ICML International Conference On Machine Learning
From Frames to Stories Workshop

WorldTrace
Addressable Memory for Video World Models

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

Project page

Motivation

Autoregressive video world models promise interactive worlds - but **visual persistence** collapses once generation exceeds the training horizon.

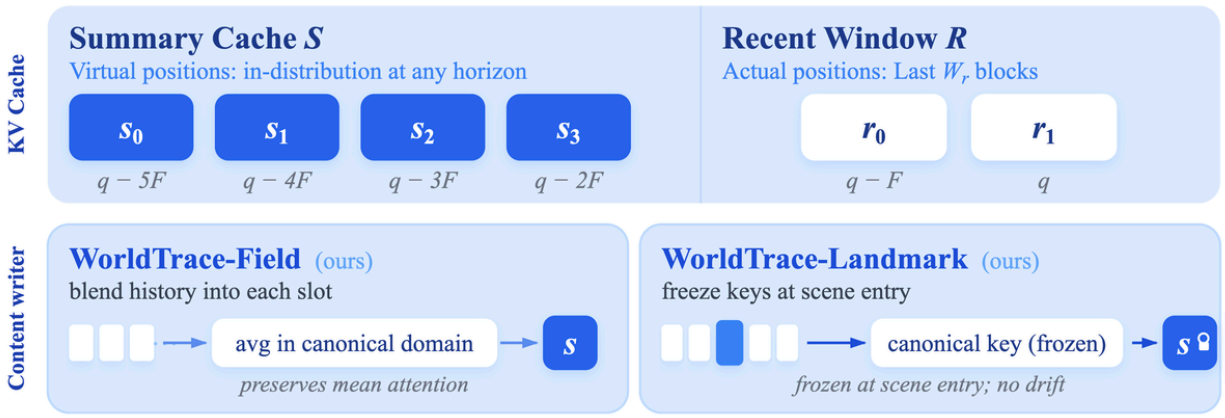
Can an autoregressive video world model reliably remember where it has been, at any generation length?

Position and Content are Two Coupled Bottlenecks

① **Addressability:** Temporal RoPE offsets exceed the training range, making cached memories unreadable even when physically present.

② **Content Fidelity:** Naive key averaging in RoPE-rotated space causes phase cancellation, destroying the signal that compressed summaries should carry.

Method: WorldTrace



Slot-Rank Position Assignment

Virtual position for summary slot s:

$$v_s = q - \left(L_{\text{train}} - 1 - s\right) \cdot F$$

q = current query · L_{train} = training context length · F = frames per AR block

In-distribution at any generation length

Horizon-stable: rank-only offsets

WorldTrace-Field

Averages *all* history into each summary slot in the **canonical (unrotated) key domain**, then re-rotates to the slot's virtual position.

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

Avoids phase cancellation by aligning keys to a shared canonical phase. **Goal:** coherence: drift-free generation at long horizons.

WorldTrace-Landmark

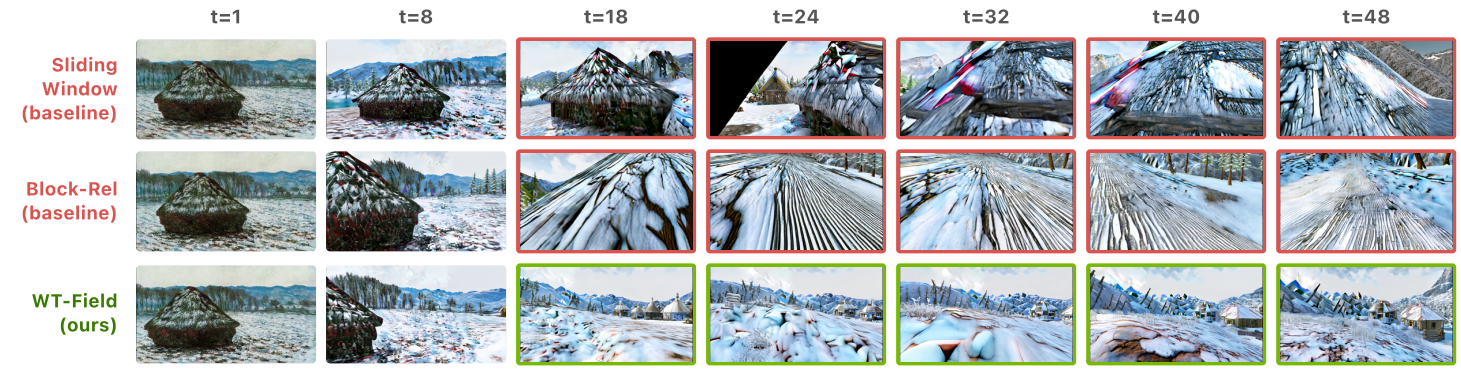
Detects scene-entry events via cosine-distance spikes in canonical-K space. Fills summary slots with **verbatim frames** at detected boundaries, frozen in canonical form.

