# MEMIX: SELECTIVE STATE PRESERVATION FOR STREAMING 3D RECONSTRUCTION

**Anonymous authors**
Paper under double-blind review

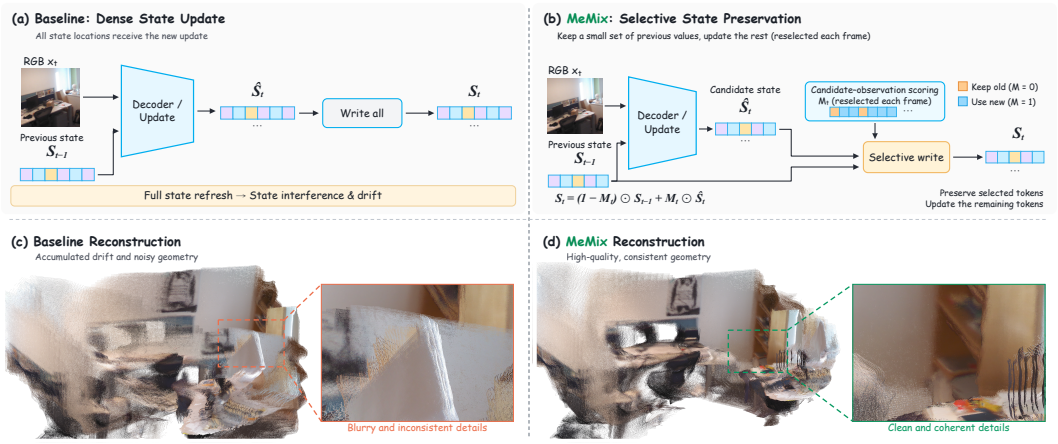


Figure 1: **MeMix.** A training-free exact-preservation transition for fixed-state streaming 3D reconstruction. Candidate/current-frame scores select 60 old-state locations to preserve ($M = 0$); the other 708 follow the base update ($M = 1$). Preservation acts after the current-frame readout. Recurrent memory is $O(1)$ in stream length at fixed token count.

## ABSTRACT

Streaming 3D reconstruction must absorb new observations while retaining historical context in bounded memory. This requires deciding not only how strongly to update a recurrent state, but which locations should temporarily avoid a write and when to reselect them. We present MeMix, a training-free module that allocates a fixed preservation budget using current-frame compatibility. At each frame, it scores candidate state locations against the observation, preserves the previous values at the highest-scoring locations, and leaves the remaining locations on the original update path. MeMix adds no learnable parameters and keeps recurrent memory constant in stream length. Matched interventions show that compatibility-guided selection and current-frame reselection improve long-horizon performance beyond gate binarization or a fixed protected subset. We evaluate MeMix across update rules sharing a pretrained CUT3R checkpoint. On 16 1,000-view 7-Scenes streams, MeMix reduces TTSA3R's accuracy and completeness errors by **29.9%** and **30.7%**. With the FILT3R Kalman update, it lowers 1,000-view ScanNet pose ATE by **17.8%** across 65 scenes.

## 1 INTRODUCTION

Streaming 3D reconstruction estimates scene geometry and camera poses as RGB observations arrive. Long-running perception requires bounded memory as the scene evolves. Fixed-state recurrent models such as CUT3R (Wang et al., 2025b) compress history into a constant-size latent representation, but the same locations must retain past information and absorb new evidence. This raises a central allocation question: *which locations should temporarily avoid a new write, and when should that choice be renewed?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Retention depends on both state capacity and its update rule. Full-history attention caches grow with sequence length (Zhuo et al., 2026), while bounded-cache methods evict or prune historical tokens (Mahdi et al., 2025; Yuan et al., 2026). In a compressed recurrent state, direct replacement and adaptive soft-gated updates (Wang et al., 2025b; Chen et al., 2026; Zheng et al., 2026) control writes to existing locations. Selective preservation adds a complementary allocation decision: which previous values should remain unchanged while the rest of the state adapts? Neither a larger state nor a new learned gate is required to make this decision.

We propose **MeMix**, a training-free module for allocating a fixed preservation budget. Inspired by routed memory mixtures (Du et al., 2025), MeMix scores each candidate state location against the current-frame feature, preserves the previous values at high-scoring locations, and sends the remaining locations through the original update. By default, 60 of 768 locations are protected while 708 remain writable; the protected subset changes with the observations rather than forming a permanent memory bank. Preservation acts after the current-frame readout, adds no learnable parameters, and retains constant recurrent memory in stream length.

Exact preservation creates a continuity path; the remaining question is how to allocate it. Our analysis characterizes this path and decomposes candidate compatibility into old-state and proposed-change alignment. Matched interventions show that compatibility-guided protection outperforms reverse, random, and periodic allocation at the same budget; candidate scoring outperforms old-state scoring; and fixing or delaying the mask weakens performance. These results identify protected-location selection and current-frame reselection as key components of long-horizon state preservation.

Within the CUT3R state family, the long-horizon gains also hold against unmodified adaptive update rules. On all 16 eligible 1,000-view 7-Scenes streams, MeMix lowers TTSA3R's accuracy and completeness errors by 29.9% and 30.7%. With FILT3R's Kalman-style update (Jin & Ye, 2026), it reduces 1,000-view ScanNet ATE by 17.8% across 65 scenes.

*Our contribution can be summarized as follows:*

1. A compatibility-guided state-preservation module that allocates continuity and adaptation within fixed recurrent memory, without retraining.
2. Evidence that compatibility-guided selection and current-frame reselection improve long-horizon performance beyond gate binarization or a fixed protected subset.
3. Long-horizon reconstruction and pose gains against unmodified update rules at fixed recurrent memory.

## 2 Related Work

**Feedforward Offline Reconstruction.** DUSt3R and MASt3R (Wang et al., 2024; Leroy et al., 2024) combine ViT (Dosovitskiy et al., 2021) encoders with cross-attention for pairwise 3D matching, using post-processing for multi-view reconstruction. VGGT (Wang et al., 2025a) uses intra-frame and global attention for multi-view consistency; later variants (Shen et al., 2025b; Wang & Xu, 2025; Sun et al., 2025) improve speed and accuracy. These offline methods require the full sequence at inference, limiting their use on arbitrarily long, bounded-resource streams, while downstream applications demand real-time perception (Selvaratnam & Bazazian, 2025; Liao et al., 2025).

**Streaming Online Reconstruction.** Online methods (Wang & Agapito, 2025; Wang et al., 2025b; Wu et al., 2025; Lan et al., 2026; Chen et al., 2025b; Jin & Ye, 2026; Liu et al., 2026) reconstruct incrementally and differ in how they retain history. Full-history KV-cache methods (Zhuo et al., 2026) retain historical features with memory growth over time, whereas Evict3R (Mahdi et al., 2025) and InfiniteVGGT (Yuan et al., 2026) bound the cache through token eviction or pruning. Fixed-state and adaptive latent-state methods such as CUT3R (Wang et al., 2025b), TTT3R (Chen et al., 2026), TTSA3R (Zheng et al., 2026), and FILT3R (Jin & Ye, 2026) maintain bounded recurrent states for constant-memory inference. CUT3R interacts input tokens with memory states via cross-attention and writes back the full candidate state; it has no explicit preservation branch. TTT3R and TTSA3R modulate state updates through training-free adaptive gates inspired by test-time learning and online-optimization views (Sun et al., 2024; Chai & Xu, 2025; Chen et al., 2026; Zheng et al., 2026), while FILT3R introduces a latent-state adaptive Kalman filter. Mem3R (Liu et al., 2026) further redesigns the CUT3R recurrent architecture with hybrid memory, using an implicit fast-weight memory for

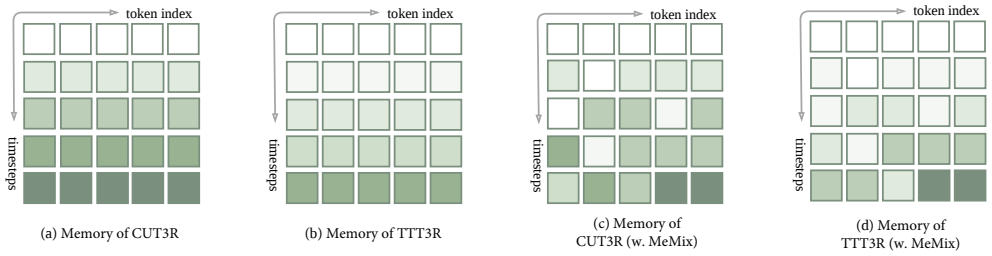


Figure 2: **Selective State Preservation with Mixture Memory.** (a) CUT3R refreshes the recurrent state at every timestep. (b) TTT3R applies a dense soft gate to modulate update magnitude, without explicitly allocating exactly preserved locations. (c–d) MeMix adds an exact preservation branch: high-compatibility memory locations are retained, while low-compatibility locations are refreshed, and the same binary preservation mask can be plugged into CUT3R (c) or combined with TTT-style gating (d). Shading schematically distinguishes contributions over time; only the MeMix preservation branch explicitly leaves selected locations unchanged.

camera tracking and an explicit fixed-size token memory for geometric mapping. These methods adjust update strength, filter the latent state, or train redesigned memory architectures, whereas MeMix adds a training-free plug-in binary exact-preservation route without introducing learnable modules. Point3R (Wu et al., 2025) anchors historical tokens to explicit 3D point positions, achieving strong recall but with memory consumption that grows linearly with the number of views.

**Memory Mixture.** In sequence modeling (Sutskever et al., 2014), linear attention and state-space models use gating to control retention: RetNet (Sun et al., 2023) employs exponential decay; Mamba (Gu & Dao, 2023) introduces input-dependent selection; DeltaNet (Yang et al., 2024) and Gated DeltaNet (Yang et al., 2025) address key collisions in additive updates; and Titans (Behrouz et al., 2024) learns a neural long-term memory at test time. Subsequent work (Li et al., 2025; Cao & Wang, 2025; Zhang et al., 2024) further analyzes gating, but continuous gates modulate update magnitude rather than guaranteeing that selected state dimensions remain unchanged. MoM (Du et al., 2025) brings routed memory mixtures into linear attention, partitioning the recurrent state into independent memory locations to reduce cross-time interference. Our work follows this direction with a training-free exact-preservation transition, evaluated across multiple inference-time update rules that share the released CUT3R checkpoint and state structure.

## 3 METHOD

MeMix separates the content of a proposed write from the decision to accept it. The recurrent reconstruction model produces a candidate state; MeMix allocates a fixed budget of locations that instead retain their previous values. This training-free transition leaves the other locations on the original update path, without fine-tuning or additional learnable modules.

### 3.1 CANDIDATE STATES AND RECURRENT WRITES

Given an RGB stream $\{\mathbf{I}_t\}_{t=1}^{T}$, online reconstruction estimates camera pose $\mathbf{T}_t$, intrinsic $\mathbf{K}_t$, and pointmap $\mathbf{P}_t$ frame by frame. Following CUT3R (Wang et al., 2025b), we use a fixed-length recurrent state:

$$\begin{aligned} \mathbf{X}_t = \texttt{Enc}(\mathbf{I}_t), \quad (\hat{\mathbf{S}}_t, \mathbf{Y}_t) = \texttt{Dec}(\mathbf{S}_{t-1}, \mathbf{X}_t), \\ (\mathbf{T}_t, \mathbf{K}_t, \mathbf{P}_t) = \texttt{Head}(\mathbf{Y}_t), \quad \mathbf{S}_t = \texttt{Update}(\mathbf{S}_{t-1}, \hat{\mathbf{S}}_t) \end{aligned} \tag{1}$$

where $\mathbf{X}_t$ are encoder tokens, $\hat{\mathbf{S}}_t$ is the candidate state, and $\mathbf{Y}_t$ are decoded tokens. The state $\mathbf{S}_t \in \mathbb{R}^{n \times d}$ starts from learned embeddings. The current prediction is computed before the recurrent write; preservation changes the state available to subsequent frames.

