FrameVGGT: Coherence-Preserving Memory for Bounded Streaming Geometry

Zhisong Xu
Institute of Industrial Science
The University of Tokyo
Tokyo, Japan
zhisongxv@cvl.iis.u-tokyo.ac.jp

Takeshi Oishi
Institute of Industrial Science
The University of Tokyo
Tokyo, Japan
oishi@cvl.iis.u-tokyo.ac.jp

Abstract

Streaming Visual Geometry Transformers such as StreamVGGT enable strong online 3D perception, but their KV-cache grows unbounded over long streams, limiting practical deployment. We study bounded-memory streaming geometry and identify a key failure mode of token-level compression: **progressive attention collapse**. Under fixed memory budgets, repeatedly retaining individually salient tokens makes memory increasingly redundant, reduces complementary cross-view evidence, and concentrates attention on a small subset of observations. We therefore propose **FrameVGGT**, a bounded-memory framework that changes the atomic memory unit from individual tokens to coherent frame-wise KV segments. Each segment is summarized by a compact key-space prototype and organized within a fixed-capacity complementary memory bank, together with a sparse anchor tier for persistent long-range references. Across long-sequence 3D reconstruction, video depth estimation, and camera pose estimation, FrameVGGT mitigates attention collapse and achieves favorable accuracy–memory trade-offs while maintaining stable long-horizon geometry.

1 Introduction

Dense 3D reconstruction from images is fundamental to robotic perception, navigation, and interaction [1, 2, 3, 4, 5]. While classical SfM/MVS pipelines achieve high fidelity through explicit geometric optimization, they are typically multi-stage and computationally expensive [6, 7, 8]. Recent feed-forward geometry models instead learn to infer cameras and dense geometry directly from multi-view observations [9, 10, 11, 12, 13, 14]. However, extending these models to online, long-horizon streams remains challenging: preserving all past context causes memory consumption and inference latency to grow with sequence length.

Long-horizon reconstruction has been addressed in part by chunk-based or SLAM-style systems. These methods improve scalability by dividing streams into chunks or submaps, selecting keyframes, and aligning local reconstructions through geometric optimization or dedicated backends [15, 16, 17]. While effective, such pipelines often introduce additional alignment stages, hand-designed keyframe policies, or iterative optimization, complicating the otherwise direct feed-forward inference process.

A more direct streaming alternative processes frames causally while maintaining a bounded state or cache. Implicit-state methods compress history into a latent representation, but this compression can weaken access to long-range evidence and lead to drift [18, 19]. Explicit-memory methods instead cache past representations for direct reuse, but unconstrained accumulation requires eviction or selection to remain tractable [20]. Under bounded resources, many cache-management strategies

Preprint.

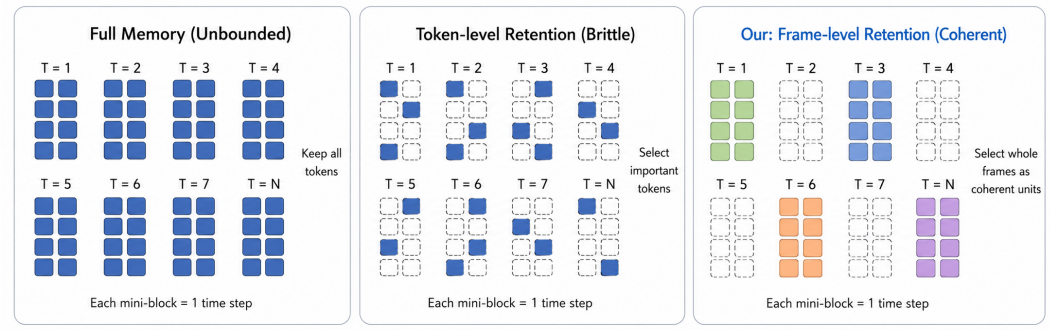


Figure 1: Memory organization strategies for bounded streaming geometry. Full memory grows unbounded over time. Under a fixed budget, token-level retention can fragment frame-wise context and progressively concentrate attention on redundant evidence. Frame-level retention keeps selected frame-wise KV segments intact, preserving more coherent observations for multi-view inference.

select retained history at token granularity using importance, diversity, or eviction heuristics [21, 22, 23, 24, 25, 26, 27].

However, bounding cache size alone does not ensure that the retained history remains useful for geometric inference. As illustrated in Fig. 1, the granularity of retention determines which structures survive cache selection. Token-level policies select individual entries, potentially retaining only scattered fragments of each observation. Because geometric tokens vary in noise and informativeness, repeated selection can favor dominant features while discarding weaker but complementary evidence. Over time, the cache may become increasingly redundant, with attention concentrated on a small subset of retained entries and reduced support from complementary viewpoints. We refer to this coupled degradation of attention distribution and geometric coverage as **progressive attention collapse**.

Retention granularity is particularly relevant to geometry because multi-view inference relies on relationships among observations, including their spatial structure and cross-view correspondences. The usefulness of a token therefore depends not only on its individual salience, but also on the context retained alongside it. Selecting isolated entries can disrupt this context even when the cache remains fully populated. By contrast, frame-level retention keeps all tokens from a selected frame together, preserving its internal structure and providing more complete observations for cross-view reasoning.

We instantiate this perspective with **FrameVGGT**, an inference-time bounded-memory framework that retains history as coherent frame-wise units. FrameVGGT maintains a fixed-capacity set of complementary memory units, each instantiated as a frame-wise KV segment summarized by a compact key-space prototype. A lightweight **sparse anchor tier** complements this memory with persistent long-range references as the retained frame set evolves. Together, the two tiers preserve structured multi-view context and mitigate progressive attention collapse under a fixed memory budget. Across long-sequence 3D reconstruction, video depth estimation, and camera pose estimation, FrameVGGT achieves favorable accuracy–memory trade-offs while maintaining more stable geometry over long streams.

Contributions.

(1) We formulate **retention granularity** as a key design axis for bounded streaming geometry, highlighting the importance of coherent geometric context beyond cache size alone.

(2) We identify and characterize **progressive attention collapse** in token-level bounded memory, where increasingly redundant retention concentrates attention while reducing spatio-temporal geometric support.

(3) We introduce **FrameVGGT**, an inference-time two-tier memory framework that preserves frame-wise KV segments as coherent units and uses sparse anchors to support long-range consistency under fixed memory budgets.

2 Related Work

2.1 Geometry-based Reconstruction

Classical 3D reconstruction relies on explicit geometric optimization. Offline pipelines such as SfM and MVS estimate camera poses and scene structure from image collections, typically involving multi-stage processing and global refinement [28, 29, 30, 31]. These approaches achieve high accuracy, but their computational and storage demands complicate deployment on long image streams. Online systems such as SLAM update motion estimates and maps incrementally [32, 33]. Nevertheless, maintaining dense geometry and long-horizon consistency under strict memory budgets remains challenging.

2.2 Learning-based Reconstruction with Multi-view Inputs

Recent learning-based approaches perform feed-forward geometric inference, predicting camera parameters, depth, or correspondences directly from multiple views [9, 10, 11, 12, 13, 14]. Joint processing provides rich cross-view context, but memory and computation grow rapidly with the number of input views. Several extensions improve scalability through submap decomposition, anchor-based representations [34], or token-level acceleration. However, these designs remain largely batch-oriented and do not directly address how historical context should be retained and updated during bounded-memory streaming.

2.3 Learning-based Reconstruction under Streaming Constraints

Streaming methods process frames online while maintaining historical context. Explicit-memory approaches cache past representations for reuse [35, 36, 37], but unconstrained accumulation leads to growing memory consumption and latency. Implicit-state approaches compress history into bounded latent states, improving efficiency but potentially weakening access to long-range evidence [18, 19]. Windowed methods restrict computation to local temporal neighborhoods [38], stabilizing runtime while limiting access to distant observations. Recent bounded-cache approaches further control memory through token selection or eviction [24, 25, 26, 27].

These approaches differ in how history is represented, compressed, or selected. Our focus is the granularity at which explicit memory is retained. Under a fixed budget, token-level retention can fragment observations and progressively concentrate attention on redundant evidence. FrameVGGT instead preserves complete frame-wise KV segments and complements them with sparse anchors, maintaining coherent multi-view context for long-horizon streaming inference.

3 Problem Definition and Analysis

We consider online geometric inference over an unbounded image stream $I = \{I_t\}_{t\geq 1}$, where predictions y_t are produced sequentially. For decoder layer l, let $\tilde{C}_t^{(l)}$ denote the retained Transformer cache at time t, and $C_{t,\mathrm{new}}^{(l)}$ the cache entries generated from the incoming observation. Without memory control,

$$\tilde{C}_t^{(l)} = \tilde{C}_{t-1}^{(l)} \oplus C_{t,\mathrm{new}}^{(l)},$$

so memory usage and attention cost grow continuously with stream length.

We instead consider bounded streaming memory,

$$\tilde{C}_t^{(l)} = \Omega\Big(\tilde{C}_{t-1}^{(l)} \oplus C_{t,\mathrm{new}}^{(l)}\Big), \qquad \text{s.t.} \quad \sum_l |\tilde{C}_t^{(l)}| \leq M,$$

where M is the fixed cache budget. Historical evidence must therefore be continually removed as new observations arrive. The central question is not only *which* entries should remain, but *at what granularity* memory should be retained so that bounded context remains geometrically useful.

Geometric Coherence and Attention Collapse. Unlike language modeling, multi-view geometry relies on spatially distributed and mutually compatible evidence across observations. Cross-attention acts as differentiable feature matching, so the utility of a geometric token depends not only on

its individual response but also on the coherent observation to which it belongs. We refer to this requirement as *geometric coherence*.

Token-level retention can progressively violate this structure. Importance estimates from correlated and noisy queries repeatedly favor tokens aligned with dominant visual modes, while weaker but complementary evidence is more easily discarded. Consequently, retained observations become fragmented, memory covers fewer regions and viewpoints, and the remaining keys become increasingly redundant. Softmax attention can further amplify this imbalance by assigning increasing probability mass to a dominant subset. We refer to this progressive concentration of retrieval onto redundant evidence as **attention collapse**. Thus, even with a fixed and fully populated cache, effective geometric support can progressively shrink.

Directional Contrast. To characterize the key-space imbalance underlying this failure, let R_k denote a retained subset associated with a dominant local mode and N_k its complement. For normalized keys $\hat{\mathbf{k}}_i$, define

$$\boldsymbol{\mu}_R = \frac{\frac{1}{|R_k|}\sum_{i\in R_k}\hat{\mathbf{k}}_i}{\left\|\frac{1}{|R_k|}\sum_{i\in R_k}\hat{\mathbf{k}}_i\right\|_2},$$

and

$$\Delta_k = \underbrace{\frac{1}{|R_k|}\sum_{i\in R_k}\cos(\hat{\mathbf{k}}_i,\boldsymbol{\mu}_R)}_{A_R} - \underbrace{\frac{1}{|N_k|}\sum_{j\in N_k}\cos(\hat{\mathbf{k}}_j,\boldsymbol{\mu}_R)}_{A_N}.$$

A larger Δ_k indicates a stronger directional advantage of the dominant subset over the remaining memory. For queries aligned with $\boldsymbol{\mu}_R$, this advantage induces a larger query–key logit gap that softmax can amplify into more concentrated attention. Thus, Δ_k characterizes the key-space bias that can be amplified into concentrated attention. We further provide empirical evidence that this bias is accompanied by increasing concentration in the downstream attention distribution, quantified using the normalized inverse-entropy metric C_{att}. The formal connection and the corresponding concentration–pose analysis are provided in Appendix C.4.

Why Frame-Level Retention? Retention granularity affects how much directional variation is preserved within an observation. Partition R_k into source-frame key segments $\{S_t\}$, and let

$$w_t = \frac{|S_t|}{|R_k|},$$

with σ_t^2 denoting the within-segment directional dispersion. Variance decomposition gives

$$\Delta_k \le 1 - \frac{1}{2}\sum_t w_t\sigma_t^2 - A_N.$$

Hence, preserving broader within-frame variation limits the directional advantage of a dominant subset.

Token-level selection can retain only highly aligned features from an observation while discarding complementary entries, reducing this variation. Frame-level retention instead keeps complete source-frame KV segments intact, preserving both strong correspondences and weaker complementary evidence. It therefore changes the atomic memory unit from isolated features to coherent geometric observations, providing a structural mechanism for mitigating attention collapse under a fixed budget. The full derivation and long-horizon diagnostics are given in Appendix C.4.

