the two variants. WORLDTRACE-FIELD and WORLDTRACE-LANDMARK are two choices of projection matrices that target different future query families. Under WORLDTRACE-FIELD (Sec. 3.3), each summary row uniformly averages keys over a disjoint temporal group, where the compressed key for group g_s is the canonical mean $\bar{K}_{(g_s)}^{(k)}$. This is near-optimal when attention varies smoothly within each compressed block: future queries care about group-level history, not individual frame identity (Eq. (6), App. B). Under WORLDTRACE-LANDMARK (Sec. 3.4), summary rows are one-hot selectors at detected scene-entry frames, recovering a verbatim canonical key $P_i K_m^{(k)} = K_{t_{\ell^* s}}^{(k)}$. This is near-optimal when recall queries concentrate on a small set of salient past scenes (Eq. (8), App. B). The two assumptions are incompatible in the failure cases we study: averaging is efficient when attention is smooth, but destroys the token a return query needs. Verbatim selection preserves episodic targets but wastes slots when coherence dominates.

3.5 Position-Content Coupling

As noted in Sec. 2, position assignment and content compression are coupled. Concretely, the virtual positions assigned to summary slots can diverge from the offset regime under which their keys were compressed, so the cached content may no longer match what the query attends to. Block-relative collapse makes compressed entries effectively unaddressable, reducing compression to sliding-window eviction once addressability collapses (Rem. 1). Averaging position-conditioned keys under one offset distribution and querying them under another changes the effective phase of cached tokens, breaking the mean-attention preservation that motivates Def. 2.

WORLDTRACE resolves this coupling: Def. 1 fixes a consistent virtual positional regime, while Sec. 3.3 and Sec. 3.4 absorb the source-to-virtual shift by unrotating keys to canonical space and re-encoding at slot-rank positions (Eq. (3), Eq. (4)). Within this framework, the choice of slot content, whether averaged frames for WORLDTRACE-FIELD or frozen landmarks for WORLDTRACE-LANDMARK, becomes a content-only design choice. We provide the detailed per-chunk cache-update algorithm in App. G.

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. $\Delta t_{\text{train}} + 1 = 6$ latent frames. We provide additional experiments for LingBot-World [Team et al., 2026] in App. E. We use $N_s=2/N_r=4$ for coherence (Sec. 4.2) and $N_s=4/N_r=2$ for episodic recall (Sec. 4.3) (latent-frame slots; $N_s + N_r = L_{\text{attn}}=6$).

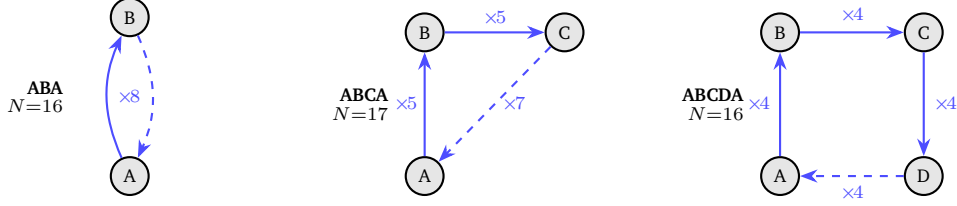


Figure 3: **Three LOOPMEM geometries.** We propose LOOPMEM, a memory benchmark with diverse geometries, including ABA (straight reversal), ABCA (approximate L-triangle with diagonal $5\sqrt{2} \approx 7$), and ABCDA (square). N is rollout length in AR chunks. Each path returns to scene A. We provide a full gallery of edge-length, orientation, and multi-revisit settings in App. F (Fig. 9).

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 TempSSIM ($\uparrow$, SSIM [Wang et al., 2004] between consecutive decoded frames) and Local Scene Drift ($\downarrow$, mean per-chunk CLIP feature distance to the preceding chunk). For *recall*, we use Position-Aligned CLIP (PAC $\uparrow$, CLIP-ViT-H/14 [Radford et al., 2021] cosine similarity between geometrically paired return- and forward-leg frames in loops). Full definitions are in Tab. 18 (App. G.4).

Baselines. We compare against MG2’s default sliding-window cache throughout, plus controlled variants that isolate position assignment (Block-relative, Centroid-linear, WORLDTRACE) and content compression (naive vs. canonical averaging, Sec. 4.2, Sec. 4.4). Episodic recall compares three retention tiers (compression, canonical-K anchoring via *Latent re-anchor*, and verbatim landmarks) on LOOPMEM and in the PAC sweep (Tab. 3, Tab. 12 in App. E.2). MemRoPE [Kim et al., 2026], YaRN [Peng et al., 2024], and Landmark+Block-relative are evaluated in App. E.1.

Benchmark. We introduce LOOPMEM, a *memory* benchmark for scene revisits in AR world models. The model traces waypoints before returning to a previously visited location, and 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. 9 (App. F).

4.2 Coherence: WORLDTRACE-FIELD

Virtual Position Assignment. WORLDTRACE-FIELD averages keys in the canonical domain, then assigns each averaged slot a virtual position, maintaining scene coherence under compression. We compare it against canonical-K averaging variants that apply this same key averaging but differ in position assignments. *Block-relative* [Yesiltepe et al., 2026] clamps offsets beyond the training horizon, so distinct older slots collapse onto a single capped position, whereas *Centroid-linear* positions

each slot by the average timestamp of the frames it summarizes, linearly mapped into $[t_{\min}^v, t_{\max}^v]$, so slots stay distinct but their positions shift as the rollout grows. Relative to the four position properties of Sec. 3.1, Centroid-linear satisfies (i) linearity and (ii) in-distribution placement but violates (iii) horizon-stability, since its virtual positions depend on N . Instead, our WORLDTRACE-FIELD assigns positions from slot rank alone, so every summary slot stays in-distribution and individually addressable no matter how long the rollout runs.

With the same compression approach but different position assignment shows that compression helps only when the summary slots remain addressable. With Block-relative positions, offsets beyond the training window receive zero softmax weight under local attention mask, so canonical averaging reduces to a plain sliding-window cache. Centroid-linear avoids this saturation by keeping positions within the attention mask, but its N -dependent assignment shifts old summaries into positional regimes the model was not trained to use as the horizon grows. Our WORLDTRACE instead assigns positions by slot rank alone and keeps summary slots in-distribution at all horizons, outperforming Block-relative by +5.9% and +2.8% TempSSIM and Centroid-linear by +9.5% and

Table 1: **Position assignment.** Canonical averaging stays the same so that only positions vary. Values are TempSSIM ($\uparrow$).

Position Assignment	$N=8$	$N=16$
Block-relative	0.390	0.530
Centroid-linear	0.377	0.479
WORLDTRACE (ours)	0.413	0.545

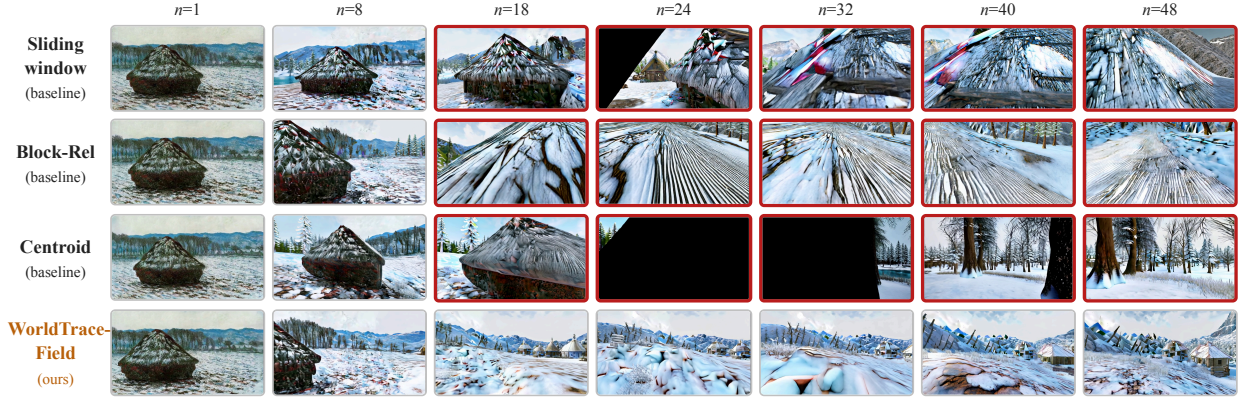


Figure 4: **WORLDTRACE-FIELD** at number of chunks $N=48$. Four KV-cache approaches conditioned on the same initial image and camera trajectory, including three baselines (sliding window, Block-relative, Centroid-linear) and WORLDTRACE-FIELD (ours). Baseline approaches start drifting at $n=18$ (red border). WORLDTRACE-FIELD remains coherent across the horizon. Camera path plotted in Fig. 12 (App. G).

Table 2: **WORLDTRACE-FIELD** achieves the highest TempSSIM at both horizons and the lowest Local Drift. At $N=32$, WORLDTRACE-FIELD has the highest TempSSIM (0.613) though centroid variants have lower Drift. At $N=48$, WORLDTRACE-FIELD is best on both metrics.

Method	$N=32$		$N=48$	
	TempSSIM $\uparrow$	Local Drift $\downarrow$	TempSSIM $\uparrow$	Local Drift $\downarrow$
Sliding window	0.571	0.0229	0.472	0.0305
Canonical averaging + Block-relative	0.585	0.0215	0.530	0.0339
Canonical averaging + Centroid-linear	0.573	0.0211	0.479	0.0297
WORLDTRACE-FIELD (ours)	0.613	0.0250	0.545	0.0295

+13.8% at $N=8$ and $N=16$ (Tab. 1).

We show this qualitatively in Fig. 4: by $n=18$, the baselines drift from the initial scene context and generate incoherent geometry, while our WORLDTRACE-FIELD stays coherent. We quantify it at two horizons ($N=32, 48$) in Tab. 2: with the same canonical averaging setting, at $N=32$, WORLDTRACE-FIELD achieves higher TempSSIM (0.613 vs. 0.571), whereas Centroid-linear reaches the lowest Drift but at 0.04 lower TempSSIM. At $N=48$, WORLDTRACE-FIELD is best on *both* dimensions (+15.5% relative gain on TempSSIM, lowest Drift), while a position assignment that depends on N (Centroid-linear: 0.479) ranks inconsistently across horizons, as its slot indices shift out-of-distribution once N exceeds the training window. These results show that horizon-independent position assignment is the bottleneck for long-horizon generation.