A scalar gate per location, broadcast over feature channels, controls the state write:

$$\mathbf{S}_t = \boldsymbol{G}_t \odot \hat{\mathbf{S}}_t + (1 - \boldsymbol{G}_t) \odot \mathbf{S}_{t-1}. \tag{2}$$

CUT3R uses $\boldsymbol{G}_t = 1$; TTT3R/TTSA3R use soft gates $\boldsymbol{\beta}_t \in [0,1]^{n\times 1}$ (Chen et al., 2026; Zheng et al., 2026). MeMix adds a separate allocation decision: set the gate to zero at selected locations and leave the native write elsewhere. Thus the base rule determines *how much to write*, while MeMix determines *where to preserve exactly*. Table 1 summarizes the compositions. Appendix A1 retains decoder and gate details.

Table 1: **Unified state-write gates.** FILT3R also updates covariance (Eq. (6)).

| Method | State-write gate |
|---|---|
| CUT3R | $\boldsymbol{G}_t = 1$ |
| TTT3R/TTSA3R | $\boldsymbol{G}_t = \boldsymbol{\beta}_t$ |
| FILT3R | $\boldsymbol{G}_t = \boldsymbol{\kappa}_t$ |
| CUT3R + MeMix | $\boldsymbol{G}_t = \mathbf{M}_t$ |
| TTT3R/TTSA3R + MeMix | $\boldsymbol{G}_t = \mathbf{M}_t \odot \boldsymbol{\beta}_t$ |
| FILT3R + MeMix | $\boldsymbol{G}_t = \mathbf{M}_t \odot \boldsymbol{\kappa}_t$ |

### 3.2 Compatibility-Guided Selective Preservation

**Scoring.** MeMix scores candidate locations against the mean projected image feature. With decoder input projection $W_{\mathrm{dec}}$ and $m$ image tokens,

$$\mathbf{g}_t = \frac{1}{m}\sum_{j=1}^{m} W_{\mathrm{dec}}\mathbf{X}_{t,j}, \qquad r_{t,i} = \langle\, \hat{\mathbf{S}}_{t,i},\ \mathbf{g}_t\rangle, \quad i = 1,\dots,n. \tag{3}$$

No spatial correspondence between image tokens and recurrent locations is required. One state token is one location by default; grouped-token scores are averaged before selection.

**Allocation.** With preservation budget $p = n - k$, MeMix protects the highest-scoring locations and refreshes their complement:

$$\begin{aligned} \mathcal{R}_t &= \operatorname*{arg\,max}_{\substack{\mathcal{Q}\subseteq\{1,\dots,n\}\\ |\mathcal{Q}|=p}} \sum_{i\in\mathcal{Q}} r_{t,i} = \mathrm{TopK}_p(r_t), \\ \mathcal{U}_t &= \{1,\dots,n\} \setminus \mathcal{R}_t = \mathrm{BottomK}_k(r_t). \end{aligned} \tag{4}$$

Thus `Bottom-k` names the refreshed set. Recomputed each frame, $\mathbf{M}_t$ is zero on $\mathcal{R}_t$ and one on $\mathcal{U}_t$. Protection is temporary, not a permanent memory bank: routing scores the candidate but preserves the *old* value.

**Composition with the native write.** For direct and soft-gated writes, MeMix composes with the native gate:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + (\mathbf{M}_t \odot \boldsymbol{\beta}_t) \odot (\hat{\mathbf{S}}_t - \mathbf{S}_{t-1}). \tag{5}$$

Here $\boldsymbol{\beta}_t = 1$ for CUT3R and is the native soft gate for TTT3R/TTSA3R. Setting $M_{t,i} = 0$ retains the old value; $M_{t,i} = 1$ leaves the native write unchanged.

FILT3R (Jin & Ye, 2026) maintains scalar per-location covariances, not cross-location entries. Given native predicted covariance $v^-_{t,i}$, measurement noise $R_{t,i}$, and gain $\kappa_{t,i}$, MeMix sets $\tilde{\kappa}_{t,i} = M_{t,i}\kappa_{t,i}$ and updates

$$\begin{aligned} \mathbf{S}_{t,i} &= \mathbf{S}_{t-1,i} + \tilde{\kappa}_{t,i}(\hat{\mathbf{S}}_{t,i} - \mathbf{S}_{t-1,i}), \\ v_{t,i} &= (1-\tilde{\kappa}_{t,i})^2 v^-_{t,i} + \tilde{\kappa}^2_{t,i} R_{t,i}. \end{aligned} \tag{6}$$

Shared statistics and gains are computed before masking. Protected values remain unchanged with $v_{t,i} = v^-_{t,i}$; other locations retain the native filter update. Appendix A4 details prediction and initialization. All compositions act after readout without learned parameters or fine-tuning.

### 3.3 Analysis of Selective Preservation

**Multi-step retention.** For direct and soft-gated writes, unrolling Eq. (2) with $G_{t,i} = M_{t,i}\beta_{t,i}$ isolates an explicit retention path:

$$\mathbf{S}_{t,i} = a_{\tau:t,i}\mathbf{S}_{\tau,i} + \sum_{u=\tau+1}^{t} G_{u,i}\, a_{u:t,i}\, \hat{\mathbf{S}}_{u,i}, \tag{7}$$

where $a_{u:t,i} = \prod_{v=u+1}^{t}(1 - G_{v,i})$ and $a_{t:t,i} = 1$. If $G_{v,i} \geq \underline{G} > 0$ for all $L = t - \tau$ steps, then $1 - x \leq e^{-x}$ yields

$$a_{\tau:t,i} \leq (1-\underline{G})^L \leq e^{-\underline{G}L}. \tag{8}$$

Each protected step contributes a factor of one; continuous protection gives $\mathbf{S}_{t,i} = \mathbf{S}_{\tau,i}$ with no intervening writes. This is the explicit carry path along a realized trajectory; candidates can also transmit history.

**Local isolation of protected values.** To examine coupling across locations, let $\mathbf{s} \in \mathbb{R}^{nd}$ stack the tokens, $F_t$ be the native transition, and $D_t$ select protected channels (diagonal entries $1 - M_{t,i}$). At fixed input, a strict selection-boundary gap in continuous scores makes $D_t$ locally constant. Where $F_t$ is differentiable, $D_t^2 = D_t$ gives

$$\begin{aligned} F_t^{\mathrm{MeMix}}(\mathbf{s}) &= D_t\mathbf{s} + (I - D_t)F_t(\mathbf{s}), \\ J_t &= D_t + (I - D_t)J_{F_t}, \\ D_tJ_t &= D_t^2 + D_t(I - D_t)J_{F_t} = D_t. \end{aligned} \tag{9}$$

Thus $D_t\mathbf{v} = 0$ implies $D_tJ_t\mathbf{v} = 0$: perturbing unprotected values cannot change protected values through this local write. $J_{F_t}$ includes candidate and gate dependence on the full state.

**Candidate scoring includes the proposed change.** With $\Delta_{t,i} = \hat{\mathbf{S}}_{t,i} - \mathbf{S}_{t-1,i}$, Eq. (3) decomposes exactly:

$$r_{t,i} = \underbrace{\langle \mathbf{S}_{t-1,i}, \mathbf{g}_t \rangle}_{\text{old-state alignment}} + \underbrace{\langle \Delta_{t,i}, \mathbf{g}_t \rangle}_{\text{proposed-change alignment}} . \tag{10}$$

Old-state scoring omits the second term; rewrite-magnitude routing uses $\|\Delta_{t,i}\|_2$ without its direction relative to $\mathbf{g}_t$. Candidate scoring uses both terms to select old values to carry forward.

Exact preservation supplies an identity path; compatibility routing determines how this path is allocated. Section 4.4 evaluates location selection, score source, and reselection timing through matched interventions. Appendices A1 and A4 give full derivations and retained-alignment analysis.

## 4 Experiments

We test three questions: does preservation improve long-horizon reconstruction, pose, and depth, which locations should be preserved, and does their current-frame allocation matter? Matched interventions isolate selection and timing. Short-sequence pose results and efficiency measurements are retained in the appendix.

**Baselines.** We evaluate MeMix with direct CUT3R writes (Wang et al., 2025b), soft-gated TTT3R/TTSA3R updates (Chen et al., 2026; Zheng et al., 2026), and the covariance-aware FILT3R update (Jin & Ye, 2026). We compare these inference-time rules using the released CUT3R checkpoint and state structure.

For broader benchmark context, we include pairwise models DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), MonST3R (Zhang et al., 2025), and Easi3R (Chen et al., 2025a); full-attention multiview models AETHER (Zhu et al., 2025) and VGGT (Wang et al., 2025a); and the KV-cache streaming Transformer StreamVGGT (Zhuo et al., 2026). Explicit-memory models Spann3R (Wang & Agapito, 2025) and Point3R (Wu et al., 2025) are compared in the short-sequence pose protocol where their public settings are directly comparable.

**Settings.** Unless a budget control is specified, MeMix preserves $p = 60$ of $n = 768$ memory locations and leaves $k = n - p = 708$ on the base update path. No variant is fine-tuned; each baseline–MeMix comparison uses the same hardware, software, inference settings, and evaluation protocol. The 300–500-view reconstruction tables, short-sequence benchmarks, and efficiency profile use a single NVIDIA A100 GPU. Appendix Table 6 reports computation costs.

**Evaluation protocol.** The CUT3R-family reconstruction evaluations use $512 \times 384$ input images. For CUT3R/TTT3R/TTSA3R pose estimation, images are resized to a 512-pixel long side and center-cropped to multiples of 16. Main-text ATE is the translation RMSE in meters after Sim(3) alignment to ground truth, fitted separately for each evaluated trajectory prefix. Means weight evaluation streams equally; paired 95% intervals resample streams (100,000 draws), not frames. The sequence sampling and eligible cohorts are specified with each comparison.

Table 2: **3D Reconstruction Results on 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022).** We test MeMix on 7-Scenes and NRGBD, with every frame sampled (Dense Sampling, **-D**). Green boxes indicate improved or unchanged performance over the base model (w/o MeMix) under the same input length.