4 Method

Sec. 3 shows that bounded streaming geometry depends not only on cache capacity, but also on *memory granularity*. FrameVGGT therefore changes the atomic memory unit from individual tokens to **frame-wise KV segments**. For frame t and layer l, let $S_t^{(l)}$ denote the corresponding KV segment. Once selected, the complete segment is retained as an indivisible observation, preserving within-frame geometric evidence while moving redundancy removal from individual tokens to observations.

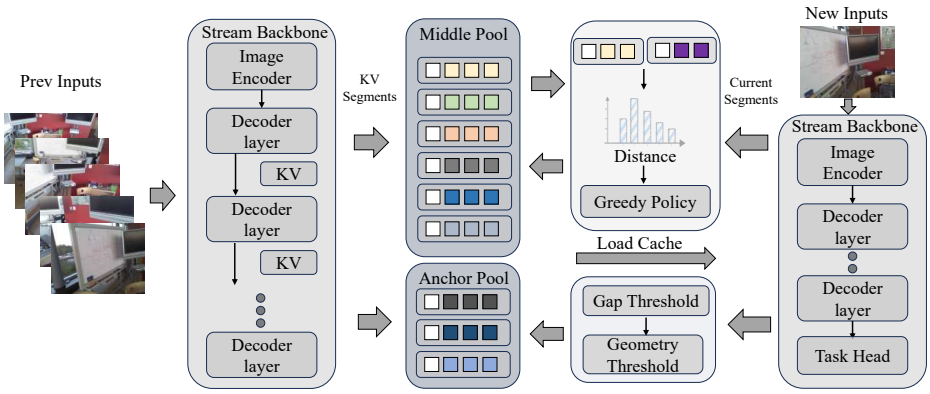


Figure 2: Pipeline of FrameVGGT. Previous inputs are encoded into frame-wise KV segments. A mid-term bank maintains complementary observations, while a sparse anchor tier preserves selected long-range references. The retained segments form the bounded cache used for subsequent streaming inference.

Under a fixed budget, FrameVGGT organizes these frame-wise units using two complementary memories. A **mid-term bank** maintains broad coverage of complementary observations, while a lightweight **sparse anchor memory** protects a small number of long-range references from repeated eviction. Figure 2 summarizes the framework.

Frame-Wise Retrieval Representation. Directly comparing complete KV segments is expensive. We therefore summarize each segment with a lightweight descriptor used only for memory selection. Let H_l denote the number of attention heads, d_l the per-head key dimension, and $\mathcal{T}_t$ the token set of frame t. We compute

$$\mathbf{v}_t^{(l)} = \frac{1}{H_l|\mathcal{T}_t|} \sum_{h=1}^{H_l} \sum_{\tau \in \mathcal{T}_t} \mathbf{k}_{t,h,\tau}^{(l)}, \tag{1}$$

followed by

$$\bar{\mathbf{v}}_t^{(l)} = \frac{\mathbf{v}_t^{(l)}}{\|\mathbf{v}_t^{(l)}\|_2}.$$

The descriptor is used only for selection; the retained memory always stores the complete KV segment. We measure segment dissimilarity in key space by

$$d\Big(S_i^{(l)}, S_j^{(l)}\Big) = 1 - \Big\langle \bar{\mathbf{v}}_i^{(l)}, \bar{\mathbf{v}}_j^{(l)} \Big\rangle. \tag{2}$$

Since future memory access is determined by query–key compatibility, this provides a direct proxy for retrieval redundancy. Frame-wise retention preserves coherence *within* observations, while this distance removes redundancy *between* observations.

Mid-Term Memory Bank. The mid-term bank maintains a compact set of observations with broad retrieval-space coverage. Given candidate segments $\mathcal{M}_t$, we seek $Q \subseteq \mathcal{M}_t$, $|Q| \le M_{\text{mid}}$, minimizing the worst-case coverage gap:

$$Q^* = \arg \min_{\substack{Q \subseteq \mathcal{M}_t \\ |Q| \le M_{\text{mid}}}} \max_{S \in \mathcal{M}_t} \min_{S' \in Q} d(S, S'). \tag{3}$$

We approximate this objective online using greedy farthest-first selection, initialized with the most recent segment. For each candidate S, define

$$m(S) = \min_{S' \in Q} d(S, S'). \tag{4}$$

Table 1: Reconstruction results on 7-Scenes and NRGBD. Best results are shown in bold.

Method	7-Scenes						NRGBD					
	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑
CUT3R	0.181	0.127	0.095	0.0327	0.525	0.531	0.322	0.237	0.128	0.0334	0.553	0.566
Point3R	0.063	0.026	0.031	0.0150	0.559	0.589	0.113	0.048	0.037	0.0060	0.621	0.696
TTT3R	0.060	0.035	0.028	0.0050	0.560	0.588	0.161	0.082	0.093	0.0147	0.602	0.654
XStreamVGGT	0.105	0.052	0.048	0.0100	0.556	0.583	0.128	0.080	0.043	0.0069	0.621	0.703
InfiniteVGGT	0.041	0.016	0.024	0.0047	0.561	0.593	0.078	0.049	0.035	0.0069	0.647	0.757
OVGGT*	0.033	0.011	0.020	0.0040	0.560	0.591	0.056	0.036	0.032	0.0069	0.635	0.731
Ours(12)	0.035	0.013	0.020	0.0048	0.562	0.594	0.054	0.036	0.031	0.0068	0.656	0.764
Ours(16)	0.033	0.010	0.019	0.0044	**0.564**	0.597	**0.053**	**0.035**	**0.030**	0.0061	0.663	0.773
Ours(20)	**0.027**	**0.008**	**0.018**	**0.0038**	0.563	0.595	**0.053**	0.036	0.033	0.0067	0.661	0.776
Ours(24)	0.028	0.009	0.019	0.0040	**0.564**	**0.598**	0.054	**0.035**	**0.030**	**0.0060**	**0.670**	**0.782**

We repeatedly select the candidate with the largest $m(S)$ and update

$$m(S) \leftarrow \min(m(S), d(S, S_{\text{new}})) \,. \tag{5}$$

Each incoming frame is added to the candidate set. When the bank exceeds capacity, the retained subset is refreshed using the above rule and reassembled into the bounded KV cache. Selection is performed independently at each layer. This preserves complementary observations and prevents the retained memory from becoming dominated by near-duplicate retrieval modes.

Sparse Anchor Memory. Coverage alone does not guarantee long-term persistence: a distant but geometrically useful observation may be evicted before its value becomes apparent. We therefore maintain a sparse anchor set A_t, with $|A_t| \leq B_A$, to protect reliable and geometrically novel observations. The first frame is always retained as a persistent global reference.

For candidate frame i, promotion is determined by temporal spacing, reliability, and novelty:

$$A_t = \begin{cases} \text{FIFO}_{B_A}(A_{t-1} \cup \{i\}) \,, & \text{if } g_i \geq G,\ \Phi(i) \geq \tau_{\text{conf}},\ \nu(i) \geq \tau_{\text{novel}}, \\ A_{t-1}, & \text{otherwise,} \end{cases}$$

where

$$g_i = i - i_{\text{last}}$$

is the gap from the most recently promoted anchor. The reliability score $\Phi(i) = q_i s_i$ combines normalized prediction confidence q_i and Laplacian sharpness s_i, while $\nu(i)$ measures novelty relative to existing anchor pose signatures. FIFO replacement applies only to historical anchors, leaving the first frame persistent.

The two tiers therefore address different failure modes: the mid-term bank maintains adaptive complementary coverage, whereas the anchor memory provides sparse protected persistence over longer horizons. Both operate on the same frame-wise memory representation under a fixed overall budget.

5 Experiments

5.1 Experimental Setup

We evaluate FrameVGGT on *online 3D reconstruction*, *video depth estimation*, and *monocular camera pose estimation* under a strict *resource-controlled protocol*: pretrained weights, frame sampling, and inference pipeline are fixed, while only the bounded-memory organization varies under comparable KV-cache budgets. Baselines marked with $*$ indicate contemporaneous token-level streaming methods. FrameVGGT is used purely at inference time without retraining or fine-tuning, and memory denotes the effective streaming KV-cache footprint. All experiments use a single NVIDIA RTX A6000 GPU.

Memory-budget sweep. We vary the mid-term capacity M, i.e., the maximum number of refreshable frame segments retained in the mid-term bank, over $M \in \{12, 16, 20, 24\}$. This tests whether preserving broader complementary coverage improves long-horizon performance, and whether gains saturate once the dominant memory structure is covered.

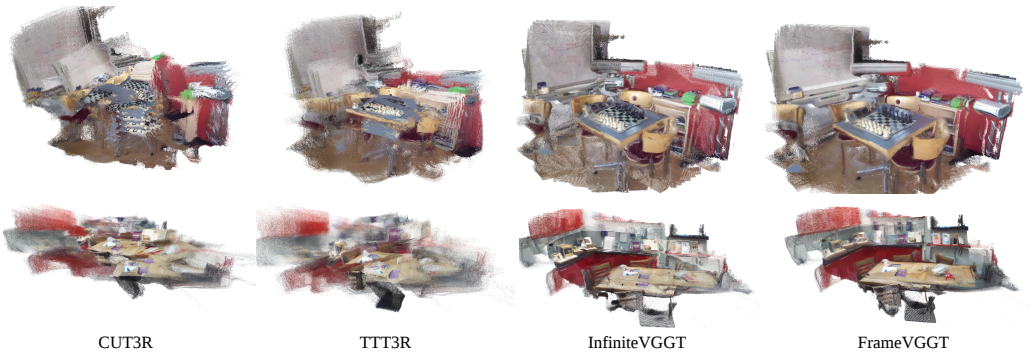


Figure 3: Reconstruction visualization comparison on the 7scenes dataset. Our method ensures robust geometric consistency over extended sequences, maintaining a highly stable and reliable result.

Table 2: Reconstruction results on Long3D. Best results are shown in bold.

Method	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑
CUT3R	3.217	2.235	1.605	1.351	0.500	0.491
TTT3R	2.964	2.310	2.265	1.805	0.507	0.514
InfiniteVGGT	2.138	1.728	**0.963**	**0.364**	0.513	0.526
OVGGT*	2.148	1.687	1.594	0.574	0.526	0.530
Ours	**1.698**	**1.331**	1.032	0.433	**0.536**	**0.540**

Timescale-separation diagnostic. To isolate the role of sparse persistent references, we compare **24 mid-term + 0 anchor** against **20 mid-term + 4 anchor** under the same total memory budget. This tests whether long-range references provide benefits beyond the primary mid-term memory bank.

Short-term continuity vs. complementary context. We evaluate a **Recent-K** variant that reserves capacity for the last K frames ($K \in \{2, 4, 6\}$) and allocates the rest to mid-term memory. This tests whether bounded memory should prioritize recent continuity or complementary context under a fixed budget.

5.2 3D Reconstruction

Protocol and metrics. We evaluate 3D reconstruction on 7-SCENES, NRGBD [39, 40], and LONG3D, reporting Accuracy (Acc, ↓), Completeness (Comp, ↓), and Normal Consistency (NC, ↑). For 7-SCENES and NRGBD, we test long-horizon sequences (up to 1000 frames, stride 2) under a unified bounded-memory protocol. For LONG3D (2k–10k frames), predicted point clouds are downsampled to 2% to align with ground-truth density.

Effect of mid-term capacity. Tab. 1, 2 and Fig. 3 demonstrate that token-level methods degrade on long sequences due to context fragmentation, which reduces surface coverage and introduces artifacts. In contrast, our coherence-aligned memory preserves multi-view context, yielding superior accuracy–memory trade-offs. By operating on frame-level prototypes, FRAMEVGGT bypasses the overhead of fine-grained token sampling, achieving significantly higher throughput than baselines. For instance, our 1.9–3.7 GB footprint (12–24 blocks) is substantially more efficient than INFINITEVGGT (6.9 GB) or XSTREAMVGGT (10.3 GB).

Global-anchor diagnostic. Tab. 3 shows that anchors enhance robustness during unreliable mid-term context (e.g., blur, occlusion, or weak parallax) by providing persistent references. While their impact is minimal in regular scenarios, anchors mitigate error accumulation in degraded cases. Anchor-only baselines (FIFO, random, uniform) consistently underperform, confirming that anchors serve as a vital robustness-oriented complement to mid-term memory.

Recent vs. mid-term evidence. Tab. 4 shows that forcing a Recent-K buffer often degrades reconstruction quality. Recency-biased buffering over-allocates memory to highly overlapping adjacent frames while displacing more complementary mid-term memory, reducing effective scene coverage

Table 3: Performance comparison across anchor configurations. Best results are shown in bold.

