Addressable Memory for Closed-Loop Video World Models

Xindi Wu[1,2] Sven Elflein[1,3,4] James Lucas[1] Olga Russakovsky[2]
Laura Leal-Taixé[1] Despoina Paschalidou[1] Jonathan Lorraine[1*] Aljoša Ošep[1*]
[1]NVIDIA [2]Princeton University [3]University of Toronto [4]Vector Institute

Abstract

An embodied agent inside a video world model closes a perception–action loop: it observes, plans, acts, and revisits. Planning over such loops requires the world model to reproduce the scene the agent left behind, not a plausible-looking alternative. Today's autoregressive video models fail this test once a rollout exceeds the training context, not because the cache forgets, but because cached-key positions go out of distribution under temporal RoPE; on the evaluated architecture, any compressed summary at out-of-distribution positions is byte-identical to a sliding window. We propose WORLDTRACE*, a training-free framework that restores closed-loop addressability by anchoring every summary slot to a fixed in-distribution position by slot rank, paired with two writers:* WORLDTRACE-FIELD *(canonical key averaging) for trajectory coherence, and* WORLDTRACE-LANDMARK *(frozen canonical landmarks committed at* scene-entry events*) for revisit recall–the passive analog of active sensing, in which the cache chooses what is worth remembering. We introduce* ***LoopMem****, a scripted-navigation benchmark with return-to-origin, multi-revisit, and orientation-closure tiers. On Matrix-Game-2,* WORLDTRACE *raises trajectory coherence by* $+15.5\%$ *at* $24\times$ *the training horizon and sustains near-perfect scene reconstruction at* $256\times$*, at* $O(1)$ *peak memory and* $< 10\%$ *wall-clock overhead–no fine-tuning. Cross-architecture replication on a pose-conditioned generator confirms the bottleneck is generic.*

1 Introduction

Closed-loop embodied agents acting inside a learned world model must *observe, plan, act, and revisit*. Autoregressive video world models [7, 8, 24, 35, 36, 38] are increasingly positioned as the visual substrate for this loop: differentiable simulators, interactive engines, and rollout for robot policies. All of these uses share a requirement ordinary text-to-video generation does not impose: when an agent returns to a previously visited location, the model must reproduce *the same world*, not a plausible-looking alternative. We call this property *closed-loop visual persistence*, and find it collapses sharply once a rollout exceeds the training context.

Standard fixes compress the linearly growing Key–Value (KV) cache into a fixed-size memory of recent context plus older summary slots [11, 53, 62, 125, 134], presuming better summarization retains more relevant past content. We show summarization cannot help: at long horizons, the temporal positional encodings [93] of cached keys go out of distribution and attention silently fails to retrieve their content. *Long-horizon failure is an addressability problem first, and a summarization problem second.* On the evaluated architecture, under Matrix-Game-2's local-attention mask [38], any compressed summary at out-of-distribution positions yields outputs byte-identical to a sliding window over recent KV entries–regardless of what was stored.

We propose WORLDTRACE, a training-free framework that restores closed-loop addressability by anchoring each summary slot to a fixed in-distribution position by slot rank. With addressability fixed, retention becomes a clean design choice. We pair the recent verbatim window with two writers: WORLDTRACE-FIELD aggregates older context in canonical (rotation-invariant) key space for trajectory coherence, and WORLDTRACE-LANDMARK commits frozen canonical keys at *scene-entry events* detected from the cached representations themselves–an event-triggered, passive analog of active perception, in which the cache decides which past observations are worth keeping for a possible future revisit. Both are drop-in to a frozen generator. On controlled return-to-origin tasks, a sliding-window cache rapidly loses structural fidelity on the return leg, while WORLDTRACE recovers the original scene (Fig. 1), extending the closed-loop horizon from seconds to minutes. We exercise it on **LoopMem**, a scripted-navigation benchmark that scores the regenerated return frame against initial-scene appearance at geometrically matched positions, organized along four difficulty axes (waypoint count, edge length, orientation closure, revisit depth) chosen to interrogate distinct closed-loop failure modes.

Contributions. (i) We diagnose *positional addressability*, not summary content, as the binding constraint on

*Co-senior authors

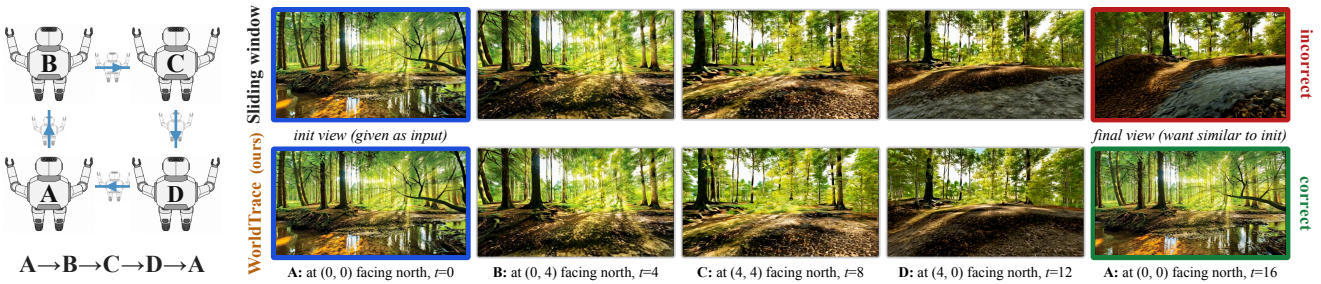


Figure 1. **Addressable memory closes the perception–action loop.** *Left:* A scripted closed-loop trajectory A→B→C→D→A. *Right:* Representative frames generated by Matrix-Game-2 along the path; the final column is the return to scene A. The sliding-window cache (top) drifts and no longer matches the original scene (red border), so an agent that revisited A on this trajectory would see a different world. WORLDTRACE with frozen landmark keys (*ours*, green border) reproduces scene A on return, restoring closed-loop visual persistence.

closed-loop visual persistence in autoregressive video world models, and show that on the evaluated architecture any compressed summary at out-of-distribution positions reduces byte-identically to a sliding window over recent context. (ii) We propose WORLDTRACE, a training-free KV-cache framework that fixes addressability by slot-rank positions, then exposes a two-knob design space for retention: WORLDTRACE-FIELD for trajectory coherence and WORLDTRACE-LANDMARK for event-triggered episodic recall. (iii) We introduce **LoopMem**, a scripted return-to-origin benchmark for closed-loop visual persistence, and show WORLDTRACE turns minutes-long rollout from forgetting into recall: $+15.5\%$ TempSSIM at $24\times$ training horizon, near-perfect PAC at $256\times$, with $<10\%$ wall-clock overhead and no fine-tuning.

2 WORLDTRACE

WORLDTRACE pairs (i) a slot-rank position scheme that keeps summary entries addressable at any rollout horizon (Sec. 2.2) with (ii) a canonical-key writer: WORLDTRACE-FIELD (Sec. 2.3) for trajectory coherence and WORLDTRACE-LANDMARK (Sec. 2.4) for revisit-time recall. We assume temporal RoPE on keys and a fixed AR KV cache; single-shot diffusion and content-agnostic encodings are out of scope (App. F).

2.1 Cache Structure

We split the cache into a *recent window* $\mathcal{R}$ of W_r verbatim AR blocks and a *summary cache* $\mathcal{S}$ of N_s compressed slots, with $N_s + W_r = L_{\text{train}}$ (F latent frames per block). Total storage matches a sliding window. The N_s/W_r split trades coherence for recall: more summary slots favour an agent expected to revisit; more recent slots favour smooth local trajectories (App. F).

2.2 Slot-Rank Virtual Positions

During training, attention is restricted to a local window $|q - t| \le \Delta t_{\text{train}}$, so cached tokens at offsets exceeding Δt_{train} enter RoPE phase regimes the model never inverted at training time. Block-relativistic RoPE [125] caps offsets but collapses every summary slot beyond the cap to the same virtual position (Rem. 1); MemRoPE [53] sidesteps this only for two slots. To keep N_s slots distinguishable and in-distribution at any horizon–so an agent's earlier observations remain individually addressable–we assign positions by slot rank.

Definition 1 (WORLDTRACE Slot Indexing)**.** The virtual position of slot s (oldest first) is

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

Slot offsets depend only on slot rank, not absolute horizon N, so positions stay in-distribution as N grows.

Remark 1 (Block-Rel saturation)**.** *Block-Rel caps look-back at $(L_{\text{train}}{-}1)F$, so any slot whose centroid exceeds the cap is clamped to the same position; under uniform averaging all centroids drift past the cap as N grows, becoming indistinguishable. Eq. (1) avoids this: per-slot offsets stay distinct and N-independent.*

With slot positions fixed in-distribution, the remaining choice is what each slot *stores*.

2.3 WORLDTRACE-FIELD: Canonical Key Averaging for Coherence

The most literal compressor for keys is averaging [6, 83]. Naively averaging RoPE-rotated keys $\bar{K}^{(k)}_{\text{naive}} = \frac{1}{M}\sum_m R(\theta_k t_m)\, K^{(k)}_{\text{cx},m}$ sums vectors at different RoPE phases; once source frames span a RoPE period, components cancel and the summary erases its own signal. WORLDTRACE-FIELD averages in canonical (un-rotated) key space and re-rotates to the slot's virtual position.

Definition 2 (WORLDTRACE-FIELD operator)**.** The compressed key at t_v for RoPE pair k is

$$K^{(k)}_{\text{field}}(t_v) \;:=\; R(\theta_k t_v)\,\frac{1}{M}\sum_{m=1}^{M} R(-\theta_k t_m)\; K^{(k)}_{t_m}. \qquad (2)$$

Each source key is unrotated to canonical content $K^{(k)}_{\text{cx},m}$, averaged, and re-encoded at t_v. Values carry no RoPE; slot values are the per-frame mean of source values. Sources

are split into N_s contiguous temporal groups, so M has no separate hyperparameter and grows linearly with N. The result is a coherence-preserving summary that anchors local trajectory continuity without committing to any particular past observation.