| Model | MeMix | Input | 7-Scenes-D | | | | | | NRGBD-D | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Acc. (m) ↓ | | Comp. (m) ↓ | | NC ↑ | | Acc. (m) ↓ | | Comp. (m) ↓ | | NC ↑ | |
| | | | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. |
| VGGT *(Offline)* (Wang et al., 2025a) | – | *300* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *400* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *500* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| StreamVGGT (Zhuo et al., 2026) | – | *300* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *400* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *500* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| CUT3R (Wang et al., 2025b) | × | *300* | 0.099 | 0.062 | 0.048 | 0.014 | 0.542 | 0.562 | 0.137 | 0.092 | 0.066 | 0.024 | 0.572 | 0.609 |
| | ✓ | | 0.076 | 0.045 | 0.039 | 0.010 | 0.549 | 0.573 | 0.113 | 0.081 | 0.060 | 0.035 | 0.578 | 0.618 |
| | × | *400* | 0.150 | 0.093 | 0.090 | 0.037 | 0.531 | 0.543 | 0.225 | 0.155 | 0.119 | 0.076 | 0.554 | 0.579 |
| | ✓ | | 0.117 | 0.071 | 0.056 | 0.015 | 0.536 | 0.552 | 0.196 | 0.128 | 0.098 | 0.062 | 0.572 | 0.609 |
| | × | *500* | 0.165 | 0.114 | 0.094 | 0.039 | 0.522 | 0.531 | 0.313 | 0.203 | 0.202 | 0.148 | 0.554 | 0.580 |
| | ✓ | | 0.146 | 0.094 | 0.067 | 0.022 | 0.528 | 0.541 | 0.273 | 0.173 | 0.162 | 0.110 | 0.568 | 0.602 |
| TTT3R (Chen et al., 2026) | × | *300* | 0.030 | 0.016 | 0.019 | **0.004** | 0.558 | 0.588 | 0.057 | 0.035 | 0.016 | **0.003** | 0.595 | 0.650 |
| | ✓ | | 0.030 | 0.016 | 0.019 | **0.004** | **0.559** | 0.589 | 0.052 | 0.032 | 0.015 | **0.003** | 0.599 | 0.656 |
| | × | *400* | 0.044 | 0.026 | 0.024 | **0.004** | 0.551 | 0.577 | 0.093 | 0.053 | 0.018 | **0.003** | 0.587 | 0.635 |
| | ✓ | | 0.039 | 0.023 | 0.025 | **0.004** | 0.552 | 0.578 | 0.078 | 0.042 | 0.016 | **0.003** | 0.592 | 0.644 |
| | × | *500* | 0.068 | 0.046 | 0.033 | 0.009 | 0.542 | 0.562 | 0.127 | 0.061 | 0.033 | **0.003** | 0.586 | 0.635 |
| | ✓ | | 0.057 | 0.039 | 0.030 | 0.008 | 0.546 | 0.568 | 0.105 | 0.048 | 0.026 | 0.004 | 0.586 | 0.633 |
| TTSA3R (Zheng et al., 2026) | × | *300* | 0.023 | 0.011 | 0.018 | **0.004** | 0.558 | 0.588 | 0.039 | **0.022** | 0.011 | **0.003** | **0.606** | **0.669** |
| | ✓ | | **0.022** | **0.009** | **0.017** | **0.004** | **0.559** | **0.588** | **0.037** | **0.022** | **0.010** | **0.003** | 0.605 | 0.668 |
| | × | *400* | 0.030 | 0.016 | 0.022 | **0.004** | 0.553 | 0.580 | 0.060 | **0.027** | **0.010** | **0.003** | **0.598** | **0.655** |
| | ✓ | | **0.025** | **0.012** | **0.021** | **0.004** | **0.554** | **0.581** | **0.059** | **0.027** | **0.010** | **0.003** | 0.596 | 0.651 |
| | × | *500* | 0.045 | 0.029 | 0.025 | **0.004** | 0.545 | 0.567 | 0.085 | 0.034 | 0.020 | **0.003** | **0.596** | **0.651** |
| | ✓ | | **0.035** | **0.021** | **0.023** | **0.004** | **0.548** | **0.571** | **0.081** | **0.032** | **0.014** | **0.003** | 0.595 | 0.649 |

### 4.1 Long-sequence 3D Reconstruction

To test whether preservation improves accumulated geometry under bounded recurrent memory, we evaluate streams of 300/400/500 consecutive frames on 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022), following the reconstruction protocols of prior work (Wang et al., 2025b; Chen et al., 2026; Zhuo et al., 2026; Yuan et al., 2026). We report accuracy and completeness errors and normal consistency.

MeMix lowers or matches 7-Scenes accuracy error at every tested length and improves completeness in most settings (Table 2). On NRGBD, mean accuracy error improves across all three update rules, while some completeness and normal-consistency entries are unchanged or worse. VGGT and StreamVGGT run out of memory at all three lengths under this protocol.

**Thousand-view reconstruction.** On all 16 official 7-Scenes test streams with at least 1,000 valid views, we evaluate the first 1,000 consecutive views. MeMix lowers TTSA3R's accuracy error from 0.08924 to 0.06254 m (29.9%) and completeness error from 0.02916 to 0.02020 m (30.7%). Section 4.4 tests alternative preservation rules, and Appendix A4 gives the protocol and statistical details.

Appendix Fig. 13 provides qualitative reconstruction comparisons, and Appendix Table 7 reports the complementary sparse-sampling evaluation.

### 4.2 Long-sequence Camera Pose Estimation

Pose evaluation tests whether the geometry gains are accompanied by lower accumulated trajectory drift. Following prior protocols (Wang et al., 2025b; Chen et al., 2026; Zheng et al., 2026; Shen et al., 2025a), Fig. 3 reports absolute trajectory error (ATE) against input length on TUM and ScanNet. Across 400–1000 views, MeMix reduces ATE by 13.2% on TUM and 10.0% on ScanNet on average across CUT3R, TTT3R, and TTSA3R.

Appendix Table 8 reports the separate short-sequence pose benchmarks.

**Does the benefit persist for thousands of views?** On five ScanNet scenes with at least 3,000 consecutive frames and complete ground-truth poses, MeMix reduces TTSA3R's mean ATE by 31.6%, 13.5%, and 8.33% at consecutive 1,000-, 2,000-, and 3,000-view prefixes (Fig. 4). The ATE

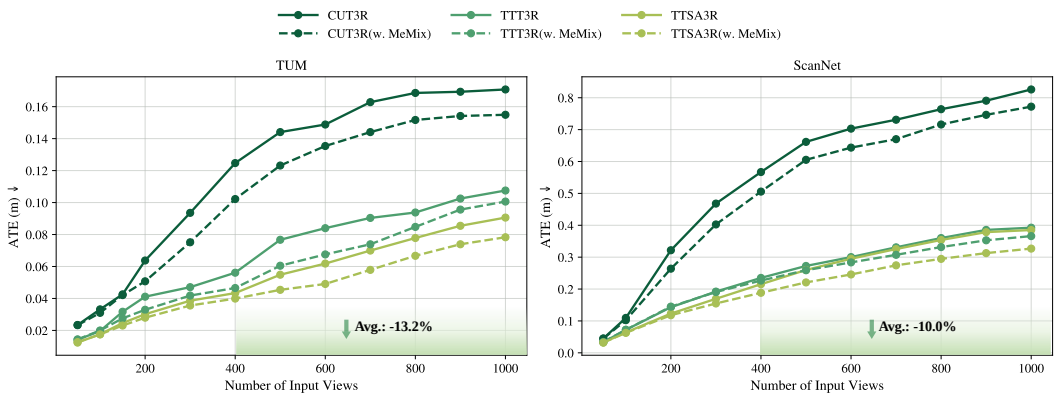


Figure 3: **Long-sequence pose estimation on TUM and ScanNet.** We compare CUT3R, TTT3R, and TTSA3R with their MeMix variants on long input streams, and report the absolute trajectory error (ATE) as the number of input views increases.

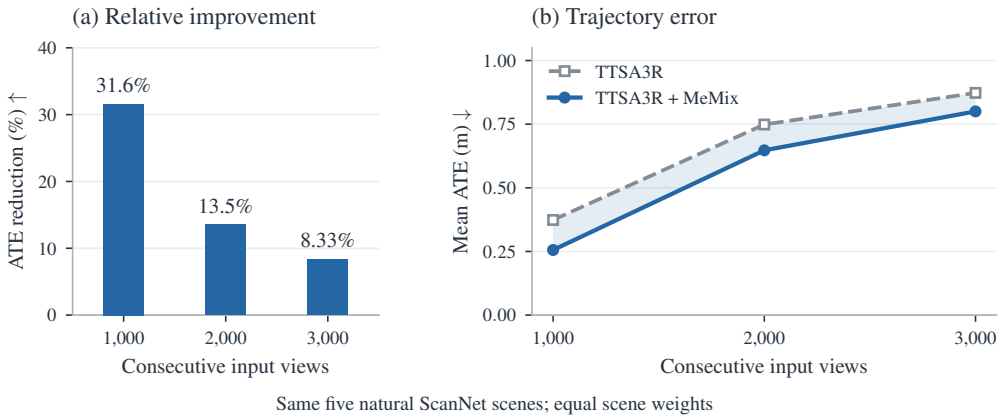


Figure 4: **Natural ScanNet streams through 3,000 views.** (a) Reduction in mean ATE relative to TTSA3R. (b) Mean ATE on the same five scenes and trajectory runs at each prefix. Both panels use consecutive natural frames and equal scene weights. Shading marks the mean difference, not uncertainty.

benefit persists at all three prefixes, with a narrowing relative gain. Appendix A5 gives the cohort, protocol, and uncertainty.

**Does preservation help a covariance-aware update?** On all 65 ScanNet scenes with at least 1,000 consecutive frames and valid ground-truth poses, adding MeMix to FILT3R lowers mean ATE from 0.28861 to 0.23719 m at 1,000 views (17.8%). The paired analysis supports the ATE benefit from 600 views onward. Appendix A4, Fig. 12, provides the full prefix curves, including early-prefix costs, together with the protocol, uncertainty, and complementary TUM and RPE results.

### 4.3 Long-sequence Video Depth Estimation

Figure 5 tests whether preservation improves dense framewise depth on KITTI (Geiger et al., 2013) over streams of up to 1,000 views. For CUT3R and TTT3R, MeMix lowers absolute relative error and raises threshold accuracy $\delta_{1.25}$ on longer streams; the TTSA3R curves remain close to the baseline.

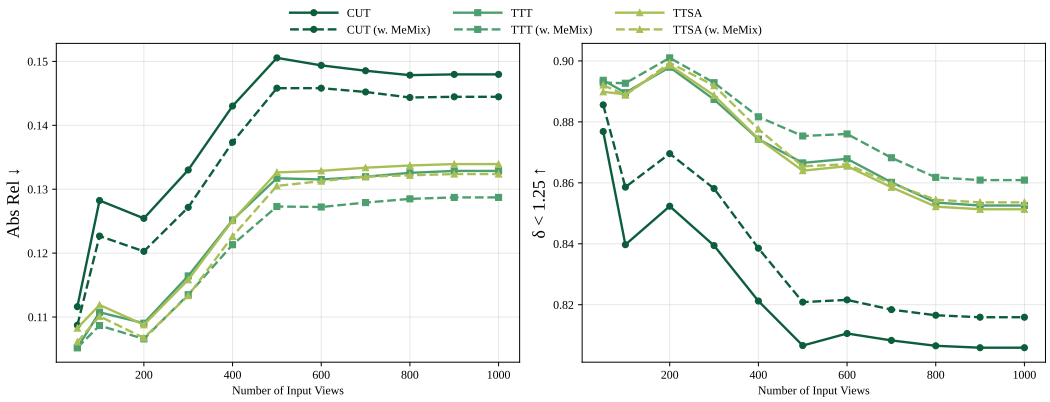


Figure 5: **Video depth estimation on KITTI versus input length.** CUT3R, TTT3R, and TTSA3R with and without MeMix. Left: absolute relative depth error (lower is better). Right: threshold accuracy $\delta_{1.25}$ (higher is better, shown as a fraction).

### 4.4 Where and When to Preserve

**Location selection matters under the same write rule.** Table 3 uses 16 7-Scenes streams at 1,000 views and eight TUM streams at $\min(1000, \text{available})$; MeMix has the lowest mean errors. Rank-hard replaces native gates with zeros at 60 locations and ones elsewhere; literal hard thresholds them at 0.5. Both change allocation and write strength. Reverse Top-k, Periodic, and Random instead vary allocation at fixed $p = 60$, retaining native soft writes elsewhere.

Table 3: **Ablations on long-horizon preservation allocation.** Shared CUT3R checkpoint and TTSA3R update rule, varying the preservation rule; 16 7-Scenes streams at 1,000 views and eight TUM streams up to 1,000 views. Errors are in meters; Random averages three seeds within stream.

| | | 7-Scenes | | TUM |
|---|---|---|---|---|
| Strategy | Preserved | Acc.↓ | Comp.↓ | ATE↓ |
| Rank-hard | 60 | 0.2199 | 0.0751 | 0.1641 |
| Literal hard@0.5 | Variable | 0.0800 | 0.0294 | 0.1086 |
| Reverse Top-k | 60 | 0.1824 | 0.0651 | 0.1643 |
| Periodic | 60 | 0.0759 | 0.0251 | 0.0868 |
| Random | 60 | 0.0800 | 0.0257 | 0.0909 |
| **MeMix** | 60 | **0.0625** | **0.0202** | **0.0767** |

Moving protection from ranks 1–60 to 709–768 at fixed $p = 60$ raises ATE from 0.0767 to 0.1643 (Fig. 6a). The highest-scoring window gives the lowest mean ATE.