Scene	Anchor=0						Anchor=4					
	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑
office/seq-09	0.0429	**0.0088**	0.0126	**0.0044**	0.5627	0.6085	**0.0275**	**0.0088**	**0.0119**	**0.0044**	**0.5758**	**0.6163**
stairs/seq-01	0.1392	0.0556	**0.0915**	**0.0077**	0.5411	0.5585	**0.1250**	**0.0521**	0.0932	0.0081	**0.5565**	**0.5726**

Table 4: Recent ablation on 7-Scenes and NRGBD under a fixed total budget (M_{total}=16). Recent-0 corresponds to Ours (M=16). Best results are shown in bold.

Method	7-Scenes						NRGBD					
	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑
Recent-0	**0.033**	**0.010**	**0.019**	**0.0044**	**0.564**	**0.597**	**0.053**	**0.035**	**0.030**	**0.0061**	**0.663**	**0.773**
Recent-2	0.037	0.013	0.020	**0.0044**	0.562	0.593	0.056	0.040	0.031	0.0070	0.656	0.768
Recent-4	0.053	0.017	0.024	0.0050	0.558	0.587	0.066	0.045	0.031	0.0073	0.637	0.734
Recent-6	0.069	0.025	0.027	0.0050	0.555	0.583	0.085	0.059	0.034	0.0080	0.623	0.707

Table 5: Video depth estimation results on Bonn. Best results are shown in bold.

Method	Abs Rel↓	$\delta < 1.25$ ↑
CUT3R	0.0831	0.9402
Point3R	0.0809	0.9513
TTT3R	0.0752	0.9587
XStreamVGGT	0.0807	0.9697
InfiniteVGGT	0.0560	**0.9801**
OVGGT*	0.0658	0.9558
Ours (12)	0.0526	0.9793
Ours (16)	0.0525	**0.9801**
Ours (20)	0.0514	0.9799
Ours (24)	**0.0512**	0.9799

Table 6: Recent ablation on Bonn under a fixed total budget (M_{total}=16). Recent-0 corresponds to Ours (M=16). Best results are shown in bold.

Sliding Window	Abs Rel↓	$\delta < 1.25$ ↑
Recent-0	**0.0525**	**0.9801**
Recent-2	0.0535	0.9794
Recent-4	0.0562	0.9788
Recent-6	0.0589	0.9780

and weakening the support needed to maintain surface completeness and geometric consistency over long horizons.

5.3 Video Depth Estimation

Protocol and metrics. We evaluate streaming video depth estimation on BONN [41] over sequences of up to 500 frames, reporting Abs Rel (↓) and $\delta < 1.25$ (↑) after per-sequence scale alignment. The evaluation protocol is fixed across all memory settings.

Effect of mid-term capacity. Tab. 5 shows that our frame-level memory maintains strong depth accuracy even under tight bounded-memory settings. Compared with reconstruction and pose, video depth is less sensitive to memory size, suggesting that much of the required support remains relatively local. Still, increasing the mid-term capacity M yields modest but consistent improvements in Abs Rel, indicating that additional mid-term context remains beneficial when local evidence becomes ambiguous or insufficient. The overall trend also suggests an earlier saturation regime: once nearby context is already strong, additional history provides diminishing returns.

Recent vs. mid-term evidence. Tab. 6 shows that reserving memory for the most recent frames does not improve depth estimation, and performance gradually declines as K increases. This suggests that, as in 3D reconstruction, bounded memory is more effective when used to preserve complementary context rather than redundant adjacent views.

5.4 Camera Pose Estimation

Protocol and metrics. We evaluate streaming camera pose estimation on TUM-DYNAMICS [42] using sequences of up to 300 frames. We report ATE, $\text{RPE}_{\text{trans}}$, and RPE_{rot} (all ↓) after Sim(3) Umeyama alignment, used only for global similarity normalization.

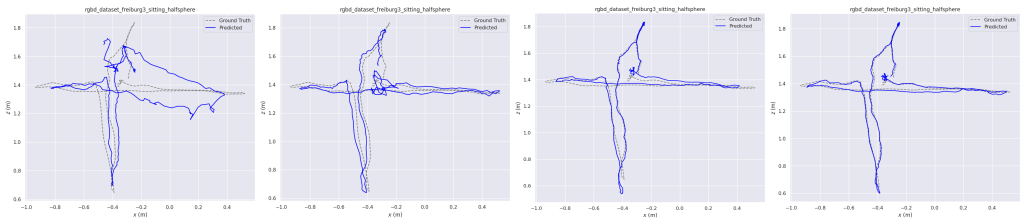


Figure 4: Pose visualization comparison on the TUM dataset. Our method maintains a robustly stable trajectory and ensures geometric integrity even over exceptionally long-duration sequences.

Table 7: Pose estimation results on TUM. Best results are shown in bold.

Method	ATE↓	RPE$_t$ ↓	RPE$_r$ ↓
CUT3R	0.1089	0.0148	0.619
Point3R	0.1387	0.0369	2.347
TTT3R	0.0620	0.0147	0.618
XStreamVGGT	0.0728	0.0262	0.580
InfiniteVGGT	0.0478	0.0138	0.350
OVGGT*	0.0561	0.0215	0.448
Ours (12)	0.0387	0.0140	0.346
Ours (16)	0.0386	0.0137	0.342
Ours (20)	0.0389	0.0135	0.339
Ours (24)	**0.0385**	**0.0133**	**0.336**

Table 8: Recent ablation on TUM under a fixed total budget (M_{total}=16). Recent-0 corresponds to Ours (M=16). Best results are shown in bold.

Sliding Window	ATE↓	RPE$_t$ ↓	RPE$_r$ ↓
Recent-0	**0.0386**	**0.0137**	0.342
Recent-2	0.0399	0.0138	**0.341**
Recent-4	0.0410	0.0140	0.342
Recent-6	0.0432	0.0140	**0.341**

Effect of mid-term capacity. Similar trends are observed for pose estimation. Tab. 7 and Fig. 4 show that token-level streaming suffers from accumulated error and drift as sequences grow. In contrast, our frame-level memory preserves complementary multi-view context over time, resulting in more stable pose estimation and lower ATE and RPE over long sequences. Increasing the mid-term capacity M consistently improves pose accuracy, with diminishing returns at larger budgets.

Recent vs. mid-term evidence. Tab. 8 shows that forcing a Recent-K buffer is insufficient for long-horizon pose stability. Consistent with 3D reconstruction, allocating more memory to recent frames reduces capacity for complementary mid-term context and weakens long-range geometric constraints.

6 Discussion

6.1 Interpretation of Results

Across all tasks, FrameVGGT achieves a favorable accuracy–memory trade-off over token-level bounded-memory baselines, suggesting that memory organization is as important as raw capacity in streaming geometry. Frame-level retention preserves coherent cross-view evidence, while token-level selection can fragment observations and progressively concentrate retrieval on redundant modes. The additional analyses support this interpretation: simple frame selectors such as Uniform, Random, and FIFO remain clearly worse under the same budget (Appendix C.6), showing that both retention granularity and complementary observation selection matter. Performance also remains stable across a broad range of capacities and continues to improve beyond the default 24-frame setting (Appendix H.2), indicating a controllable accuracy–memory trade-off rather than a narrow operating point.

The sparse anchor tier provides a complementary long-range role, improving difficult cases involving revisits, occlusion, and repeated geometry without increasing the total memory budget (Appendix D). Moreover, long-horizon diagnostics show that token-level baselines exhibit increasing attention concentration together with growing pose error, whereas FrameVGGT substantially suppresses both trends (Appendix C.4). Full-length TUM evaluation further confirms that these gains persist over complete sequences with bounded memory and practical runtime (Appendix G.1).

Acknowledgments and Disclosure of Funding

This work was supported in part by JSPS KAKENHI Grant Numbers JP24K21173 and JP24H00351, the UTokyo-SMBC Forest GX Project, and the MEXT Supporting Pioneering Research through AI for 1,000 Discovery Challenges Program (SPReAD), Japan, under Grant Number JPMXP1726304006.

References

[1] Xingdong Sheng, Shijie Mao, Yichao Yan, and Xiaokang Yang. Review on slam algorithms for augmented reality. *Displays*, 84:102806, 2024.

[2] Ioan Andrei Bârsan, Peidong Liu, Marc Pollefeys, and Andreas Geiger. Robust dense mapping for large-scale dynamic environments. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7510–7517. IEEE, 2018.

[3] Mehmet Turan, Yusuf Yigit Pilavci, Ipek Ganiyusufoglu, Helder Araujo, Ender Konukoglu, and Metin Sitti. Sparse-then-dense alignment-based 3d map reconstruction method for endoscopic capsule robots. *Machine Vision and Applications*, 29(2):345–359, 2018.

[4] Ruben Mascaro and Margarita Chli. Scene representations for robotic spatial perception. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):351–377, 2025.

[5] Sonia Raychaudhuri and Angel X Chang. Semantic mapping in indoor embodied ai–a comprehensive survey and future directions. *arXiv preprint arXiv:2501.05750*, 3, 2025.

[6] Richard A Andersen and David C Bradley. Perception of three-dimensional structure from motion. *Trends in cognitive sciences*, 2(6):222–228, 1998.

[7] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[8] Tanveer Hussain, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C De Albuquerque. A comprehensive survey of multi-view video summarization. *Pattern Recognition*, 109:107567, 2021.

[9] 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)*, pages 20697–20709, June 2024.

[10] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025.

[11] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[12] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[13] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[14] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

[15] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025.

[16] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[17] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16651–16662, 2025.

[18] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[19] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

[20] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

[21] Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

[22] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

[23] 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. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

[24] 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.

[25] Zunhai Su, Weihao Ye, Hansen Feng, Keyu Fan, Jing Zhang, Dahai Yu, Zhengwu Liu, and Ngai Wong. Xstreamvggt: Extremely memory-efficient streaming vision geometry grounded transformer with kv cache compression. *arXiv preprint arXiv:2601.01204*, 2026.

[26] Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3r: Training-free token eviction for memory-bounded streaming visual geometry transformers. *arXiv preprint arXiv:2509.17650*, 2025.

[27] Si-Yu Lu, Po-Ting Chen, Hui-Che Hsu, Sin-Ye Jhong, Wen-Huang Cheng, and Yung-Yao Chen. Ovggt: O (1) constant-cost streaming visual geometry transformer. *arXiv preprint arXiv:2603.05959*, 2026.

[28] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[30] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[31] Fan Zhu, Ziyu Chen, Chunmao Jiang, Liwei Xu, Shijin Zhang, Biao Yu, and Hui Zhu. Slm-slam: a visual slam system based on segmented large-scale model in dynamic scenes and zero-shot conditions. *Measurement Science and Technology*, 35(8):086315, 2024.

[32] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[33] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ transactions on computer vision and applications*, 9(1):16, 2017.

[34] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. Sail-recon: Large sfm by augmenting scene regression with localization. *arXiv preprint arXiv:2508.17972*, 2025.

[35] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[36] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.

[37] 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*, pages 5273–5284, 2025.

[38] Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool. *arXiv preprint arXiv:2509.05296*, 2025.

[39] Hunter Blanton, Connor Greenwell, Scott Workman, and Nathan Jacobs. Extending absolute pose regression to multiple scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–39, 2020.

[40] 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*, pages 6290–6301, 2022.

[41] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019.

[42] 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*, pages 573–580, 2012.

A AI use statement

In this work, we used generative AI tools to assist with manuscript language polishing, including improving grammar, readability, and clarity of technical descriptions. Generative AI tools were also used for limited code editing and debugging assistance during implementation. All AI-assisted text and code were carefully reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this work and for the correctness of the implementation and reported results.

B Limitations and Future Work.

Our current formulation uses a fixed, training-free memory policy to isolate the effect of memory granularity and organization. Although the method is stable across memory capacities and anchor settings, its allocation does not explicitly adapt to scene complexity, motion, or visibility. Future work may explore learned observation-level scoring, scene-adaptive memory allocation, and more flexible long-range persistence while retaining the frame-wise memory formulation introduced here.

C Additional Diagnostics for Token-Level Compression in Geometric Streaming

This appendix complements Sec. 3 with additional diagnostics for the granularity effects discussed in the main text. We introduce lightweight proxies that illustrate several degradation patterns observed under bounded token-level retention, while keeping the notation consistent with Sec. 3.

C.1 Problem Setup

Consider a video stream of length T, where frame t produces a set of token vectors

$$X_t = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N_t}\}, \qquad \mathbf{x}_{t,i} \in \mathbb{R}^d.$$

Here, T, N_t, and d are scalars, X_t is a set, and $\mathbf{x}_{t,i}$ is a vector. Under a bounded-memory policy, a retained subset $\widetilde{X}_t \subseteq X_t$ is maintained under the global KV-cache budget

$$\sum_{t=1}^{T} |\widetilde{X}_t| \leq M,$$

where M is the total memory budget.

