# STREAMCACHEVGGT: STREAMING VISUAL GEOMETRY TRANSFORMERS WITH CROSS-LAYER SCORING AND UNCERTAINTY-GATED COMPACTION

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

## ABSTRACT

Reconstructing dense 3D geometry from continuous video is severely bottlenecked by the unbounded growth of streaming KV caches. We present StreamCacheVGGT, a constant-budget visual geometry transformer equipped with two training-free modules. Cross-Layer Consistency-Enhanced Scoring (CLCES) tracks rank stability across Transformer layers to yield robust token importance scores and uncertainty estimates. Uncertainty-Gated Deferred Compaction (UGDC) then routes candidates through either immediate retain, merge, and evict compaction or a small deferred pool for short-term verification. This deferred pool operates strictly within the fixed per-layer budget ($B_l = M_l + D_l$), ensuring the persistent memory footprint remains completely unchanged. Across diverse indoor, outdoor, long-sequence, and depth benchmarks, our method significantly improves geometric completeness and long-horizon stability while transparently reporting the expected throughput trade-off of deferred verification.

## 1 INTRODUCTION

Reconstructing dense and consistent 3D geometry from continuous video streams remains a fundamental and challenging problem in computer vision. Classical Structure from Motion and Multi View Stereo pipelines rely on iterative optimization and remain vulnerable to error propagation over time. Geometric Foundation Models (Leroy et al., 2024; Wang et al., 2024; 2025a) offer a powerful alternative by regressing geometry directly using a unified Transformer architecture. Despite their impressive reconstruction quality, models such as VGGT (Wang et al., 2025a) exhibit quadratic memory complexity. This immense computational burden restricts their application strictly to short offline video sequences.

To overcome these offline limitations, researchers introduced causal attention variants like StreamVGGT (Zhuo et al., 2025). These models enable continuous single pass streaming inference without requiring global sequence alignment. Unfortunately, their key and value caches grow linearly with each processed frame. This unbounded memory accumulation eventually exhausts GPU resources and prevents long horizon deployment. Enforcing a strictly constant $\mathcal{O}(1)$ cache budget is therefore mandatory. Maintaining this rigid budget requires compressing the token stream, which immediately exposes two critical bottlenecks for long sequence geometric stability.

The first major bottleneck lies in accurately estimating token importance. Modern architectures leverage FlashAttention (Dao et al., 2022), which avoids materializing attention weights to conserve memory. This optimization forces streaming models to rely on inexpensive local proxies for token evaluation, such as the magnitude of individual Feed Forward Network residual vectors (Bebis & Georgiopoulos, 2002). However, these localized proxies suffer from representational bias and layer specific activation noise. For example, a texture token suppressed in early processing stages might become absolutely essential for global spatial reasoning in deeper layers. Relying on an isolated snapshot from a single layer often leads the model to misjudge true geometric salience, causing it to prematurely evict vital features.

The second critical bottleneck stems from the destructive nature of hard eviction itself. Tokens from textureless regions including smooth walls, flat floors, and subtle depth gradients frequently fall below retention thresholds when evaluated individually. Collectively, however, these exact



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.