**Selected-write strength.** At fixed selection rule and $p = 60$, we vary Eq. (2) via

$$G_{t,i}^{(\lambda)} = [M_{t,i} + \lambda(1 - M_{t,i})]\beta_{t,i}, \qquad 0 \le \lambda \le 1. \tag{11}$$

Here $\lambda = 0$ preserves exactly and $\lambda = 1$ restores native writes. Zero minimizes tested TUM ATE and improves both 7-Scenes errors over full writes; zero-versus-10% differences on 7-Scenes remain unresolved (Appendix Table 5).

**Candidate compatibility provides a useful allocation signal.** Candidate scores include proposed-change alignment (Eq. (10)). At $p = 60$, old-state scoring raises TUM ATE $0.07671 \to 0.11188$ under TTSA3R and $0.13634 \to 0.18146$ under CUT3R (Appendix Fig. 9). Protecting the 60 largest candidate displacements gives TTSA3R ATE 0.09565 (Appendix A4).

**Current-frame reselection improves on fixed and stale protection.** With candidate scoring, the TTSA3R update, and $p = 60$ fixed, we next vary when protection is reselected. Fixing the initial mask raises ATE $0.07671 \to 0.09787$ m; using the previous step's mask gives 0.10252 m. Reselection every 16 or 64 writes also raises ATE (Fig. 6b).

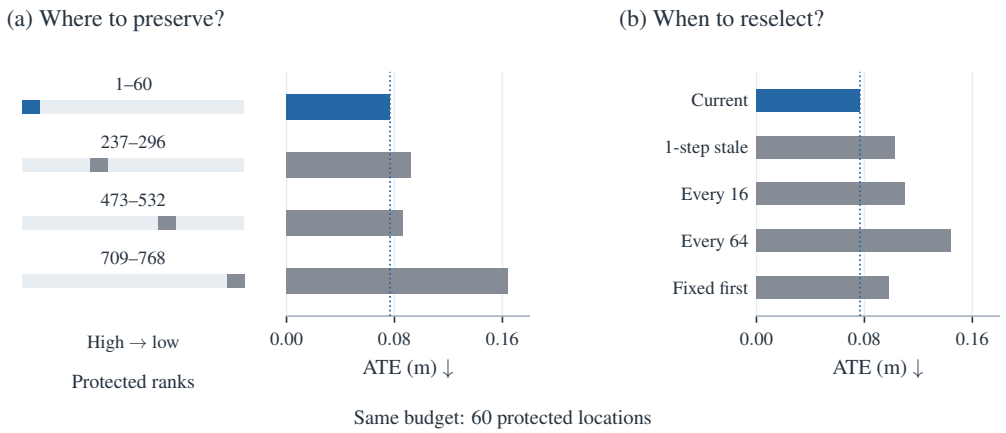


Figure 6: **Where to preserve and when to reselect.** (a) Protecting 60 of 768 locations in different candidate-compatibility rank windows. (b) Varying the mask schedule. Both panels show mean ATE over eight TUM streams at up to 1,000 views; blue and dotted lines mark MeMix.

Complementing these task-level controls, Fig. 7 shows cumulative refresh counts per memory location. In these examples, `Top-k` repeatedly updates a subset while leaving others rarely refreshed; `Bottom-k` spreads writes more evenly. Thus, retaining selected old values at each step still allows broad state refresh over time. Appendices A2 and A4 give further ablations and protocol details.

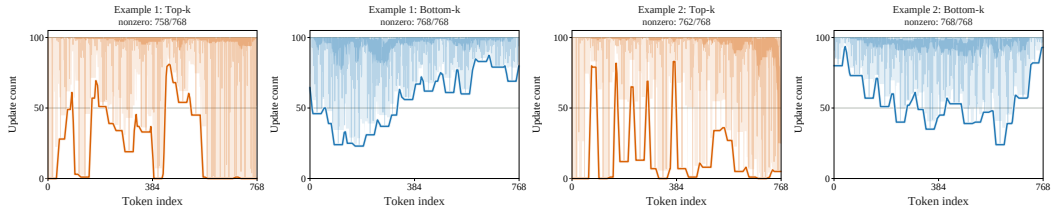


Figure 7: **Refresh counts per memory location.** Two representative CUT3R sequences from 7-Scenes, each comparing `Top-k` and `Bottom-k` routing.

## 5 Conclusion

The central finding is that long-horizon performance depends on both where recurrent-state writes are withheld and when the protected subset is reselected. Matched interventions favor compatibility-guided, current-frame selection over gate binarization, score-agnostic allocation, and fixed or stale protection. MeMix implements this allocation without retraining or increasing recurrent memory, improving reconstruction and pose; the pose benefit also extends to the covariance-aware FILT3R update on ScanNet.

**Limitations.** Although MeMix improves long-horizon reconstruction and pose estimation, kilometer-scale trajectories remain unevaluated. Exploratory longer runs also show a mean ATE regression relative to the baseline. Moreover, the `Bottom-k` selection rule uses compatibility as an empirical proxy for long-term memory utility, without explicitly accounting for scene geometry. Future work may develop more principled, geometry-aware, or scene-adaptive refresh policies for selective state preservation.

### AI USE STATEMENT

Generative AI tools were used to polish the language and improve the readability of the manuscript, and to assist with code development. The authors take full responsibility for the final manuscript, code, and reported results.

## REPRODUCIBILITY STATEMENT

Section 3.2 specifies the routing rule and its composition with recurrent updates. Appendix A1 provides pseudocode, theoretical derivations, and their assumptions. Appendix A4 details the controls, preservation budgets, paired resampling, and FILT3R protocol; Appendix A5 gives the 3,000-view ScanNet protocol and uncertainty.

## REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6290–6301, 2022.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012.

Yuan Cao and Dong Wang. SAGA: Selective adaptive gating for efficient and expressive linear attention. *arXiv preprint arXiv:2509.12817*, 2025.

Wenhao Chai and Weili Xu. View transformer layers from online optimization perspective. *Preprint*, 2025. URL `https://wenhaochai.com/assets/file/blog/mesa.pdf`.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9158–9168, October 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3R: Long-sequence streaming 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5273–5284, 2025b.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Jusen Du, Weigao Sun, Disen Lan, Jiaxi Hu, and Yu Cheng. MoM: Linear sequence modeling with mixture-of-memories. *arXiv preprint arXiv:2502.13685*, 2025.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Seonghyun Jin and Jong Chul Ye. FILT3R: Latent state adaptive Kalman filter for streaming 3D reconstruction, 2026. URL https://arxiv.org/abs/2603.18493.

Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1611–1621, 2021. doi: 10.1109/CVPR46437.2021.00166.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STREAM3R: Scalable sequential 3D reconstruction with causal transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91, Cham, 2024. Springer. ISBN 978-3-031-73219-5. doi: 10.1007/978-3-031-73220-1_5.

Yingcong Li, Davoud Ataee Tarzanagh, Ankit Singh Rawat, Maryam Fazel, and Samet Oymak. Gating is weighting: Understanding gated linear attention through in-context learning. *arXiv preprint arXiv:2504.04308*, 2025.

Liewen Liao, Weihao Yan, Wang Xu, Ming Yang, Songan Zhang, and Hongtei Eric Tseng. Learning-based 3D reconstruction in autonomous driving: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Changkun Liu, Jiezhi Yang, Zeman Li, Yuan Deng, Jiancong Guo, and Luca Ballan. Mem3R: Streaming 3D reconstruction with hybrid memory via test-time training, 2026. URL https://arxiv.org/abs/2604.07279.

Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3R: Training-free token eviction for memory-bounded streaming visual geometry transformers, 2025. URL https://arxiv.org/abs/2509.17650.

Dharmendra Selvaratnam and Dena Bazazian. 3D reconstruction in robotics: A comprehensive review. *Computers & Graphics*, 130:104256, 2025.

Guole Shen, Tianchen Deng, Xingrui Qin, Nailin Wang, Jianyu Wang, Yanbo Wang, Yongtao Chen, Hesheng Wang, and Jingchuan Wang. MUT3R: Motion-aware updating transformer for dynamic 3D reconstruction. *arXiv preprint arXiv:2512.03939*, 2025a.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025b.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773.

Xianbing Sun, Zhikai Zhu, Zhengyu Lou, Bo Yang, Jinyang Tang, Liqing Zhang, He Wang, and Jianfu Zhang. AVGGT: Rethinking global attention for accelerating VGGT. *arXiv preprint arXiv:2512.02541*, 2025.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89, 2025. doi: 10.1109/3DV66043.2025.00013.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. CUT3R: Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Zipeng Wang and Dan Xu. FlashVGGT: Efficient and scalable visual geometry transformers with compressed descriptor attention. *arXiv preprint arXiv:2512.01540*, 2025.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025. ICLR 2025 Spotlight.

Yu Zhang, Songlin Yang, Rui-Jie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, et al. Gated slot attention for efficient linear-time sequence modeling. *Advances in Neural Information Processing Systems*, 37:116870–116898, 2024.

Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T. Freeman. Structure and motion from casual videos. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13693 of *Lecture Notes in Computer Science*, pp. 20–37, Cham, 2022. Springer. ISBN 978-3-031-19827-4. doi: 10.1007/978-3-031-19827-4_2.

Zhijie Zheng, Xinhao Xiang, and Jiawei Zhang. TTSA3R: Training-free temporal-spatial adaptive persistent state for streaming 3D reconstruction. *arXiv preprint arXiv:2601.22615*, 2026.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546, October 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. StreamVGGT: Streaming visual geometry transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.

# SUPPLEMENTARY MATERIAL

## A1. METHOD, ALGORITHM, AND THEORETICAL ANALYSIS

**Readout details.** The image stream starts from projected encoder tokens and interacts with the state stream through the dual-stream decoder. Its final tokens $\mathbf{Y}_t$ feed the prediction head before the state write.

**Inference procedure.** Blue lines are base-model operations; red lines add routing and selective writing. The pseudocode shows direct CUT3R updates. For TTT3R/TTSA3R, replace $M_t$ in the final write by $M_t \odot \beta_t$. The routing feature $\mathbf{g}_t$ is defined in Sec. 3.2.

Constant recurrent memory excludes storage of all outputs in $\mathcal{O}$; they may instead be emitted frame by frame.

**MeMix inference pseudocode.**

**Require:** Input stream $\{I_t\}_{t=1}^{T}$, base model $f$, refresh budget $k = 708$

$S_0 \leftarrow \texttt{InitialState}(f)$
$\mathcal{O} \leftarrow \emptyset$
**for** $t = 1$ to $T$ **do**
  $X_t \leftarrow \texttt{Tokenize}(I_t)$
  $\hat{S}_t, Y_t \leftarrow \texttt{Decoder}(S_{t-1}, X_t)$
  $g_t \leftarrow \texttt{Mean}(W_{\text{dec}} X_t)$
  $r_t \leftarrow \texttt{RouteScore}(\hat{S}_t, g_t)$
  $\mathcal{U}_t \leftarrow \texttt{BottomK}_k(r_t)$
  $M_t \leftarrow \texttt{Gate}(\mathcal{U}_t)$
  $(T_t, K_t, P_t) \leftarrow \texttt{Head}(Y_t)$
  $\mathcal{O} \leftarrow \mathcal{O} \cup \{(T_t, K_t, P_t)\}$
  $S_t \leftarrow M_t \odot \hat{S}_t + (1 - M_t) \odot S_{t-1}$
**end for**
**return** $\mathcal{O}$

The full readout, including self-attention, cross-attention, residual connections, normalization, and MLPs, is denoted by $F_Y$:

$$\mathbf{Y}_t = F_Y(\mathbf{S}_{t-1}, \mathbf{X}_t). \tag{12}$$

The decoded tokens are fed to the prediction head:

$$(\mathbf{T}_t,\ \mathbf{K}_t,\ \mathbf{P}_t) = \texttt{Head}(\mathbf{Y}_t). \tag{13}$$

**Base-model candidate and gate details.** The candidate-state and readout streams are produced by an $L$-layer dual-stream Transformer before the final recurrent write. To illustrate the interaction, the state-stream cross-attention sublayer has the form

$$\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \tag{14}$$

where queries come from state features, keys/values from image features, and $d_h$ is the head dimension. The full decoder also contains self-attention, normalization, residual connections, and MLPs. Writing its complete state proposal as $F_S$ avoids equating a cross-attention sublayer with the entire decoder:

$$\hat{\mathbf{S}}_t = F_S(\mathbf{S}_{t-1}, \mathbf{X}_t) = \mathbf{S}_{t-1} + \Delta\mathbf{S}_t, \qquad \mathbf{S}_t = \hat{\mathbf{S}}_t. \tag{15}$$

This full refresh can overwrite reliable state content with new-frame evidence over long horizons.