As emphasized in the main text, the key question is not only how many entries are retained, but whether the retained subset preserves sufficient *geometrically usable context* for downstream inference. For geometric prediction, the utility of retained memory depends not only on individual entries, but also on whether they remain organized as coherent observations that jointly support matching, fusion, and long-range consistency.

Geometric-context proxy. To make this intuition explicit, we introduce a qualitative scalar proxy $s(\cdot)$ that reflects the amount of usable geometric context preserved by a retained subset. Possible instantiations include spatial coverage, reprojection consistency, or other task-dependent measures. We emphasize that $s(\cdot)$ is not uniquely defined and is used only as a conceptual diagnostic.

We then define the context-damage proxy

$$d_t = s(X_t) - s(\widetilde{X}_t),$$

and, for $s(X_t) > 0$, the relative context ratio

$$\rho_t = \frac{s(\widetilde{X}_t)}{s(X_t)}.$$

These quantities are not used for formal guarantees, but serve as diagnostics of how compression affects geometrically usable context.

C.2 Context Thinning Under Bounded Memory

Under a fixed global budget, retained memory must be distributed across an increasing number of frames as the sequence grows. As a result, the average number of retained observations per frame tends to decrease with longer horizons.

In practice, this leads to progressive *context thinning*: each frame contributes fewer retained entries, reducing the density of usable geometric evidence. For geometric tasks, where stable estimation relies on redundant multi-view observations, such thinning weakens the effective context available for depth, pose, and reconstruction.

Importantly, degradation is not determined solely by the number of retained entries. Different retention policies with similar budgets can preserve very different amounts of usable geometric context, depending on how the retained memory is distributed.

C.3 Context Fragmentation Across Space and Time

Geometric inference depends not only on how many observations are retained, but also on whether they form coherent multi-view context across frames.

Under token-level retention, selection is performed independently at the token level and is not explicitly constrained to preserve within-frame structure or cross-view compatibility. As a result, retained entries may no longer correspond to mutually compatible observations of the same underlying structure.

For example, different frames may retain entries from unrelated regions or surfaces, leading to memory that remains populated but lacks stable correspondences. We refer to this effect as *context fragmentation*, since the retained geometric context becomes scattered across both space and time.

This effect is particularly pronounced in regimes with already limited redundancy, such as low-parallax motion, blur, occlusion, or texture-poor regions. In such cases, fragmentation further reduces the availability of jointly usable observations and makes inference more sensitive to local inconsistencies.

C.4 Directional Concentration, Attention Concentration, and Retention Granularity

Context thinning and context fragmentation affect not only what is retained, but also how retained memory is used during attention. For a given layer l, let $\hat{\mathbf{k}}_i \in \mathbb{R}^{d_l}$ denote an ℓ_2-normalized key vector, where d_l is the key dimension of layer l. For a query step k, we partition retained keys into a local subset R_k and a non-local subset N_k, where R_k contains keys from a short temporal neighborhood of the query.

Assuming the local mean is nonzero, we define the local representative direction $\boldsymbol{\mu}_R \in \mathbb{R}^{d_l}$ by

$$\boldsymbol{\mu}_R = \frac{\frac{1}{|R_k|}\sum_{i\in R_k}\hat{\mathbf{k}}_i}{\left\|\frac{1}{|R_k|}\sum_{i\in R_k}\hat{\mathbf{k}}_i\right\|_2},$$

and measure directional contrast by

$$\Delta_k = \frac{1}{|R_k|}\sum_{i\in R_k}\cos(\hat{\mathbf{k}}_i, \boldsymbol{\mu}_R) - \frac{1}{|N_k|}\sum_{j\in N_k}\cos(\hat{\mathbf{k}}_j, \boldsymbol{\mu}_R). \tag{6}$$

A larger Δ_k indicates that keys in the local subset are, on average, more strongly aligned with their dominant direction than the rest of memory. As in the main text, we use Δ_k as a diagnostic of relative directional dominance under bounded memory.

To relate key geometry to retrieval, consider a simplified dot-product attention model with normalized keys, absorbing the attention scale into the query $\mathbf{q} \in \mathbb{R}^{d_l}$. The resulting attention weights are

$$\alpha_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)}, \qquad s_i = \mathbf{q}^\top \hat{\mathbf{k}}_i.$$

The softmax converts query–key score differences into unequal retrieval weights. Whether a directional imbalance in key space produces such a score gap depends on the incoming query. The following proposition formalizes this effect in a simplified two-group model.

Directional contrast and attention concentration under a two-group model. Let the retained memory at query step k be partitioned into a local subset R_k and a non-local subset N_k, with cardinalities

$$m = |R_k|, \qquad n = |N_k|.$$

Assume a simplified score model in which all keys in R_k share a common similarity score a, and all keys in N_k share a common similarity score b, with $a > b$. The softmax attention weights are

$$\alpha_i = \begin{cases} \dfrac{e^a}{me^a + ne^b}, & i \in R_k, \\ \dfrac{e^b}{me^a + ne^b}, & i \in N_k. \end{cases}$$

Define the score gap

$$\delta = a - b > 0.$$

Proposition C.1. *For fixed m and n, the attention entropy*

$$h_{\mathrm{att}}(\boldsymbol{\alpha}) = -\sum_i \alpha_i \log \alpha_i$$

is monotonically decreasing in δ. Consequently, the effective memory size

$$s_{\mathrm{eff}} = \exp\big(h_{\mathrm{att}}(\boldsymbol{\alpha})\big)$$

is also monotonically decreasing in δ.

Proof. Let

$$p_R = \sum_{i \in R_k} \alpha_i = \frac{me^a}{me^a + ne^b} = \frac{me^\delta}{me^\delta + n}, \qquad p_N = 1 - p_R = \frac{n}{me^\delta + n}.$$

Since all weights are tied within each group, the entropy can be written as

$$h_{\mathrm{att}}(\boldsymbol{\alpha}) = -p_R \log \frac{p_R}{m} - p_N \log \frac{p_N}{n}.$$

Equivalently,

$$h_{\mathrm{att}}(\boldsymbol{\alpha}) = h_{\mathrm{bin}}(p_R) + p_R \log m + (1 - p_R) \log n,$$

where

$$h_{\mathrm{bin}}(p) = -p \log p - (1 - p) \log(1 - p)$$

denotes the binary entropy.

Now

$$\frac{dp_R}{d\delta} = \frac{mne^\delta}{(me^\delta + n)^2} > 0.$$

By the chain rule,

$$\frac{dh_{\mathrm{att}}}{d\delta} = \frac{dh_{\mathrm{att}}}{dp_R} \frac{dp_R}{d\delta}.$$

Using

$$h_{\mathrm{att}}(p_R) = h_{\mathrm{bin}}(p_R) + p_R \log m + (1 - p_R) \log n,$$

we obtain

$$\frac{dh_{\mathrm{att}}}{dp_R} = \log \frac{1 - p_R}{p_R} + \log m - \log n = \log \frac{m(1 - p_R)}{np_R}.$$

Substituting

$$p_R = \frac{me^\delta}{me^\delta + n}, \qquad 1 - p_R = \frac{n}{me^\delta + n},$$

gives

$$\frac{m(1 - p_R)}{np_R} = \frac{m \cdot \frac{n}{me^\delta + n}}{n \cdot \frac{me^\delta}{me^\delta + n}} = e^{-\delta}.$$

Hence

$$\frac{dh_{\mathrm{att}}}{dp_R} = \log(e^{-\delta}) = -\delta < 0.$$

Since $\frac{dp_R}{d\delta} > 0$, it follows that

$$\frac{dh_{\mathrm{att}}}{d\delta} < 0.$$

Therefore $h_{\mathrm{att}}(\boldsymbol{\alpha})$ decreases monotonically with δ, and so does $s_{\mathrm{eff}} = \exp(h_{\mathrm{att}}(\boldsymbol{\alpha}))$. □

This proposition shows that, under the fixed two-group model, a larger score gap reduces attention entropy and the effective number of keys contributing to retrieval. The link from directional contrast to this score gap requires an additional query condition. For example, if $\mathbf{q} = \lambda \boldsymbol{\mu}_R$ with $\lambda > 0$, the difference between the average local and non-local scores is exactly $\lambda \Delta_k$. For general queries, components orthogonal to $\boldsymbol{\mu}_R$, key magnitudes, and within-group score variation can also affect the realized distribution. Accordingly, Δ_k is a diagnostic of a potential retrieval imbalance, rather than a direct estimator of attention entropy.

Retention granularity and local directional dispersion. The same diagnostic also clarifies why retention granularity matters. Define

$$A_R = \frac{1}{|R_k|} \sum_{i \in R_k} \cos(\hat{\mathbf{k}}_i, \boldsymbol{\mu}_R), \qquad A_N = \frac{1}{|N_k|} \sum_{j \in N_k} \cos(\hat{\mathbf{k}}_j, \boldsymbol{\mu}_R),$$

so that

$$\Delta_k = A_R - A_N.$$

Because both $\hat{\mathbf{k}}_i$ and $\boldsymbol{\mu}_R$ are unit-normalized,

$$\|\hat{\mathbf{k}}_i - \boldsymbol{\mu}_R\|_2^2 = 2 - 2\cos(\hat{\mathbf{k}}_i, \boldsymbol{\mu}_R),$$

and therefore

$$A_R = 1 - \frac{1}{2|R_k|} \sum_{i \in R_k} \|\hat{\mathbf{k}}_i - \boldsymbol{\mu}_R\|_2^2. \tag{7}$$

Now partition the local retained set R_k into non-empty frame groups S_t. For each frame group S_t, define its mean key vector $\bar{\mathbf{k}}_t \in \mathbb{R}^{d_l}$ by

$$\bar{\mathbf{k}}_t = \frac{1}{|S_t|} \sum_{i \in S_t} \hat{\mathbf{k}}_i,$$

and its intra-frame directional dispersion by

$$\sigma_t^2 = \frac{1}{|S_t|} \sum_{i \in S_t} \|\hat{\mathbf{k}}_i - \bar{\mathbf{k}}_t\|_2^2.$$

Proposition C.2. *Let $R_k = \bigsqcup_t S_t$ be a partition of the local retained keys into frame groups, and define*

$$w_t = \frac{|S_t|}{|R_k|}.$$

Then

$$A_R \le 1 - \frac{1}{2} \sum_t w_t \sigma_t^2. \tag{8}$$

Consequently,

$$\Delta_k \le 1 - \frac{1}{2} \sum_t w_t \sigma_t^2 - A_N. \tag{9}$$

Proof. Starting from (7),

$$A_R = 1 - \frac{1}{2|R_k|} \sum_{i \in R_k} \|\hat{\mathbf{k}}_i - \boldsymbol{\mu}_R\|_2^2.$$

We decompose the sum over frame groups:

$$\frac{1}{|R_k|}\sum_{i\in R_k}\|\hat{\mathbf{k}}_i-\boldsymbol{\mu}_R\|_2^2=\frac{1}{|R_k|}\sum_t\sum_{i\in S_t}\|\hat{\mathbf{k}}_i-\boldsymbol{\mu}_R\|_2^2.$$

For each group S_t, variance decomposition gives

$$\frac{1}{|S_t|}\sum_{i\in S_t}\|\hat{\mathbf{k}}_i-\boldsymbol{\mu}_R\|_2^2=\frac{1}{|S_t|}\sum_{i\in S_t}\|\hat{\mathbf{k}}_i-\bar{\mathbf{k}}_t\|_2^2+\|\bar{\mathbf{k}}_t-\boldsymbol{\mu}_R\|_2^2=\sigma_t^2+\|\bar{\mathbf{k}}_t-\boldsymbol{\mu}_R\|_2^2,$$

where the cross term vanishes because

$$\sum_{i\in S_t}(\hat{\mathbf{k}}_i-\bar{\mathbf{k}}_t)=0.$$

Hence

$$\frac{1}{|S_t|}\sum_{i\in S_t}\|\hat{\mathbf{k}}_i-\boldsymbol{\mu}_R\|_2^2\geq\sigma_t^2.$$

Multiplying by $|S_t|/|R_k|=w_t$ and summing over t yields

$$\frac{1}{|R_k|}\sum_{i\in R_k}\|\hat{\mathbf{k}}_i-\boldsymbol{\mu}_R\|_2^2\geq\sum_t w_t\sigma_t^2.$$

Substituting this into (7) gives

$$A_R\leq 1-\frac{1}{2}\sum_t w_t\sigma_t^2,$$

which proves (8). Finally, since $\Delta_k=A_R-A_N$, we immediately obtain

$$\Delta_k\leq 1-\frac{1}{2}\sum_t w_t\sigma_t^2-A_N.$$

□

This bound makes the role of retention granularity explicit. If memory is controlled at excessively fine granularity, token-level selection can retain only a few highly aligned entries from the same local mode, causing the corresponding intra-frame dispersion σ_t^2 to collapse toward zero. In that case, A_R can approach 1, enlarging the directional contrast between R_k and N_k. By contrast, frame-level retention preserves each frame's KV contribution as a coherent frame-wise context unit. Under the mild assumption that retained frame groups exhibit non-degenerate intra-frame directional dispersion, namely

$$\sigma_t^2\geq\epsilon>0,$$

we obtain

$$A_R\leq 1-\frac{\epsilon}{2},\qquad \Delta_k\leq 1-\frac{\epsilon}{2}-A_N.$$

Under this non-degeneracy assumption, preserving frame groups yields a conditional upper bound on local directional alignment. Frame-level retention preserves the within-frame variation that token selection may discard, but it does not by itself guarantee a positive dispersion bound for every observation or a universal ordering between retention policies. The analysis therefore motivates the empirical comparison of retention granularity rather than establishing unconditional protection against attention collapse.