4.3 Episodic Recall: WORLDTRACE-LANDMARK

While Sec. 4.2 evaluates WORLDTRACE-FIELD on open-ended rollouts, episodic recall asks whether the cache can restore a *specific* past scene at revisit points. We therefore test WORLDTRACE-LANDMARK (Sec. 3.4) on LOOPMEM (Fig. 3, Sec. 4.1) and score each return frame against scene A at matched poses with PAC. Fig. 5 shows the qualitative effect: the sliding-window baseline fails to reconstruct scene A at revisits, whereas WORLDTRACE-LANDMARK matches it across all three geometries.

Tab. 3 evaluates WORLDTRACE-LANDMARK across four difficulty tiers: *topology* (ABA, ABCA, ABCDA), *edge length* (ABA at $L \in \{4, 8, 16\}$), *camera-orientation* (pan $90^\circ/180^\circ/360^\circ$ with the agent fixed at A), and *multi-revisit* (ABABA, ABCBA, ABCDBA). All runs use $N_s=4$, $N_r=2$.

Across tiers, WORLDTRACE-LANDMARK improves PAC over the sliding-window baseline in every reported condition, with the largest relative gains when the KV gap to scene A is longest (Tier 2) and when multiple waypoints separate departure and return (Tier 1). Tier 3 is the hardest setting, where gains shrink at a full

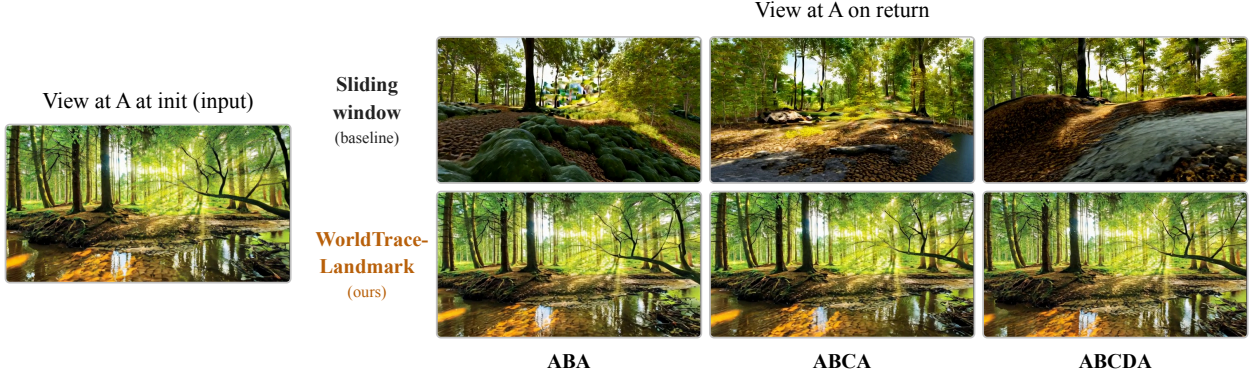


Figure 5: **LOOPMEM qualitative results** on the three revisit geometries in Fig. 3 (ABA, ABCA, and ABCDA). The sliding-window baseline fails to reconstruct scene A when revisiting. **WORLDTRACE-LANDMARK** matches scene A in all three, keeping memory addressable across waypoints.

Table 3: **WORLDTRACE-LANDMARK improves episodic recall across topology, edge length, camera orientation, and multi-revisit depth.** Loop geometries and difficulty axes are in Fig. 9 (App. F). Each tier groups methods (rows) by loop configuration (columns), and entries are PAC / TempSSIM. N is rollout length in AR chunks, and $n=100$ is the number of distinct initial scenes per condition (Tab. 6). Tier 4 ($R=2$, row 4 of Fig. 9): ABABA revisits scene A, while ABCBA and ABCDBA revisit waypoint B on palindrome/shortcut paths.

Tier 1: varying topology			
	ABA ($K=1, N=16$)	ABCA ($K=2, N=17$)	ABCDA ($K=3, N=16$)
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	$0.864_{\pm 0.009} / 0.786$	$0.792_{\pm 0.011} / 0.758$	$0.799_{\pm 0.011} / 0.771$
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	$0.922_{\pm 0.004} / 0.789$	$0.864_{\pm 0.009} / 0.786$	$0.825_{\pm 0.009} / 0.787$
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	$0.861_{\pm 0.007} / 0.493$	$0.781_{\pm 0.010} / 0.486$	$0.577_{\pm 0.015} / 0.467$
Tier 4: multi-revisit			
	ABABA ($N=32$)	ABCBA ($N=20$)	ABCDBA ($N=20$)
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	$0.941_{\pm 0.005} / 0.789$	$0.863_{\pm 0.009} / 0.771$	$0.876_{\pm 0.009} / 0.775$

360° pan, and Tier 4 shows a narrower margin on palindrome returns that revisit waypoint B before scene A. At extended horizons, the PAC sweep (Tab. 12, App. E.2) shows only verbatim landmark recall sustains $\text{PAC} \approx 0.99$ at $N=256$ (0.610 for canonical-K anchoring). We provide further comparisons with MemRoPE and YaRN, history-selection ablations, and slot-count sensitivity in App. E.

4.4 Ablation

We isolate two design axes. (1) **Position assignment** (WORLDTRACE vs. Block-relative vs. Centroid-linear): 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. We use the ABA recall protocol with a fixed cache budget.

(1) **Slot-rank positions keep summary slots addressable** (Tab. 4). Canonical averaging with Block-relative positions collapses *exactly* to the sliding-window baseline, meaning OOD offsets receive zero softmax weight.

Table 4: **Position assignment & canonical compression ablation (Axes 1 & 2).** PAC ($\uparrow$) and TempSSIM ($\uparrow$) on ABA recall.

Method	PAC $\uparrow$				TempSSIM ($N=16$) $\uparrow$
	$N=16$	$N=64$	$N=128$	$N=256$	
<i>Axes 1 & 2: position assignment & canonical compression</i>					
Sliding window (baseline)	0.540	0.412	0.504	0.631	0.9945
+ Naive averaging (Block-relative)	0.570	0.443	0.486	0.598	0.9944
+ Field averaging (Block-relative)	0.540	0.412	0.504	0.631	0.9945
+ WORLDTRACE (WORLDTRACE-FIELD, ours)	0.555	0.442	0.495	0.602	0.9948

Naive averaging in RoPE-rotated space scores marginally higher on PAC at $N=16$ (0.570 vs. 0.540), but with the same Block-relative positions it cannot reflect addressable recall: canonical averaging matches the sliding-window baseline byte-for-byte at every horizon (Tab. 4), and both stay in the compression tier (≈ 0.4 to 0.6 PAC) far below anchoring or verbatim recall ($\approx 0.95+$, Tab. 12). Only WORLDTRACE keeps every summary slot in-distribution: WORLDTRACE-FIELD recovers PAC=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. 5). Holding positions fixed at Block-relative, 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 each additional summary slot mixes keys from frames at more diverse timestamps amplifying RoPE phase cancellation, while canonical averaging is unaffected (0.224 to 0.257). Averaging in canonical (unrotated) space is therefore necessary to preserve the low-frequency content of compressed summary Keys.

 Table 5: **Phase cancellation: Naive vs. canonical key averaging.** LatentDiff ($\downarrow$) at $N=16$ rollout chunks for varying numbers of summary-cache slots N_s . Canonical averaging avoids the phase cancellation that corrupts low frequencies under naive RoPE-space averaging.

N_s	Naive LatentDiff $\downarrow$	Canonical LatentDiff $\downarrow$
1	0.257	0.224
2	0.261	0.257
4	0.312	0.233

5 Conclusion

Long-horizon failure in autoregressive video world models is not only a matter of retaining more past content, but of keeping that content addressable. We show that temporal RoPE extrapolation can make compressed memories effectively unreadable once rollouts move beyond the training horizon, and introduce WorldTrace, a training-free cache mechanism that keeps distant traces at fixed in-distribution virtual positions. Within this addressable cache, WorldTrace-Field provides a coherence-oriented summary of history, while WorldTrace-Landmark preserves verbatim traces for episodic recall. Empirically, WorldTrace-Field improves TempSSIM by +15.5% at $N=48$, and WorldTrace-Landmark raises PAC by +19.5% on LoopMem ABA loops, extending visually persistent generation from seconds toward minute-scale return-to-origin rollouts.

More broadly, WorldTrace suggests that memory for video world models should be designed around two questions: how to ensure past content remains addressable, and what kind of visual traces should be stored for future queries. The two variants studied here represent different points in this design space: smooth summaries support temporal coherence, while sparse landmarks support recall of specific places. This perspective also clarifies the limits of the current method. WorldTrace is designed for temporal-RoPE autoregressive models with a fixed KV-cache budget, and its current content writers instantiate two simple structured projections of history: WorldTrace-Field favors coherence but can blur specific scene details, while WorldTrace-Landmark preserves specific places but depends on detecting and retaining the right scene entries. Future memory systems may need to choose these projections more adaptively, especially important in interactive settings where agents move freely, revisit locations, and query the past from changing viewpoints. In this view, WorldTrace-Field and WorldTrace-Landmark are not endpoints, but two structured approximations to the same underlying problem: projecting a long visual history into a small addressable cache.