TTT3R (Chen et al., 2026) instead derives a token-wise soft update gate $\boldsymbol{\beta}_t \in [0,1]^{n\times 1}$ from the decoder attention:

$$\boldsymbol{\beta}_{t,i} = \sigma\left(\frac{1}{LHm\sqrt{d_h}} \sum_{\ell,h,j} Q^{(\ell,h)}_{S_{t-1},i} \cdot K^{(\ell,h)^T}_{X_t,j}\right), \quad i = 1, \ldots, n, \tag{16}$$

where $H$ is the number of heads and $m$ the number of image tokens. Its state update is

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{\beta}_t \odot (\hat{\mathbf{S}}_t - \mathbf{S}_{t-1}). \tag{17}$$

This dense gate modulates update magnitude without explicitly selecting an exactly preserved subset.

**Local identity path at protected locations.** For the direct and soft-gated rules, stack the state-token vectors into $\mathbf{s} \in \mathbb{R}^{nd}$, and let $F_t(\mathbf{s})$ be the complete native state transition at fixed input frame $t$, including candidate generation and the native gate. Let $D_t$ be the diagonal protection matrix, with each selected-location indicator repeated across its $d$ channels. The MeMix transition is

$$F_t^{\text{MeMix}}(\mathbf{s}) = D_t\mathbf{s} + (I - D_t)F_t(\mathbf{s}). \tag{18}$$

At a state where the score ranking has a strict gap between the $p$-th and $(p+1)$-th locations, continuous scores give a neighborhood with unchanged selection. Wherever $F_t$ is differentiable in this neighborhood, its Jacobian $J_{F_t}$ and the MeMix Jacobian $J_t$ satisfy

$$J_t = D_t + (I - D_t)J_{F_t}, \qquad D_t J_t = D_t. \tag{19}$$

The second identity follows from $D_t^2 = D_t$ and $D_t(I - D_t) = 0$. Thus the protected rows transmit their own input values through an identity path. For a perturbation $\mathbf{v}$ supported only outside the protected locations, $D_t\mathbf{v} = 0$ implies $D_t J_t \mathbf{v} = 0$: it cannot enter the protected values through this local state write. Candidate dependence on the full previous state is included in $J_{F_t}$, rather than held fixed. The identity applies while selection is unchanged; recomputing the mask can change which rows are protected.

**Conditional transition analysis.** For location $i$, let $G_{t,i} \in [0,1]$ be the scalar refresh gate in Eq. (2), broadcast over feature channels, and write $\Delta_{t,i} = \hat{\mathbf{S}}_{t,i} - \mathbf{S}_{t-1,i}$. Then

$$\mathbf{S}_{t,i} = \mathbf{S}_{t-1,i} + G_{t,i}\Delta_{t,i}.$$

Holding the realized candidate sequence fixed, repeated substitution gives

$$\mathbf{S}_{t,i} = a_{\tau:t,i}\mathbf{S}_{\tau,i} + \sum_{s=\tau+1}^{t} b_{s:t,i}\hat{\mathbf{S}}_{s,i},$$

$$a_{\tau:t,i} = \prod_{v=\tau+1}^{t}(1 - G_{v,i}), \qquad b_{s:t,i} = G_{s,i}\prod_{v=s+1}^{t}(1 - G_{v,i}).$$

The explicit carry coefficient $a_{\tau:t,i}$ describes this decomposition, not total historical information. For example, if $\hat{\mathbf{S}}_{t,i} = \mathbf{S}_{t-1,i}$ and $G_{t,i} = 1$, the explicit carry coefficient is zero although the state is unchanged. Candidates themselves depend on previous states, so changing a gate also changes future candidates; the fixed-candidate identity is not a causal comparison between two recurrent rollouts.

If $G_{v,i} \geq \underline{G} > 0$ for $L = t - \tau$ steps, then

$$a_{\tau:t,i} \leq (1 - \underline{G})^L \leq e^{-\underline{G}L}.$$

If $G_{v,i} = 0$ throughout the interval, then $a_{\tau:t,i} = 1$, all $b_{s:t,i} = 0$, and $\mathbf{S}_{t,i} = \mathbf{S}_{\tau,i}$ exactly. A full candidate replacement cuts this explicit path; a nonzero soft write attenuates it. Under dynamic routing, the exact carry lasts while the location remains protected.

For the same pre-state, candidate, and base gate $\beta_{t,i}$, the local write-energy reduction from MeMix is exactly

$$\mathcal{W}_t^{\text{base}} - \mathcal{W}_t^{\text{MeMix}} = \frac{1}{n}\sum_{i \in \mathcal{R}_t} \beta_{t,i}^2 \|\Delta_{t,i}\|_2^2,$$

where $\mathcal{W}_t = n^{-1}\sum_i \|\mathbf{S}_{t,i} - \mathbf{S}_{t-1,i}\|_2^2$ and $\beta_{t,i} = 1$ for direct CUT3R writes. This identity quantifies the write energy removed at protected locations. With $p = 60$ and $n = 768$, MeMix protects 7.8% of locations and leaves 92.2% on the base update path. This sparse preservation keeps most locations writable; Appendix A4 evaluates its task-level effect through matched interventions.

### A2. Implementation-level Ablations

Table 4 compares score functions, update strategies, feature sources, and routing direction under the short-sequence protocol. These choices complement the main preservation-allocation controls in Table 3; their different update rules and sequence protocols require within-table comparisons.

**Memory-location selection strategy.** `Top-k` refreshes the $k$ highest-score memory locations; `Bottom-k` refreshes the $k$ lowest and preserves the high-score complement of size $p = n - k$. These names refer to the refreshed set, not the preserved set. When a memory location groups multiple state tokens, their scores are averaged before selection. By default, each of the 768 memory locations corresponds to one state token.

**Scoring function.** We evaluate cosine similarity, dot product (default), and attention-derived scores as routing signals:

$$r_{t,i}^{\cos} = \left\langle \frac{\hat{\mathbf{S}}_{t,i}}{\|\hat{\mathbf{S}}_{t,i}\|_2}, \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|_2} \right\rangle, \quad r_{t,i}^{\text{dot}} = \langle \hat{\mathbf{S}}_{t,i}, \mathbf{g}_t \rangle, \quad i = 1, \ldots, n. \tag{20}$$

Table 4: **Ablations on preservation routing and score design.** The Default row corresponds to `Bottom-k` refresh routing + `Dot` + score($\hat{\mathbf{S}}_t, \mathbf{g}_t$), which preserves the high-score complement; other rows change one component at a time. We use KITTI 110-frame depth, TUM-dynamics 90-frame pose, and NRGBD-S 300-frame reconstruction settings. ATE, translational RPE, accuracy, and completeness are in meters; rotational RPE is in degrees and $\delta_{1.25}$ in percent.

| | Depth@KITTI | | Camera@TUM | | | 3R@NRGBD | | |
|---|---|---|---|---|---|---|---|---|
| **Variant** | AbsRel↓ | $\delta_{1.25}$↑ | ATE↓ | $\text{RPE}_\text{t}$↓ | $\text{RPE}_\text{r}$↓ | Acc↓ | Comp↓ | NC↑ |
| Default (TTT3R + MeMix) | 0.103 | **92.1** | **0.028** | **0.013** | 0.376 | **0.099** | **0.020** | **0.616** |
| **Location selection** | | | | | | | | |
| Top-k | **0.102** | 91.8 | 0.068 | 0.021 | 0.595 | 0.178 | 0.049 | 0.587 |
| **Scoring function** | | | | | | | | |
| Cosine ($r_t^{\cos}$) | 0.105 | 91.6 | **0.028** | **0.013** | **0.375** | **0.099** | 0.023 | **0.616** |
| Attn ($r_t^{\text{attn}}$) | 0.107 | 90.7 | 0.030 | 0.014 | 0.418 | 0.105 | 0.025 | 0.611 |
| **Update strategy** | | | | | | | | |
| Layer-wise routing | 0.108 | 91.4 | 0.035 | 0.015 | 0.443 | 0.127 | 0.034 | 0.604 |
| No-update | 0.114 | 89.0 | 0.162 | 0.066 | 1.624 | 0.461 | 0.672 | 0.528 |
| **Routing score** | | | | | | | | |
| score($\mathbf{S}_{t-1}, \mathbf{g}_t$) | 0.107 | 91.1 | 0.039 | 0.016 | 0.437 | 0.111 | 0.027 | 0.608 |
| score($\mathbf{S}_{t-1}, \bar{\mathbf{Y}}_t$) | 0.124 | 85.9 | 0.041 | 0.016 | 0.426 | 0.249 | 0.055 | 0.575 |
| score($\hat{\mathbf{S}}_t, \bar{\mathbf{Y}}_t$) | 0.128 | 84.7 | 0.046 | 0.046 | 0.465 | 0.308 | 0.112 | 0.573 |

Here $\hat{\mathbf{S}}_{t,i}$ is the candidate state at memory location $i$ and $\mathbf{g}_t$ is the image-level routing feature obtained by projecting and averaging the current-frame encoder tokens; these location-wise scores form $r_t \in \mathbb{R}^n$. For TTT3R-style attention routing (Chen et al., 2026), we use aggregated decoder cross-attention.

**Refresh strategy.** The default single-update rule acts after the decoder and retains the native gate $\boldsymbol{\beta}_t$:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + (\mathbf{M}_t \odot \boldsymbol{\beta}_t) \odot (\hat{\mathbf{S}}_t - \mathbf{S}_{t-1}). \tag{21}$$

Here $\boldsymbol{\beta}_t$ is the TTT3R gate for the default in Table 4; setting it to one recovers direct CUT3R writes. We compare this with layer-wise routing, whose intra-decoder masking operation is

$$\mathbf{S}_t^{(\ell+1)} = \mathbf{M}_t^{(\ell)} \odot \hat{\mathbf{S}}_t^{(\ell+1)} + (1 - \mathbf{M}_t^{(\ell)}) \odot \mathbf{S}_t^{(\ell)}, \tag{22}$$

together with a no-update (freeze) control.

**Feature source for routing.** Let $\mathbf{S}_{t-1}$ be the pre-update state, $\hat{\mathbf{S}}_t$ the decoder-final candidate state, $\mathbf{g}_t$ the current-frame routing feature used by the default score, and $\bar{\mathbf{Y}}_t$ the mean-pooled decoded image feature considered in routing-score ablations. We compare score($\mathbf{S}_{t-1}, \mathbf{g}_t$), score($\hat{\mathbf{S}}_t, \bar{\mathbf{Y}}_t$), score($\hat{\mathbf{S}}_t, \mathbf{g}_t$), and score($\mathbf{S}_{t-1}, \bar{\mathbf{Y}}_t$).

**Preservation budget.** The preservation budget trades continuity against adaptation. Fig. 8 plots the equivalent refresh budget $k = n - p$: increasing $p$ protects more old-state locations but leaves fewer available for new observations. The fixed default $p = 60$ keeps most locations writable. In contrast, the rank intervention in Fig. 6 holds $p = 60$ fixed to isolate which locations should be preserved.