Connection to the empirical diagnostics. Fig. 5 measures the same key-space diagnostic $\Delta_k=A_R-A_N$ during streaming inference. A larger value indicates stronger relative alignment of the local subset with its representative direction. The propositions explain how this imbalance can favor concentrated retrieval under suitable query conditions and why preserving within-frame variation can limit local alignment. They do not imply that Δ_k alone determines the attention distribution or geometric error. We therefore complement this diagnostic with a post-softmax concentration statistic and a pose-error study below.

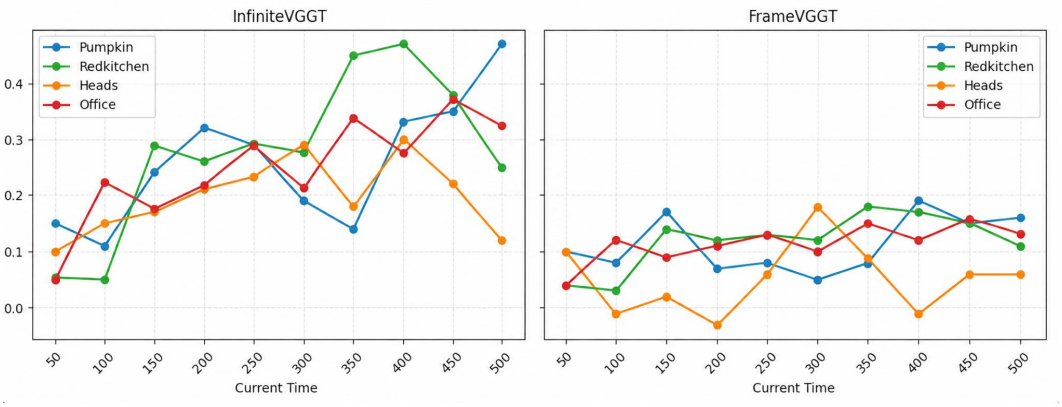


Figure 5: Temporal evolution of directional contrast Δ_k under bounded memory. Larger values indicate a stronger local directional advantage in retained key space. Token-level retention exhibits a higher and increasing contrast, whereas FrameVGGT remains more stable.

C.5 Does Attention Collapse Co-Evolve with Pose Degradation?

The analysis in Sec. 3 uses the directional contrast Δ_k to characterize a key-space condition that can promote attention collapse. However, Δ_k is a *mechanistic diagnostic*: it is computed from the geometry of retained keys before query interaction and depends on how the dominant subset R_k and its complement N_k are identified. It therefore describes the tendency of bounded memory to develop a dominant retrieval mode, rather than the concentration of the attention distribution actually realized during inference.

To measure the latter directly, we use a separate post-softmax statistic, denoted by C_{att}. Let

$$\mathbf{p} = (p_1, \ldots, p_N), \qquad \sum_{i=1}^{N} p_i = 1,$$

be the attention distribution over the $N > 1$ retained memory entries, with $0 \log 0 = 0$. We define attention concentration as

$$C_{\text{att}} = 1000 \left(1 - \frac{H(\mathbf{p})}{\log N}\right), \tag{10}$$

where

$$H(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i \tag{11}$$

is the attention entropy. Since $\log N$ is the maximum entropy of a distribution over N entries, this normalization expresses concentration relative to the uniform-attention case and facilitates comparison across memory sizes. The factor 1000 is used only for numerical readability. A uniform distribution gives $C_{\text{att}} = 0$, while larger values indicate that attention mass is concentrated on a smaller effective subset of memory.

Why use C_{att} rather than Δ_k for the empirical study? The two statistics are intentionally used for different purposes. Directional contrast Δ_k isolates the *upstream key-space imbalance* associated with retention granularity, whereas C_{att} measures the *realized retrieval concentration* after the current query interacts with all retained keys and softmax is applied. Using Δ_k as the final empirical concentration metric would require a particular partition into R_k and N_k, which is less suitable for comparing methods with different memory organizations. In contrast, C_{att} is computed directly from the complete attention distribution and does not require such a partition.

Accordingly, Δ_k and C_{att} are neither numerically comparable nor expected to coincide pointwise. They instead characterize consecutive stages of the same failure mechanism:

$$\underbrace{\Delta_k}_{\text{key-space imbalance}} \longrightarrow \underbrace{\text{query–key logit advantage}}_{\text{query-dependent}} \longrightarrow \underbrace{C_{\text{att}}}_{\text{post-softmax concentration}} .$$

Table 9: Evolution of attention concentration C_{att} and local pose error from frame 100 to 700 on TUM RGB-D. Higher C_{att} indicates more concentrated post-softmax attention.

Method	C_{att} 100	C_{att} 700	Local error 100	Local error 700
FrameVGGT	**5.30**	**5.64** (+6.4%)	**0.0725**	**0.0838** (+15.6%)
InfiniteVGGT	5.46	9.24 (+69.2%)	0.0761	0.1317 (+73.1%)
XStreamVGGT	6.66	10.88 (+63.3%)	0.0794	0.1572 (+98.0%)
O-VGGT	7.55	11.23 (+48.7%)	0.0765	0.1431 (+87.1%)

Table 10: Frame-selection comparison on TUM RGB-D under the same 24-frame budget.

Selector	ATE (m) ↓	Rot. RPE (deg) ↓
Uniform	0.0588	0.3634
Random (5 seeds)	0.0693 ± 0.0040	0.3648
FIFO/Recent-K	0.0826	0.4634
FrameVGGT	**0.0385**	**0.3360**

The two-group analysis in Sec. C.4 formalizes the logit-gap-to-concentration relation: for a fixed partition, increasing the dominant query–key logit gap decreases attention entropy and therefore increases C_{att}. The connection from Δ_k to the logit gap additionally depends on query alignment, which is why we treat Δ_k as a diagnostic of the underlying memory geometry rather than as a direct estimator of C_{att}.

Long-horizon evaluation. We evaluate all eight TUM RGB-D sequences every 100 frames from frame 100 to 700, where all sequences remain available. At each checkpoint, attention concentration and local pose error are evaluated over the preceding 100-frame causal window, with C_{att} summarized by the median. This protocol measures whether increasing concentration of the realized attention distribution is accompanied by degradation in local geometric estimation as the streaming horizon grows.

From frame 100 to 700, the token-level bounded-memory baselines exhibit 48.7%–69.2% increases in C_{att}, accompanied by 73.1%–98.0% increases in local pose error. FrameVGGT limits these increases to 6.4% and 15.6%, respectively. At frame 700, FrameVGGT achieves 39.0%–49.8% lower attention concentration and 36.4%–46.7% lower local pose error than the token-level baselines.

These results complement the key-space diagnostic with measurements of realized attention. The concurrent increases in C_{att} and local pose error are consistent with a loss of effective geometric support as retrieval becomes more concentrated. FrameVGGT suppresses both trends under the same bounded streaming setting. This co-evolution supports the proposed interpretation, but does not by itself establish a causal relation between attention concentration and pose error.

C.6 Does Frame Selection Matter Beyond Retention Granularity?

Frame-wise retention preserves complete observations, but it does not by itself determine which observations should remain under a fixed budget. We therefore test whether the gains of FrameVGGT can be reproduced using simple frame-level selection policies once token fragmentation is removed.

We replace only the selector for the non-anchor memory slots with **Uniform**, **Random**, or **FIFO/Recent-K** retention. Inputs, total retained-frame budget, anchor handling, backbone weights, and evaluation pipelines remain unchanged.

For a fixed-capacity online bank, FIFO is equivalent to Recent-K: each incoming observation removes the oldest retained entry.

C.6.1 Pose Estimation on TUM RGB-D

Random sampling is averaged over five pre-defined seeds, producing an ATE range of $[0.065, 0.074]$ m. Even the best random realization remains worse than Uniform and substantially worse than FrameVGGT.

Table 11: Frame-selection comparison on 7-Scenes under the same retained-frame budget.

Selector	Acc. ↓	Comp. ↓	NC ↑
Uniform	0.1355	0.0200	0.5399
Random (5 seeds)	0.1920 ± 0.0278	0.0290 ± 0.0066	0.5426 ± 0.0131
FIFO/Recent-K	0.2320	0.0266	0.5413
FrameVGGT	**0.0273**	**0.0190**	**0.5640**

Relative to Uniform, the strongest simple baseline, FrameVGGT reduces ATE by 34.5% and rotational RPE by 7.5%. Relative to Random and FIFO/Recent-K, it reduces ATE by 44.4% and 53.4%, respectively.

C.6.2 3D Reconstruction on 7-Scenes

Across the five random seeds, Acc., Comp., and NC lie within $[0.1610, 0.2310]$, $[0.0215, 0.0380]$, and $[0.5260, 0.5590]$, respectively. Even the best random run remains worse than FrameVGGT on all three metrics.

Relative to Uniform, FrameVGGT reduces Acc. by 79.8%, improves Comp. from 0.0200 to 0.0190, and increases NC from 0.5399 to 0.5640.

These results separate two aspects of FrameVGGT. Frame-level granularity preserves coherent evidence *within* an observation, while observation-level selection determines the diversity of evidence preserved *across* observations. Complete-frame retention avoids within-frame token fragmentation in this formulation, while the choice of retained frames remains important for maintaining broad geometric support.

C.7 Memory Capacity and Diminishing Returns

These observations motivate examining both memory capacity and its organization. When the budget is very small, increasing capacity can recover missing geometric support. At larger budgets, additional observations may be redundant or only weakly informative, so the marginal benefit depends on the dataset, sequence horizon, and retention policy. This does not imply a universal finite saturation point or an intrinsic limit at the default 24-frame setting. The capacity sweeps in Sec. H.2 distinguish behavior near the default budget from scaling on longer streams.

Why key space? In attention-based memory, Keys and Values play asymmetric roles: Keys determine how memory is addressed by future queries, while Values provide the retrieved content. Accordingly, redundancy for memory selection is most naturally measured in key space. By summarizing each frame's incremental KV contribution into a prototype $v_t^{(l)}$, we obtain a compact descriptor of its retrieval direction while preserving block-level coherence.

D Anchor Tier: Long-Range Reference Selection

To complement the bounded mid-term bank, we maintain a sparse *anchor tier* that preserves a small number of long-range reference frames. While the mid-term bank provides the primary bounded memory over recent-to-mid horizons, anchors serve as persistent geometric references when local context becomes unreliable, e.g., under blur, occlusion, weak parallax, or rotation-dominant motion. This separation allows the model to retain both dense local evidence and sparse global references under a fixed memory budget.

The mid-term bank maintains adaptive coverage of complementary observations, but coverage alone does not guarantee persistence. A distant observation may be removed before its geometric value becomes apparent, particularly in repetitive, texture-poor, occluded, or revisited scenes. Frames from distinct viewpoints can appear similar in retrieval space while still providing useful long-baseline constraints. Once evicted from bounded memory, such evidence cannot be recovered.

Sparse anchors address this delayed-utility failure mode by protecting a small number of reliable and geometrically novel observations from frequent replacement.

D.1 Anchor Promotion

Anchor promotion is regulated by a temporal sparsity constraint

$$g_t = t - t_{\text{last}} \geq G, \tag{12}$$

where t, t_{last}, and G are scalars. This prevents over-concentration in locally redundant intervals. Among eligible candidates, we prioritize frames that are both geometrically reliable and non-redundant.

Frame-level reliability. We define the anchor reliability score as

$$\phi(i) = q_i \, s_i, \tag{13}$$

where $q_i \in \mathbb{R}$ denotes a frame-level model-confidence score and $s_i \in \mathbb{R}$ measures image sharpness.