Remark 2 (Mean-attention preservation, informal)**.** *The compressed key in Def. 2 preserves the mean attention logit that the source keys would receive after being reassigned to the shared virtual position t_v (logits only; softmax nonlinearity precludes the same statement for weights). Formal statement and proof in App. C (Prop. 1).*

2.4 WORLDTRACE-LANDMARK: Event-Triggered Active Retention

WORLDTRACE-FIELD cannot recover specific scenes once they have been averaged away. WORLDTRACE-LANDMARK instead stores verbatim frames at detected *scene-entry events* as frozen canonical keys at slot-rank positions (in the spirit of Landmark Attention [76]). We mark a scene-entry when the cosine distance between consecutive canonical-K signatures (Eq. (2)) spikes above τ; the most recent fill the N_s slots (FIFO). This is the passive analog of active perception: the agent does not choose which views to acquire, but the cache chooses which views are worth keeping in case the agent returns. Pose- or task-conditioned event detectors are a natural extension (App. G.1); the threshold-on-canonical-K rule already suffices to demonstrate the regime.

Frozen canonical keys. Slots shift toward the older end as new chunks arrive; under standard summary updates each shift unrotates and re-rotates a stored key, accumulating bfloat16 phase error [101]. WORLDTRACE-LANDMARK stores each landmark's canonical key once and applies a single fresh rotation to the current virtual position at every shift:

$$K^{(k)}_{\text{land}}\left(t_v^{(s)}\right) \;:=\; R(\theta_k t_v^{(s)})\; R(-\theta_k t_{\ell^*})\; K^{(k)}_{t_{\ell^*}}, \qquad (3)$$

where t_{ℓ^*} is the landmark timestamp and $K^{(k)}_{t_{\ell^*}}$ its source key; the form matches Eq. (2) with $M{=}1$, with canonical key reused across shifts.

2.5 Position–Content Coupling

Position and content are not independent: writers accumulate statistics at write-time offsets but are queried at read-time offsets, and when these disagree the stored statistic no longer matches the query. WORLDTRACE handles both: Def. 1 fixes positions and Def. 2 writes in canonical space, which absorbs the source-to-virtual shift; slot content is then orthogonal. Storage stays at L_{train} blocks ($O(1)$ in N, identical to a sliding window) with $<10\%$ per-chunk overhead up to $N{=}102$ (App. F, Tab. 15; full procedure in Alg. 1).

3 Experiments

We organize our experiments around three claims tied to closed-loop visual persistence. **(Q1, Sec. 3.1)** Is position, not content, the binding constraint? **(Q2, Sec. 3.2)** Does WORLDTRACE-FIELD improve *trajectory coherence* over compression baselines? **(Q3, Sec. 3.3)** Does WORLDTRACE-LANDMARK sustain *revisit recall* on **LoopMem** at extended horizons?

Setup. We evaluate on Matrix-Game-2 (MG2-1.3B) [38], a distilled 1.3B-parameter autoregressive video world model based on Wan 1.3B T2V [100] with 3D-RoPE, training-time KV cache extent $L_{\text{train}}{=}6$ AR blocks ($F{=}3$ latent frames each), and local attention window of two blocks ($\Delta t_{\text{train}}{+}1{=}6$ latent frames). We use $N_s{=}2$, $W_r{=}4$ for coherence and $N_s{=}4$, $W_r{=}2$ for revisit recall. Baselines: the MG2 default sliding window; Naive+Block-Rel (averaging in rotated key space without RoPE disentanglement); canonical averaging + Block-Rel (canonical-domain compression with Block-Rel positions, isolating position contribution); Landmark+Block-Rel (verbatim-recall ablation); concurrent training-free MemRoPE [53] and YaRN [81]. Metrics: TempSSIM [102] and Local Scene Drift (CLIP feature distance to preceding chunk) for trajectory coherence; Position-Aligned CLIP (PAC, ViT-H/14 cosine between geometrically paired return- and forward-leg frames) for revisit recall. Full definitions in App. F.5; cross-architecture replication on a pose-conditioned generator (LingBot-Fast [86]) in App. D.

LoopMem: A Closed-Loop Recall Benchmark. We introduce **LoopMem**, a scripted-navigation benchmark for closed-loop visual persistence in autoregressive world models. The model executes a scripted path that returns to a previously visited location; the regenerated return frame is scored against the original scene appearance at geometrically matched positions, requiring no external reference. LoopMem organizes difficulty along four axes that interrogate distinct closed-loop failure modes (App. E): waypoint count K (how many intermediate scenes between departure and return), edge length L (how long the agent stays away before returning), camera-orientation closure (whether the agent returns facing the original direction), and multi-revisit depth R (whether earlier revisits stabilize the cache for later ones). Each axis isolates a property a closed-loop agent would care about.

3.1 Position is the Binding Constraint

Holding the content writer fixed at canonical key averaging (Eq. (2)) and varying only the position scheme, Block-Rel offsets beyond the training window receive zero softmax weight under MG2's local-attention mask, collapsing canonical averaging to byte-identical sliding-window eviction; Centroid linear avoids saturation but uses an N-dependent

Table 1. TempSSIM (↑): canonical averaging fixed; only positions vary.

Position	$N{=}8$	$N{=}16$
Block-Rel	0.390	0.530
Centroid	0.377	0.479
WORLDTRACE	**0.413**	**0.545**

formula that becomes unstable past the training window. Only WORLDTRACE assigns positions by slot rank alone (Eq. (1)), keeping every summary slot in-distribution at every horizon (Tab. 1; +5.9% over Block-Rel at N=8, +2.8% at N=16). The practical consequence for closed-loop agents: until positions are fixed, no smarter content writer can pay off.

3.2 Trajectory Coherence: WORLDTRACE-FIELD

WORLDTRACE-FIELD combines WORLDTRACE slot-rank positions with canonical key averaging. Against the sliding-window baseline at N=48 (24× training horizon, ∼36 s of decoded video), WORLDTRACE-FIELD improves TempSSIM from 0.472 to 0.545 (+15.5% relative) and lowers Local Scene Drift from 0.0305 to 0.0295–the only method to lead on both (Tab. 6, App. D.1). Canonical-K averaging variants with N-dependent positions cluster non-monotonically as horizon grows (0.479 to 0.449 at N=48), confirming that stacking content-domain heuristics cannot recover what an unstable position scheme loses; qualitative comparison in Fig. 2.

3.3 Revisit Recall: WORLDTRACE-LANDMARK on LoopMem

Tab. 2 (N_s=4, W_r=2). WORLDTRACE-LANDMARK improves PAC across all four tiers: +17–20% on **waypoint count** (ABA/ABCA/ABCDA), with the gap widening on **edge length** from +7.3% at N=8 to +31.6% at N=32 (the longer the agent is away from a scene, the more addressable memory matters). **Orientation closure** recovers +3.9%/+16.4% at 90°/180° (smaller 360° from approximate yaw scaling), and **multi-revisit depth** sustains the lift (ABABA: 0.941 vs. 0.892). At N=256 (∼ 192 s), WORLDTRACE-LANDMARK sustains PAC=0.989 while the best baseline drops below 0.65 (App. D).

Concurrent baselines and cross-architecture replication. MemRoPE [53] (Block-Rel + dual-rate EMA) survives OOD positions via decay but is beaten by Landmark+Block-Rel, and transplanting WORLDTRACE positions into MemRoPE degrades it (p<0.001, App. D.2), confirming the position–content coupling of Sec. 2.5. YaRN [81] uses $O(N)$ memory and OOMs by N≈100; WORLDTRACE-LANDMARK exceeds it (0.964 vs. 0.490 at N=32) in $O(1)$ memory and stays near-constant through N=512. On the pose-conditioned LingBot-Fast [86], WORLDTRACE-LANDMARK improves PAC from 4× training horizon onward (App. D.3), so the addressability bottleneck is architecture-generic, not MG2-specific.

Discussion. The addressability framing factorizes long-horizon memory into a horizon-stable position scheme and a content writer; position-only fixes (Block-Rel, YaRN) and content-only fixes (EMA, naive averaging) each plateau because they absorb only one failure mode. With both axes addressed, the N_s/W_r split exposes a clean trade-off: more summary slots favour revisit-dominant agents; more recent slots favour smooth local trajectories. The framework requires temporal RoPE on keys and a known training context L_{train}. Pose-, action-, or task-conditioned scene-entry detectors that promote the current passive event-triggering into actively-selected retention are a natural next step (App. G.1).

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

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

Ablation. Three axes isolate the design choices at fixed cache budget N_s+W_r=6 (Tab. 7, App. D): position assignment is the binding constraint (canonical averaging with Block-Rel collapses to sliding-window byte-for-byte); canonical vs. naive averaging cuts LatentDiff by −25.3%; and frozen landmark keys prevent rotation-drift accumulation, with the gap widening across summary updates (+0.040 at N=64 to +0.046 at N=128).

4 Conclusion

Closed-loop visual persistence in autoregressive video world models is bottlenecked by *positional addressability*, not summary content: once RoPE offsets exceed the training window, attention fails to read cached observations regardless of how they were compressed, breaking any planner that relies on revisits. WORLDTRACE fixes addressability by slot-rank positions, paired with two writers: WORLDTRACE-FIELD (canonical key averaging) for trajectory coherence, and WORLDTRACE-LANDMARK (frozen canonical landmarks committed at scene-entry events) for revisit recall. Both are training-free, drop-in, and $O(1)$ in horizon, turning minutes-long generation from forgetting into recall and bringing video world models a step closer to viable closed-loop substrates for embodied agents. Pose- or task-conditioned scene-entry detectors that promote the current passive event-triggering to actively-selected retention are a natural next step. Limitations and broader impact: App. G.1, App. G.3.

