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

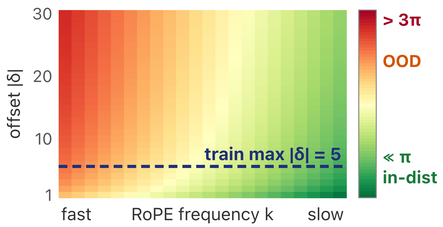
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Autoregressive video world models promise interactive worlds, but **visual persistence degrades** beyond the training horizon.

Two Coupled Bottlenecks

① Where: Addressability
Old frames get RoPE offsets past the trained range: stored ≠ reliably retrievable.



Bracket labels are schematic phase bands.

② What: Informativeness
Averaging keys with different RoPE rotations can partially cancel their components, distorting retrieval scores.

WorldTrace-Field

Unrotate keys, average in canonical space, re-rotate to virtual slot v.

$$K_{\text{field}} = R_v \frac{1}{M} \sum_i R_i^{-1} K_i$$

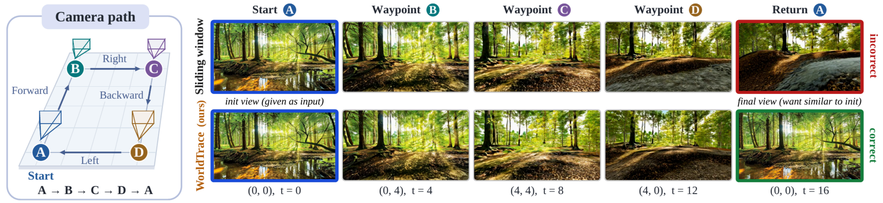
WorldTrace-Landmark

Keep verbatim frames ℓ at scene boundaries, frozen in canonical form.

$$K_{\text{land}} = R_v R_i^{-1} K_i$$

Can an autoregressive video world model reliably remember the past?

Visual Persistence on Revisit



The camera loops A→B→C→D→A. A sliding-window cache forgets scene A on return; WorldTrace regenerates it, **without retraining**.

WorldTrace: a training-free memory that keeps compressed history addressable and informative

Addressable Memory Bank

Summary cache *S*
compressed history at fixed virtual positions

s0	*s1*	*s2*	*s3*
q − 5	q − 4	q − 3	q − 2

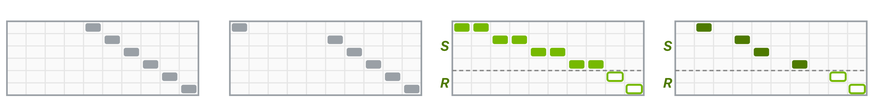
Recent window *R*
last frames, verbatim

r0	*r1*
q − 1	q

Field
merge compatible frames
□□□ → canonical avg → *s*

Landmark
keep distinctive frames verbatim
□■□ → frozen key → *s★*

Memory compression as **structured sparse attention**

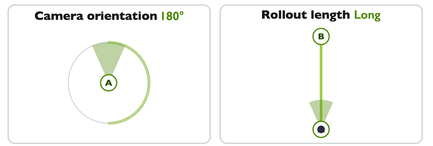


Sliding window | **+ attention sink** | **WorldTrace-Field** | **WorldTrace-Landmark**

rows: cache slots (**S** summary cache, filled / **R** recent window, outlined) · columns: past frames

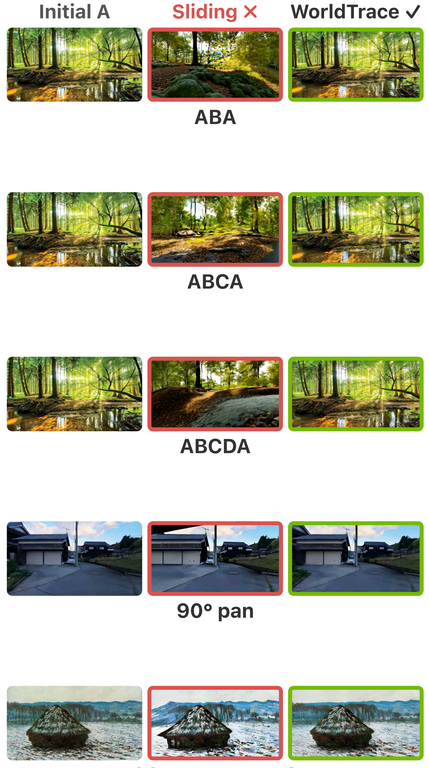
LoopBench

The camera **returns to scene A** after a long detour.



Qualitative Results

View at A on Return

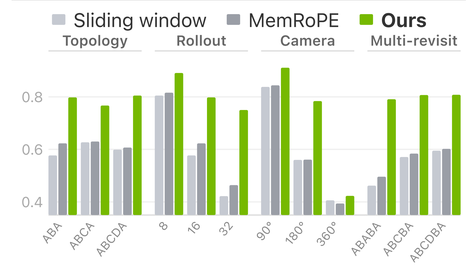


Quantitative Results

+32.6% mean recall vs. sliding window
0.587 → **0.778**

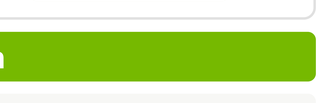
LoopBench Recall

DINOv2 ↑, n = 51



Fields + Landmarks

mean DINOv2 ↑



Conclusion

1. **Memory needs to stay addressable**
Old frames drift beyond the temporal range the model was trained on.

2. **Memory needs to stay informative**
Averaging keys with different RoPE rotations can partially cancel their components, distorting retrieval scores.

Takeaway: memory needs to be both **addressable** and **informative**. We view memory compression as **structured sparse attention**.