### A3. Preservation Budget

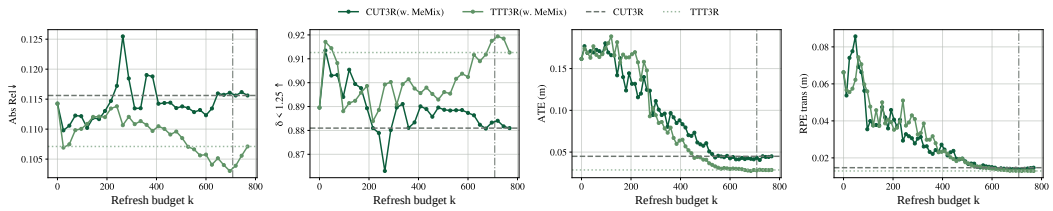


Figure 8: **Sensitivity to preservation budget.** We sweep the refresh budget $k = n - p$ for CUT3R and TTT3R with/without MeMix. First two panels: KITTI depth (110-frame streams), reporting AbsRel (lower is better) and threshold accuracy $\delta_{1.25}$ (higher is better). Last two panels: TUM-dynamics pose (90-frame streams), reporting ATE and translational RPE (lower is better). Vertical dash-dotted lines mark the default $k = 708$, which preserves 60 of 768 locations.

### A4. Long-horizon Controls and Mechanism Diagnostics

The controls below distinguish allocation, score source, routing cadence, and write strength. TUM endpoint results use $\min(1000, \text{available})$ views per stream. Paired uncertainty is estimated using 95% percentile bootstraps over evaluation streams (100,000 resamples), with streams rather than frames as the sampling unit. Random seeds are averaged within stream before resampling. Prefix comparisons are assessed pointwise.

**Hardness, allocation, and budget.** Rank-hard sets the 60 smallest final TTSA3R gate values to zero and the other 708 to one. Literal hard sets each gate to one if it is at least 0.5 and to zero otherwise. These are binary replacements of the soft write gate. In contrast, MeMix, Reverse Top-k, Periodic, and Random multiply the original soft gate by a binary preservation mask, keeping the update strength at writable locations. Reverse Top-k protects the 60 lowest compatibility scores. Periodic protects a contiguous cyclic block of 60 locations and advances it by 60 at each routed write, modulo 768. Random draws a fresh size-60 subset without replacement at each routed write, using seeds 0, 1, and 2. Thus these three controls isolate allocation under the MeMix budget and write rule, whereas the hard-gate controls also change write strength.

For reference, the unmodified soft TTSA3R update on the same long-horizon protocols yields 7-Scenes accuracy/completeness errors of $0.0892/0.0292$ m and TUM ATE of $0.0951$ m. MeMix reduces these by 29.9%/30.7% and 19.3%, respectively.

The paired controls in Table 3 test gate binarization and alternative allocations; Eq. (11) separately varies selected-write strength without replacing native gates. Unlike the 60-location Rank-hard control, literal thresholding at 0.5 preserves 750.9/768 TUM locations per frame on average, so it tests gate binarization with an input-dependent budget.

The useful preservation budget varies by task. For $p = 30/60/120$, mean TUM ATE is $0.08454/0.07671/0.08301$, while 1,000-view 7-Scenes accuracy error is $0.0725/0.0625/0.0613$. The fixed default $p = 60$ is best on TUM and lies on the $p = 60$–$120$ plateau on 7-Scenes, supporting a small preserved subset.

**Interpretation of exact preservation.** The conditional carry decomposition and local write-energy identity in Appendix A1 describe a tradeoff between continuity and adaptation. Exact protection removes selected candidate-write terms, retaining historical values at the cost of adapting those locations to the current observation. This can be harmful when a protected old value has become stale. The diagnostics below examine how compatibility scoring and current-frame reselection affect this tradeoff.

**Which state should supply the compatibility score?** Candidate-state scoring outperforms old-state scoring under both soft-gated TTSA3R and direct CUT3R writes (Fig. 9). The comparison holds the preservation budget and evaluator fixed and changes only the state supplying the score. Its replication under direct writes shows that this advantage does not require a soft gate. All variants use the same pretrained checkpoint, preservation budget $p = 60$, and eight paired TUM streams. At the

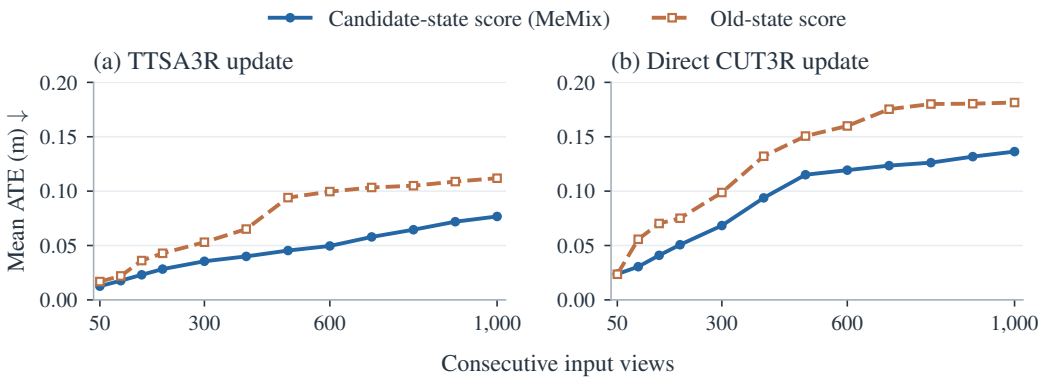


Figure 9: **Candidate versus old-state scoring.** Mean ATE on eight matched TUM streams over consecutive prefixes capped by available views. Within each update rule, only the state supplying the score changes; the checkpoint, $p = 60$, and evaluator are fixed.

longest evaluated prefix, candidate/old-state mean ATE is $0.07671/0.11188$ m under TTSA3R and $0.13634/0.18146$ m under direct CUT3R.

**Conditional retained-alignment bound.** Let $o_{t,i} = \langle \mathbf{S}_{t-1,i}, \mathbf{g}_t \rangle$ denote the old-state alignment actually retained when location $i$ is protected, and let $\mathcal{P}_r = \mathrm{TopK}_p(r_t)$. If, conditional on the observed score vector, $\mathbb{E}[o_{t,i} \mid r_t] = h(r_{t,i})$ for a location-independent nondecreasing function $h$, then $\mathcal{P}_r$ maximizes $\mathbb{E}[\sum_{i\in\mathcal{Q}} o_{t,i} \mid r_t]$ over every size-$p$ set $\mathcal{Q}$—equivalently, it minimizes the expected old-state alignment overwritten by the complementary fixed-budget write. A deterministic stability form makes the assumption explicit: if $|o_{t,i} - h(r_{t,i})| \le \varepsilon_t$ for every location and $\mathcal{P}_o = \mathrm{TopK}_p(o_t)$ is the old-alignment oracle, then

$$0 \le \sum_{i\in\mathcal{P}_o} o_{t,i} - \sum_{i\in\mathcal{P}_r} o_{t,i} \le 2p\varepsilon_t. \tag{23}$$

The upper bound follows by replacing $o_{t,i}$ with $h(r_{t,i})$ on both sets and using the top-$p$ optimality of $\mathcal{P}_r$ for the middle sum. This bounds an old-alignment surrogate, not reconstruction accuracy. Direct old-state scoring maximizes that surrogate on each frame, yet candidate scoring performs better in the task intervention; the surrogate is therefore not a sufficient account of task utility. Across 7,023 TUM frames, candidate/old compatibility has mean per-frame Pearson $0.693$, the candidate-selected set has higher old-state compatibility on $79.3\%$ of frames, and candidate/old Top-60 overlap is above random ($0.125$ versus $0.041$).

The observed candidate-to-old association is not strictly monotone and does not certify the location-independent monotone-$h$ premise or a uniform $\varepsilon_t$ bound; the guarantee remains explicitly conditional.

**Compatibility or direct rewrite magnitude?** The state analysis shows larger would-be candidate displacements at protected locations. To test whether compatibility simply selects the largest rewrites, Fig. 10 compares it with protecting the 60 largest or 60 smallest values of $\|\hat{\mathbf{S}}_{t,i} - \mathbf{S}_{t-1,i}\|_2$, holding the TTSA3R rule, $p = 60$, eight TUM streams, and evaluator fixed. At the longest evaluated prefix, compatibility routing achieves mean ATE of $0.07671$ m, compared with $0.09565$ m for protecting the largest rewrites and $0.08868$ m for protecting the smallest. Thus suppressing the largest candidate displacements does not reproduce the compatibility-routing benefit.

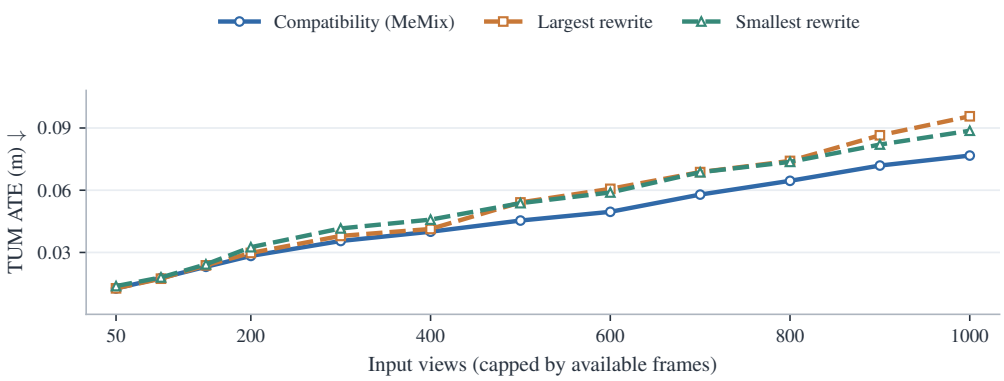


Figure 10: **Compatibility versus direct rewrite-magnitude routing.** Mean ATE over eight matched TUM streams at every evaluated prefix. Only the criterion selecting 60 protected locations changes: candidate compatibility, largest rewrite, or smallest rewrite.

**How protected locations evolve.** The protected set changes slowly between frames but covers most memory locations over time. In a 200-frame analysis, adjacent protected sets have mean Jaccard 0.701, yet protection reaches 391/768 locations. Over all eight TUM streams (7,023 frames), adjacent Jaccard is 0.737 and 754/768 locations are protected at least once. In the final progress quartile, adjacent Jaccard remains 0.756 and candidate/old compatibility correlation is 0.710. Thus protection is locally persistent while its allocation moves across the state over time.

To test whether this adaptation contributes to the task result, we change only the mask schedule, holding the TTSA3R rule, $p = 60$, streams, and evaluator fixed. Fig. 11 compares current-frame recomputation with masks recomputed every 16 or 64 routed writes and a fixed mask selected at the first routed update. At the longest evaluated prefix, mean ATE is $0.07671/0.10971/0.14435/0.09787$ m, respectively. A separate one-step-stale intervention yields 0.10252 m. Current-frame routing gives the lowest mean ATE among these schedules.

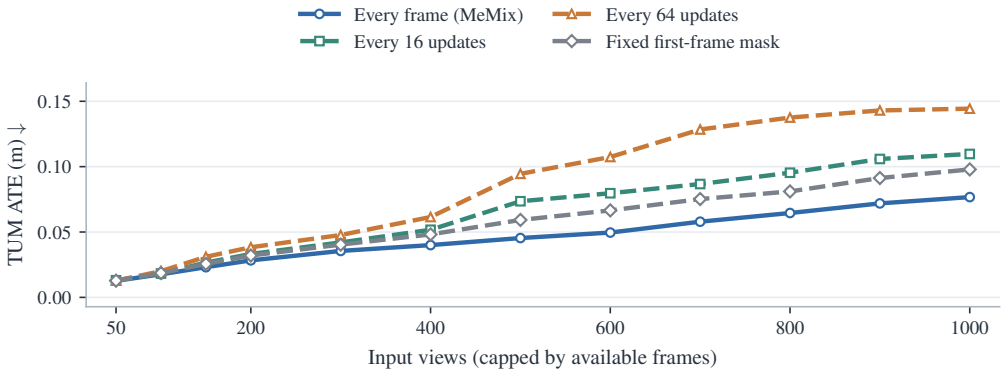


Figure 11: **Current-frame routing versus tested slower and fixed schedules.** Mean ATE over eight matched TUM streams. The compatibility mask is recomputed every frame, every 16 or 64 routed updates, or fixed after the first update. All four variants use the same checkpoint, TTSA3R rule, and 60-location budget.