Acknowledgements

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

References

- Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. 1
- 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. 31
- 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. 3, 28, 30
- Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, et al. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026. 1, 31
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 30
- Jesse Bettencourt, Xindi Wu, Matan Atzmon, James Lucas, and Jonathan Lorraine. Variance reduction for expectations with diffusion teachers. *arXiv preprint arXiv:2605.21489*, 2026. 41
- 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, 4, 30
- 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 generation models as world simulators. OpenAI Technical Report, <https://openai.com/research/video-generation-models-as-world-simulators>, 2024. 31
- 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. 31
- Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 30
- Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyang 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=y6XJZ1EC2x>. 31
- 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, 4, 30
- 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. 31, 35
- Hanno 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. 31

- 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. 31
- Kaijin Chen, Dingkan 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. 32
- 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. 30
- Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhui Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026d. 31
- 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>. 32
- 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. 30
- 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. 30
- 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. 31
- 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. 31
- 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. 29
- 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, 31
- 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. 30
- 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. 32, 42
- Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 569–578, 2011. 28
- 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. 32
- Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what’s important. *arXiv preprint arXiv:2405.18719*, 2024. 30
- 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. 31

- 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. 1, 31, 42
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. 30
- Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 31
- 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. 31
- 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. 32
- David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 31
- Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. 1, 31
- Junhui He, Junna Xing, Nan Wang, Rui Xu, Shangyu Wu, Peng Zhou, Qiang Liu, Chun Jason Xue, and Qingan Li. A²ATS: 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. 30
- 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, 7, 31
- 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. 32
- 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. 30
- 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. 31
- 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. 30
- 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. 32
- 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. 31
- 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. [31](#)
- 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. [31](#), [35](#)
- Ziqi Huang, Yanan 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. [40](#)
- 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. [30](#)
- 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. [32](#)
- 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. [30](#)
- 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. [30](#)
- 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](#), [4](#), [7](#), [8](#), [31](#), [32](#), [33](#), [41](#)
- Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. [6](#), [25](#)
- 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. [32](#)
- 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. [31](#)
- 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. [31](#)
- 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. [31](#)
- 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. [31](#)
- 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. [31](#)
- Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026e. [30](#)

- 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. [32](#)
- 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. [31](#)
- 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](#), [4](#), [30](#), [41](#)
- 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. [32](#)
- 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. [31](#)
- 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. [30](#)
- Jonathan Lorraine. *Scalable nested optimization for deep learning*. PhD thesis, University of Toronto, 2024. Ph.D. thesis; arXiv:2407.01526. [42](#)
- Jonathan Lorraine and Safwan Hossain. JacNet: Learning functions with structured Jacobians. In *ICML Workshop on Invertible Neural Networks and Normalizing Flows (INNF)*, 2019. [29](#)
- Jonathan Lorraine, Nihesh Anderson, Chansoo Lee, Quentin De Laroussilhe, and Mehadi Hassen. Task selection for AutoML system evaluation. *arXiv preprint arXiv:2208.12754*, 2022a. [35](#)
- 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. [28](#)
- 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. [41](#)
- Yuxiao 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. [31](#)
- 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. [30](#)
- 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. [32](#)
- 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 (EDFM)*, 2024. [41](#)
- 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. [31](#)

- Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16300. 6, 30
- Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024. 30
- NVIDIA. KVPress: A compression library for transformer KV caches. <https://github.com/NVIDIA/kvpress>, 2024. 30, 41
- NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 31
- 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. 31
- 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. 3, 8, 28, 30, 32, 33
- 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. 30
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 8
- 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, 4, 29
- 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. 41
- 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. 31
- Sand AI. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 31
- 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. 30
- 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. 31
- 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. 35, 41
- Sebastian Stapf, Pablo Acuviva, 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>. 31
- 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, 30

- 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. 31
- 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. 30
- 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. 30
- 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, 7, 33, 34, 42
- 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. 30
- 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, 31
- Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 7, 28, 31, 35
- 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. 6, 30
- 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. 41
- 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. 8, 39, 40
- Zile Wang, Zexiang Liu, Jiaying 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. 32, 42
- 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. 32
- 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. 30
- 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. 32
- 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)*, 2025a. arXiv:2506.05284, project page <https://spmем.github.io>. 31

- Xindi Wu, Uriel Singer, Zhaojiang Lin, Andrea Madotto, Xide Xia, Yifan Xu, Paul Crook, Xin Luna Dong, and Seungwhan Moon. Corgi: Cached memory guided video generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4585–4594. IEEE, 2025b. 32
- 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)*, 2026b. arXiv:2601.08828. 41
- Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 29
- 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. 30
- 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)*, 2024. 24, 30
- 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. 30
- 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. 32
- 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. 41
- 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. 31
- 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. 30, 32, 41
- Yan Team. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025. 31
- 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*, 2026. 31
- 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. 32, 42
- 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, 5, 8, 31, 32, 38
- 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. 31, 32

- 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. 31, 35
- 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. 32
- 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. 30, 32, 41
- 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. 31
- 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. 32
- 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. 41
- 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. H₂O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 4, 30, 41
- 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. 30
- 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. 31
- 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. 32, 42
- Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI, 2025a. 40
- 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. 30
- Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026. 31

Appendices

A	Notation	22
B	WORLDTRACE as Structured Sparse Attention	22
B.1	Full and Projected Attention	23
B.2	Structured Attention Approximations Unified by P	24
B.3	Approximation Quality and Error Decomposition	25
B.4	Reduction to Nonnegative Matrix Factorization	25
B.5	When Are WORLDTRACE Projections Close to Optimal?	27
C	WORLDTRACE-FIELD: Additional Properties	28
C.1	Temporal RoPE at Long Horizons	28
D	Related Work	29
D.1	Memory-Augmented Transformers	29
D.2	KV Cache Compression for Large Language Models	30
D.3	Position Extrapolation for RoPE	30
D.4	Autoregressive Video World Models	31
D.5	Memory and KV Cache for Video World Models	31
E	Additional Experimental Results	32
E.1	Concurrent Training-Free Baselines	32
E.2	PAC Episodic Recall Sweep	33
E.3	Cross-Architecture Experiments	33
E.4	Slot Allocation	34
F	LOOPMEM: Loop Memory Benchmark	34
G	Implementation Details	35
G.1	Model Architecture and Self-Forcing Background	35
G.2	WORLDTRACE Algorithm and Architectural Baselines	38
G.3	Hyperparameters, Compute, and Evaluation Protocol	38
G.4	Evaluation Metrics	40
H	Discussion	40
H.1	Limitations	40
H.2	Future Directions	41
H.3	Broader Impact	42

A Notation

Symbols used throughout the paper. Tab. 6 lists model and cache notation, indices, virtual positions, and LOOPMEM parameters. Tab. 7 lists RoPE and key representations.

B WORLDTRACE as Structured Sparse Attention

Standard scaled dot-product attention over T past frames scales as $O(T^2)$ in time and memory, and video world models typically approximate this with a sliding window. Our WORLDTRACE instead introduces *structured* sparse approximations that compress the full T -frame history to $L_{\text{attn}} \ll T$ effective tokens via a projection matrix $P \in \mathbb{R}^{L_{\text{attn}} \times T}$, where each row defines which frames contribute to one compressed slot. Throughout this section, K denotes the canonical key matrix $K_m^{(k)}$ for simplicity, and all keys are in canonical space with in-distribution slot-rank positions already assigned.

Table 6: **Core notation.** Methods, cache structure, indices, virtual positions, scene-boundary threshold, conventions.

Symbol	Description
<i>Cache Structure</i>	
N	Total number of AR chunks in a rollout (generation length)
N_s	Number of summary slots in the KV cache (latent frames)
N_r	Number of recent-window slots (verbatim latent frames)
L_{attn}	Local attention / inference KV window in latent frames ($N_s + N_r = L_{\text{attn}}$)
L_{train}	Self-Forcing rollout cache extent in AR blocks (distinct from L_{attn})
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: $T_{\text{old}} = \max(0, NF - N_r)$ for long rollouts (linear in N)
T	Total latent frames in the full history at a given chunk
s	Summary slot index, $s = 0, \dots, N_s - 1$ (0 = oldest)
$\mathcal{R}$	Recent-window cache: ordered set of verbatim KVs for the newest N_r latent frames
$\mathcal{S}$	Summary cache: N_s compressed slots indexed by s
K_*	Latent frame just evicted from $\mathcal{R}$ when the window overflows
SE	Scene-entry indicator: SE = 1 if cosine-distance spike on canonical- K exceeds τ
J	Number of detected scene-entry landmarks at the current horizon ($J \leq N_s$)
<i>Indices</i>	
k	RoPE frequency-pair index
m	Source-frame iteration index (inside $\sum_{m=1}^M$)
n	AR-chunk index (0-based or 1-based as stated in context, distinct from replication count n below)
f	Intra-block frame index, $f \in \{0, \dots, F-1\}$ (per-frame slot of an AR block)
μ	Algorithm mode selector ($\mu \in \{\text{WORLDTRACE-FIELD}, \text{WORLDTRACE-LANDMARK}\}$, Alg. 1)
<i>Virtual Position Assignment (Slot Indexing)</i>	
q^a	Absolute timestamp of the current query frame
t	Absolute timestamp of a cached key frame
$t_{\min}^v$	In-distribution lower bound: $\max(0, q - \Delta t_{\text{train}})$
$t_{\max}^v$	In-distribution upper bound for summary slots: $q - N_r$
t_s^v	Generic virtual position (no slot index)
t_s^v	Virtual position of slot s : $q - (L_{\text{attn}} - 1 - s)$
<i>LOOPMEM Benchmark</i>	
K	Number of intermediate waypoints between departure and return to scene A
L	Edge length: AR chunks per directed navigation edge (distinct from L_{train})
R	Multi-revisit depth: revisits to the repeated waypoint in the pattern (scene A in ABABA, waypoint B in ABCBA/ABCDBA)
<i>Scene-Boundary Detection</i>	
τ	Scene-boundary detection threshold ($\tau=0.15$, used by WORLDTRACE-LANDMARK)
<i>Experimental Conventions</i>	
n	Number of distinct initial scenes (test instances) per evaluation condition
$N_s + N_r = L_{\text{attn}}$	Capacity constraint: total cache fills the local attention window

^a The glyph q here is a frame timestamp. The single query vector in App. B (q) reuses the same letter, but the two never co-occur: positions are kept implicit in that appendix.

B.1 Full and Projected Attention

For a single query q , the standard attention output is:

$$\mathcal{A}(q, K, V) = \text{softmax}(q K^\top) V = \alpha_q V,$$

which is a weighted sum over all T value vectors, with $\alpha_q \in \Delta_T$.

The *projected* attention approximation replaces the T -token sequence with $L_{\text{attn}} \ll T$ prototype tokens defined by a row-stochastic matrix $P \in \mathbb{R}^{L_{\text{attn}} \times T}$:

$$\hat{\mathcal{A}}(q, K, V; P) = \text{softmax}(q (PK)^\top) PV, \quad (5)$$

where $\hat{\alpha}_q = \text{softmax}(q (PK)^\top) \in \Delta_{L_{\text{attn}}}$ are the prototype attention weights and $P_i V$ is the compressed value for slot i .