For q_i, we use a lightweight frame-level confidence proxy derived from the model's own geometric predictions. Specifically, let $\{c_{i,u}\}_{u=1}^{N_i}$ denote per-pixel or per-token confidence values at frame i, where $c_{i,u} \in \mathbb{R}$ and N_i is the number of pixels or tokens used for aggregation. Depending on the backbone, these may correspond to matching confidence, visibility confidence, or prediction confidence. We aggregate them into a single frame-level score by

$$q_i = \frac{1}{N_i} \sum_{u=1}^{N_i} c_{i,u}, \tag{14}$$

followed by min–max normalization within a sliding temporal buffer. Intuitively, q_i favors frames whose predicted geometry is more self-consistent and reliable for later reuse.

For s_i, we use a blur-sensitive image-quality proxy based on the variance of the Laplacian:

$$s_i = \operatorname{Var}\left(\nabla^2 I_i\right), \tag{15}$$

where I_i denotes frame i, and $\nabla^2 I_i$ is its grayscale Laplacian response. This quantity is high for sharp images with rich local structure and low for blurred or low-detail frames. We normalize s_i to $[0, 1]$ within the same temporal buffer before combining it with q_i.

Anchor novelty. To encourage diversity, we additionally require novelty with respect to the current anchor set A_t. Let $\bar{\mathbf{p}}_i \in \mathbb{R}^{d_p}$ denote the normalized pose signature of candidate frame i, where d_p is the pose-signature dimension. We define the novelty score as

$$\nu(i) = \min_{a \in A_t} \left(1 - \langle \bar{\mathbf{p}}_i, \bar{\mathbf{p}}_a \rangle\right). \tag{16}$$

This discourages promoting frames that are geometrically redundant with already retained anchors.

A candidate frame i is promoted only if

$$g_t \geq G, \qquad \phi(i) \geq \tau_{\text{conf}}, \qquad \nu(i) \geq \tau_{\text{novel}}, \tag{17}$$

and the anchor tier retains at most M_{anc} frames using FIFO eviction, except for the first frame, which is always preserved as a persistent global reference.

From a retrieval perspective, the anchor tier extends coverage beyond the effective horizon of the mid-term bank by preserving sparse but non-local geometric evidence. This complements adaptive mid-term coverage with protected long-range persistence. In the same-budget ablations below, anchor slots are taken from the mid-term allocation rather than added to the total budget.

D.2 Same-Budget Anchor Ablation

We isolate the anchor contribution under the same 24-frame total memory budget. The anchor variant allocates four slots to sparse anchors, whereas the non-anchor variant returns these four slots to the mid-term bank. Thus, the comparison changes only how the fixed budget is organized. The aggregate reconstruction results are reported in Tab. 12, with representative challenging sequences detailed in Tab. 13.

On easier sequences, anchors have limited influence when the mid-term bank already preserves sufficient context. The absolute Acc. difference between the two variants is below 0.003 on $13/18$

Table 12: Reconstruction results on 7-Scenes and NRGBD. The two variants retain the same total number of frames.

Method	7-Scenes						NRGBD					
	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑	Acc↓	Acc_{med} ↓	Comp↓	Comp_{med} ↓	NC↑	NC_{med} ↑
Ours (Without Anchor)	0.028	0.009	0.019	0.0040	0.564	0.598	0.054	0.035	0.030	0.0060	0.670	0.782
Ours (With Anchor)	0.026	0.009	0.018	0.0040	0.571	0.607	0.053	0.034	0.031	0.0057	0.682	0.793

Table 13: Sparse-anchor ablation on representative challenging sequences. Both variants use the same total 24-frame memory budget.

Scene	Setting	Acc. ↓	Comp. ↓	NC ↑
office/seq-02	Non-anchor	0.11794	0.01329	0.56566
	Anchor	**0.09061**	**0.01043**	**0.56626**
office/seq-06	Non-anchor	0.08753	0.06194	0.55701
	Anchor	**0.06688**	**0.02664**	**0.57206**
redkitchen/seq-03	Non-anchor	0.05017	**0.01103**	**0.56792**
	Anchor	**0.04203**	0.01128	0.56526
staircase	Non-anchor	0.09819	0.03055	0.72019
	Anchor	**0.08896**	**0.02984**	**0.72325**

7-Scenes sequences and $8/9$ NRGBD sequences. The benefit becomes more pronounced under viewpoint changes, revisits, occlusion, clutter, and repeated geometry.

On `office/seq-02`, anchors reduce Acc. and Comp. by 23.2% and 21.5%, respectively. On `office/seq-06`, the corresponding reductions reach 23.6% and 57.0%, together with improved normal consistency. For `redkitchen/seq-03`, Acc. improves by 16.2% with small trade-offs in the remaining metrics. All three metrics improve on `staircase`.

The same trend appears in the main-paper ablations: anchors reduce Acc. by 35.9% on `office/seq-09`, while on `stairs/seq-01`, Acc. and Acc._{med} improve by 10.2% and 6.3%, respectively.

D.3 Anchor Hyperparameter Sensitivity

We further evaluate the sensitivity of the anchor promotion policy under the first-1,000-frame TUM RGB-D protocol. The default configuration contains 20 mid-term frames and four sparse anchors.

Confidence and novelty thresholds are highly stable: varying confidence from 0.25 to 0.45, or novelty from 0.10 to 0.30, changes paired geometric-mean ATE by less than 0.2%.

The temporal-gap setting has a larger effect. Changing the reported gap setting from 500 to 100 frames increases promotion frequency from 2.55 to 10.60 anchors per 1,000 frames and worsens ATE by 53.4%. The 250-frame setting increases promotions to 4.43 and worsens ATE by 11.1%. The disabled-gap setting performs nearly identically to the 500-frame default.

The same-budget gains and the sensitivity to excessive promotion are consistent with the role of anchors as *sparse protected references*. They indicate that the benefit cannot be attributed to additional memory capacity and does not require frequent replacement of the protected references.

E Additional Reconstruction Visualization

We further examine the long-horizon behavior of reconstructed geometry. Compared with per-frame depth maps, reconstruction renderings provide a more direct view of whether the retained memory preserves *global structural consistency* as evidence accumulates over time.

Fig. 6 presents additional qualitative reconstruction results on 7-Scenes and NRGBD. Across extended streaming horizons, the reconstructed geometry remains visually coherent: dominant planes, major scene boundaries, and overall layout structure are stably preserved. Although regions with weaker multi-view context may exhibit mild local uncertainty as the sequence grows, we do not observe

Table 14: Sensitivity to anchor promotion parameters on TUM RGB-D. The paired ratio is the geometric-mean ATE ratio relative to the default.

Setting	ATE (m) ↓	Paired ratio	Anchors / 1k
Gap 100	0.05906	1.534	10.60
Gap 250	0.04277	1.111	4.43
Gap 500	**0.03850**	**1.000**	2.55
Gap disabled	0.03854	1.001	1.07
Conf. 0.25	0.03850	0.9999	2.55
Conf. 0.35	**0.03850**	**1.0000**	2.55
Conf. 0.45	0.03857	1.0017	2.55
Novelty 0.10	0.03856	1.0016	2.55
Novelty 0.20	**0.03850**	**1.0000**	2.55
Novelty 0.30	0.03843	0.9982	2.55

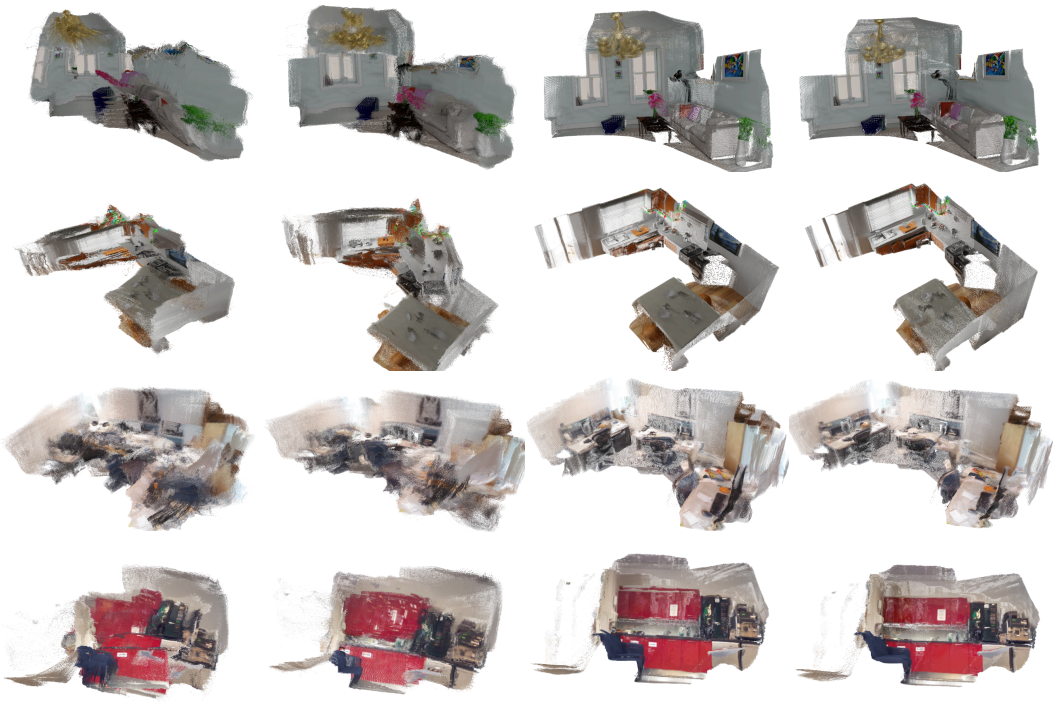


Figure 6: Additional reconstruction visualizations on 7-Scenes and NRGBD. We show representative renderings from FrameVGGT over extended streaming sequences. These examples complement the quantitative reconstruction results in the main paper by illustrating long-horizon structural consistency under bounded memory.

large-scale structural collapse or severe global distortion. This suggests that the retained memory continues to preserve sufficiently coherent geometric context even in prolonged streaming regimes.

To further isolate the role of memory structure, Figs. 7–9 compare three retention regimes: **recent-only**, **mid-term**, and the **global anchor tier**.

Recent-only memory (Fig. 7). When memory is concentrated on only the most recent observations, reconstruction is driven primarily by short-range temporal continuity. This often suffices in early segments, where adjacent views still provide dense local context. However, as the horizon extends, the lack of longer-range geometric witnesses makes the reconstruction increasingly vulnerable to accumulated drift, duplication artifacts, and weakened global consistency. In our framework, this



Figure 7: Additional reconstruction visualizations with **recent-only** memory on 7-Scenes and NRGBD.

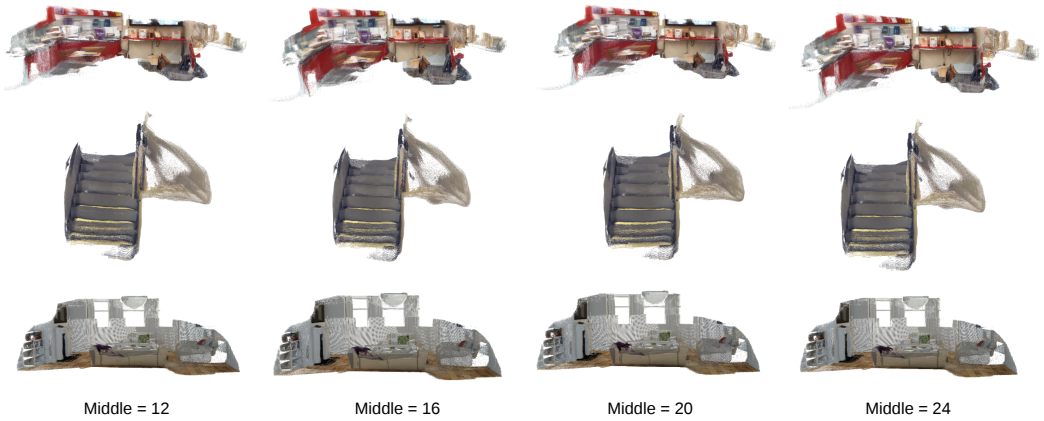


Figure 8: Additional reconstruction visualizations with **mid-term** memory on 7-Scenes and NRGBD.

corresponds to a regime where local context remains available but the broader geometric skeleton of the scene is no longer sufficiently preserved.

Mid-term memory (Fig. 8). Retaining mid-horizon context substantially improves long-range stability by preserving additional support from intermediate temporal neighborhoods. Compared with the recent-only regime, dominant structures are more consistently maintained and drift accumulation is visibly reduced. This indicates that memory aligned with intermediate-range context is more effective at sustaining global reconstruction coherence. Nevertheless, when the sequence extends well beyond the effective temporal span of the mid-term bank, residual artifacts can still emerge, reflecting the finite coverage of a bounded but non-global memory tier.