**Detached image-key support.** Protected and updated memory locations differ in their detached image-key support. Across eight complete TUM streams, protected locations have, on average, more concentrated support (lower normalized entropy) and place more mass on depth discontinuities, while

placing less mass on RGB edges. These observer-only measurements characterize an association between routing and image structure.

**How much of the protected write must be suppressed?** Equation (11) keeps the compatibility selection rule, $p = 60$, and native gate computation unchanged, scaling only the native write at the selected locations. We test $\lambda \in \{0, 0.1, 0.25, 0.5, 1\}$ on eight TUM streams and $\lambda \in \{0, 0.1, 1\}$ on 16 7-Scenes streams, with comparisons paired within each dataset-specific sweep (Table 5). Each variant recomputes its selection and native gates from its own evolving state. Restoring the full write ($\lambda = 1$) is worse than exact protection on TUM ATE and both 7-Scenes reconstruction errors. On 7-Scenes, $\lambda = 0.1$ is slightly better in the mean than $\lambda = 0$, but the paired analysis does not resolve a difference. Across both tasks, strong suppression of selected writes improves on full native writes.

Table 5: **Protected-location write strength across pose and reconstruction.** Mean errors over eight TUM streams (up to 1,000 views) and 16 7-Scenes streams (1,000 views). $\lambda$ scales the native write at the selected 60 locations; the other 708 retain their native updates. Bold marks the lowest mean per column; — denotes untested settings.

| Selected-write scale $\lambda$ | TUM | 7-Scenes | |
|---|---|---|---|
| | ATE (m) ↓ | Acc. (m) ↓ | Comp. (m) ↓ |
| 0 (exact preservation) | **0.0793** | 0.0625 | 0.0202 |
| 0.1 | 0.0846 | **0.0600** | **0.0201** |
| 0.25 | 0.0940 | — | — |
| 0.5 | 0.0894 | — | — |
| 1 (full native write) | 0.0945 | 0.0892 | 0.0292 |

**Covariance-consistent preservation with FILT3R.** Equation (6) follows the scalar-per-location covariance representation in the official FILT3R (Jin & Ye, 2026) implementation. The scalar $v_{t,i}$ and gain are broadcast over a token's feature channels; no cross-location covariance terms are represented. Candidate features can still couple locations through the decoder, and the process-noise calculation uses a shared normalization statistic.

For each valid filtering step after initialization, let $\delta_{t,i} = \|\hat{\mathbf{S}}_{t,i} - \hat{\mathbf{S}}_{t-1,i}\|_2$ compare the current and previous valid candidates, not the preserved state values. The native prediction and gain are

$$
\begin{aligned}
b_t &= \max\left\{(1-\eta)b_{t-1} + \frac{\eta}{n}\sum_{i=1}^{n}\delta_{t,i},\ b_{\min}\right\}, \\
Q_{t,i} &= q_{\min} + (q_{\max} - q_{\min})\,\sigma\left(\alpha\left[\frac{\delta_{t,i}}{b_t + \epsilon} - \tau\right]\right), \\
v^{-}_{t,i} &= \gamma v_{t-1,i} + Q_{t,i}, \quad \kappa_{t,i} = \mathrm{clip}\left(\frac{v^{-}_{t,i}}{v^{-}_{t,i} + R_{t,i} + \epsilon}, \kappa_{\min}, \kappa_{\max}\right).
\end{aligned}
\tag{24}
$$

Here $\sigma$ is the sigmoid; $b_t$ is initialized from the current mean $\delta_{t,i}$ with the same floor. We retain the official settings $\eta = 0.05$, $b_{\min} = 0.01$, $(q_{\min}, q_{\max}) = (0.02, 0.5)$, $\alpha = 20$, $\tau = 3$, $\gamma = 1$, $R_{t,i} = 1$, $(\kappa_{\min}, \kappa_{\max}) = (0.01, 0.99)$, and $\epsilon = 10^{-6}$. All locations contribute to $b_t$ before masking; MeMix changes neither this calculation nor the candidate-history update.

The effective gain $\tilde{\kappa}_{t,i} = M_{t,i}\kappa_{t,i}$ is used in both the state and covariance updates in Eq. (6), with the native covariance floor $\epsilon$ retained. Thus a protected location receives no measurement write and takes $v_{t,i} = v^{-}_{t,i}$ rather than the covariance contraction associated with an assimilated observation. The other 708 locations follow the native filter update. At sequence initialization, the candidate write is masked directly and all covariances are initialized to the official value 1.5; native reset handling is unchanged. After initialization, a skipped frame leaves the carried state, covariance, normalization statistic, and previous-candidate buffer unchanged.

On eight TUM streams at $\min(1000, \text{available})$ views (7,023 views in total), adding MeMix to FILT3R increases mean ATE by 2.13% and lowers translational RPE by 8.42%. Paired analysis supports the translational-RPE improvement, while the ATE and rotational-RPE differences remain unresolved.

**FILT3R ScanNet long-horizon protocol and uncertainty.** The comparison in Fig. 12 includes all 65 scenes with at least 1,000 consecutive RGB frames and complete, finite ground-truth poses for those views. Both methods use the official FILT3R implementation with the same released checkpoint, Kalman hyperparameters, inputs, and evaluator; only the covariance-consistent $p = 60$ preservation described above changes. Each 1,000-view trajectory is evaluated at 50, 100, 150, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 views. Bootstrap intervals use 100,000 paired scene resamples and the pointwise convention above.

At 1,000 views, MeMix reduces mean ATE by $0.05142$ m. Paired analysis supports the ATE benefit from 600 views onward. At 50 views, mean ATE is $0.000183$ m higher, while differences at 100–500 views remain unresolved. At 1,000 views, translational RPE changes from $0.02042$ to $0.02002$ m, while rotational RPE changes from $0.67150°$ to $0.87528°$. Both RPE contrasts remain unresolved.

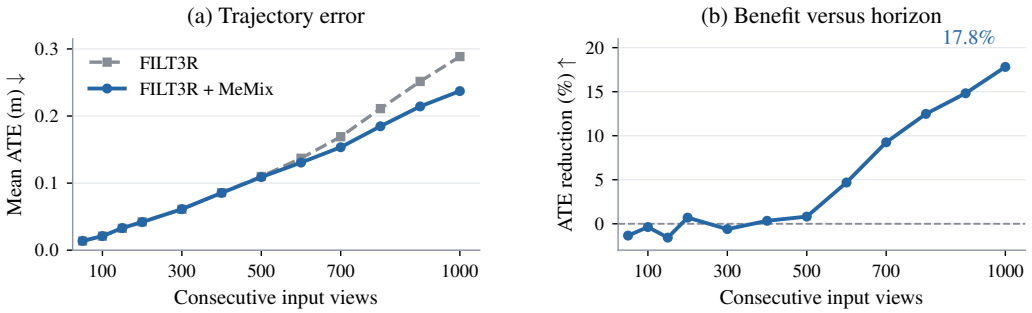


Figure 12: **Long-horizon preservation with FILT3R on ScanNet.** Left: mean ATE at all 12 evaluated prefixes on the same 65 scenes. Right: percentage reduction in mean ATE relative to FILT3R. Both methods share the checkpoint and evaluator. Early-prefix costs and unresolved RPE contrasts are reported above.

**Efficiency.** Table 6 reports the KITTI efficiency profile on a single NVIDIA A100 GPU. Reported peak GPU memory is unchanged; the small FPS differences reflect profiling variance rather than a systematic speedup.

Table 6: **Efficiency.** Inference FPS and peak GPU memory with/without MeMix on KITTI; small FPS differences reflect profiling variance.

| **Method** | **FPS** | | **GPU (GB)** | |
|---|---|---|---|---|
| **w. MeMix** | × | ✓ | × | ✓ |
| **CUT3R** | 14.39 | 14.13 | 5.31 | 5.31 |
| **TTT3R** | 12.72 | 12.81 | 6.96 | 6.96 |
| **TTSA3R** | 12.58 | 12.78 | 6.63 | 6.63 |

## A5. LONG-HORIZON SCANNET PROTOCOL

The five-scene cohort in both panels of Fig. 4 contains `scene0739_00`, `scene0747_00`, `scene0757_00`, `scene0766_00`, and `scene0785_00`. Each scene has at least 3,000 consecutive RGB frames and complete, finite ground-truth poses for those views. TTSA3R and TTSA3R+MeMix receive identical consecutive prefixes, without repeated frames or concatenated scenes. MeMix reduces mean ATE by 31.6%, 13.5%, and 8.33% at 1,000, 2,000, and 3,000 views, respectively. A paired scene-bootstrap analysis with 100,000 resamples supports the ATE reductions at all three horizons. Rotational-RPE differences remain unresolved at all three horizons, and translational RPE is mixed.

The mean ATE pairs (TTSA3R / TTSA3R+MeMix) are $0.37367/0.25566$, $0.74873/0.64730$, and $0.87290/0.80016$ m at the three horizons. The percentage reductions are computed from these same five-scene means. Each prefix is evaluated with its own Sim(3) alignment; all frames in that prefix contribute to the per-scene translation RMSE, and scenes receive equal weight in the reported mean. Both methods use the same checkpoint, inputs, and evaluator.

## A6. 3D RECONSTRUCTION

Table 7: **3D Reconstruction Results on 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022).** We test MeMix on 7-Scenes and NRGBD, with one frame sampled every two frames (Sparse Sampling, **-S**). Green boxes indicate improved or unchanged performance over the base model (w/o MeMix) under the same input length.