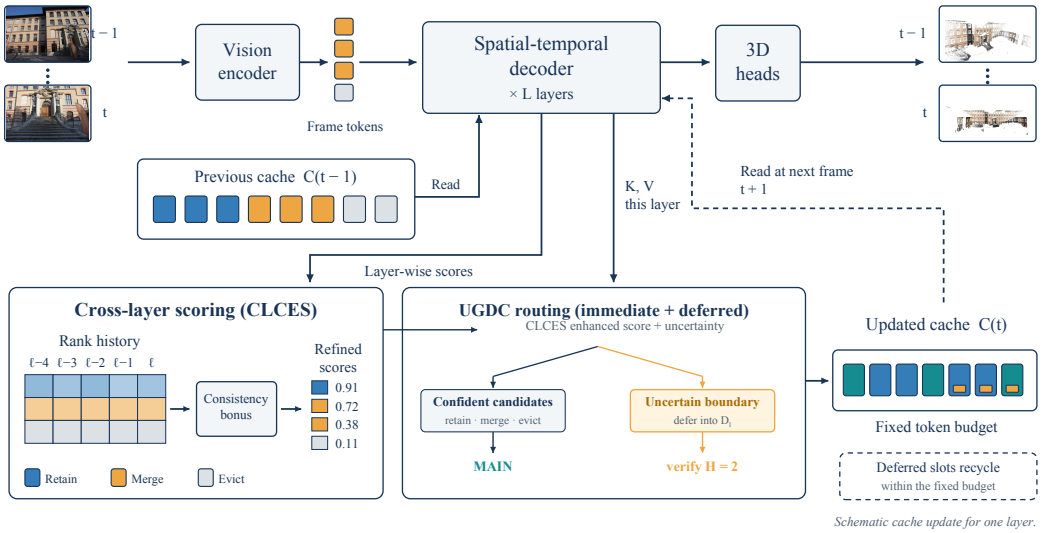



Figure 1: StreamCacheVGGT overview. Consecutive frames share a bounded streaming cache. CLCES produces an enhanced score and uncertainty estimate; UGDC routes confident candidates through an immediate retain, merge, and evict compaction process while deferring uncertain boundary candidates for verification over two frames. Cache updates are schematic; the photographs and point maps are original manuscript materials.

same tokens encode the indispensable background context required to maintain global spatial scale, ensure surface continuity, and facilitate accurate camera pose tracking. Unlike redundant words in language modeling, permanently deleting visual geometric context causes severe accumulating alignment errors. Over extended streaming periods, this progressive information loss inevitably leads to complete reconstruction collapse, even if the initial token importance estimates were perfectly accurate.

To resolve these vulnerabilities directly, we present **StreamCacheVGGT**, a constant cost $\mathcal{O}(1)$ streaming 3D reconstruction framework illustrated in Figure 1. Our architecture integrates two novel modules that require no additional training. Cross Layer Consistency Enhanced Scoring refines activation based importance estimates by tracking rank stability across a window of Transformer layers. Uncertainty Gated Deferred Compaction then executes the crucial cache routing decisions. Highly confident candidates are immediately routed to a retain, merge, or evict pipeline. Meanwhile, uncertain boundary candidates are temporarily deferred and verified over a short temporal window. Importantly, this deferred pool operates entirely within the fixed persistent memory budget.

In summary, our main contributions are as follows:

- We introduce Cross Layer Consistency Enhanced Scoring to neutralize single layer evaluation noise. By tracking rank consistency across a sliding window of Transformer layers, this module filters transient activation spikes and yields a robust metric for geometric token salience that remains fully compatible with FlashAttention.
- We incorporate a retain, merge, and evict compaction strategy into the immediate branch of our cache manager. This module assigns moderately important tokens to their closest retained representatives in cosine space before applying a score weighted feature fusion, preserving distributed geometric context without expanding the cache budget.
- We introduce Uncertainty Gated Deferred Compaction to strategically delay decisions for borderline tokens. This module isolates uncertain eviction candidates and verifies their geometric relevance over a two frame window without increasing the overall memory.
- We conduct extensive evaluations across indoor, outdoor, long sequence, and video depth benchmarks. Our experiments quantify both the significant geometric accuracy gains and the explicit throughput cost associated with the deferred verification modules under a strict constant memory protocol.

# 2 Related Work

## 2.1 Geometric Foundation Models

Visual perception is shifting toward universal representation learning (Arnold et al., 2022; Huang et al., 2018; Shen et al., 2025; Reizenstein et al., 2021; Ji et al., 2024). In 3D vision, feed-forward models (Li & Snavely, 2018; Wang et al., 2025c;a) replace iterative optimization with direct geometry synthesis. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) established this by regressing pointmaps and features from image pairs. However, scaling to many views requires costly global alignment (Wang et al., 2024; Leroy et al., 2024). VGGT (Wang et al., 2025a) and Fast3R (Yang et al., 2025) advanced this by jointly processing all views via spatial and global attention. Despite their quality, the quadratic memory of global attention confines them to offline batch processing.

## 2.2 Streaming 3D Reconstruction

To handle unbounded videos, recent methods introduce sequential processing. Spann3R (Wang & Agapito, 2025) and Point3R (Wu et al., 2025) utilize external spatial memories, CUT3R (Wang et al., 2025b) uses temporal causal attention, and TTT3R (Chen et al., 2025a) employs constant-size recurrent states. StreamVGGT (Zhuo et al., 2025) enabled streaming via temporal causal attention with a KV cache. While preserving representational capacity, its linearly growing cache increases computational costs monotonically, eventually triggering memory exhaustion during extended stream.