Table 7: **RoPE and key representations.** Parameters, rotations, and key/query variants used across the equations.

Symbol	Description
<i>RoPE Parameters</i>	
θ	RoPE base frequency ($\theta=10000$ 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)
D_{true}	Block-diagonal matrix applying the true absolute-position RoPE rotation to each row of the full T -frame key matrix
D_v	Block-diagonal matrix applying slot-rank virtual-position rotations to the L compressed keys: $\text{blockdiag}(R(\theta_k t_0^v), \dots, R(\theta_k t_{L-1}^v))$
<i>Key / Query Variants</i>	
$K_{t_m}^{(k)}$	RoPE-rotated key at absolute position t_m , frequency pair k
$K_m^{(k)}$	Canonical (unrotated) key content: $R(-\theta_k t_m) K_{t_m}^{(k)}$
$\bar{K}^{(k)}$	Canonical mean across M source frames (implicit operator output of WORLDTRACE-FIELD)
$\bar{K}_{\text{naive}}^{(k)}$	Naive RoPE-space average of M rotated keys
$K_{\text{field}}^{(k)}(t^v)$	WORLDTRACE-FIELD compressed key at virtual position t^v (Def. 2)
$K_{\text{land}}^{(k)}(t_v^{(s)})$	WORLDTRACE-LANDMARK frozen canonical key at slot s (Eq. (4))
t_{ℓ^*}	Original timestamp of a selected landmark frame
$K_{t_{\ell^*}}^{(k)}$	Landmark source key at t_{ℓ^*} , frequency pair k
$Q_q^{(k)}$	RoPE-rotated query at q , frequency pair k
<i>Attention Quantities</i>	
$\delta_{q,t}$	Query-key temporal offset: $q - t$
Δt_{train}	Largest temporal offset within the local attention window during training (latent frames, e.g. $\Delta t_{\text{train}}=5$ for MG2, so $L_{\text{attn}}=\Delta t_{\text{train}}+1$)
$a_k(q, t)$	Attention-score contribution from frequency pair k at offset $\delta_{q,t}$
$A_{q,t}^{(k)}$	Query-key content term in canonical coordinates for $a_k(q, t)$
α_q	True softmax attention weights for query q over all T past frames: $\alpha_q \in \Delta_T$
<i>Math Operators</i>	
$\text{Re}(\cdot)$	Real part of a complex expression (used in $a_k(q, t)$)
e, i	Euler’s number and imaginary unit (italic, complex exponentials in Eq. (3))
sp	Spatial-axis label superscript ($e^{i\theta^{\text{sp}} \cdot (h,w)}$, App. C)
$s(\cdot)$	Row-wise softmax operator
Δ_n	Probability simplex in $\mathbb{R}^n$: $\{x \in \mathbb{R}^n : x_i \geq 0, \sum_i x_i = 1\}$
$\mathcal{Q}$	Finite set of query frames used to evaluate approximation quality

B.2 Structured Attention Approximations Unified by P

The four cache structures discussed in the paper correspond to four structured choices of P , illustrated in Fig. 6.

Sliding window attention. $P = [0 \ I_{L_{\text{attn}}}]$: each row is a one-hot vector selecting one of the most recent L_{attn} frames, discarding all earlier history (Fig. 6a). Adding an *attention sink anchor* is a simple modification (Fig. 6b): row 0 is pinned to the first (sink) frame (the *attention sink anchor*) and the remaining $L_{\text{attn}}-1$ rows select the most recent $L_{\text{attn}}-1$ frames verbatim, always keeping a global context token in cache regardless of generation length. This anchor prevents the attention distribution from collapsing onto only local tokens at long generation horizons [Xiao et al., 2024].

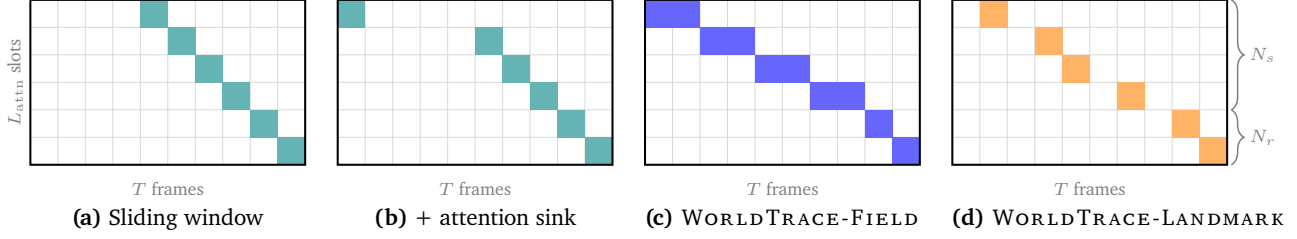


Figure 6: **Projection matrices P for the four cache structures.** (a) *Sliding window*: last L_{attn} frames verbatim. (b) *Sliding window with attention sink*: initial frame held as an attention sink, remaining slots verbatim recent. (c) *WORLDTRACE-FIELD*: recent rows verbatim newest, summary slots each average a group of earlier frames. (d) *WORLDTRACE-LANDMARK*: summary rows are one-hot at scene-entry frames, recent rows verbatim.

WORLDTRACE-FIELD. Rows corresponding to the N_r recent slots are identity rows as above. For each summary slot s , one row shares a uniform average over the M frames in source group g_s (Fig. 6c):

$$P_{i,t} = \begin{cases} 1/M & t \in g_s, \\ 0 & \text{otherwise.} \end{cases}$$

WORLDTRACE-LANDMARK. Summary rows are one-hot selectors at the detected scene-entry frame t_{ℓ^*s} (Fig. 6d):

$$P_{i,t} = \begin{cases} 1 & t = t_{\ell^*s}, \\ 0 & \text{otherwise,} \end{cases}$$

so each summary row simply copies one stored key, $P_i K = K[t_{\ell^*s}]$, matching the frozen key of Eq. (4). The rows update as generation proceeds: a new scene-entry frame overwrites a summary row, and once all rows are full the oldest landmark is dropped to make room. Fig. 8 shows these choices in the broader spectrum of P -row structures, from sparse to dense.

B.3 Approximation Quality and Error Decomposition

For a single query q , let $\hat{\alpha}_q = \text{softmax}(q(PK)^\top) \in \Delta_{L_{\text{attn}}}$ be the prototype attention weights under the compressed cache, and define the output residual $r_q(P) = \hat{\mathcal{A}}(q, K, V; P) - \mathcal{A}(q, K, V)$.

Proposition 1 (Approximation bound).

$$\|r_q(P)\|_2 \leq \underbrace{\|\hat{\alpha}_q P - \alpha_q\|_1}_{\text{distribution mismatch}} \cdot \|V\|_{\infty, \text{row}},$$

where $\|V\|_{\infty, \text{row}} = \max_t \|V[t]\|_2$.

Proof. $r_q(P) = (\hat{\alpha}_q P - \alpha_q) V$. Applying Hölder’s inequality with the ℓ_1 -norm on the weight vector and the ℓ_∞ -norm on the value rows gives the bound. $\square$

The bound motivates minimizing distribution mismatch $\|\hat{\alpha}_q P - \alpha_q\|_1$ independently of value geometry. For **WORLDTRACE-FIELD**, Prop. 2 ensures that the compressed key $K_{\text{field}}^{(k)}(t^v)$ reproduces the mean attention score of all source keys reassigned to the shared virtual position. This limits per-slot mismatch when query preferences vary smoothly across the group. For **WORLDTRACE-LANDMARK**, using a verbatim scene-entry frame concentrates $P_i K$ at a high-salience key, which minimises the worst-case ℓ_1 mismatch for queries targeting that scene-entry event. Fig. 7 measures this distribution mismatch empirically on a synthetic key pool, confirming that the projection families order as predicted by the bound.

B.4 Reduction to Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) approximates a nonnegative matrix as the product of two low-rank nonnegative factors [Lee and Seung, 1999], which form a compact summary of the original data. We show that

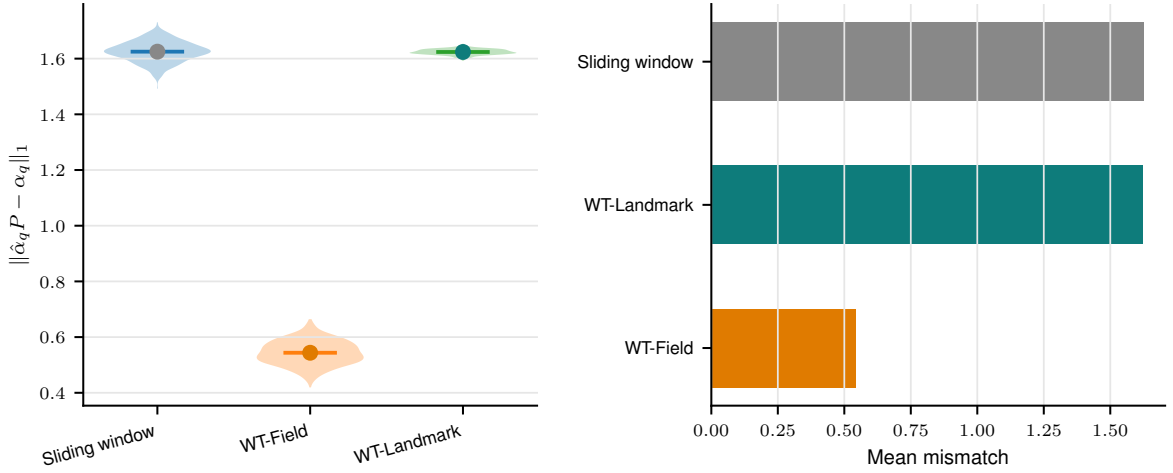


Figure 7: **Distribution mismatch on a $T=32$ random-walk canonical-key pool.** Left: per-query ℓ_1 mismatch $\|\hat{\alpha}_q P - \alpha_q\|_1$ over 512 random queries. Right: mean mismatch ranks WORLDTRACE-LANDMARK lowest, consistent with the ordering in Sec. 4.3.