References

[1] Dmitry Akulov, Mohamed Sana, Antonio De Domenico, Tareq Si Salem, Nicola Piovesan, and Fadhel Ayed. Kv-compose: Efficient structured kv cache compression with composite tokens. *arXiv preprint arXiv:2509.05165*, 2025. 12

[2] 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. 14

[3] 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. 14, 20

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 12

[5] Jesse Bettencourt, Xindi Wu, Matan Atzmon, James Lucas, and Jonathan Lorraine. Variance reduction for expectations with diffusion teachers. *arXiv preprint arXiv:2605.21489*, 2026. 27

[6] 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. 2, 12

[7] 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. 1, 14

[8] 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. 1, 14

[9] Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 11

[10] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.21058, OpenReview: https://openreview.net/forum?id=y6XJZlEC2x. 15

[11] 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, 12

[12] 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. 14, 20

[13] Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past- and future-informed kv cache policy with salience estimation in autoregressive video diffusion. *arXiv preprint arXiv:2601.21896*, 2026. 15

[14] 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*, 2026. 15

[15] Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026. 15

[16] 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. 14

[17] Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026. 15

[18] Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025. Project page https://sites.google.com/view/vrag. 15

[19] 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)*, 2025. arXiv:2410.21216. 14

[20] 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. 13

[21] 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*, 2026. 15

[22] 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)*, 2026. arXiv:2510.02283. 14

[23] 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. 11

[24] 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, 14

[25] 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. 14

[26] 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. 16, 27

[27] 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. 16

[28] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024. 14

[29] 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. 14

[30] 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. 14, 27

[31] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. 11

[32] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 14, 15, 16

[33] 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. 14

[34] 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. 15

[35] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 14

[36] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. 1, 14

[37] Junhui He, Junna Xing, Nan Wang, Rui Xu, Shangyu Wu, Peng Zhou, Qiang Liu, Chun Jason Xue, and Qingan Li. A^2ATS: Retrieval-based KV cache reduction via windowed rotary position embedding and query-aware vector quantization. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025. arXiv:2502.12665. 13

[38] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. 1, 3, 14

[39] 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. 15

[40] 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. 14

[41] 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. 15

[42] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.06654. 13

[43] 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. 14

[44] 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*, 2025. 15

[45] 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*, 2026. 14

[46] 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)*, 2026. arXiv:2505.14357. 14

[47] 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)*, 2025. Spotlight; arXiv:2506.08009. 14, 20

[48] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 25

[49] 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. 11

[50] 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. 15

[51] 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. 11

[52] 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. 14

[53] 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, 2, 3, 4, 15, 16, 18, 27

[54] 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*, 2025. arXiv:2502.20694. 16

[55] 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*, 2026. 15

[56] 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*, 2025. 14

[57] 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*, 2026. 15

[58] 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*, 2026. 15

[59] 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)*, 2025. arXiv:2506.18903. 15

[60] 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*, 2026. 15

[61] 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*, 2025. 15

[62] 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, 12, 26, 27

[63] 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. 16

[64] 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. 15

[65] 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. 13

[66] Jonathan Lorraine. *Scalable nested optimization for deep learning*. PhD thesis, University of Toronto, 2024. Ph.D. thesis; arXiv:2407.01526. 27

[67] Jonathan Lorraine and Safwan Hossain. JacNet: Learning functions with structured Jacobians. In *ICML Workshop on Invertible Neural Networks and Normalizing Flows (INNF)*, 2019. 17

[68] Jonathan Lorraine, Nihesh Anderson, Chansoo Lee, Quentin De Laroussilhe, and Mehadi Hassen. Task selection for AutoML system evaluation. *arXiv preprint arXiv:2208.12754*, 2022. 20

[69] 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)*, 2022. 20

[70] 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. 27

[71] Yuexiao Ma, Xuzhe Zheng, Jing Xu, Xiwei Xu, Feng Ling, Xiawu Zheng, Huafeng Kuang, Huixia Li, Xing Wang, Xuefeng Xiao, Fei Chao, and Rongrong Ji. Flow caching for autoregressive video generation. *arXiv preprint arXiv:2602.10825*, 2026. 15

[72] 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*, 2026. 13

[73] 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*, 2026. 15

[74] Nikhil Mehta, Jonathan Lorraine, Steve Masson, Ramanathan Arunachalam, Zaid Pervaiz Bhat, James Lucas, and Arun George Zachariah. Improving hyperparameter optimization with checkpointed model weights. In *ECCV Workshop on Efficient Deep Learning Foundation Models (EFM)*, 2024. 27

[75] 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. 14

[76] Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16300. 3, 11

[77] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024. 11

[78] NVIDIA. KVPress: A compression library for transformer KV caches. https://github.com/NVIDIA/kvpress, 2024. 12, 13, 27

[79] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 14

[80] 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. 15

[81] 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, 4, 14, 18, 20

[82] 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. 14

[83] 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. 2, 11

[84] Suraj Ranganath, Vaishak Menon, and Anish Patnaik. KV cache quantization for self-forcing video generation: A 33-method empirical study. *arXiv preprint arXiv:2603.27469*, 2026. 13

[85] 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. 27

[86] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. 3, 4, 18, 19, 27

[87] 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. 15

[88] Sand AI. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 14

[89] 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. 14

[90] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.06764. 14

[91] 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)*, 2025. arXiv:2410.23274. 20, 26

[92] Sebastian Stapf, Pablo Acuaviva, Aram Davtyan, and Paolo Favaro. Composition of memory experts for diffusion world models. In *International Conference on Learning Representations (ICLR)*, 2026. OpenReview: https://openreview.net/forum?id=sUEdpZCHdp. 15

[93] 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, 14

[94] 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*, 2025. 15

[95] 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. 14

[96] 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)*, 2025. 11

[97] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context LLM inference. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2406.10774. 12

[98] 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. 12

[99] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025. 14

[100] Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 14, 20

[101] Haonan Wang, Qian Liu, Chao Du, Tongyao Zhu, Cunxiao Du, Kenji Kawaguchi, and Tianyu Pang. When preci-

sion meets position: BFloat16 breaks down RoPE in long-context training. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2411.13476. 3, 14

[102] 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. 3, 24, 26

[103] 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. 27

[104] 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*, 2026. 15

[105] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026. 14, 27

[106] 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. 14

[107] Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026. 15

[108] Shunlong Wu, Hai Lin, Shaoshen Chen, Tingwei Lu, Yongqin Zeng, Shaoxiong Zhan, Hai-Tao Zheng, and Hong-Gee Kim. Semanticache: Efficient kv cache compression via semantic chunking and clustered merging. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19562–19566. IEEE, 2026. 12

[109] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.05284, project page https://spmem.github.io. 15

[110] 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)*, 2026. arXiv:2601.08828. 27

[111] Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 11

[112] Haocheng Xi, Shuo Yang, Yilong Zhao, Muyang Li, Han Cai, Xingyang Li, Yujun Lin, Zhuoyang Zhang, Jintao Zhang, Xiuyu Li, Zhiying Xu, Jun Wu, Chenfeng Xu, Ion Stoica, Song Han, and Kurt Keutzer. Quant VideoGen: Auto-regressive long video generation via 2-bit KV-cache quantization. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.02958. 13

[113] 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. 14

[114] Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.04617. 13

[115] 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. 12

[116] 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)*, 2025. arXiv:2410.10819. 13

[117] 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)*, 2025. 14

[118] 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. 27

[119] 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*, 2026. 15

[120] 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*, 2026. 14, 15, 27

[121] Yan Team. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025. 14

[122] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.22622. 14

[123] 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. 15

[124] 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. 16, 27

[125] 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. 1, 2, 15, 16, 24

[126] 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. 15, 16

[127] 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. 14, 20

[128] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025. 15

[129] 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. 14, 15, 27

[130] 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*, 2025. 15

[131] 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. 14

[132] 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. 26

[133] Xuan Zhang, Cunxiao Du, Chao Du, Tianyu Pang, Wei Gao, and Min Lin. SimLayerKV: A simple framework for layer-level KV cache reduction. *arXiv preprint arXiv:2410.13846v1*, 2024. 12

[134] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 12, 26, 27

[135] 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. 14

[136] 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. 15

[137] 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. 16, 27

[138] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI, 2025. 25

[139] 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*, 2025. 13

Appendices

A Notation

Symbols used throughout the paper. Tab. 3 lists model and cache notation, indices, and virtual positions; Tab. 4 lists RoPE and key representations. Evaluation metrics: Tab. 16 (App. F.5); LoopMem parameters: App. E.

A.1 Core Notation

A.2 RoPE and Key Representations

B Related Work

B.1 Memory-Augmented Transformers

WorldTrace inherits the two-tier (recent verbatim window plus compressed older history) pattern from a lineage of memory-augmented transformers. Transformer-XL [23] pairs segment-level recurrence with a relative position encoding so recurrence does not corrupt absolute indices; Compressive Transformers [83] add a compressed memory of older tokens on top, and the two-tier structure we adopt. Memorizing Transformers [111] attach an external kNN-retrieved memory; Recurrent Memory Transformer [9] and Block-Recurrent Transformers [49] pass learned memory tokens between segments; Infini-attention [77] fuses a sliding window with a compressive long-term memory in one block. Landmark Attention [76] 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–rerotate drift in 3D RoPE. The broader full-attention-to-compressed-state spectrum spans linear attention [51], Mamba [31], and Test-Time Training [96]; we target the fixed-budget, training-free regime in AR video diffusion.

Table 3. **Core notation.** Methods, cache structure, indices, virtual positions, scene-boundary threshold, conventions.