## 2.3 Token Compression for Transformers

Efficient token management is crucial across domains. Language models like $H_2O$ (Zhang et al., 2023), AdaKV (Feng et al., 2024), and SnapKV (Li et al., 2024) compress caches via attention-based eviction. In vision, token pruning (Rao et al., 2021) and merging (Bolya et al., 2022; Choi et al., 2025) reduce costs, with ToMe (Bolya et al., 2022) demonstrating that merging retains collective information better than hard pruning. In streaming geometry, FFN-residual-based scoring (Lu et al., 2026) estimates token importance. We advance this by unifying eviction and merging, introducing cross-layer consistency as a robust scoring prior broadly applicable to bounded-memory transformers.

# 3 Method

As shown in Fig. 1, we propose a unified framework for bounded-memory streaming visual geometry transformers. Our approach builds upon the foundational causal streaming architecture (Zhuo et al., 2025; Lu et al., 2026) but rethinks scoring and cache management as two training-free modules: Cross-Layer Consistency-Enhanced Scoring (CLCES) and Uncertainty-Gated Deferred Compaction (UGDC). Within UGDC, confident candidates take an immediate retain/merge/evict path, while uncertain boundary candidates take a deferred verification path.

## 3.1 Preliminaries

The core objective of streaming 3D reconstruction is to perform continuous geometric inference over temporal video streams while adhering to a fixed memory budget. Unlike global reconstruction models that require all-to-all attention, streaming architectures prioritize temporal consistency and memory efficiency through causal attention and token-wise cache management. We formalize this pipeline below to establish the foundation for our proposed enhancements.

**Streaming Formulation.** At time step $t$, the input frame $I_t$ is processed by a frozen DINOv2 (Oquab et al., 2023) backbone into $M$ tokens, including patch, camera, and register tokens. The tokens are then passed through $L$ Transformer blocks, where each block $l$ performs temporal causal attention over a Key-Value (KV) cache $\mathcal{C}_t^{(l)}$ storing tokens from previous frames:

$$\mathcal{C}_t^{(l)} = [\mathcal{C}_{t-1}^{(l)};\ (\mathbf{K}_t^{(l)}, \mathbf{V}_t^{(l)})], \tag{1}$$

where $[\cdot;\cdot]$ denotes the concatenation operation. The total memory footprint $\mathrm{Mem}(\mathcal{C}_T)$ after $T$ frames grows as $O(L \cdot T \cdot M)$, which leads to unavoidable memory exhaustion for long-duration sequences.

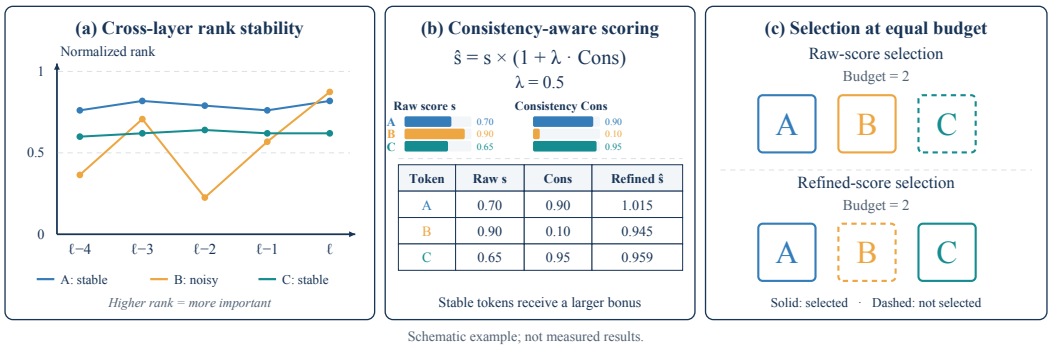



Figure 2: CLCES mechanism. (a) Rank trajectories over the five most recent layers distinguish stable importance from transient spikes. (b) A multiplicative consistency bonus enhances raw scores. (c) At the same two-token budget, the illustrative refined scores select stable tokens A and C instead of A and B. Trajectories and scores are schematic examples, not measured results.