$$K^{(k)}_{\text{land}}(t_v) = R(\theta_k t_v)\,R(-\theta_k t_{\ell^*})\,K^{(k)}_{t_{\ell^*}}$$

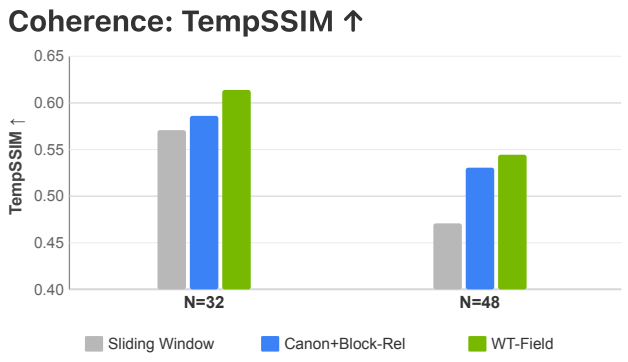
Key stored once at landmark time; single rotation per shift; no float error accumulation. **Goal:** episodic recall after a long detour.

Experiments

WorldTrace-Field



Both baselines (Sliding Window, Block-Rel) diverge from t=18 ✕; WT-Field remains coherent through t=48 ✓.

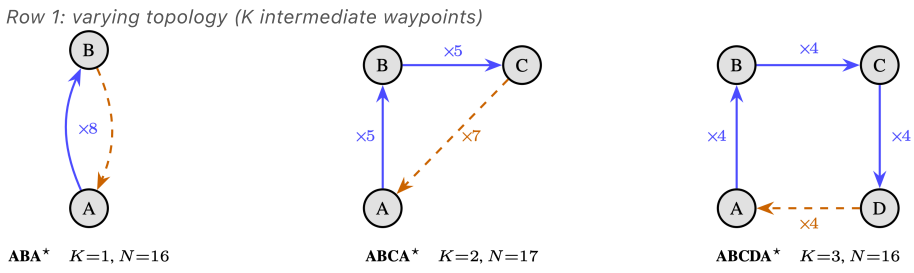


+15.5% relative TempSSIM at N=48; WT-Field leads on both horizons.

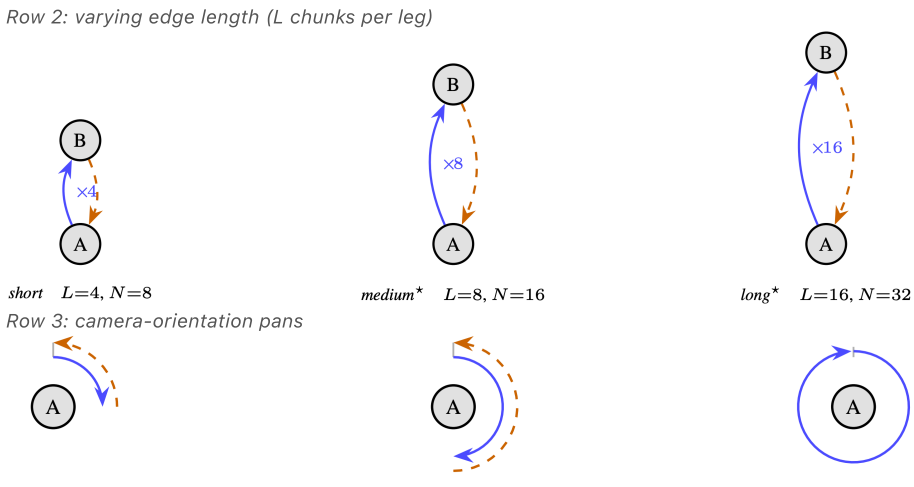
WorldTrace-Landmark

LoopMem Benchmark Topologies

Row 1: varying topology (K intermediate waypoints)