Symbol	Description
Method Names	
WORLDTRACE	Training-free KV-cache framework (Sec. 2); assigns each summary slot a fixed slot-rank virtual position
WORLDTRACE-FIELD	WORLDTRACE variant that compresses history by canonical key averaging (coherence; Sec. 2.3, Def. 2)
WORLDTRACE-LANDMARK	WORLDTRACE variant keeping verbatim scene-entry frames with frozen canonical keys (recall; Sec. 2.4, Eq. (3))
Cache Structure	
N	Total number of AR chunks in a rollout (generation length)
N_s	Number of summary slots in the KV cache
W_r	Number of recent-window slots (verbatim, newest W_r chunks)
L_{train}	Training context length in AR blocks; $N_s + W_r = L_{\text{train}}$
F	Latent frames per AR block (chunk size)
M	Source frames compressed into one summary slot under WORLDTRACE-FIELD
T_{old}	Frames outside the recent window: $T_{\text{old}} = (N - W_r)F$ for $N > L_{\text{train}}$ (linear in N)
s	Summary slot index, $s = 0, \ldots, N_s - 1$ (0 = oldest)
$\mathcal{R}$	Recent-window cache: ordered set of verbatim KVs for the newest W_r blocks
$\mathcal{S}$	Summary cache: N_s compressed slots indexed by s
K_*	Block just popped from $\mathcal{R}$ when a new chunk is appended; $K^{(k)}_{\text{cx},K_*,f}$ denotes its canonical key at intra-block frame f
SE	Scene-entry indicator: SE $= 1$ if cosine-distance spike on canonical-K exceeds τ
J	Number of detected scene-entry landmarks at the current horizon ($J \leq N_s$)
Indices (dummy / iteration)	
k	RoPE frequency-pair index
m	Source-frame iteration index (inside $\sum_{m=1}^{M}$)
n	AR-chunk iteration index
f	Intra-block frame index, $f \in \{0, \ldots, F-1\}$ (per-frame slot of an AR block; Alg. 1)
μ	Algorithm mode selector ($\mu \in \{$WORLDTRACE-FIELD, WORLDTRACE-LANDMARK$\}$, Alg. 1)
Virtual Position Assignment (Slot Indexing, Def. 1)	
q	Absolute timestamp of the current query frame
t	Absolute timestamp of a cached key frame
$t^v_{\min}$	In-distribution lower bound: $\max(0,\ q - (L_{\text{train}} - 1)F)$
$t^v_{\max}$	In-distribution upper bound for summary slots: $q - W_r F$
t_v	Generic virtual position (no slot index)
$t^{(s)}_v$	Virtual position of slot s: $q - (L_{\text{train}} - 1 - s)F$
Scene-Boundary Detection	
τ	Scene-boundary detection threshold (τ=0.15; used by WORLDTRACE-LANDMARK)
Experimental Conventions	
n	Number of distinct initial scenes (test instances) per evaluation condition
seed	Global random seed controlling initial scene selection and denoising noise
$N_s + W_r = L_{\text{train}}$	Capacity constraint: total cache always fills the training window

B.2 KV Cache Compression for Large Language Models

Window and eviction methods. Window methods [4] retain the most recent w tokens; StreamingLLM [115] augments windows with initial "sink" tokens. Token eviction methods select high-importance tokens via accumulated attention mass (H_2O [134]), key L2-norm (KnormPress [78]), pooled attention (SnapKV [62]), per-layer budget (PyramidKV [11]), lazy-layer drop (SimLayerKV [133]), or query-aware page criticality (Quest [97]).

Token merging methods. Merging methods combine similar tokens rather than evicting them: ToMe [6], KVCompose [1] (composite tokens), SemantiCache [108] (semantic clusters), and KeepKV [98] (lossless via attention-score adjustment).

Table 4. **RoPE and key representations.** Parameters, rotations, and key/query variants used across the equations.

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

Pre-RoPE Q and K vectors concentrate around stable directional centers in LLMs; TriAttention [72] scores each token by angular alignment plus magnitude and retains the top-scoring ones verbatim.

Quantization, architectural, and retrieval alternatives. KIVI [65] quantizes the cache (orthogonal to and composable with our position-content axis); on the video side, Quant VideoGen [112] achieves 2-bit KV quantization with semantic-aware smoothing for AR video diffusion, and Ranganath et al. [84] systematically benchmark 33 quantization variants under self-forcing rollouts; quantization is empirically decoupled from the position-OOD failure mode WORLDTRACE addresses. DuoAttention [116] partitions heads into retrieval (full-cache) and streaming (constant-size + sinks) families, parallel to our verbatim-recent vs. compressed-summary split; InfLLM [114] combines sliding-window attention with memory-unit retrieval for training-free long-context extrapolation. RULER [42] is the standard synthetic retrieval/aggregation benchmark on the LLM side; our LoopMem (App. E) extends this paradigm to closed-loop video generation.

Position handling in KV compression. Most LLM-side methods operate only on content and do not address what virtual position to assign a summary representing multiple merged tokens. EliteKV [139] identifies per-head frequency preferences for low-rank compression but compresses individual tokens, so position reassignment does not arise; A^2ATS [37] decouples positional dependency for retrieval rather than reassigning positions for summaries. The unrotate–rerotate primitive appears in FINCH [20] and the KeyRerotationPress module of KVPress [78], both applied per-retained-token after compression; our contribution is its use as an *averaging* operator over multiple canonical keys (WORLDTRACE-FIELD), coupled with the slot-rank positions of WORLDTRACE.

Why KV compression differs in our setting. The KV context spans separately denoised AR blocks rather than a single token stream, and the RoPE-OOD failure mode is benign for LLM deployments operating within the trained window (the orthogonal long-context-extension problem is in the next subsection). The angular-concentration prerequisite of TriAttention

does not hold under denoising-timestep-conditioned Q: on MG2, TriAttention collapses to a norm-based eviction that matches canonical key averaging or underperforms both it and sliding-window eviction. Layer-allocation methods (PyramidKV, SimLayerKV) reallocate budget across layers but do not reassign positions.

B.3 Position Extrapolation for RoPE

LLM-side analysis. A long line of work addresses the same mechanism we exploit: RoPE [93] is not robust to relative offsets beyond the training distribution. Position Interpolation [16] down-scales position indices; YaRN [81] refines this with NTK-by-parts scaling; LongRoPE [25] reaches 2M-token context via non-uniform per-dimension interpolation; LongRoPE2 [89] attributes residual OOD to undertrained higher RoPE dimensions, aligned with the per-frequency severity we measure in App. F.2. ALiBi [82] replaces RoPE with a fixed linear distance bias; xPos [95] adds an exponential-decay term; Kazemnejad et al. [52] (NoPE) show position-free transformers can outperform RoPE/ALiBi/APE on length generalization (clarifying that our RoPE-OOD argument concerns models *already* trained with RoPE). CoPE [28] makes positions content-conditional; Barbero et al. [3] dissect which RoPE frequencies carry position vs. semantics. HoPE [19] and FoPE [43] perform per-frequency RoPE-OOD analysis at the LLM scale via cascade-failure and Non-Uniform Discrete Fourier Transform theory, respectively, anchoring our per-frequency view (App. F.2).

Vision and video. RoPE-ViT [40] adapts RoPE to vision transformers via a 2D RoPE-Mixed split; VideoRoPE [106] introduces a 3D RoPE structure with low-frequency temporal allocation and the V-NIAH-D long-context retrieval test for video LLMs; ViewRope [113] replaces screen-space coordinates with camera-ray geometry to maintain loop-closure consistency in pose-conditioned video world models; RIFLEx [135] achieves training-free length extrapolation in video diffusion via per-frequency intrinsic-frequency reduction.

Position-content coupling. Wang et al. [101] show position couples to numerical precision (bf16 deviates RoPE from its intended relative encoding over long contexts), consistent with our position-content coupling thesis (Sec. 2.5). On the video side, UCM [120] reassigns 3D positional encodings via time-aware warping for camera-controlled world models, and MosaicMem [129] introduces "Warped RoPE" that reprojects positional encodings of memory patches into the target view. Both are virtual-position-assignment operators in our sense, but their warping is camera/geometry-driven and paired with trained content writers; ours is slot-rank-driven and training-free.

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

B.4 Autoregressive Video World Models

Lineage and training paradigms. Recent autoregressive video diffusion models [7, 8, 24, 29, 30, 33, 38, 56, 99, 121] generate interactively via AR chunk prediction. The dominant training paradigm Self-Forcing [47] (built on Diffusion Forcing [12], DFoT [90], and CausVid [127]) rolls out on the student's own KV cache, inheriting a local attention window not designed for arbitrarily long inference. Self-Forcing++ [22] extends this to multi-minute video via long-rollout self-distillation; FAR [32] reformulates AR video as next-frame prediction over an unbounded cache; Vid2World [46] causalizes bidirectional video diffusion into interactive world models; LIVE [45] stabilizes long horizons via a cycle-consistency objective. Our work is complementary, operating purely at inference time on top of an already-trained model. Open foundation backbones [79, 88, 100] provide the capacity that makes long-horizon AR rollout viable.

World-model and RL lineage. The world-model framing traces to Ha and Schmidhuber [35] through the Dreamer line [36]; transformer-based world models like IRIS [75] and DIAMOND [2] establish the autoregressive transformer as a sample-efficient world-model backbone, and open foundation platforms [79, 100] provide the capacity our inference-time intervention runs on.

B.5 Memory and KV Cache for Video World Models

Existing memory approaches. Existing systems use one of three cache approaches. *Sliding-window KV caches* (MG2 [38], LongLive [122]) bound memory at $O(1)$ but discard history and leave cached positions OOD. *Memory re-encoding* (MG3 [105]) selects past frames via field-of-view similarity and re-encodes at each AR step, sidestepping both problems at the cost of roughly doubled per-step latency; FramePack [131] compresses input frame contexts by frame-wise importance, and WorldMem [117] attaches a state-aware memory of pose-tagged frames re-injected via auxiliary attention. *Architectural memory modules*

include StreamingT2V [39], which pairs a short conditional-attention window with a long-term appearance-preservation module anchored on the first chunk.