### 3.2 Cross-Layer Consistency-Enhanced Scoring

To enforce an $O(1)$ constant-cost budget $B$, recent works (Lu et al., 2026) evaluate the geometric salience of each token $i$ at layer $l$ using the magnitude of the Feed-Forward Network (FFN) residual:

$$s_i^{(l)} = \left\| \lambda_2^{(l)} \cdot \mathrm{FFN}\big(\mathrm{LN}(\mathbf{h}_i^{(l)})\big) \right\|_2 . \tag{2}$$

Standard practice then performs "hard eviction" by permanently discarding tokens with the lowest scores $s_i^{(l)}$. As identified in section 1, this mechanism is inherently flawed due to: (1) *context destruction* from binary deletion, and (2) *evaluation noise* from relying on a localized, single-layer observation. Our proposed modules, CLCES and UGDC, replace this fragile eviction logic with a robust, information-preserving compression scheme.

As shown in Figure 2, to mitigate the evaluation noise inherent in *single-layer* FFN residuals (Challenge 1), CLCES treats the importance of a token as a multi-stage trajectory: given the raw score vector $\mathbf{s}^{(l)} = [s_1^{(l)}, \ldots, s_N^{(l)}]^\top \in \mathbb{R}^N$ at block $l$, we derive an enhanced score $\hat{s}_i^{(l)}$ that rewards rank-stability across the network depth, distinguishing sporadic activation spikes from sustained geometric salience.

**Windowed Rank Statistics.** At each layer $l$ we convert the raw scores $\mathbf{s}^{(l)}$ into normalized ranks $\tilde{R}_i^{(l)} \in [0, 1]$ via a double `argsort`, and track them over the sliding window $\mathcal{W}_l$ of the $W$ most recent global attention layers (formal construction in Appendix B). Let $\mu_i$ and $\sigma_i$ denote the mean and standard deviation of token $i$'s normalized ranks within $\mathcal{W}_l$. Calibrating $\sigma_i$ against the uniform-rank null, whose reference deviation is $\sigma_{\mathrm{ref}} = 1/\sqrt{12}$, yields the consistency metric and the complementary uncertainty supplied to UGDC:

$$\mathrm{Cons}_i = \max\Big(0,\ 1 - \sigma_i \sqrt{12}\Big) \in [0, 1], \qquad u_i = 1 - \mathrm{Cons}_i . \tag{3}$$

$\mathrm{Cons}_i = 1$ indicates perfect cross-layer rank stability, while 0 indicates dispersion at least as large as the uniform (maximum-entropy) reference.

**Enhanced Score Fusion.** The final enhanced importance score multiplicatively fuses the raw single-layer score with the consistency reward:

$$\hat{s}_i^{(l)} = s_i^{(l)} \cdot \left(1 + \lambda \cdot \mathrm{Cons}_i\right), \tag{4}$$

where $\lambda \geq 0$ controls the strength of the consistency prior; $\lambda = 0$ recovers standard single-layer scoring. The multiplicative form ensures that consistency modulates rather than overrides the raw score, preventing consistently low-ranked tokens from receiving artificially inflated scores. The added cost is an $\mathcal{O}(N \log N)$ argsort plus $\mathcal{O}(WN)$ windowed statistics per layer, negligible next to the $\mathcal{O}(N^2 d)$ attention, with no persistent memory overhead.