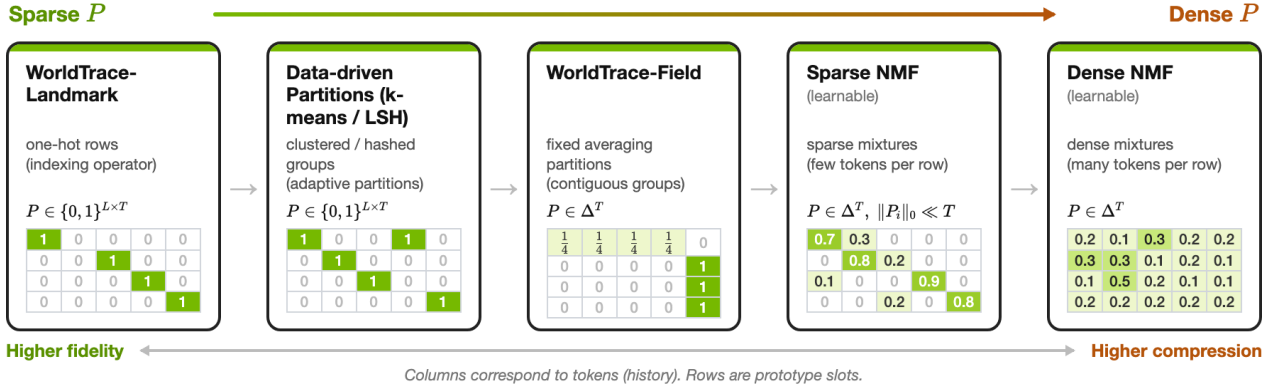


Figure 8: **Spectrum of memory compression under the unified P -matrix view.** Moving left to right, sparsity decreases and expressive capacity increases. The ordering reflects joint row sparsity and learnability.

compressing attention through P is an instance of this problem: the rows of P play the role of one nonnegative factor, and the cache structures of Sec. 3.3 and Sec. 3.4 correspond to constrained NMF solutions.

Treating $\hat{\alpha}_q \in \Delta_{L_{\text{attn}}}$ as a free variable (decoupling it from the softmax in Eq. (5)) turns the aggregate distribution-mismatch objective into a bilinear problem over non-negative matrices with row-sum constraints:

$$\min_{\substack{P: P_{i,:} \in \Delta_T \\ \{\hat{\alpha}_q\}: \hat{\alpha}_q \in \Delta_{L_{\text{attn}}}}} \sum_{q \in \mathcal{Q}} \|\hat{\alpha}_q P - \alpha_q\|_1.$$

Stacking rows α_q into $\mathbf{A} \in \mathbb{R}^{|\mathcal{Q}| \times T}$ and rows $\hat{\alpha}_q$ into $\mathbf{B} \in \mathbb{R}^{|\mathcal{Q}| \times L_{\text{attn}}}$, this becomes

$$\min_{\mathbf{B} \geq 0, P \geq 0} \|\mathbf{B}P - \mathbf{A}\|_1 \quad \text{subject to row-sum constraints on } \mathbf{B} \text{ and } P,$$

which is **NMF with an ℓ_1 fidelity term** (robust NMF). Each structural constraint on P corresponds to a WORLDTRACE variant and a known clustering algorithm (Tab. 8), spanning the spectrum of Fig. 8.

WORLDTRACE-FIELD uses uniform-partition rows where each group g_s covers a disjoint temporal segment. This is equivalent to k -means with cluster-mean prototypes, where the time-ordered partition provides a fixed assignment. WORLDTRACE-LANDMARK uses 0/1 rows with data-driven scene-entry placement. When the prototype weights $\hat{\alpha}_q$ are additionally constrained to be 1-sparse (hard assignment to one prototype), this

Table 8: **Structural constraints on P** . Each row structure corresponds to a `WORLDTRACE` variant and a known clustering algorithm.

Structure of P	Corresponding classic algorithm
Dense rows in Δ_T	NMF / dictionary learning
Uniform-partition rows (<code>WORLDTRACE-FIELD</code>)	k -means with cluster-mean prototypes
0/1 rows / token selection (<code>WORLDTRACE-LANDMARK</code>)	k -medoids in TV distance

reduces to the k -medoids problem on attention distributions:

$$\min_{S \subset [T], |S|=N_s} \sum_{q \in \mathcal{Q}} \min_{j \in S} \|\alpha_q - \mathbf{e}_j\|_1,$$

where $\mathbf{e}_j$ is the j -th standard basis vector. The canonical-key cosine-distance rule (Sec. 3.4) is a lightweight, single-pass $O(1)$ -per-chunk approximation to this oracle selection.

Remark 3 (Reintroducing the softmax coupling). Fixing $\hat{\alpha}_q = \text{softmax}(q(PK)^\top)$ couples $\mathbf{B}$ to P through the softmax, resembling conditional NMF. The relaxed formulation above provides the theoretical grounding. The softmax-coupled version is what runs at inference. This two-stage view mirrors the position-content decoupling of the main text: positions are fixed first for addressability, leaving the choice of P free to be determined by any content criterion.

B.5 When Are `WORLDTRACE` Projections Close to Optimal?

The NMF objective above is deliberately broad: without assumptions on the query distribution, no fixed L_{attn} -row projection can be uniformly optimal for every possible future query. A compressed cache must decide which parts of history to preserve. `WORLDTRACE` makes two different choices, and each is near-optimal for a different query regime. We state the regimes in terms of the relaxed mismatch objective

$$\mathcal{J}(P) = \sum_{q \in \mathcal{Q}} \min_{\gamma_q \in \Delta_{L_{\text{attn}}}} \|\gamma_q P - \alpha_q\|_1,$$

which lower-bounds the softmax-coupled objective used at inference. The gap between $\mathcal{J}(P)$ and inference performance is the softmax-coupling slack discussed above.

Coherence regime: group-level attention. Suppose the old history is partitioned into N_s temporal groups $g_1, \dots, g_{N_s}$ plus a recent identity window. Let u_i be the uniform distribution on group g_i . `WORLDTRACE-FIELD` is close to optimal when future queries care about *group-level* history but not the identity of individual frames inside each group: for every query q , there exists a distribution $a_q \in \Delta_{N_s}$ such that

$$\left\| \alpha_q - \sum_i a_{q,i} u_i \right\|_1 \leq \varepsilon_q. \quad (6)$$

This is exactly the assumption that attention is smooth within each compressed temporal block. Under Eq. (6), the `WORLDTRACE-FIELD` projection P_{field} satisfies

$$\mathcal{J}(P_{\text{field}}) \leq \sum_{q \in \mathcal{Q}} \varepsilon_q. \quad (7)$$

Indeed, choose $\gamma_q = a_q$ in the relaxed objective. Since $\mathcal{J}(P) \geq 0$ for any projection, `WORLDTRACE-FIELD` is within $\sum_q \varepsilon_q$ of the best possible projection in this regime. The canonical averaging operator of Def. 2 is the key implementation detail: it realizes the uniform rows u_i in a RoPE-invariant domain, so the group mean remains addressable after being placed at the slot-rank virtual position.

Recall regime: sparse scene-entry attention. For episodic recall, the relevant query distribution is not smooth over time. A return query often wants one specific past scene. Let $\mathcal{S}$ be the set of stored landmarks,

$|\mathcal{S}| \leq N_s$, and suppose every recall query has a matching stored scene-entry frame $m(q) \in \mathcal{S}$ such that

$$\|\alpha_q - \mathbf{e}_{m(q)}\|_1 \leq \delta_q, \quad (8)$$

where $\mathbf{e}_{m(q)}$ is the one-hot distribution on that frame. Then the `WORLDTRACE-LANDMARK` projection P_{landmark} satisfies

$$\mathcal{J}(P_{\text{landmark}}) \leq \sum_{q \in \mathcal{Q}} \delta_q. \quad (9)$$

The proof is again by construction: choose γ_q to put all mass on the landmark slot containing $m(q)$. Thus `WORLDTRACE-LANDMARK` is close to the best one-hot-token projection when the scene-entry detector forms a δ -cover of the recall targets. The FIFO scene-entry rule used in our experiments is a streaming approximation to this oracle cover. It is reliable when the number of simultaneously relevant scenes is at most N_s and scene transitions are sharp enough to trigger the detector.

A unified view and future directions. The P -matrix view is the unifying framework: `WORLDTRACE-FIELD` and `WORLDTRACE-LANDMARK` are two structured instantiations of the same NMF objective, differing only in the sparsity of their projection rows. Rather than separate methods, they occupy complementary positions in a continuous design space, with uniform-partition rows capturing smooth group-level history and one-hot rows preserving verbatim scene-entry frames for episodic recall. The NMF framing also points to natural extensions: online k -means or streaming coresets [Feldman and Langberg, 2011] could adapt partition boundaries to the observed attention distribution at inference time, soft-sparse rows could interpolate between averaging and verbatim selection, and a directly optimized P could jointly balance coherence and recall across the full query mix.

C WORLDTRACE-FIELD: Additional Properties

This section gives temporal-RoPE background for the addressability failure in Sec. 2.1 and expands Eq. (3), summarized at the end of Sec. 3.3.

C.1 Temporal RoPE at Long Horizons

We analyze which temporal RoPE frequencies stay in-distribution at long horizons and which become positional noise on MG2-1.3B.

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 $k = 0, \dots, c_t-1$, giving $\theta_k=1.0$ at $k=0$ (fastest rotating) down to $\theta_k \approx 1.5 \times 10^{-4}$ at $k=c_t-1$ (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 9: **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), and high frequencies ($k\geq 10$) stay near in-distribution.

Freq k	θ_k	Train max ($ \delta_{q,t} \leq 5$)	Inference ($ \delta_{q,t} =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

Mean attention preservation. The WORLDTRACE-FIELD compression preserves the mean attention score: a single summary token $K_{\text{field}}^{(k)}(t^v)$ reproduces exactly the average of the scores 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 $Q_q^{(k)}$ and every temporal RoPE frequency k , so no query can distinguish a slot that was built from one frame versus many.