Concurrent training-free cache work. Closest to ours, four concurrent training-free works share our diagnosis (temporal RoPE OOD as the long-horizon bottleneck) but pair it with a different content writer: Infinity-RoPE [125] uses Block-Rel offsets with KV Flush (sink + most recent block only); FAR [32] introduces FlexRoPE for temporal-decay extrapolation on an unbounded $O(T)$ cache; Deep Forcing [126] dedicates $\sim$50% of the window to "Deep Sink" tokens with re-aligned RoPE phases; MemRoPE [53] pairs Block-Rel positions with a dual-rate EMA summary cache. The first three either operate on existing tokens (verbatim or sink-aligned) or scale to unbounded caches; WORLDTRACE, instead, assigns positions to *compressed summaries* that represent multiple merged frames within a fixed budget. Deep Forcing's importance-pruning is content-side and composes with slot-rank. Tab. 8 shows that transplanting WORLDTRACE positions into MemRoPE without retuning its EMA content regime regresses performance, confirming position and content must be jointly designed. PackCache [58] is also training-free and uses a "spatially preserving position embedding" for cache removal; it applies virtual position assignment per-token rather than per-summary-slot, paired with cross-frame decay rather than canonical-key averaging or verbatim landmarks. FlowCache [71] introduces chunk-wise denoising-step caching and an importance–redundancy KV compression scheme for autoregressive video; its goal is per-step compute reduction rather than long-horizon recall, but the chunk-aware cache structure is similar to the recent–summary split we adopt. TempCache [87] merges near-duplicate cached keys across AR frames via approximate-nearest-neighbor temporal correspondence, paired with sparse self- and cross-attention; merging operates at per-token similarity level rather than reassigning summary-slot virtual positions in canonical space.

Closest concurrent training-time analogs. Two contemporaneous works share WORLDTRACE's diagnosis and propose related fixes through training-time mechanisms. Grounded Forcing [14] introduces a Dual Memory KV Cache (Local Temporal Memory plus Global Consistency Memory) coupled with Dual-Reference RoPE Injection that stores raw pre-RoPE keys and injects fixed $t{=}0$ for global anchors and relative offsets for local frames; their dual-reference indexing is the training-time analog of our slot-rank virtual positions. Anchor Forcing [123] partitions the cache into sink, junction, and local regions, each with its own RoPE reference origin capped at the pretrained limit, and learns the assignment via RoPE re-alignment distillation. WORLDTRACE differs in being inference-time-only with no distillation or retraining, in decomposing retention into orthogonal coherence (WORLDTRACE-FIELD) and verbatim recall (WORLDTRACE-LANDMARK) operators, selectable per task, and in pinning summary-slot position rather than content as the binding constraint. Composition of Memory Experts [92], the long-term spatial-memory framework of Wu et al. [109], and Mixture of Contexts [10] provide training-time decompositions: a contrastive product-of-experts over short-term, long-term episodic, and spatial memory; point-cloud spatial memory plus sparse episodic keyframes; and sparse top-k routing with mandatory text/intra-shot anchors. Context Forcing [17] addresses the same student-teacher mismatch by training a long-context student under a long-context teacher with a Slow-Fast Memory architecture that bounds attention cost while preserving 20 s+ effective context. Stable Video Infinity [61] bridges the train-test hypothesis gap via Error-Recycling Fine-Tuning that injects historical errors as supervisory prompts so the DiT learns to correct its own drift. Hybrid Forcing [60] combines a linear-attention summary state for evicted tokens with block-sparse local attention through a decoupled-distillation pipeline. All three are training-time analogs that complement WORLDTRACE's inference-time-only operation; WORLDTRACE can, in principle, compose with any of them at the cache level.

Concurrent trained video-memory architectures. A second cluster of concurrent works addresses the same long-horizon memory problem but with trained components. RELIC [41] stores compressed historical latents in the KV cache with both relative-action and absolute-camera-pose annotations (camera-aware memory); MosaicMem [129] pairs Warped RoPE (geometry-driven virtual position assignment) with Warped Latent injection; UCM [120] reassigns 3D positional encodings via time-aware warping for camera-controlled world models; PackForcing [73] partitions the cache into sink/mid-compressed/recent and applies a Continuous Temporal RoPE Adjustment to re-align position gaps left by dropped tokens (zero-shot or 5s-clip-trained). Compared to these, WORLDTRACE pursues the strict zero-fine-tune regime, depends on slot-rank rather than camera/geometry signals, and decomposes the cache update into orthogonal coherence (WORLDTRACE-FIELD) and recall (WORLDTRACE-LANDMARK) operators that can be selected per task without retraining.

Other concurrent video-cache and memory-bank work. Other concurrent training-free work spans position-side fixes (LoL [21], FLEX [57]), content-side selection (PaFu-KV [13], Rolling Sink [55], Relax Forcing [136], Rolling Forcing [64]), world-model architectures (WorldPack [80], WorldPlay [94], VideoSSM [130]), and memory-bank designs (Memory Forcing [44], VMem [59], Context as Memory [128], AnchorWeave [104], Infinite-World [107], HyDRA [15], MemFlow [50], Memorize-When-Needed [34], VRAG [18]). Sparse Forcing [119] learns native block-sparse attention plus persistent spatiotemporal anchors via a Persistent Block-Sparse Attention kernel, observing the same "implicit spatiotemporal memory" on persistent KV slots that motivates WORLDTRACE-LANDMARK's scene-entry detection but training the sparsity

Table 5. **KV cache comparison.** WORLDTRACE makes compressed memory addressable; WORLDTRACE-FIELD uses a field vs. WORLDTRACE-LANDMARK retaining selected landmarks.

Method	Fixes pos. OOD?	Bounded memory?	Preserves history?	Training-free?
Full KV	No	No ($O(N)$)	Yes (exact)	Yes
Sliding window	No	Yes ($O(1)$)	Recent only	Yes
KV Flush + Block-Rel [125] (KV-Flush variant)	Yes	Yes ($O(1)$)	No (amnesic)	Yes
Naive + Block-Rel	Partial	Yes ($O(1)$)	Corrupted (Sec. 1)	Yes
MemRoPE [53]	Partial	Yes ($O(1)$)	EMA-diluted	Yes
FAR (FlexRoPE) [32]	Yes	No ($O(T)$)	Yes (exact)	Yes
Deep Forcing [126]	Partial	Yes ($O(1)$)	Sink + importance-pruned recent	Yes
KnormEvict + WORLDTRACE	Yes	Yes ($O(1)$)	Partial (1 token/slot)	Yes
WORLDTRACE-FIELD (ours)	**Yes**	**Yes** ($O(1)$)	**Yes (compressed field)**	**Yes**
WORLDTRACE-LANDMARK (ours)	**Yes**	**Yes** ($O(1)$)	**Yes (verbatim landmarks)**	**Yes**

end-to-end. MemCam [27] pairs a context-compression module with co-visibility-based historical-frame selection for camera-controlled long video generation; the selection is geometry-driven rather than canonical-K-spike-driven, and the framework is trained. MemFlow retrieves the most relevant historical frames per chunk and activates only top-k tokens in attention; Memorize-When-Needed adds a decoupled memory branch with camera-aware gating that conditions generation on memory only when meaningful historical references exist; VRAG augments interactive video generation with explicit global-state retrieval to reduce compounding errors. The memory-bank designs store geometry-anchored or retrieval-indexed frames *outside* the standard KV cache and re-inject them via auxiliary attention; WORLDTRACE instead compresses history into a constant-size summary *within* the standard KV cache, supporting both coherence (WORLDTRACE-FIELD) and verbatim recall (WORLDTRACE-LANDMARK) under a single $O(1)$ budget without re-encoding or retrieval.

Evaluation benchmarks for world generation. WorldScore [26], MIND [124], and VBench-2.0 [137] provide unified long-horizon evaluation; WorldModelBench [54] explicitly positions video generation as world modeling. Closer to our setting, Lian et al. [63] constructs a Minecraft loop-navigation benchmark that assesses spatial consistency during revisits to previously seen locations. Our LoopMem benchmark (App. E) targets the orthogonal axis of *closed-loop* episodic recall (return-to-origin trajectories) on top of pretrained autoregressive video world models, complementing rather than replacing these broader quality benchmarks.

Differentiation summary. Tab. 5 summarizes four axes: whether positional OOD is fixed, whether memory is bounded, whether intermediate history is preserved, and whether retraining is required.

C WORLDTRACE-FIELD: Additional Properties

This section expands Eq. (2), summarised at the end of Sec. 2.3.

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

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

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

By Eq. (2), $K^{(k)}_{\text{field}}(t_v) = R(\theta_k t_v)\,\bar{K}^{(k)}_{\text{cx}}$ with $\bar{K}^{(k)}_{\text{cx}} = \frac{1}{M}\sum_m K^{(k)}_{\text{cx},m}$; linearity of the inner product in its second argument then gives $\langle Q^{(k)}_q, K^{(k)}_{\text{field}}(t_v)\rangle = \langle Q^{(k)}_q, R(\theta_k t_v)\,\bar{K}^{(k)}_{\text{cx}}\rangle = \frac{1}{M}\sum_m\langle Q^{(k)}_q, R(\theta_k t_v)\,K^{(k)}_{\text{cx},m}\rangle$. The summary token contributes the mean attention logit that the source keys would produce if their content were first moved to the shared virtual position t_v.

Concretely, $K^{(k)}_{\text{field}}(t_v)$ acts as if all M source frames were simultaneously repositioned to t_v: no individual timestamp t_m appears in the compressed representation, only the canonical content average $\bar{K}^{(k)}_{\text{cx}}$. This fully decouples the summary content from its temporal placement, enabling WORLDTRACE to freely reassign virtual positions (Def. 1) without distorting the stored signal.

Characterization. Eq. (2) is the linear compression operator that (a) maps M RoPE-encoded keys to a single token at virtual position t_v and (b) satisfies the mean attention preservation property of Prop. 1: unrotate each key to canonical space, average, and re-encode at t_v. The content of the compressed token is fully determined by property (b); any linear operator satisfying

Table 6. **WORLDTRACE-FIELD leads on** TempSSIM **at both horizons and on Local Drift.** At N=32, WORLDTRACE-FIELD has the highest TempSSIM (0.613) though centroid variants have lower Drift; at N=48, WORLDTRACE-FIELD leads on both metrics. *Centroid* replaces Block-Rel's saturating offset by linearly mapping each slot's mean source-frame timestamp into $[t^v_{\min}, t^v_{\max}]$; *norm* rescales the canonical mean to preserve per-frame norm; *geom* uses geometric group sizes; *knorm* weights frames by their canonical-key L2 norm.