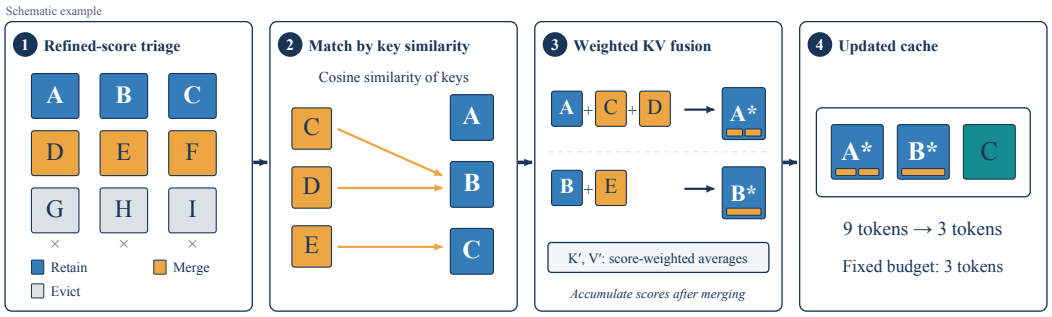



Figure 3: UGDC's immediate compaction branch in four steps: score-based triage, assignment by cosine key similarity, score-weighted KV fusion, and cache update. Confident candidates are retained, merged, or evicted; deferred candidates are handled by the two-frame branch in Fig. 4.

### 3.3 Uncertainty-Gated Deferred Compaction

UGDC owns the cache decision after CLCES. For every candidate, UGDC combines the CLCES uncertainty with proximity to the keep cutoff:

$$g_i = u_i - |b_i - q|, \tag{5}$$

where $u_i$ is the cross-layer uncertainty, $b_i$ is the candidate's score percentile, and $q$ is the keep cutoff. Confident candidates with $g_i \leq 0$ take the immediate retain/merge/evict branch; uncertain boundary candidates with $g_i > 0$ enter the deferred pool. This split makes the retain/merge/evict operation part of UGDC rather than a third method module.

#### 3.3.1 Immediate Retain/Merge/Evict Branch

The immediate branch uses the robustified scores $\hat{\mathbf{s}}^{(l)} \in \mathbb{R}^N$ from CLCES to compact confident candidates without waiting for future evidence.

**Three-Tier Token Triage.** Two thresholds $\tau_{\text{merge}} < \tau_{\text{evict}}$ partition the candidate set $\mathcal{U}$ into an evict set $\mathcal{E} = \{i \mid \hat{s}_i < \tau_{\text{merge}}\}$, a merge set $\mathcal{M} = \{i \mid \tau_{\text{merge}} \leq \hat{s}_i < \tau_{\text{evict}}\}$, and a retain set $\mathcal{R} = \{i \mid \hat{s}_i \geq \tau_{\text{evict}}\}$. If $d_l$ deferred descriptors are currently resident at layer $l$, the thresholds are dynamically determined for the available MAIN capacity:

$$|\mathcal{R}| + |\text{merged}(\mathcal{M})| + d_l \leq B_l, \tag{6}$$

so a deferred descriptor never consumes memory outside the fixed budget.

**Assignment and Fusion.** Each merge candidate $i \in \mathcal{M}$ is assigned to its most compatible retained token by cosine similarity on the key manifold, and absorbed via a score-weighted convex combination:

$$j^* = \arg\max_{j \in \mathcal{R}} \frac{\langle \mathbf{k}_i, \mathbf{k}_j \rangle}{\|\mathbf{k}_i\|_2 \, \|\mathbf{k}_j\|_2}, \qquad \mathbf{k}_{j^*}^{\text{new}} = \frac{\hat{s}_{j^*} \, \mathbf{k}_{j^*} + \hat{s}_i \, \mathbf{k}_i}{\hat{s}_{j^*} + \hat{s}_i}, \quad \mathbf{v}_{j^*}^{\text{new}} = \frac{\hat{s}_{j^*} \, \mathbf{v}_{j^*} + \hat{s}_i \, \mathbf{v}_i}{\hat{s}_{j^*} + \hat{s}_i}. \tag{7}$$

When multiple candidates map to the same target, fusion is applied sequentially with the target's score updated after each merge, $\hat{s}_{j^*} \leftarrow \hat{s}_{j^*} + \hat{s}_i$, so the accumulated representation reflects the total information content of all absorbed tokens; a Bayesian minimum-variance interpretation of this weighting is given in Appendix B. The updated immediate cache is $\tilde{\mathcal{C}}_t^{(l)} = \mathcal{R} \cup \text{merged}(\mathcal{M})$ with $\mathcal{E}$ discarded, and its occupancy is bounded by Eq. 6.

#### 3.3.2 Deferred Verification Branch

Existing bounded-cache methods make irreversible compression decisions from a single-frame snapshot. Tokens near the keep/evict cutoff are precisely where these decisions are least reliable, because their evidence is ambiguous and many reappear in subsequent frames. UGDC therefore sends only candidates with $g_i > 0$ to a small per-layer deferred pool.