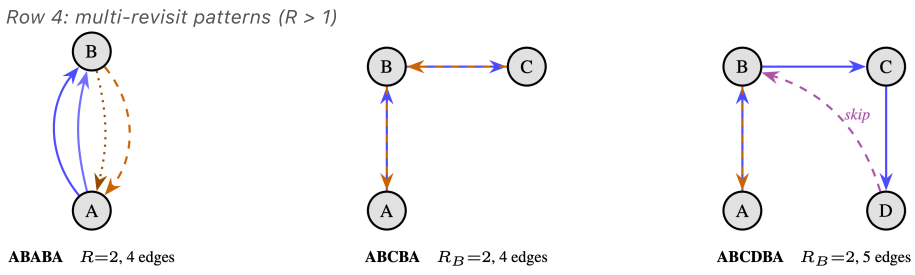


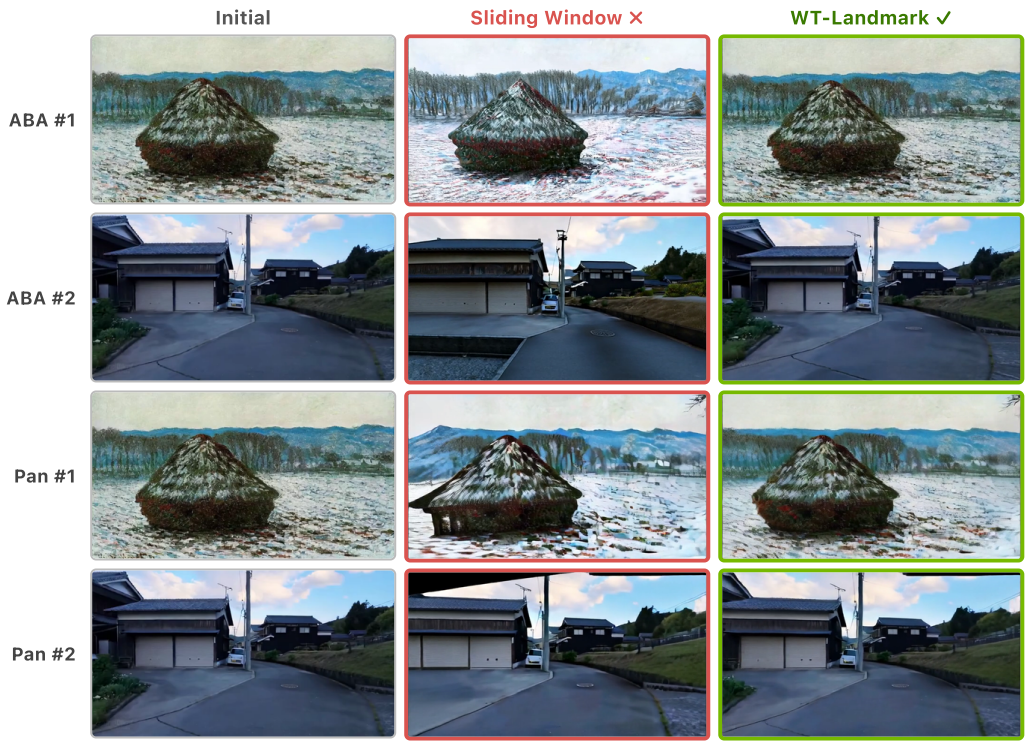
Row 2: varying edge length (L chunks per leg)



Row 3: camera-orientation pans

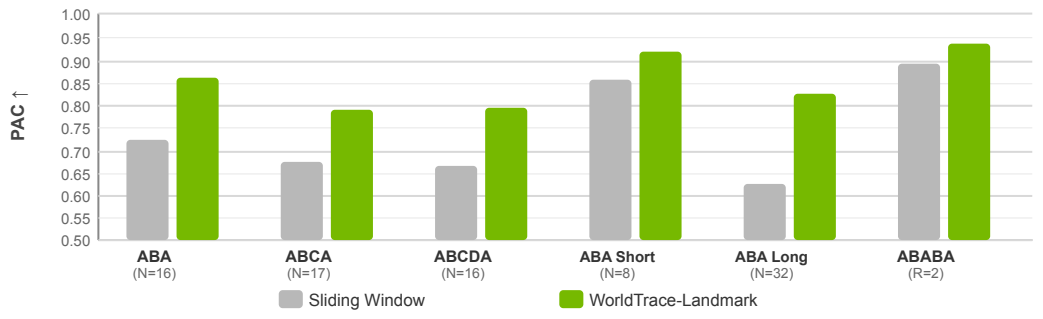
Row 4: multi-revisit patterns (R > 1)





WT-Landmark faithfully restores scene A across diverse environments.

LoopMem PAC ↑ — Sliding Window vs. WT-Landmark



Gap widens with context distance (+6.3% to +19.8%). Consistent gains across revisit depth.

Conclusion

Long-horizon failure is fundamentally a problem of *addressability*.

OOD Positions Are the Binding Constraint
Content compression alone cannot restore recall once temporal RoPE offsets go out-of-distribution.

Training-Free
Only the KV cache changes: no fine-tuning, no model modification. Applies to any temporal-RoPE AR video model.

WT-Field **+15.5% TempSSIM**
Canonical key averaging improves long-range coherence at 24× generation horizon.

WT-Landmark **+19.3–20.0% PAC**
Frozen canonical landmark keys restore episodic recall across all LoopMem topologies.