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

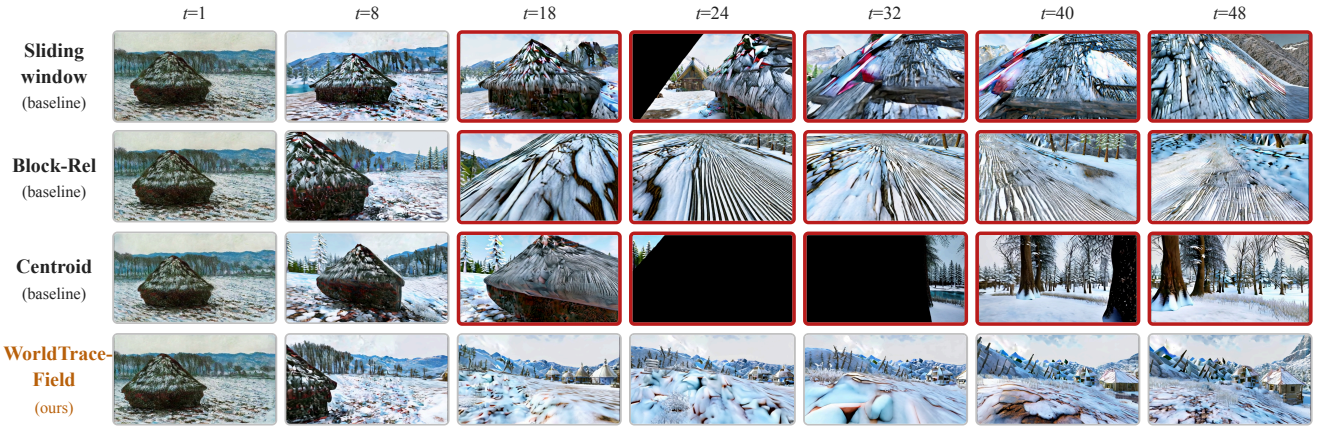


Figure 2. **WORLDTRACE-FIELD qualitative comparison at** N=48**.** Four position-encoding methods generated from the same t=1 conditioning frame: sliding window, Block-Rel, Centroid, and WORLDTRACE-FIELD (ours). Once a baseline diverges, it does not recover. Diverged frames are bordered in red. WORLDTRACE-FIELD retains memory of earlier chunks and remains coherent across the horizon. The camera path across the seven columns is plotted in Fig. 6.

it must produce the same canonical average $\bar{K}^{(k)}_{\text{cx}}$, since the property pins the output attention logit for all queries. Only t_v remains free, and WORLDTRACE chooses it (Sec. 2.2). Pinning a learned operator by a structural property of its derivative (here, the block-diagonal phase rotation of RoPE that Eq. (2) inverts) has antecedents in structured-Jacobian parameterizations [67]; WORLDTRACE-FIELD is the closed-form, training-free analog for the temporal RoPE block. Non-linear alternatives (*e.g.* selecting the highest-norm token verbatim) satisfy neither property and lose phase coherence across compressed frames. We confirm this empirically: KnormEvict+WORLDTRACE (single highest-norm canonical key per slot) achieves PAC 0.553 vs. 0.574 for WORLDTRACE-FIELD (-3.7%, p=0.074, n=10, paired at N=16), with TempSSIM also significantly lower (p=0.007), consistent with information loss from discarding all history except the single highest-norm frame.

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

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

D Additional Experimental Results

D.1 Coherence and Ablation Tables

Tab. 6 reports TempSSIM and Local Scene Drift for WORLDTRACE-FIELD and three canonical-averaging baselines at N=32 and N=48. Fig. 2 shows a qualitative four-method comparison along a fixed camera path at N=48 on Matrix-Game-2. Tab. 7 isolates the three design choices of WORLDTRACE on the ABA recall protocol.

Table 7. **Ablation.** PAC (↑) on ABA recall with $N_s{+}W_r{=}6$ slots. Top section (Axes 1 & 2): isolates position assignment and canonical vs. naive compression. Bottom section (Axes 1 & 3): isolates frozen landmark keys and position assignment in the verbatim tier. Field averaging with Block-Rel matches sliding-window eviction byte-for-byte (OOD offsets suppress summary-slot attention).

Method	PAC ($N{=}64$) ↑	PAC ($N{=}128$) ↑	PAC ($N{=}256$) ↑
Axes 1 & 2: position assignment & canonical compression (compression tier)			
Sliding window (baseline)	0.412	0.504	0.631
+ Naive averaging (Block-Rel)	0.443	0.486	0.598
+ Field averaging (Block-Rel)	0.412	0.504	0.631
+ WorldTrace **(WorldTrace-Field, ours)**	0.442	0.495	0.602
Axis 3: frozen vs. unfrozen landmark keys (verbatim recall):			
Landmark + Block-Rel *(unfrozen)*	0.936	0.940	0.965
WorldTrace-Landmark **(ours, frozen)**	**0.976**	**0.986**	**0.989**

Table 8. **Concurrent training-free baselines.** PAC (↑) at two horizons (ABA). YaRN [81] uses $O(N)$ memory; only at $N{=}32$. MemRoPE+WorldTrace swaps Block-Rel for WorldTrace; dual-rate EMA unchanged.

Method	Position	Content	$N{=}32$ ($16\times$)	$N{=}48$ ($24\times$)
$O(N)$ *cache:*				
Sliding window (baseline)	actual	Sliding window	0.401	0.388
YaRN [81]	NTK	Sliding window	0.490	0.412
$O(1)$ *cache, Block-Rel positions:*				
MemRoPE [53]	Block-Rel	dual EMA	0.651	0.706
Landmark + Block-Rel	Block-Rel	verbatim	0.929	0.934
$O(1)$ *cache,* WorldTrace *positions:*				
MemRoPE + WorldTrace positions	WorldTrace	dual EMA	0.592	0.662
WorldTrace-Landmark **(ours)**	WorldTrace	verbatim	**0.964**	**0.972**

D.2 Concurrent Training-Free Baselines

Tab. 8 compares MemRoPE [53] and YaRN [81] on the Sec. 3.3 ABA loops, factoring the position scheme (Block-Rel vs. WorldTrace) from the content writer at two horizons.

MemRoPE [53] (Block-Rel + 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-Rel: verbatim canonical keys retain stronger query-key similarity than smoothed averages, and WorldTrace-Landmark adds in-distribution positions on top. MemRoPE+WorldTrace (Block-Rel swapped for WorldTrace, EMA kept) degrades ($p{<}0.001$): write/read offsets disagree.

YaRN [81] rescales temporal RoPE frequencies to bring offsets back in-distribution, gaining $+22\%$ over the sliding-window baseline at $N{=}32$ ($p{<}0.001$) and confirming position is the primary bottleneck; but it needs $O(N)$ memory (OOM by $\sim N{=}100$) and the rescale diverges with horizon. WorldTrace-Landmark exceeds it by a wide margin at $N{=}32$ (0.964 vs. 0.490) in $O(1)$ memory and stays near-constant through $N{=}512$.

D.3 Cross-Architecture Experiments

Tab. 9 reports PAC at $2\times$–$8\times$ training horizon on LingBot-Fast [86] 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 9. **LingBot-Fast [86] 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**

D.4 Slot Allocation

Summary/recent split sensitivity. The main experiments use $N_s{=}4$ summary slots and $W_r{=}2$ recent-window slots (total $N_s{+}W_r{=}6$, matching the training context size). To test sensitivity to this split, we swept five allocations ($N_s{+}W_r{=}6$, ABA loops at $N{=}32$ and $N{=}64$); results are in Tab. 10. $N_s{\leq}2$ 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 10. **Slot Allocation.** PAC at $N{=}32$ and 64 as a function of N_s (total budget fixed at $N_s{+}W_r{=}6$). $N_s{\leq}2$ collapses toward the sliding-window baseline; four slots is the critical boundary for retaining the scene-A landmark through the B-side traversal.

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

Per-slot depth ablation. To isolate which slot carries the recall signal, we run WORLDTRACE-LANDMARK with only one summary slot active at a time (inactive summary slots revert to sliding-window eviction); results are in Tab. 11. Slot 3 alone recovers most of the four-slot PAC at $N{=}32$ and $N{=}48$ (Tab. 11: 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 11. **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

D.5 Phase Cancellation Validation

Tab. 12 verifies that canonical key averaging avoids the phase-cancellation failure of Sec. 1: as N_s grows, naive averaging in RoPE-rotated space loses progressively more low-frequency content (measured by $\mathrm{LatentDiff}\downarrow$), while canonical averaging in unrotated space is unaffected.

Table 12. **Phase cancellation: Naive vs. canonical key averaging.** LatentDiff (↓) at N=16 for varying N_s. Canonical averaging avoids the phase cancellation that corrupts low frequencies under naive RoPE-space averaging.

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

E LoopMem: Scripted Navigation Benchmark

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

The benchmark organizes difficulty along four axes: (1) **waypoint count** (K): number of intermediate locations between departure and return to A (ABA: K=1; ABCDA: K=3); (2) **edge length** (L): AR chunks per directed edge, controlling KV-context distance; (3) **camera orientation**: heading on arrival at A relative to departure, testing viewpoint-invariant recall; and (4) **multi-revisit depth** (R): how many times a waypoint is re-entered within a single rollout. The four-axis decomposition follows the broader principle that benchmark suites should isolate task properties capable of producing distinguishable rankings between candidate systems [68], here projected onto the position–content axes that the cache mechanism is hypothesized to probe. Fig. 3 shows representative configurations; those evaluated here are marked *.