Deferred tokens retain a compact descriptor and are verified against the next $H = 2$ frames via descriptor recurrence, optionally restricted to a pose-projected spatial neighbourhood. Tokens that

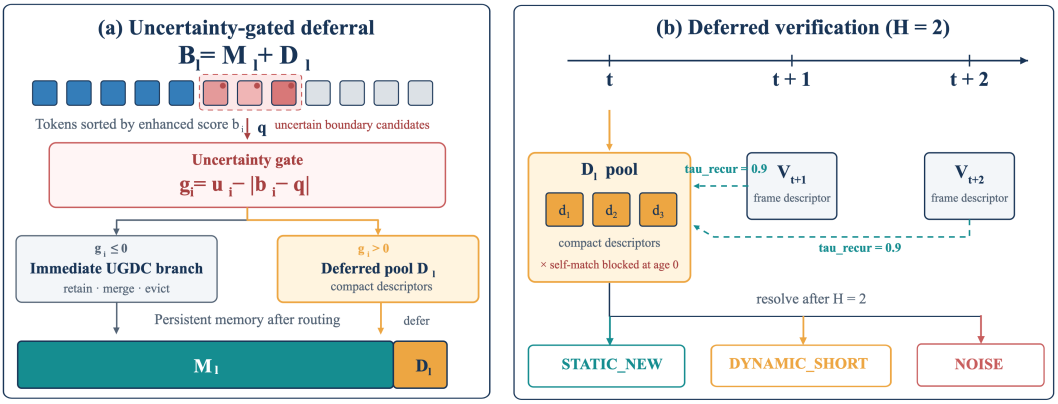



Figure 4: UGDC converts irreversible one-shot eviction into a gated, two-frame deferred decision at zero additional persistent memory. (a) Confident candidates follow immediate retain/merge/evict compaction, while uncertain boundary candidates enter $D_l$. (b) Deferred descriptors are verified over $H = 2$ frames and resolved as STATIC_NEW, DYNAMIC_SHORT, or NOISE; out-of-view tokens age without being treated as noise.

reappear stably are promoted back into the main cache (STATIC_NEW); tokens that recur but exhibit large motion-compensated residuals are retained briefly with a time-to-live (DYNAMIC_SHORT); tokens that never recur are permanently discarded (NOISE). Leaving the field of view counts as ageing, never as noise.

The mechanism is byte-exact: every layer satisfies

$$B_l = M_l + D_l, \tag{8}$$

where $M_l$ and $D_l$ are the occupied MAIN and deferred regions, respectively. The immediate branch operates on the remaining MAIN capacity after accounting for $D_l$. When a deferred descriptor is resolved or expires, its slot is returned to MAIN immediately, so total persistent memory is identical to the eviction baselines.

**End-to-end routing.** The complete pipeline computes raw FFN residual scores, applies CLCES to obtain enhanced scores and uncertainty, evaluates the UGDC gate, immediately compacts confident candidates, and verifies deferred descriptors over $H = 2$ future frames. No branch introduces persistent memory beyond the fixed per-layer budget.

## 4 EXPERIMENTS

We evaluate StreamCacheVGGT on streaming 3D reconstruction, camera pose, and video depth estimation against published external baselines; controlled same-backbone ablations isolate each module's contribution.

### 4.1 EXPERIMENTAL SETUP

**Datasets and metrics.** We evaluate indoor reconstruction on 7-Scenes (Dai et al., 2017) and NRGBD (Azinović et al., 2022), outdoor reconstruction on ETH3D (Schops et al., 2017), long-horizon reconstruction on Long3D (Chen et al., 2025b), camera pose on TUM-dynamics (Sturm et al., 2012) and Sintel (Butler et al., 2012), and video depth on Bonn (Palazzolo et al., 2019) and KITTI (Geiger et al., 2013). We report Accuracy (Acc ↓), Completeness (Comp ↓), Normal Consistency (NC ↑), ATE RMSE ↓, and Absolute Relative depth error (Abs Rel ↓).