Global anchor tier (Fig. 9). Introducing the global anchor tier further strengthens reconstruction stability by preserving sparse but persistent long-horizon reference frames. Qualitatively, anchors help suppress duplication artifacts, stabilize scene layout, and better maintain a consistent global structure over extended horizons. In effect, the anchor tier acts as a sparse geometric skeleton that complements the online mid-term pathway with globally distributed reference points. The aggregate results in Tab. 12 show generally comparable or improved reconstruction quality with small metric trade-offs. The challenging-sequence analysis in Tab. 13 further illustrates where protected long-range references are most useful.

Overall, these additional visualizations suggest that long-horizon reconstruction quality depends not only on the amount of retained evidence, but more fundamentally on the *temporal structure* of that evidence. Structured retention better preserves the memory organization required for stable geometry, which in turn yields stronger long-range reconstruction consistency. These observations complement the depth and trajectory analyses below.

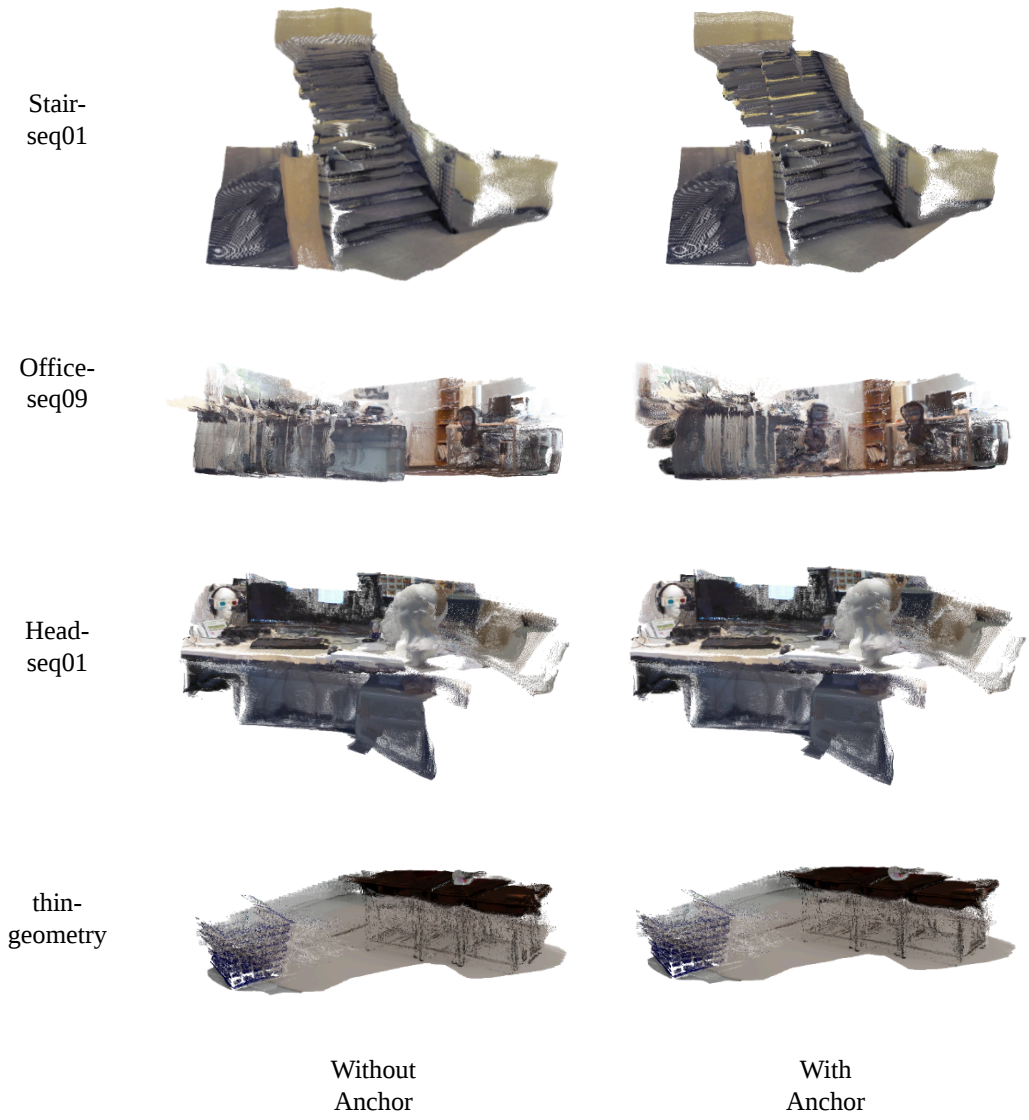


Figure 9: Additional reconstruction visualizations with the **global anchor tier** on 7-Scenes and NRGBD.

F Additional Depth Visualization

Before analyzing pose trajectories, we first examine the long-horizon behavior of predicted depth. Depth stability provides a direct proxy for whether the retained memory continues to preserve sufficiently coherent local geometric context over time.

Unlike pose trajectories, depth visualizations across different methods are not strictly comparable in a fully fair manner because of scale-shift ambiguity, normalization differences, and colormap choices. We therefore focus on the temporal behavior of FrameVGGT itself rather than presenting side-by-side method comparisons.

Fig. 10 shows representative depth maps across extended streaming sequences on BONN. In early segments, the predicted geometry is sharp and spatially coherent. As time progresses, regions with weaker multi-view context may exhibit mild local uncertainty; however, large-scale structural collapse, catastrophic distortion, or obvious degeneration of dominant layouts is not observed. Major

Table 15: Full-length TUM RGB-D pose accuracy and cross-sequence robustness.

Method	Mean ATE (m) ↓	Seq. std. (m) ↓	Wins	Paired gain
InfiniteVGGT	0.0665	0.0445	–	–
FrameVGGT	**0.0447**	**0.0303**	**8/8**	**31.5%**

planes and scene boundaries remain temporally stable over long horizons, indicating that the retained memory continues to provide sufficiently coherent geometric context for depth inference.

These qualitative observations are consistent with the intended role of structured memory retention: preserving informative mid-horizon context helps maintain local geometric coherence even as the stream becomes progressively longer. This depth-level stability complements the trajectory behavior analyzed next.

G Additional Pose Visualization

G.1 Full-Length Pose Robustness

The main TUM RGB-D evaluation uses a first-1,000-frame protocol for controlled comparison. We additionally process every available frame from all eight TUM sequences, with sequence lengths ranging from 703 to 1,258 frames. All methods use identical hardware, frame order, model weights, and inference settings.

FrameVGGT uses deterministic memory selection under fixed inputs and settings. We therefore do not treat repeated identical inference as independent algorithmic trials. We therefore evaluate robustness through sequence-level consistency, win rate, and paired significance.

FrameVGGT improves ATE on all $8/8$ sequences and yields a 31.5% paired geometric-mean gain. An exact two-sided sign test gives $p = 0.0078$. Rotational RPE improves by 10.4% with $7/8$ wins.

These values differ from the first-1,000-frame protocol because the full-length evaluation processes every available frame, but both protocols produce the same ordering between FrameVGGT and InfiniteVGGT.

G.2 Trajectory Visualizations

To further characterize long-horizon pose behavior beyond aggregate metrics, we provide additional qualitative trajectory visualizations on TUM in Fig. 11, together with focused comparisons of different retention regimes in Figs. 12a and 12b. These figures serve as qualitative evidence of how bounded-memory structure affects long-range trajectory stability.

Fig. 11 compares representative long-sequence trajectories across methods. A consistent pattern is that deviations from the global trajectory shape accumulate over time: many baselines remain reasonably stable in earlier segments but exhibit increasingly visible drift in later portions of the sequence. Although the precise failure signatures vary—for example, gradual drift versus abrupt deviation—the common trend is the progressive erosion of global consistency under long-horizon bounded streaming.

Figs. 12a and 12b isolate the role of temporal retention structure. The *mid-term* regime preserves intermediate-range history, which provides complementary constraints once motion extends beyond immediate neighbors. This additional context helps stabilize the trajectory against drift that cannot be corrected using only near-adjacent evidence. In contrast, the *recent-only* regime concentrates capacity on short-range observations. While this often preserves local continuity, it provides weaker constraints on long-horizon global consistency, especially when viewpoint diversity or baseline growth requires support from a broader temporal neighborhood.

Overall, these qualitative examples align with the quantitative results in the main paper: preserving informative mid-horizon context improves long-range trajectory stability by better maintaining the geometric context structure needed for consistent pose estimation over extended streams.

The attention–pose co-evolution study in Sec. C.5 provides a quantitative complement to these trajectory visualizations.

Table 16: Sensitivity to mid-term memory capacity on TUM RGB-D. Four anchor slots are retained in all configurations.

Mid-term capacity	ATE (m) ↓	Paired ratio
12	0.04247	1.103
16	0.03858	1.002
20	**0.03850**	**1.000**
28	0.03889	1.010
36	0.03746	0.973

H KV-Cache Budget, Scaling, and Computational Efficiency

H.1 KV-Cache Budget and Memory Accounting

For controlled memory-policy comparisons, methods operate under comparable KV-cache budget settings. In particular, memory usage is measured by the effective footprint of the retained Transformer KV cache during streaming inference.

For a Transformer layer l, the KV cache size is proportional to

$$\mathrm{Mem}_{KV}^{(l)} \propto H_l \, T_l \, D_l \times 2, \tag{18}$$

where H_l is the number of attention heads, T_l the number of retained tokens, and D_l the feature dimension per head; the factor of 2 accounts for keys and values. Total KV-cache memory is obtained by summing this quantity across all layers.

Although different methods organize memory differently (e.g., token-level vs. frame/block-level retention), the reported memory footprint always reflects the actual retained KV tensors used during inference. KV-cache footprints should be distinguished from peak total GPU allocations, which also include model parameters, activations, temporary tensors, and other runtime costs.

The KV-only measurement excludes overheads such as dataset loading, visualization buffers, and evaluation pipelines. The peak GPU and CPU measurements in Tab. 18 are reported separately as system-level efficiency metrics.

H.2 Scaling with the Memory Budget

The default 24-frame configuration is a practical operating point rather than an architectural limit. We evaluate both smaller capacities around the default setting and larger memory budgets on longer streams.

H.2.1 Sensitivity Around the Default Budget

For TUM RGB-D, the default configuration contains 20 mid-term frames and four anchors, for a total budget of 24 frames. During the following capacity sweep, the four anchor slots remain fixed while the mid-term capacity is varied.

Mid-term capacities from 16 to 36 produce similar performance, with paired ratios between 0.973 and 1.010. Only the 12-frame setting exhibits a noticeable degradation, increasing ATE by 10.3%. This suggests an under-capacity regime rather than narrow sensitivity around the default operating point.

H.2.2 Scaling Beyond 24 Retained Frames

We further evaluate total retained-frame budgets of 36 and 48 on Long3D. For variants with anchors, four slots are assigned to the sparse anchor tier while the total retained-frame budget remains unchanged. Tab. 17 reports total retained-frame budgets; for configurations with four anchors, those slots are included in the stated total.

Increasing the budget consistently improves all six reconstruction metrics. From 24 to 36 retained frames, Acc. and $\mathrm{Acc}_{\mathrm{med}}$ decrease by 6.5% and 6.4%, respectively. Increasing from 36 to 48 frames yields further reductions of 6.6% and 6.5%.

Table 17: Long3D reconstruction under larger retained-frame budgets.

Memory	Acc. ↓	Acc._{med} ↓	Comp. ↓	$\text{Comp.}_{\text{med}}$ ↓	NC ↑	NC_{med} ↑
24	1.698	1.331	1.032	0.433	0.536	0.540
36	1.587	1.246	0.968	0.407	0.564	0.569
48	1.483	1.165	0.914	0.384	0.585	0.590
36 w/ 4 anchors	1.498	1.201	0.944	0.398	0.574	0.579
48 w/ 4 anchors	**1.451**	**1.141**	**0.896**	**0.377**	**0.593**	**0.599**

Table 18: End-to-end efficiency on all eight full-length TUM RGB-D sequences.