F Implementation Details

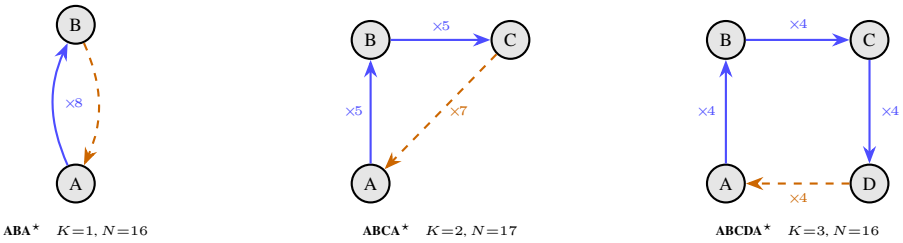
F.1 Model Architecture and Self-Forcing Background

We evaluate on MG2-1.3B, a 1.3B-parameter distilled autoregressive video world model based on Wan 1.3B T2V [100], 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 × 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 [47] (rollout on the student's own KV cache rather than teacher-forced context; *cf*. Diffusion Forcing [12], CausVid [127], and one-step diffusion distillation [91]), attention is restricted to a local window of `local_attn_size`= 6 frames (2 AR blocks), so cross-frame temporal offsets stay $\leq \Delta t_{\text{train}}$=5. At inference, the rolling KV cache grows without bound; with a constant $O(1)$ budget, our method caps it at $L_{\text{train}} \times 2640 \times 12 \times 256 \times 30 \times 2 \approx 2.79$ GB per batch element (L_{train}=6 chunks, 2640 tokens/chunk, fp16; *cf*. Tab. 14), independent of generation length. At long horizons, the OOD mismatch is not the presence of cached keys at large offsets, but the temporal RoPE rotation through which they are read: keys are stored verbatim, but the model was never trained to invert the phase rotations they accumulate beyond Δt_{train}.

F.2 Rotary Position Embeddings in 3D Video Attention

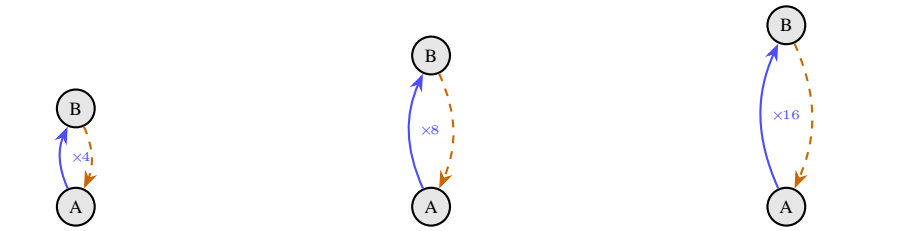
We inherit the 3D RoPE split of the Wan 2.1 backbone [100]: the 128-dimensional per-head embedding is split across three position axes, $2c_t$=44 dimensions for temporal position, and $2c_h$=$2c_w$=42 dimensions each for spatial height and width, with shared base frequency $\theta = 10000$. The k-th temporal frequency component follows $\theta_k = \theta^{-k/c_t}$ for $k = 0, \ldots, c_t-1$, giving $\theta_0 = 1.0$ (fastest rotating) down to $\theta_{c_t-1} \approx 1.5 \times 10^{-4}$ (slowest). The attention score contribution from frequency k between the query at temporal position q and the key at t is proportional to $\cos(\theta_k(q-t))$. When $\theta_k|\delta_{q,t}| \ll \pi$, the cosine is near 1, and the component is position-invariant, carrying semantic content; when $\theta_k|\delta_{q,t}| \gg \pi$, it oscillates incoherently and constitutes positional noise. This wavelength-vs-context view follows YaRN's NTK-by-parts framing [81] and is consistent with the mechanistic finding that high-frequency RoPE components carry positional structure while low-frequency components carry semantics [3]. 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 [69]: 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.

Row 1: varying topology (K intermediate waypoints)



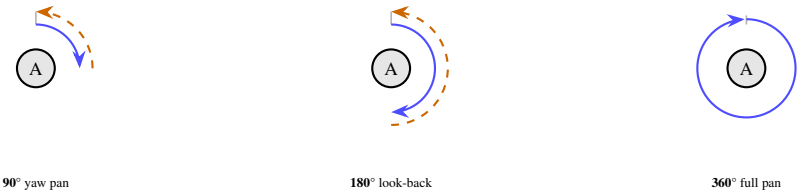
ABA* $K{=}1, N{=}16$ **ABCA*** $K{=}2, N{=}17$ **ABCDA*** $K{=}3, N{=}16$

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



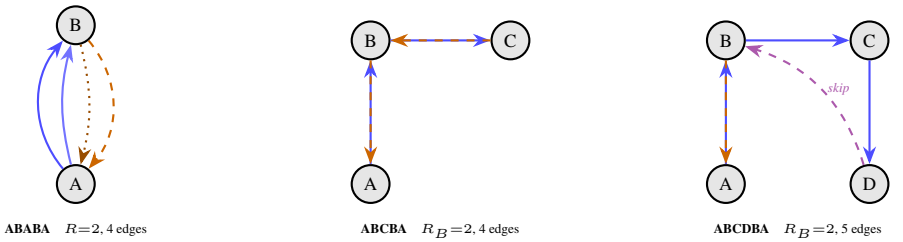
short $L{=}4, N{=}8$ *medium** $L{=}8, N{=}16$ *long** $L{=}16, N{=}32$

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



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

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



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

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

Figure 4. **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.

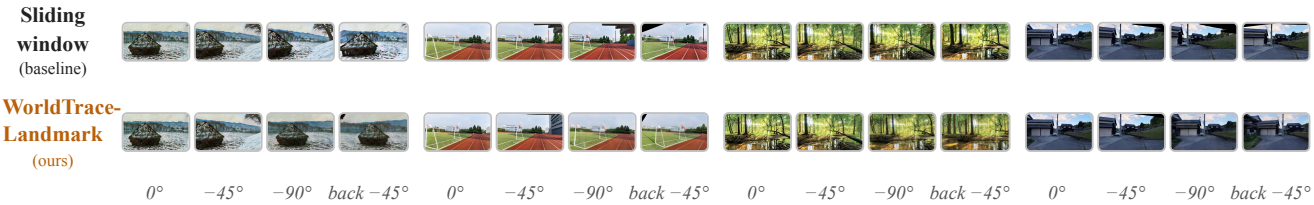


Figure 5. **Pan** $90°$ **qualitative results.** Camera-orientation Tier 3 (N=4): the agent is fixed at A while the camera pans right by $\sim$90° then returns. Each sample group of four frames shows keyframes at chunks 0 (0°) / 1 ($\sim$45°) / 2 ($\sim$90°) / 3 (back $\sim$45°); the two rows compare the sliding-window baseline (top) vs. WORLDTRACE-LANDMARK (bottom). WORLDTRACE-LANDMARK restores the initial-view appearance on the return half-pan; the sliding window does not.

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

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

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

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

F.3 WORLDTRACE Algorithm and Architectural Baselines

WORLDTRACE cache update. Each autoregressive chunk appends fresh keys and values to the recent window; when that window overflows, the evicted block is merged into the summary tier using either uniform-bucket averaging (WORLDTRACE-FIELD) or scene-entry landmark insertion (WORLDTRACE-LANDMARK), as defined in the main text. Alg. 1 presents the full control flow in pseudocode, including the shared attention-time step that applies per-slot rotations to canonical summary keys using the virtual positions from Def. 1.

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

F.4 Hyperparameters, Compute, and Evaluation Protocol

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

WORLDTRACE-FIELD hyperparameters. Short-horizon experiments (N=8) use N_s=2 summary slots and W_r=4 recent-window slots ($N_s + W_r = 6 = L_{\text{train}}$). Long-horizon PAC and ablation experiments (Secs. 3.3–3.3) use N_s=4, W_r=2 (same total capacity) for both WORLDTRACE-FIELD and WORLDTRACE-LANDMARK, as described in the Setup section. The scripted loop evaluation (Sec. 3.3) uses N_s=5, W_r=1 for WORLDTRACE-LANDMARK (one additional landmark slot) and

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

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

Append new KVs to $\mathcal{R}$ {recent window is verbatim}
if $|\mathcal{R}| > W_r$ **then**
  $K_* \leftarrow \text{POPOLDEST}(\mathcal{R})$
  if $\mu = \text{WORLDTRACE-FIELD}$ **then**
    $s^* \leftarrow$ summary slot whose source bucket now contains K_* {uniform temporal grouping}
    $\mathcal{S}[s^*] \leftarrow$ canonical mean of s^*'s source frames {Def. 2}
  else if $\mu = \text{WORLDTRACE-LANDMARK}$ **then**
    $\text{SE} \leftarrow \bigvee_{f=0}^{F-1} \left[\text{cosdist}(K^{(k)}_{\text{cx},K_*,f}, K^{(k)}_{\text{cx,prev},f}) > \tau \right]$ {per-frame check inside popped block}
    if SE **then**
      $\mathcal{S} \leftarrow \text{SHIFTLEFT}(\mathcal{S})$ {drop $\mathcal{S}[0]$}
      $\mathcal{S}[N_s{-}1] \leftarrow K^{(k)}_{\text{cx},K_*}$ {canonical landmark; cf. Eq. (3)}
    end if
    (Init.) If fewer than N_s landmarks are stored, fill empty slots with the oldest.
  end if
end if
Attention time: for $s{=}0, \ldots, N_s{-}1$, apply $R(\theta_k t_v^{(s)})$ to $\mathcal{S}[s]$ with $t_v^{(s)}$ from Def. 1.

$N_s{=}4$, $W_r{=}2$ for WORLDTRACE-FIELD. Compression uses uniform temporal grouping: the T_{old} oldest frames are split into N_s equal groups; each group's keys are unrotated to canonical space (fp64 precision), averaged, and re-rotated at the group's virtual position in the original dtype. **Position assignment:** Tab. 12 reports canonical key averaging with Block-Rel virtual positions [125], where each summary slot is assigned position $\max(0, q - (L_{\text{train}} - 1)F)$, for a fair comparison against the Naive+Block-Rel baseline. Tab. 6 reports the long-horizon conditions (Sliding Window, Block-Rel, centroid linear, fullcombo) at $N{=}16$. The recent-window slot j keeps its absolute position across schemes.