Proposition 2 (Mean attention preservation). *Fix a temporal RoPE pair k and a virtual position t^v . For any query $Q_q^{(k)}$,*

$$\langle Q_q^{(k)}, K_{\text{field}}^{(k)}(t^v) \rangle = \frac{1}{M} \sum_{m=1}^M \langle Q_q^{(k)}, R(\theta_k t^v) K_m^{(k)} \rangle.$$

Proof. By Eq. (3), $K_{\text{field}}^{(k)}(t^v) = R(\theta_k t^v) \bar{K}^{(k)}$ with $\bar{K}^{(k)} = \frac{1}{M} \sum_m K_m^{(k)}$. Linearity of the inner product in its second argument then gives $\langle Q_q^{(k)}, K_{\text{field}}^{(k)}(t^v) \rangle = \langle Q_q^{(k)}, R(\theta_k t^v) \bar{K}^{(k)} \rangle = \frac{1}{M} \sum_m \langle Q_q^{(k)}, R(\theta_k t^v) K_m^{(k)} \rangle$. The summary token contributes the mean attention score that the source keys would produce if their content were first moved to the shared virtual position t^v . $\square$

Concretely, $K_{\text{field}}^{(k)}(t^v)$ 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 $\bar{K}^{(k)}$. 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) maps M RoPE-encoded keys to one token at virtual position t^v while satisfying the mean attention preservation of Prop. 2: unrotate to canonical space, average, and re-encode at t^v . This property fully determines the compressed content. Any linear operator satisfying it yields the same canonical average $\bar{K}^{(k)}$, since it pins the output logit for all queries, leaving only t^v free, which WORLDTRACE sets (Sec. 3.1). Pinning a learned operator by a structural property of its derivative (the block-diagonal RoPE phase rotation that Eq. (3) inverts) has antecedents in structured-Jacobian parameterizations [Lorraine and Hossain, 2019]; WORLDTRACE-FIELD is its closed-form, training-free analog for the temporal RoPE block. Non-linear alternatives (e.g. keeping the highest-norm token verbatim) satisfy neither property and lose phase coherence across compressed frames.

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 could counteract this in principle. We use unscaled canonical averaging throughout.

D Related Work

D.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, giving the two-tier structure we adopt. Memorizing Transformers [Wu et al., 2022] attach an external k NN-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 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-and-rotate drift in 3D RoPE. The broader full-attention-to-compressed-state spectrum spans linear attention [Katharopoulos et al., 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.

D.2 KV Cache Compression for Large Language Models

LLM KV-cache methods fall into three families: *window/eviction* (sliding windows [Beltagy et al., 2020], sink-augmented StreamingLLM [Xiao et al., 2024], and importance-based eviction such as H₂O [Zhang et al., 2023], SnapKV [Li et al., 2024], and PyramidKV [Cai et al., 2025]); *merging* (ToMe [Bolya et al., 2023], KeepKV [Tian et al., 2026]); and *quantization* (KIVI [Liu et al., 2024]), which is orthogonal to the position-ODD failure WORLDTRACE targets. DuoAttention [Xiao et al., 2025a] splits heads into retrieval vs. streaming families, parallel to our recent-window vs. summary-slot split.

Most of these methods change *content* only: which tokens to keep or merge, not which virtual position to assign a summary spanning multiple merged frames. EliteKV [Zhou et al., 2025b] and A²ATS [He et al., 2025a] touch positional structure for per-token compression or retrieval, but not summary-slot reassignment. FINCH [Corallo and Papotti, 2024] and KVPress [NVIDIA, 2024] unrotate and rotate individual retained keys. WORLDTRACE-FIELD instead averages multiple canonical keys and pairs the result with slot-rank positions.

Why KV compression differs in our setting. LLM KV compression is typically analyzed on a single growing token stream queried at inference time. Our cache instead accumulates keys across independently denoised AR video blocks, so retention heuristics must rank importance under denoising-timestep-conditioned queries rather than stationary next-token prediction. Angular-concentration scores such as TriAttention [Mao et al., 2026a] assume stable pre-RoPE key geometry that does not hold in this regime. Layer-budget reallocations (PyramidKV, SimLayerKV) tune which layers keep tokens, not how compressed summary slots are addressed at read time. Positional reassignment is treated in Sec. D.3.

D.3 Position Extrapolation for RoPE

RoPE [Su et al., 2024] degrades past trained relative offsets. Remedies span *index rescaling* (Position Interpolation [Chen et al., 2023], YaRN [Peng et al., 2024], LongRoPE [Ding et al., 2024], LongRoPE2 [Shang et al., 2025]), *alternative distance biases* (ALiBi [Press et al., 2022], xPos [Sun et al., 2023]), *position-free training* (Kazemnejad et al. [2023]), and *frequency-/content-aware encodings* (CoPE [Golovneva et al., 2024]; Barbero et al. [2025]; HoPE [Chen et al., 2025b]; FoPE [Hua et al., 2025]). NoPE clarifies that our argument targets models *already* trained with RoPE. *Vision and video* extensions add 2D/3D splits (RoPE-ViT [Heo et al., 2024], VideoRoPE [Wei et al., 2025]), camera-relative attention encodings (PRoPE [Li et al., 2026e]; ViewRoPE [Xiang et al., 2026]), and training-free diffusion extrapolation (RIFLEX [Zhao et al., 2025]).

Most prior work *remaps* positions for individually retained tokens or fixed-length inputs, not virtual offsets for *compressed summary slots*. Wang et al. [2025] tie position to numerical precision, consistent with our coupling thesis (Sec. 3.5). UCM [Xu et al., 2026b] and MosaicMem [Yu et al., 2026] warp 3D positions for camera-controlled worlds as geometry-driven, trained operators. WORLDTRACE instead uses slot-rank reassignment without retraining.

Why RoPE extrapolation differs in our setting. We accumulate keys across independently denoised AR blocks, so the cache must be *compressed* rather than only re-mapped, and each merged summary needs an explicit virtual position absent when tokens keep separate identities. WORLDTRACE assigns slot-rank offsets for that role (Eq. (2)).

D.4 Autoregressive Video World Models

Interactive AR video world 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 via chunk-wise diffusion prediction. The dominant *Self-Forcing* paradigm [Huang et al., 2025b] (Diffusion Forcing [Chen et al., 2024], DFoT [Song et al., 2025a], CausVid [Yin et al., 2025]) aligns training rollouts with the model’s own KV cache and inherits a local attention window not sized for arbitrarily long inference. *Training-time horizon extensions* include Self-Forcing++ [Cui et al., 2026b], FAR [Gu et al., 2025], Vid2World [Huang et al., 2026b], LIVE [Huang et al., 2026a], and SANA-WM [Zhu et al., 2026]. *Foundation backbones* [Wan Team, 2025, NVIDIA, 2025, Sand AI, 2025] and the broader *world-model lineage* from Ha and Schmidhuber [2018] through Dreamer [Hafner et al., 2025], IRIS [Micheli et al., 2023], and DIAMOND [Alonso et al., 2024] to closed-loop autonomous-vehicle simulators such as OmniDreams [Basant et al., 2026] motivate interactive rollouts beyond passive long-form generation.

Most prior lines above extend horizons by *retraining* the generator or changing its architecture. WORLDTRACE instead plugs into a pretrained Self-Forcing checkpoint as-is: no retraining, no architectural change, and only inference-time cache read/write is modified. The limitation we target is inherent to that training contract: the objective aligns per-chunk generation with a rolling KV cache but does not optimize long-horizon addressability of compressed history, so that gap must be closed at inference via cache read/write rather than new distillation or rollout training.

D.5 Memory and KV Cache for Video World Models

Inference-time KV cache. Closest to WORLDTRACE, inference-time methods modify read/write on the standard KV cache without retraining the generator. *Position-side* fixes remap or cap temporal offsets: Infinity-RoPE [Yesiltepe et al., 2026] pairs Block-relative RoPE with KV Flush (sink plus most recent block only); LoL [Cui et al., 2026a] and FLEX [Li et al., 2026b] adjust position assignment at inference; PackCache [Li et al., 2026c] applies per-token virtual positions for cache removal with cross-frame decay. *Content-side* writers decide what to merge, prune, or retain: 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-relative positions with a dual-rate EMA summary cache; PaFu-KV [Chen et al., 2026a] and Rolling Sink [Li et al., 2026a] importance-prune retained tokens; TempCache [Samuel et al., 2026] merges near-duplicate keys via temporal correspondence. FlowCache [Ma et al., 2026] targets per-step compute reduction with chunk-wise denoising-step caching rather than long-horizon recall, but its recent summary structure parallels ours. WORLDTRACE couples horizon-stable slot-rank positions with canonical-domain summary writes (WORLDTRACE-FIELD compresses, WORLDTRACE-LANDMARK freezes), keeping every slot addressable inside a fixed $O(1)$ cache.

Training-time extensions. A second cluster retrains or distills the generator for longer effective context. *Forcing and distillation* methods include Grounded Forcing [Chen et al., 2026b] (Dual Memory KV Cache plus Dual-Reference RoPE Injection), Anchor Forcing [Yang et al., 2026] (sink/junction/local partitions with RoPE re-alignment distillation), Context Forcing [Chen et al., 2026d] (Slow-Fast Memory under a long-context teacher), Hybrid Forcing [Li et al., 2026d] (linear-attention summary state plus block-sparse local attention), Stable Video Infinity [Li et al., 2025d] (Error-Recycling Fine-Tuning), Relax Forcing [Zhao et al., 2026], and Rolling Forcing [Liu et al., 2026]. *Memory decomposition* approaches split retention by function: Composition of Memory Experts [Stapf et al., 2026], the spatial-memory framework of Wu et al. [2025a], and Mixture of Contexts [Cai et al., 2026]. *Architectural long-horizon models* such as FAR [Gu et al., 2025], WorldPack [Oshima et al., 2025], WorldPlay [Sun et al., 2025a], VideoSSM [Yu et al., 2025b], and SANA-WM [Zhu et al., 2026] train for extended rollouts rather than patching a fixed checkpoint at inference. Sparse Forcing [Xu et al., 2026a] learns block-sparse attention with persistent spatiotemporal anchors end-to-end. WORLDTRACE needs no retraining: it restores addressability by changing cache read/write on a fixed checkpoint, with WORLDTRACE-FIELD and WORLDTRACE-LANDMARK as interchangeable content writers.