Method	Peak GPU (GiB) ↓	Peak CPU (GiB) ↓	Latency mean / p95 (ms) ↓	Streaming FPS ↑	E2E FPS ↑	Time / seq. (s) ↓	Selection cost (ms/frame; %) ↓
InfiniteVGGT	28.0	12.63	186.13 / 194.77	5.38	5.13	181.62	56.73; 30.48%
XStreamVGGT	30.0	**12.35**	**96.43 / 102.39**	**10.37**	**9.50**	**98.04**	**6.76; 7.01%**
O-VGGT	27.0	12.96	129.56 / 136.36	7.74	7.24	128.98	17.83; 13.76%
FrameVGGT	**23.0**	12.88	125.80 / 132.50	7.95	7.46	125.20	7.12; **5.66%**

Sparse anchors remain complementary at larger capacities. Under the same total memory budget, reserving four slots for anchors reduces Acc. by 5.6% at 36 frames and by 2.2% at 48 frames.

These experiments show that 24 frames is not an intrinsic saturation point. FrameVGGT instead exposes a controllable accuracy–memory trade-off: additional capacity can preserve broader cross-view support, while the frame-wise formulation remains effective across a wide range of budgets.

H.3 End-to-End Computational Efficiency

Using the same eight full-length TUM RGB-D sequences described in Sec. G.1, we measure runtime three times per sequence, totaling 24 runs and 22,353 processed frames per method.

Compared with InfiniteVGGT, FrameVGGT reduces mean latency by 32.4%, end-to-end runtime by 31.1%, and peak GPU allocation by 17.9%, while increasing end-to-end throughput by 45.4%.

Memory selection requires 7.12 ms per frame, compared with 56.73 ms for InfiniteVGGT. Including descriptor construction and all memory-management operations, FrameVGGT incurs approximately 16.0 ms/frame of total memory-update overhead. Anchor maintenance itself requires only 0.0014 ms/frame.

XStreamVGGT achieves higher throughput, but its full-length TUM ATE is 0.1210 m, compared with 0.0447 m for FrameVGGT, corresponding to a $2.7\times$ larger pose error.

H.4 Bounded Update Complexity

FrameVGGT never revisits the complete stream history during memory selection. When the mid-term bank overflows, selection operates only on the currently retained mid-term segments together with the newly arriving segment.

Consequently, the candidate set is bounded by the configured memory capacity rather than sequence length. Segment representatives and pairwise distances are evaluated with batched GPU operations. For a fixed memory budget, memory-management cost therefore remains bounded as the stream grows.

FrameVGGT thus bounds both the amount of historical KV memory retained and the amount of history examined during each memory update.

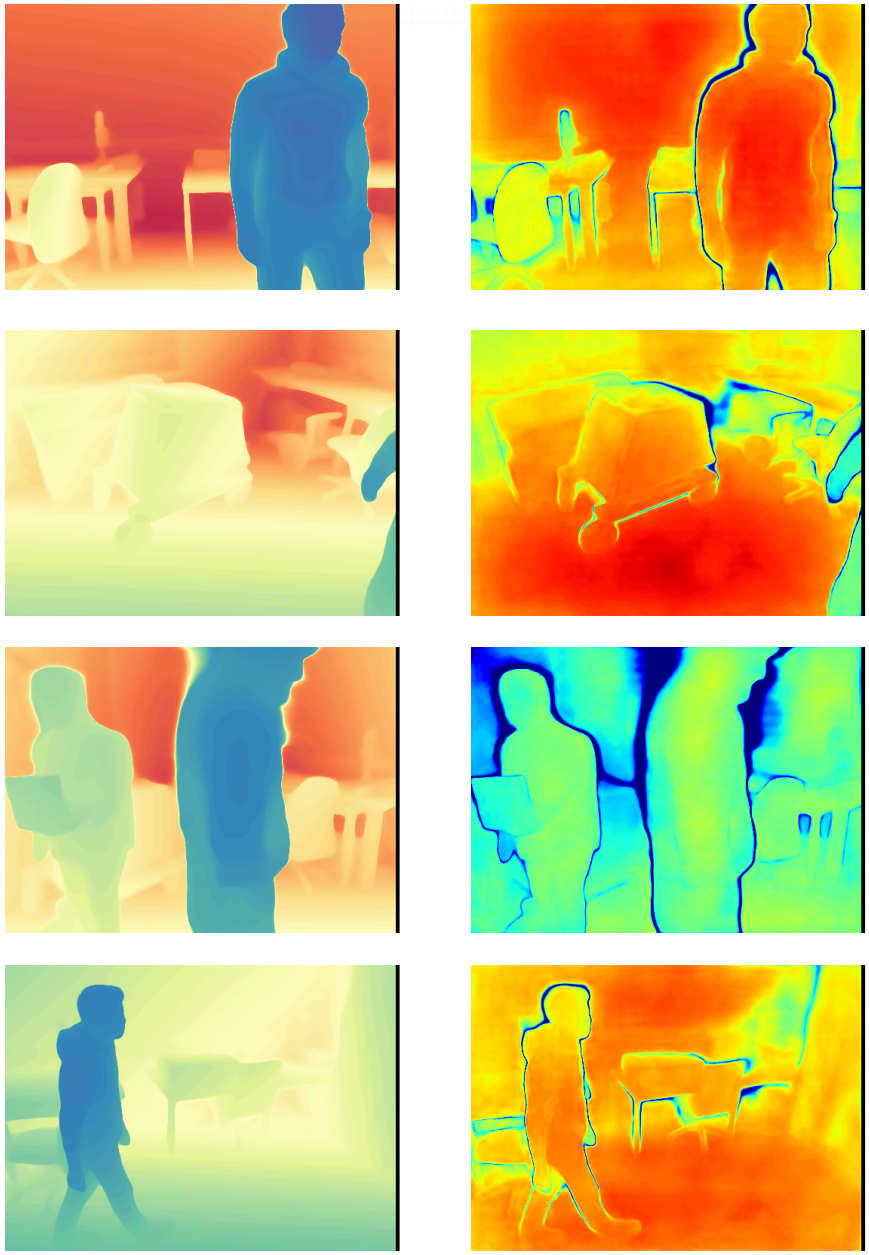
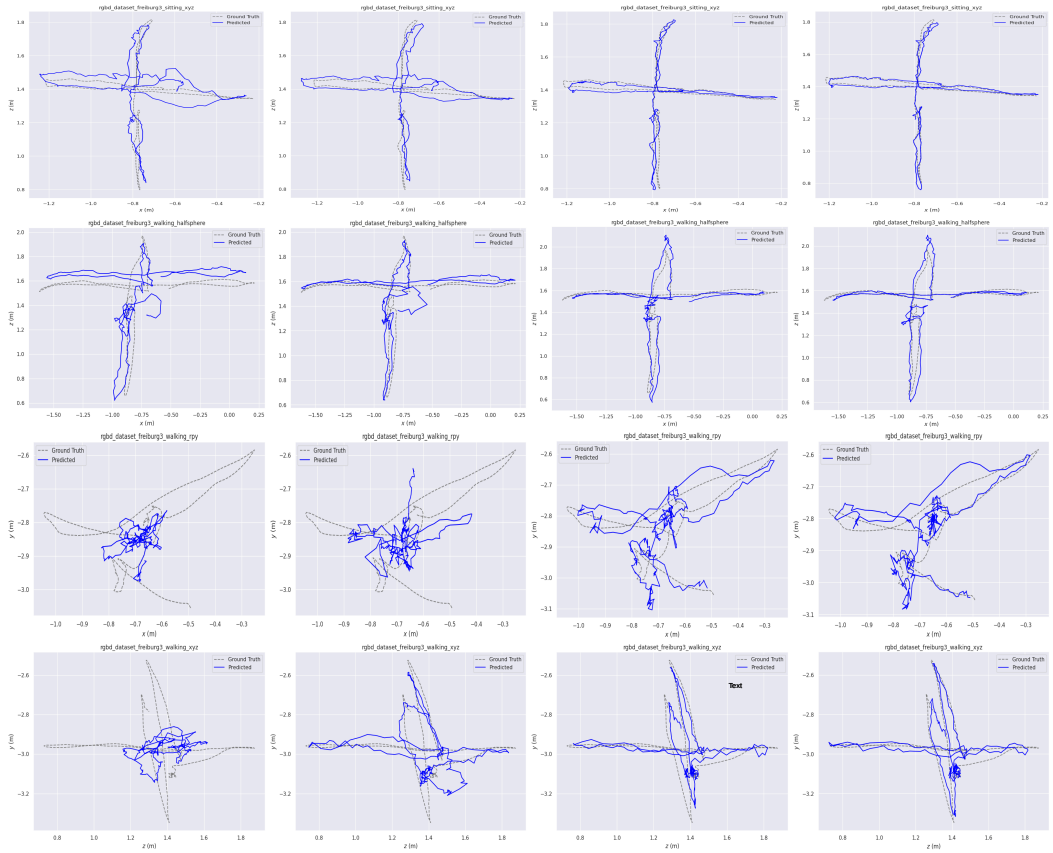
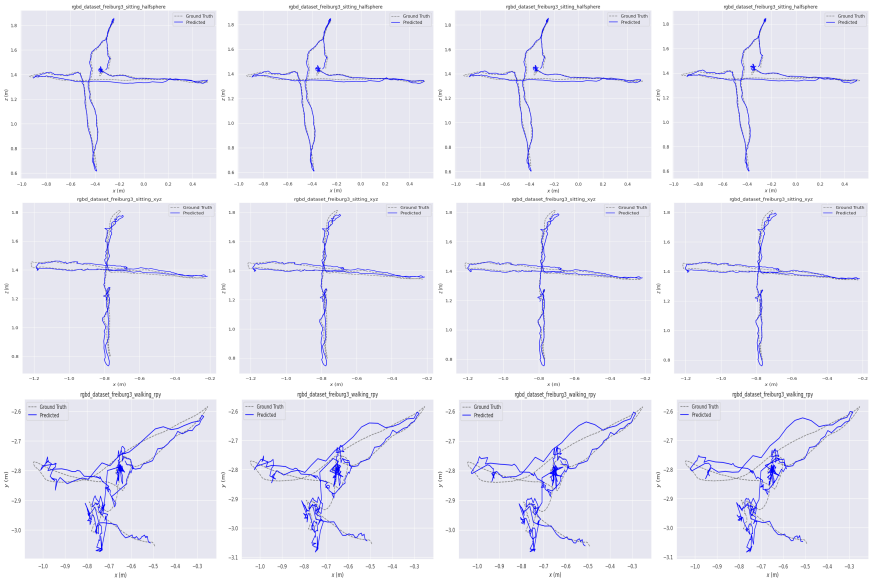


Figure 10: Additional depth visualization results on BONN. We visualize predicted depth maps from FrameVGGT over extended streaming sequences.

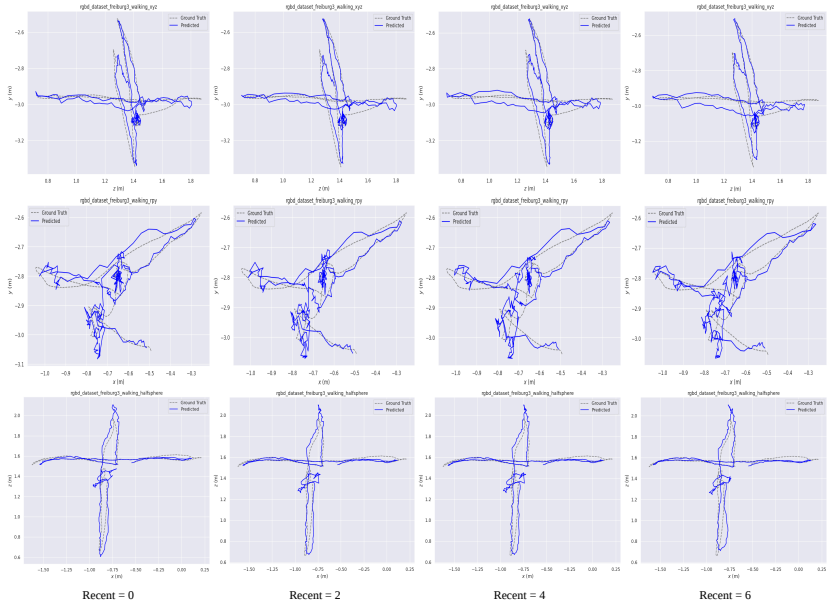


CUT3R TTT3R InfiniteVGGT FrameVGGT

Figure 11: Additional pose visualization results on the TUM dataset. The figure shows representative trajectory comparisons over longer sequences.



(a) Additional visualization focusing on the behavior of the mid-term memory regime. Preserving intermediate-range history provides stronger temporal context beyond the immediate local context.



(b) Additional visualization focusing on the behavior of the recent-memory regime. Retaining only recent observations is effective for short-range continuity but offers limited context over extended horizons.

Figure 12: Effect of temporal memory organization on pose estimation. Mid-term memory preserves intermediate-range observations, whereas recent-only memory emphasizes short-range continuity.