Evaluation protocol. Each method is evaluated on 100 videos generated from initial frames. TempSSIM is computed on decoded RGB frames using SSIM [102] 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. 2 (Sec. 3.2) is generated along a single fixed camera path on Matrix-Game 2, played out over $N{=}48$ AR chunks ($\sim$36 s of decoded video). Fig. 6 shows that path as a top-down plan: the camera leaves an initial scene (start, $t{=}1$), explores a roughly counter-clockwise loop through novel territory across $t \in \{8, 18, 24, 32, 40\}$, and arrives at a distinct end pose ($t{=}48$). Because the world model is run autoregressively on the same control inputs across baselines, every method in Fig. 2 is evaluated at exactly the same intended camera pose at each timestep; differences between columns therefore reflect the cache mechanism, not the trajectory.

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

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

Per-chunk runtime overhead. Tab. 15 reports wall-clock time per AR chunk on one A100 80 GB at batch size 1 (3 distilled denoising steps, 3 latent frames per chunk, $352{\times}640$ resolution). At short horizons ($N{=}8$), all methods are within 6% of the

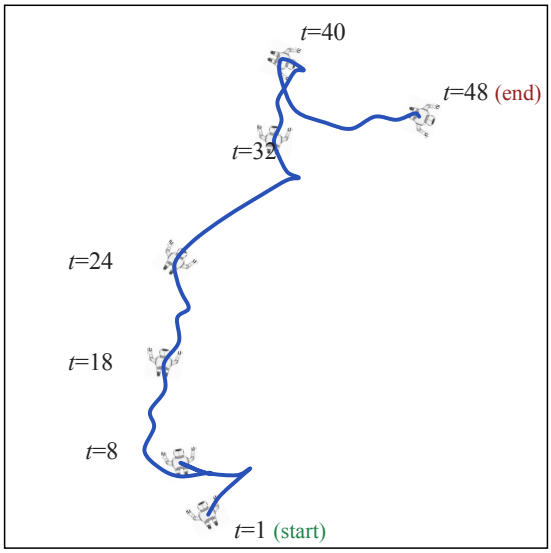


Figure 6. **Camera trajectory used for Fig. 2.** Top-down plan of the camera (x, z) path executed by the AR rollout for the WORLDTRACE-FIELD qualitative panel at $N{=}48$ ($\sim$36 s). Green: start ($t{=}1$); dark red: end ($t{=}48$). Orange ticks mark the chunk indices shown as columns in Fig. 2. The path leaves the initial scene, traverses novel territory, and lands at a different pose, so every column $t{>}1$ is out-of-window for the sliding-window baseline.

sliding-window baseline because the forward pass dominates and the cache is small. At $N{=}102$ ($\sim$77 s of decoded video), WORLDTRACE-FIELD adds $+9.4\%$ per chunk as the canonical unrotate/rerotate of the growing source set takes a larger share of each step, consistent with the $O(T_{\text{old}} \cdot c_t \cdot n_\ell)$ per-token form above. WORLDTRACE-LANDMARK's $+27.2\%$ at $N{=}102$ reflects per-step kernel overhead from bookmark accumulation rather than additional FLOP. The long-horizon column uses the `skip_decode` setting (no VAE decode during generation), so absolute times are lower than the $N{=}8$ column with VAE decode.

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

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

F.5 Evaluation Metrics

Open-ended video world models have no unique ground truth: a generated rollout legitimately diverges from any reference within a few chunks, so reference-based metrics (PSNR, SSIM vs. GT, LPIPS) and single-window benchmarks (VBench [48], PAI-Bench [138]) penalize valid creative divergence and do not test long-horizon recall. We therefore evaluate with a suite of reference-free metrics organized into two groups, *quality* (frame-to-frame coherence) and *consistency* (long-range episodic recall), with one diagnostic (LatentDiff) as a cross-check (Tab. 16).

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

Metric	Role	Dir.	Definition
TempSSIM	Coherence	↑	SSIM [102] between consecutive decoded frames, averaged over the rollout.
Local Scene Drift (SceneDrift)	Coherence	↓	Mean per-chunk CLIP feature distance to the preceding chunk.
PAC	Recall	↑	CLIP-ViT-H/14 cosine similarity between geometrically paired return- and forward-leg frames in ABA loops; PAC averages the final $N/8$ return chunks closest to scene A (Sec. 3.3).
Return CLIP	Recall	↑	CLIP cosine similarity between the final return frame and the scene-A reference frame. Used for non-ABA topologies where position alignment is undefined.
LatentDiff	Diagnostic	↓	MSE between consecutive latent frames (pre-decode). Confounded: favors slowly-varying output. Used alongside pixel-domain metrics as a fast sanity check only.

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

G Discussion

G.1 Limitations

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

Architecture and orthogonal cache mechanisms. WORLDTRACE assumes temporal RoPE on keys, and a fixed AR KV cache; single-shot generators and combinations with LLM-side eviction methods such as SnapKV [62] or H_2O [134] are out of scope here. The null PAC result for WORLDTRACE-FIELD on LingBot-Fast is consistent with our recall/coherence split (Sec. 2.5): Plücker conditioning already supplies the recall signal, but a matched return-SSIM comparison under Plücker would test whether pose conditioning also saturates MG2's coherence gains.

Metrics and headline numbers. Episodic recall is measured with paired CLIP cosine similarity on scripted return legs (Sec. 3); 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 [132]. Human studies and pixel-level visitation metrics could further provide more insights.

Training paradigm and comparative scope. MG2 inherits a Self-Forcing-style autoregressive training distribution with bounded local attention at write time (App. F.1); inference-time KV stitching does not alter that mismatch if future models train with different cache semantics. "Training-free" here excludes fine-tuning the generator (Sec. 2.1); methods that distill long-rollout consistency, enlarge context by training, or change attention masks lie outside this protocol. Multi-student distillation [91] and other one-step generator regimes change the per-block denoising budget and therefore the effective length over which an AR rollout accumulates RoPE OOD; how cache-side interventions like WORLDTRACE compose with such distillation pipelines is an interesting open direction.

Deployment overhead. WORLDTRACE adds canonical-domain key transforms and bookkeeping relative to sliding-window eviction; although peak cache memory scales as $O(1)$ in horizon (Sec. 2), wall-clock latency and bandwidth to move updated keys through fused attention kernels are not modeled here (Tab. 15 reports per-configuration timings under our reference stack).

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

G.2 Future Directions

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

Geometry-aware canonical keys. The canonical mean of Eq. (2) aggregates by screen coordinate. Coupling the unrotate/rerotate primitive with camera-pose warping (*e.g.* the Plücker-conditioned and Warped-RoPE writers of MosaicMem [129] and UCM [120]) would let the same operator average over scene coordinates instead, removing the ego-motion aliasing above and unifying the strict zero-fine-tune regime with the trained camera-aware family.

Learned scene-entry policies. Replacing the canonical-key spike or gradient-onset rule with a small policy trained on action discontinuities, agent-pose deltas, or scene-segmentation logits would enable WORLDTRACE-LANDMARK to commit landmarks during continuous motion. Active recall, in which the policy decides *when* to commit and *which* of multiple stored landmarks to re-rotate at a given query, is a natural extension. The cache layout itself (N_s, W_r, scene-entry threshold) is fixed in our experiments; treating it as an autotunable schedule under a checkpoint-conditioned surrogate [74] could let downstream practitioners adapt it per backbone or per horizon without re-running the full slot-sensitivity sweep of App. D.4.

Composition with content-side eviction. Because slot-rank positions are independent of which canonical content fills the slot, eviction heuristics from the LLM literature (H_2O [134], SnapKV [62], KnormPress [78]) 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 [53] comparison in App. D.2 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 [70], LATTE3D [118]) and LLM-conditioned mesh generation (LLaMA-Mesh [103]) consume teacher gradients whose Monte Carlo variance, rather than long-horizon recall, dominates compute; compute-aware estimators for those gradients (CARV [5]) and the corresponding analyses for non-vision teachers [85] are orthogonal axes to the position–content factorization studied here. On the data side, motion-attribution methods [110] ask which training clips improved the temporal dynamics of a generator, complementing the cache-side question of which past slots a generator can still address at inference. Establishing whether the slot-rank virtual-position primitive transfers to these adjacent regimes is an open direction.

Fine-tuning extensions. WORLDTRACE stays within the training context length L_{train} (Sec. 2.1) precisely because it is training-free: the constraint $N_s + W_r = L_{\text{train}}$ keeps every summary slot at an in-distribution offset the generator was trained on. Two light fine-tuning paths would relax this budget without retraining the generator from scratch. (i) *Context-extension fine-tuning* on synthetic long rollouts would let WORLDTRACE allocate either more recent slots for coherence or more summary slots for longer recall horizons at the same in-distribution attention cost. (ii) *Position-aware fine-tuning* that exposes the model to the slot-rank offsets of Eq. (1) 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 [66] of a frozen large inner model paired with a light outer adapter for the cache schedule, so the position/content factorization is preserved through the outer loop. Both are compatible with the WORLDTRACE cache layout and leave the position/content factorization intact.

Multi-tier and cross-architecture transfer. WORLDTRACE uses a two-tier split (recent verbatim, summary canonical). Adding intermediate tiers, with progressively larger source buckets and progressively deeper slot-rank offsets, could trade sharper recall for longer effective horizons. Porting the same factorization to KV-cache-bearing variants of MG3 [105], Genie3 [30], or LingBot-Fast [86] would test whether the position/content split is architecture-specific or generic.

Scaling LoopMem. LoopMem (App. E) exercises all four difficulty tiers on MG2-1.3B; camera-orientation Tier 3 uses MG2's mouse-yaw control, where chunk counts are nominal magnitudes rather than calibrated angles. Pose-conditioned generators with explicit pitch/roll would extend Tier 3, and broader cross-architecture leaderboards on Tiers 1, 2, and 4 would let the community compare position/content trade-offs at controlled compression ratios, complementing the broad video-generation benchmarks of WorldScore [26], MIND [124], and VBench-2.0 [137]. The Pan 360° failure of WORLDTRACE-LANDMARK on visually continuous trajectories (Tab. 2) suggests learned or rule-based scene-entry policies as a concrete next step.

G.3 Broader Impact

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