| Model | MeMix | Input | 7-Scenes-S | | | | | | NRGBD-S | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Acc. (m) ↓ | | Comp. (m) ↓ | | NC ↑ | | Acc. (m) ↓ | | Comp. (m) ↓ | | NC ↑ | |
| | | | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. | Mean | Med. |
| VGGT *(Offline)* (Wang et al., 2025a) | – | *300* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *400* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *500* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| StreamVGGT (Zhuo et al., 2026) | – | *300* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *400* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| | – | *500* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* | *OOM* |
| CUT3R (Wang et al., 2025b) | × | *300* | 0.141 | 0.096 | 0.076 | 0.034 | 0.543 | 0.564 | 0.234 | 0.139 | 0.074 | 0.018 | 0.575 | 0.614 |
| | ✓ | | 0.106 | 0.076 | 0.053 | 0.019 | 0.550 | 0.575 | 0.186 | 0.086 | 0.050 | 0.009 | 0.595 | 0.651 |
| | × | *400* | 0.178 | 0.121 | 0.115 | 0.069 | 0.532 | 0.546 | 0.342 | 0.227 | 0.127 | 0.067 | 0.567 | 0.603 |
| | ✓ | | 0.147 | 0.100 | 0.076 | 0.039 | 0.540 | 0.559 | 0.321 | 0.180 | 0.099 | 0.031 | 0.562 | 0.590 |
| | × | *500* | 0.190 | 0.138 | 0.090 | 0.033 | 0.530 | 0.543 | 0.359 | 0.264 | 0.173 | 0.081 | 0.560 | 0.591 |
| | ✓ | | 0.167 | 0.119 | 0.077 | 0.026 | 0.533 | 0.547 | 0.328 | 0.218 | 0.161 | 0.040 | 0.560 | 0.590 |
| TTT3R (Chen et al., 2026) | × | *300* | 0.040 | 0.025 | 0.024 | 0.005 | 0.567 | 0.602 | 0.101 | 0.044 | 0.025 | 0.005 | 0.610 | 0.678 |
| | ✓ | | 0.034 | 0.020 | 0.023 | 0.005 | 0.567 | 0.603 | 0.099 | 0.037 | 0.020 | **0.004** | 0.616 | 0.692 |
| | × | *400* | 0.052 | 0.031 | 0.027 | 0.005 | 0.558 | 0.588 | 0.143 | 0.065 | 0.071 | 0.012 | 0.600 | 0.658 |
| | ✓ | | 0.043 | 0.025 | 0.026 | 0.005 | 0.560 | 0.590 | 0.146 | 0.066 | 0.070 | 0.018 | 0.602 | 0.665 |
| | × | *500* | 0.066 | 0.039 | 0.031 | 0.006 | 0.551 | 0.577 | 0.166 | 0.092 | 0.087 | 0.021 | 0.593 | 0.647 |
| | ✓ | | 0.059 | 0.032 | 0.030 | 0.005 | 0.553 | 0.580 | 0.183 | 0.094 | 0.094 | 0.031 | 0.595 | 0.650 |
| TTSA3R (Zheng et al., 2026) | × | *300* | 0.036 | 0.020 | 0.035 | 0.006 | 0.566 | 0.600 | 0.090 | 0.036 | 0.020 | **0.004** | 0.620 | 0.696 |
| | ✓ | | **0.026** | **0.013** | **0.021** | **0.004** | **0.568** | **0.604** | **0.086** | **0.031** | **0.015** | **0.004** | **0.626** | **0.709** |
| | × | *400* | 0.036 | 0.019 | 0.024 | **0.004** | **0.561** | 0.592 | 0.104 | 0.045 | 0.035 | 0.006 | **0.618** | **0.692** |
| | ✓ | | **0.030** | **0.015** | **0.023** | **0.004** | **0.561** | **0.593** | **0.100** | **0.042** | **0.031** | **0.005** | 0.617 | **0.692** |
| | × | *500* | 0.042 | 0.021 | 0.024 | **0.004** | 0.556 | 0.585 | 0.121 | 0.054 | 0.050 | **0.006** | 0.613 | 0.684 |
| | ✓ | | **0.033** | **0.016** | **0.023** | **0.004** | **0.558** | **0.587** | **0.114** | **0.050** | **0.040** | 0.007 | **0.615** | **0.687** |

Table 7 gives the sparse-sampling 7-Scenes and NRGBD reconstruction results for 300, 400, and 500 input frames, sampled every other frame, complementing the dense-sampling results in Table 2.

Direct CUT3R replacement shows the clearest gains, while several TTT3R/TTSA3R NRGBD and normal-consistency entries remain mixed.

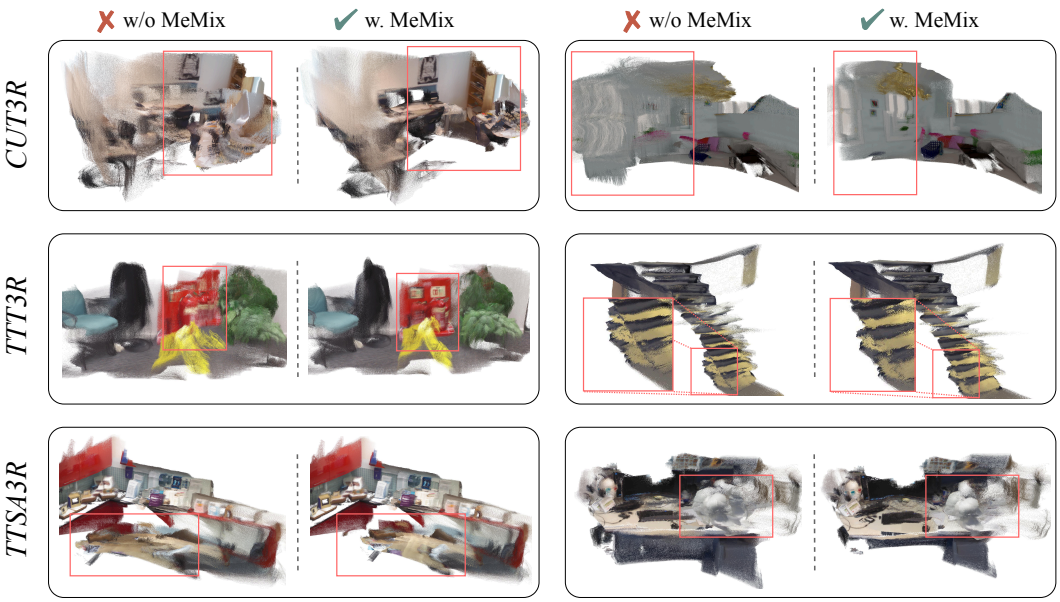


Figure 13: **Qualitative results of 3D reconstruction.** We compare CUT3R, TTT3R, and TTSA3R with their MeMix variants on long input streams. In these examples, MeMix reduces drift and recovers more complete, sharper surfaces. Red boxes highlight surface tearing, missing geometry, and ghosting.

## A7. Pose Estimation

Table 8: **Evaluation on Short-Sequence Pose Estimation.** We evaluate three datasets using clips shorter than 100 frames. ATE and translational RPE are in meters; rotational RPE is in degrees. Bold marks the lowest reported value in each offline/online group, including ties. Green boxes indicate improved or unchanged performance over the corresponding base model (w/o MeMix).

| | | TUM-dynamics (90 frames) | | | ScanNet (90 frames) | | | Sintel (50 frames) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Online** | ATE ↓ | RPE trans ↓ | RPE rot ↓ | ATE ↓ | RPE trans ↓ | RPE rot ↓ | ATE ↓ | RPE trans ↓ | RPE rot ↓ |
| Robust-CVD (Kopf et al., 2021) | × | 0.153 | 0.026 | 3.528 | 0.227 | 0.064 | 7.374 | 0.360 | 0.154 | 3.443 |
| CasualSAM (Zhang et al., 2022) | × | 0.071 | **0.010** | 1.712 | 0.158 | 0.034 | 1.618 | 0.141 | **0.035** | 0.615 |
| DUSt3R (Wang et al., 2024) | × | 0.083 | 0.017 | 3.567 | 0.081 | 0.028 | 0.784 | 0.417 | 0.250 | 5.796 |
| MASt3R (Leroy et al., 2024) | × | 0.038 | 0.012 | 0.448 | 0.078 | 0.020 | 0.475 | 0.185 | 0.060 | 1.496 |
| MonST3R (Zhang et al., 2025) | × | 0.098 | 0.019 | 0.935 | 0.077 | 0.018 | 0.529 | 0.111 | 0.044 | 0.869 |
| Easi3R (Chen et al., 2025a) | × | 0.105 | 0.022 | 1.064 | 0.061 | 0.017 | 0.525 | **0.110** | 0.042 | 0.758 |
| AETHER (Zhu et al., 2025) | × | 0.092 | 0.012 | 1.106 | 0.176 | 0.028 | 1.204 | 0.189 | 0.054 | 0.694 |
| VGGT (Wang et al., 2025a) | × | **0.012** | **0.010** | **0.310** | **0.035** | **0.015** | **0.377** | 0.172 | 0.062 | **0.471** |
| Spann3R (Wang & Agapito, 2025) | ✓ | 0.056 | 0.021 | 0.591 | 0.096 | 0.023 | 0.661 | 0.329 | 0.110 | 4.471 |
| Point3R (Wu et al., 2025) | ✓ | 0.075 | 0.029 | 0.642 | 0.106 | 0.035 | 1.946 | 0.351 | 0.128 | 1.822 |
| StreamVGGT (Zhuo et al., 2026) | ✓ | 0.061 | 0.033 | 3.209 | 0.161 | 0.057 | 3.647 | 0.251 | 0.149 | 1.894 |
| CUT3R (Wang et al., 2025b) | ✓ | 0.045 | 0.015 | 0.443 | 0.096 | 0.022 | 0.600 | 0.210 | **0.069** | 0.628 |
| CUT3R(w. MeMix) | ✓ | 0.043 | 0.014 | 0.424 | 0.090 | 0.022 | 0.604 | **0.190** | 0.075 | **0.627** |
| TTT3R (Chen et al., 2026) | ✓ | 0.029 | **0.013** | 0.380 | 0.065 | **0.021** | 0.640 | 0.208 | 0.093 | 0.725 |
| TTT3R(w. MeMix) | ✓ | 0.028 | **0.013** | 0.376 | 0.065 | **0.021** | 0.677 | 0.210 | 0.083 | 0.733 |
| TTSA3R (Zheng et al., 2026) | ✓ | 0.026 | **0.013** | **0.372** | 0.058 | **0.021** | **0.561** | 0.210 | 0.084 | 0.738 |
| TTSA3R(w. MeMix) | ✓ | **0.025** | **0.013** | **0.372** | **0.057** | **0.021** | 0.569 | 0.209 | 0.084 | 0.763 |

The short-sequence pose benchmarks use Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012), and ScanNet (Dai et al., 2017), following prior protocols (Wang et al., 2025b; Chen et al., 2026; Shen et al., 2025a; Zheng et al., 2026). Inputs contain 50 frames on Sintel and 90 frames on TUM-dynamics and ScanNet. We report ATE, translational RPE, and rotational RPE.

MeMix largely preserves and often improves pose accuracy on these short inputs (Table 8), complementing the long-sequence evaluation in Sec. 4.2.

## A8. VISUALIZATION

The long-sequence examples in Fig. 14 show where trajectory differences accumulate. MeMix stays closer to ground truth in the displayed later segments and larger-motion portions, complementing the aggregate ATE results in Sec. 4.2.

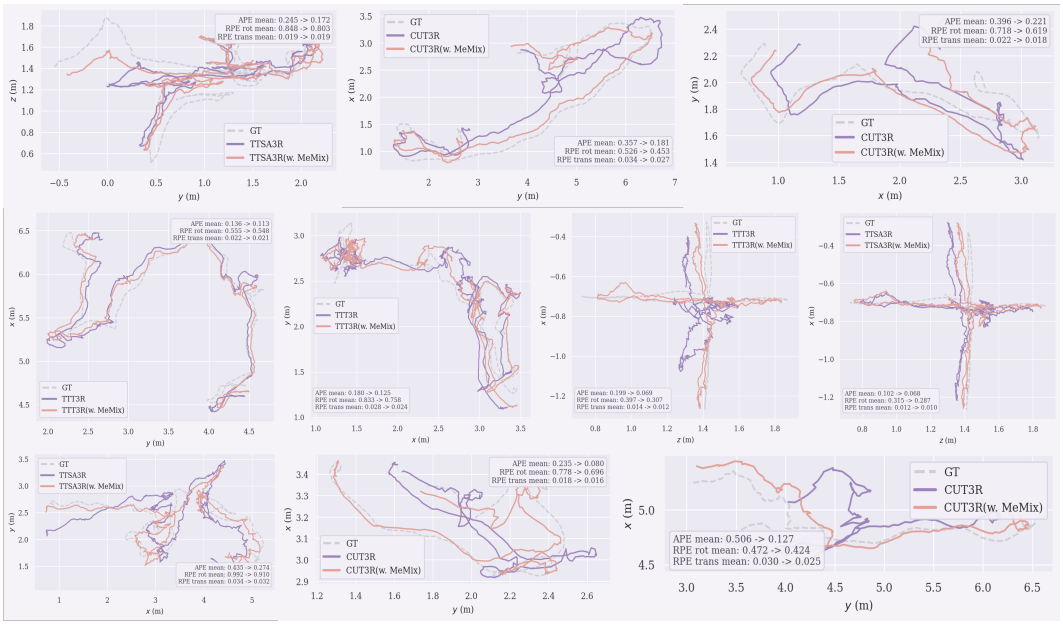


Figure 14: **Visualization of Estimated Camera Trajectories – Long Sequence.** We compare • GT, • baseline trajectories, and • their corresponding MeMix variants across three update rules: CUT3R, TTT3R, and TTSA3R. Across these rules, the MeMix variants generally stay closer to the ground-truth trajectories and exhibit reduced drift over long sequences.