Geometry-conditioned memory. Trained methods tie retention to camera pose, actions, or scene geometry rather than slot rank alone. RELIC [Hong et al., 2025] annotates compressed historical latents with relative

Table 10: **Memory approach comparison.** Selective baselines and inference-time cache methods on four axes.

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-relative [Yesiltepe et al., 2026]	Yes	Yes ($O(1)$)	No (amnesic)	Yes
Naive avg. + Block-relative	Partial	Yes ($O(1)$)	Corrupted (Sec. 2.2)	Yes
MemRoPE [Kim et al., 2026]	Partial	Yes ($O(1)$)	EMA-diluted	Yes
Deep Forcing [Yi et al., 2025]	Partial	Yes ($O(1)$)	Sink plus pruned recent	Yes
WORLDTRACE-FIELD (ours)	Yes	Yes ($O(1)$)	Yes (compressed field)	Yes
WORLDTRACE-LANDMARK (ours)	Yes	Yes ($O(1)$)	Yes (verbatim landmarks)	Yes

actions and absolute camera pose. MosaicMem [Yu et al., 2026] pairs Warped RoPE with Warped Latent injection. UCM [Xu et al., 2026b] warps 3D positional encodings for camera-controlled worlds. PackForcing [Mao et al., 2026b] applies Continuous Temporal RoPE Adjustment after sink/mid-compressed/recent partitioning (zero-shot or short-clip-trained). MemCam [Gao et al., 2026] selects history by co-visibility. WorldMem [Xiao et al., 2025b] re-injects pose-tagged frames via auxiliary attention. Memorize-When-Needed [Guo et al., 2026] gates a decoupled memory branch on camera-aware relevance. WORLDTRACE is geometry-free: slot-rank positions and canonical cache writes replace pose-, action-, or co-visibility-driven retention.

External memory banks and retrieval. Rather than compressing inside the KV cache, these systems store or retrieve frames *outside* it and re-inject via auxiliary attention or re-encoding. Examples include MG3 [Wang et al., 2026a] (FOV-based frame selection with per-step re-encoding), FramePack [Zhang et al., 2025] (importance-weighted frame contexts), StreamingT2V [Henschel et al., 2025] (appearance-preservation module anchored on the first chunk), Corgi [Wu et al., 2025b] (cached latent memory bank for multi-scene video), 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], and VRAG [Chen et al., 2025a] (explicit global-state retrieval for interactive generation). WORLDTRACE keeps memory in the KV cache: one $O(1)$ recent-plus-summary budget, with no external bank, re-encoding, or retrieval.

Comparison. Tab. 10 qualitatively compares memory approaches on four axes: whether positional OOD is fixed, whether memory is bounded, whether intermediate history is preserved, and whether retraining is required. For evaluation, WorldScore [Duan et al., 2025], MIND [Ye et al., 2026], VBench-2.0 [Zheng et al., 2025], and WorldModelBench [Li et al., 2025a] assess long-horizon quality. Lian et al. [2025] targets revisit consistency in Minecraft. We further propose LOOPMEM (App. F) with closed-loop episodic recall on AR video world models.

E Additional Experimental Results

E.1 Concurrent Training-Free Baselines

Tab. 11 compares MemRoPE [Kim et al., 2026] and YaRN [Peng et al., 2024] on the Sec. 4.3 ABA loops, factoring the position scheme (Block-relative vs. WORLDTRACE) from the content writer at two horizons.

MemRoPE [Kim et al., 2026] (Block-relative + dual-rate EMA with $\alpha_{\text{long}}=0.01$ and $\alpha_{\text{short}}=0.1$, matching the original release) survives OOD positions via EMA decay but is still beaten by Landmark+Block-relative: verbatim canonical keys retain stronger query-key similarity than smoothed averages, and WORLDTRACE-LANDMARK adds in-distribution positions on top. MemRoPE+WORLDTRACE (Block-relative 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. However, 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 $N=512$.

Table 11: **Concurrent training-free baselines.** PAC ($\uparrow$) at two horizons (ABA). YaRN [Peng et al., 2024] uses $O(N)$ memory; only at $N=32$. MemRoPE+WORLDTRACE swaps Block-relative for WORLDTRACE; dual-rate EMA unchanged.

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

Table 12: **Episodic recall splits into three tiers. Only verbatim recall scales.** ABA loops, PAC ($\uparrow$). Bold = best per column.

Method	$N=16$	$N=32$	$N=48$	$N=64$	$N=128$	$N=256$
<i>Compression only:</i>						
Sliding window (baseline)	0.540	0.401	0.388	0.412	0.504	0.631
Naive + Block-relative	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
<i>Canonical-K anchoring:</i>						
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
<i>Verbatim recall:</i>						
WORLDTRACE-LANDMARK (ours)	0.959	0.964	0.972	0.976	0.986	0.989

E.2 PAC Episodic Recall Sweep

Tab. 12 gives the full PAC breakdown across $N \in \{16, 32, 48, 64, 128, 256\}$ on ABA loops ($N_s=4$, $N_r=2$) 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 $N=32$ to 64 ($p<0.001$, paired t -test). Anchoring lifts recall by 40 to 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.

E.3 Cross-Architecture Experiments

Tab. 13 reports PAC at $2\times$ to $8\times$ training horizon on LingBot-World [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 $4\times$ onward (+8.9%, +14.1%, +7.3% at $4\times/6\times/8\times$), with no significant gain at $2\times$ (-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 13: **LingBot-World** [Team et al., 2026] PAC at extended horizons. **WORLDTRACE-LANDMARK** improves over the sliding-window baseline from $4\times$ 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	$2\times$	0.657	0.668	0.651
28	$4\times$	0.624	0.619	0.680
42	$6\times$	0.591	0.620	0.674
56	$8\times$	0.632	0.648	0.678

E.4 Slot Allocation

Summary/recent split sensitivity. The main experiments use $N_s=4$ summary slots and $N_r=2$ recent-window slots (total $N_s+N_r=6$, matching the training context size). To test sensitivity to this split, we swept five allocations ($N_s+N_r=6$, ABA loops at $N=32$ and $N=64$). Results are in Tab. 14. $N_s\leq 2$ collapses toward the sliding-window baseline: with one slot the B $\rightarrow$ 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 14: **Slot Allocation.** PAC at $N=32$ and 64 as a function of N_s (total budget fixed at $N_s+N_r=6$). $N_s\leq 2$ 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	N_r	PAC $N=32$	PAC $N=64$
Sliding window		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

E.4.1 Per-Slot Depth Ablation

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. 15. Slot 3 alone recovers most of the four-slot PAC at $N=32$ and $N=48$ (Tab. 15: 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 15: **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

F LOOPMEM: Loop Memory Benchmark

LOOPMEM evaluates episodic *memory* in autoregressive world models: the model follows a fixed 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 because the forward path generates the target and the return path is scored against it.

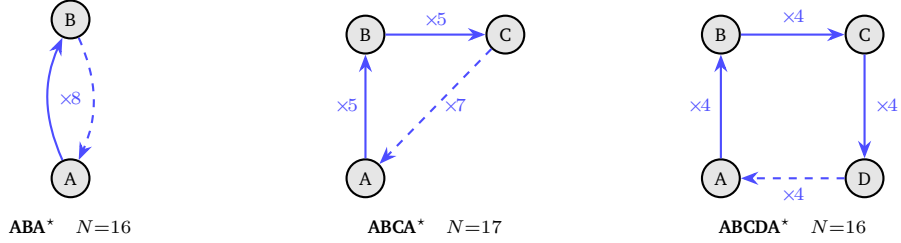
The benchmark organizes difficulty along four axes: (1) **waypoint count** (K): number of intermediate locations between departure and return to A (ABA: $K=1$; ABCDA: $K=3$); (2) **edge length** (L): AR chunks per directed edge, controlling KV-context distance; (3) **camera orientation**: heading on 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 and content axes that the cache mechanism is hypothesized to probe. Fig. 9 shows representative configurations. Those evaluated here are marked *. Fig. 10 and Fig. 11 show qualitative reconstructions on the ABA (axis 1) and pan 90° (axis 3) settings.

G Implementation Details

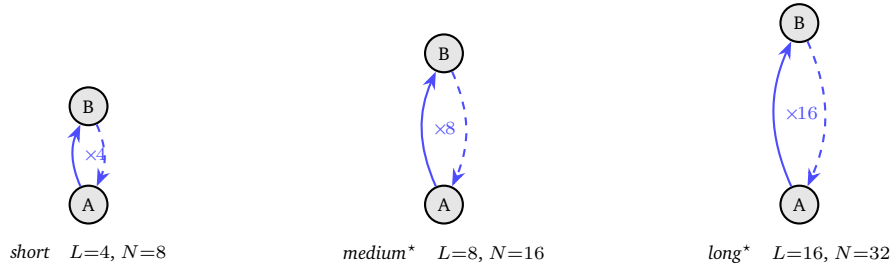
G.1 Model Architecture and Self-Forcing Background

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 $L_{\text{train}} \times 2640 \times 12 \times (128 \times 2) \times 30 \times 2 \approx 2.92$ GB per batch element ($L_{\text{train}}=6$ chunks, 2640 tokens/chunk, 12 heads, head dim 128, $\times 2$ for K and V, 30 layers, fp16 at 2B, c.f. Tab. 16), 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} (App. C.1).

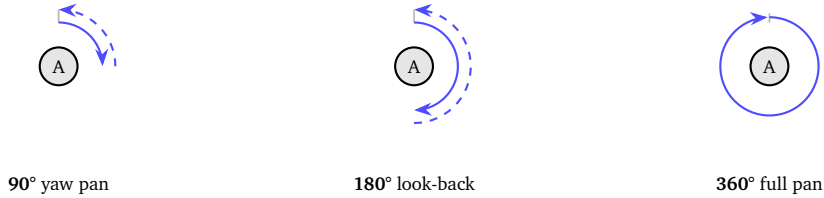
Row 1: varying topology (K intermediate waypoints)



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



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



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

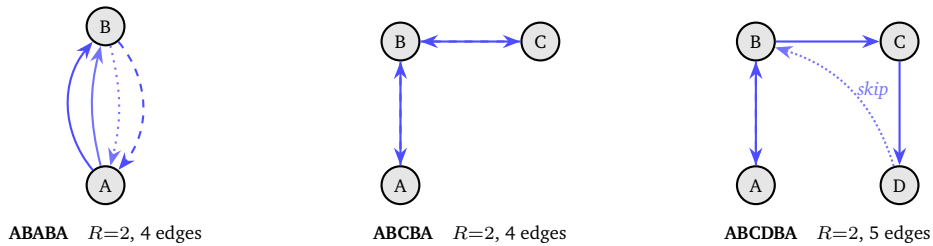


Figure 9: **LOOPMEM benchmark gallery**. Solid/dashed blue: outward/return; lighter arcs: second traversal; dotted: shortcut omitting intermediate waypoints. **Row 1**: topology (ABA/ABCA/ABCD). **Row 2**: ABA edge length (longer edges push RoPE further OOD). **Row 3**: camera orientation (agent at A; solid = pan away, dashed = return). **Row 4**: multi-revisit patterns ($R > 1$; R counts revisits to the repeated waypoint (A in ABABA, B in ABCBA/ABCDBA): palindromes (ABCBA), shortcuts (ABCDBA).



Figure 10: **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.

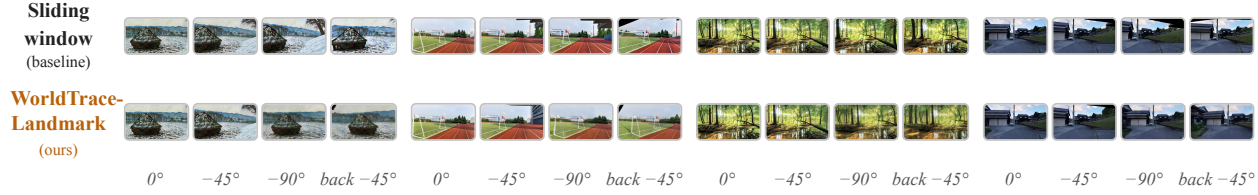


Figure 11: **Pan 90° qualitative results.** Camera-orientation Tier 3 ($N=4$): the agent is fixed at A while the camera pans right by $\sim 90^\circ$ then returns. Each sample group of four frames shows keyframes at chunks 0 (0°) / 1 ($\sim 45^\circ$) / 2 ($\sim 90^\circ$) / 3 (back $\sim 45^\circ$). 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.

Table 16: **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
<i>Architecture</i>	
Transformer layers	30
Attention heads	12
Head dimension	128
Temporal RoPE complex pairs (c_t)	22
<i>Video & Latent Space</i>	
Decoded output FPS	16
VAE temporal compression	$4\times$
VAE spatial compression	$8\times$ (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
<i>Training Context</i>	
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)	≈ 486 MB
<i>Inference Horizon</i>	
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)

Algorithm 1 WORLDTRACE cache update (per AR chunk). $\mathcal{S}$ stores canonical (unrotated) keys; $R(\theta_k t_s^v)$ is applied per slot at attention time via Def. 1, so positions recompute automatically as q advances (shared by both variants). Index $f \in \{0, \dots, F-1\}$ 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 $(\mathcal{R}, \mathcal{S})$

```

1: Append new KVs to  $\mathcal{R}$  ▷ recent window is verbatim
2: if  $|\mathcal{R}| > N_r$  then
3:    $K_* \leftarrow \text{POPOLDEST}(\mathcal{R})$ 
4:   if  $\mu = \text{WORLDTRACE-FIELD}$  then
5:      $s^* \leftarrow$  summary slot whose source bucket now contains  $K_*$  ▷ uniform temporal grouping
6:      $\mathcal{S}[s^*] \leftarrow$  canonical mean of  $s^*$ ’s source frames ▷ Def. 2
7:   else if  $\mu = \text{WORLDTRACE-LANDMARK}$  then
8:      $\text{SE} \leftarrow \bigvee_{f=0}^{F-1} [\text{cosdist}(K_{K_*,f}^{(k)}, K_{\text{prev},f}^{(k)}) > \tau]$  ▷ per-frame check inside popped block
9:     if SE then
10:       $\mathcal{S} \leftarrow \text{SHIFTLEFT}(\mathcal{S})$  ▷ drop  $\mathcal{S}[0]$ 
11:       $\mathcal{S}[N_s-1] \leftarrow K_{K_*}^{(k)}$  ▷ canonical landmark; c.f. Eq. (4)
12:     end if
13:     (Init.) If fewer than  $N_s$  landmarks are stored, fill empty slots with the oldest.
14:   end if
15: end if
16: Attention time: for  $s=0, \dots, N_s-1$ , apply  $R(\theta_k t_s^v)$  to  $\mathcal{S}[s]$  with  $t_s^v$  from Def. 1.

```

G.2 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), 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 $L_{\text{attn}} \times 880$ tokens (one token row per latent frame), 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.

G.3 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 $N_r=4$ recent-window slots ($N_s + N_r = L_{\text{attn}}=6$). Long-horizon PAC, LOOPMEM, and ablation experiments (Secs. 4.3–4.4) use $N_s=4$, $N_r=2$ (same total capacity) for both WORLDTRACE-FIELD and WORLDTRACE-LANDMARK, as described in the Setup section. 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. 5 reports canonical key averaging with Block-relative virtual positions [Yesiltepe et al., 2026], where each summary slot is assigned position $\max(0, q - \Delta t_{\text{train}})$, for a fair comparison against the Naive+Block-relative baseline. Tab. 2 reports the long-horizon conditions (Sliding Window, Block-relative, Centroid-linear, fullcombo) at $N=16$. The recent-window slot j keeps its absolute position across schemes.

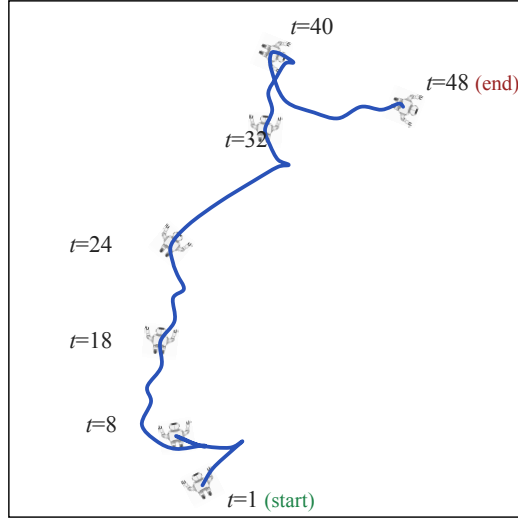


Figure 12: **Camera trajectory used for Fig. 4.** Top-down plan of the camera (x, z) path executed by the AR rollout for the WORLDTRACE-FIELD qualitative panel at $N=48$ (~ 36 s). Green: start ($n=1$); dark red: end ($n=48$). Orange ticks mark the chunk indices shown as columns in Fig. 4. The path leaves the initial scene, traverses novel territory, and lands at a different pose, so every column $n>1$ is out-of-window for the sliding-window baseline.

Evaluation protocol. Each method is evaluated on $n=100$ videos generated from distinct initial frames (n as in Tab. 6). 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 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 (~ 36 s of decoded video). Fig. 12 shows that path as a top-down plan: the camera leaves an initial scene (start, $n=1$), explores a roughly counter-clockwise loop through novel territory across $n \in \{8, 18, 24, 32, 40\}$, and arrives at a distinct end pose ($n=48$). 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 ~ 2.5 s per chunk. The full ablation set required approximately 100 GPU-hours on single A100 80 GB GPUs. At ~ 250 W typical A100 draw and ~ 0.4 kg $\text{CO}_2\text{-eq/kWh}$, this is approximately 10 kg $\text{CO}_2\text{-eq}$.

Memory and compute. Cache storage is L_{attn} latent frames of KVs (~ 2.92 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 (~ 112 s ≈ 1.9 min of decoded video; c.f. Tab. 16). 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. 17). 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. 17 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, 352×640 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 $N=102$ (~ 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 17: **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%

G.4 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. 18).

Table 18: **Metric definitions.** $\uparrow$: higher is better; $\downarrow$: lower is better.

Metric	Role	Dir.	Definition
TempSSIM	Coherence	$\uparrow$	SSIM [Wang et al., 2004] between consecutive decoded frames, averaged over the rollout.
Local Scene Drift (SceneDrift)	Coherence	$\downarrow$	Mean per-chunk CLIP feature distance to the preceding chunk.
PAC	Recall	$\uparrow$	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	$\downarrow$	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.

H Discussion

H.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-World’s Plücker, App. G) supply ego-motion as side information.

Metrics and headline numbers. Episodic recall is measured with paired CLIP cosine similarity on 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. G.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.

H.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/rotate 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 , N_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. E.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 (H₂O [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. E.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 and content factorization studied here. On the data side, motion-attribution methods [Wu et al., 2026b] 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 local attention window L_{attn} (Sec. 3.1) precisely because it is training-free: the constraint $N_s + N_r = L_{\text{attn}}$ 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.

Training memory policies from revisit supervision. Closed-loop trajectories in controllable simulators make it possible to collect return-to-origin sequences with direct return-leg supervision and fine-tune via SFT on episodes where addressable memory succeeds. SFT teaches *what* correct recall looks like on demonstrated paths, but long-horizon memory is ultimately a search-and-retrieval problem: next-chunk imitation does not by itself teach how to navigate attention over a compressed cache under novel horizons or layouts. A complementary RL stage, with terminal rewards on revisit frames (e.g. PAC at scene A) while WORLDTRACE keeps summaries in-distribution, could optimize a policy over what to store, compress, evict, and retrieve. Revisit cycles mark when preserved content must be queried, giving structured credit assignment across the long gap between encoding at departure and recall on return. In analogy to RLVR for multi-step reasoning, RL here would target the memory *policy* operating on condensed cache state rather than the generative backbone itself.

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-World [Team et al., 2026] would test whether the position/content split is architecture-specific or generic.

Scaling LOOPMEM. LOOPMEM (App. F) 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. 3) suggests learned or rule-based scene-entry policies as a concrete next step.

H.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.