**Baselines and implementation.** We compare with StreamVGGT (Zhuo et al., 2025), Evict3R (Deng et al., 2025), its budget-matched variant Evict3R$^\dagger$, InfiniteVGGT (Yuan et al., 2026), and non-streaming systems Spann3R (Wang & Agapito, 2025), CUT3R (Wang et al., 2025b), Point3R (Wu et al., 2025), and TTT3R (Chen et al., 2025a). Unless stated otherwise, our persistent cache budget is 200K tokens, full implementation details are in Appendix C.

### 4.2 Reconstruction Quality

Tables 1 and 2 report indoor reconstruction quality. On 7-Scenes (Table 1), StreamCacheVGGT obtains the lowest Accuracy and Completeness errors among the listed methods: its mean Accuracy of 0.0232 improves on the strongest streaming baseline (Evict3R†, 0.037) by 37% and on the best non-streaming system (TTT3R, 0.027) by 14%, and the same ordering holds for the median errors. The Completeness gap is consistent as well (0.0210 vs. 0.023 for the best baseline), indicating that the bounded cache retains surface coverage that hard-eviction schemes lose, while Normal Consistency remains on par with the strongest baseline.

Table 1: 3D reconstruction on 7-Scenes (200 frames).

| Method | Acc ↓ | | Comp ↓ | | NC ↑ | |
|---|---|---|---|---|---|---|
| | Mean | Med. | Mean | Med. | Mean | Med. |
| Spann3R (Wang & Agapito, 2025) | 0.215 | 0.131 | 0.122 | 0.063 | 0.535 | 0.550 |
| CUT3R (Wang et al., 2025b) | 0.087 | 0.048 | 0.045 | 0.014 | 0.566 | 0.601 |
| Point3R (Wu et al., 2025) | 0.041 | 0.019 | 0.023 | 0.006 | 0.579 | 0.622 |
| TTT3R (Chen et al., 2025a) | 0.027 | 0.015 | 0.023 | 0.005 | 0.582 | 0.627 |
| StreamVGGT (Zhuo et al., 2025) | 0.038 | 0.014 | 0.029 | 0.007 | 0.583 | 0.628 |
| Evict3R† (Deng et al., 2025) | 0.037 | 0.013 | 0.027 | 0.007 | 0.584 | 0.631 |
| InfiniteVGGT (Yuan et al., 2026) | 0.046 | 0.016 | 0.031 | 0.008 | 0.582 | 0.627 |
| **StreamCacheVGGT** | **0.0232** | **0.0080** | **0.0210** | **0.0049** | **0.5865** | **0.6343** |

On NRGBD (Table 2), StreamCacheVGGT improves all reported 100-frame metrics, and the advantage persists at 300 frames, where Accuracy and Completeness both improve over Evict3R† (0.035 vs. 0.042 and 0.014 vs. 0.017) while Normal Consistency stays close to InfiniteVGGT (0.643 vs. 0.646). The widening margin from 100 to 300 frames suggests that cross-layer scoring and uncertainty-gated merging become more beneficial as longer sequences put the cache under increasing pressure.

Table 2: 3D reconstruction on NRGBD at two sequence lengths.

| Method | Seq. Len. | Acc ↓ | | Comp ↓ | | NC ↑ | |
|---|---|---|---|---|---|---|---|
| | | Mean | Med. | Mean | Med. | Mean | Med. |
| StreamVGGT (Zhuo et al., 2025) | 100 | 0.024 | 0.014 | 0.013 | 0.003 | 0.663 | 0.777 |
| Evict3R (Deng et al., 2025) | 100 | 0.025 | 0.015 | 0.013 | 0.003 | 0.664 | 0.781 |
| Evict3R† (Deng et al., 2025) | 100 | 0.031 | 0.020 | 0.013 | 0.003 | 0.665 | 0.791 |
| InfiniteVGGT (Yuan et al., 2026) | 100 | 0.035 | 0.022 | 0.014 | 0.003 | 0.669 | 0.787 |
| **StreamCacheVGGT** | 100 | **0.0222** | **0.0140** | **0.0117** | **0.0029** | **0.6819** | **0.8088** |
| Evict3R† (Deng et al., 2025) | 300 | 0.042 | 0.026 | 0.017 | 0.004 | 0.640 | 0.739 |
| InfiniteVGGT (Yuan et al., 2026) | 300 | 0.053 | 0.031 | 0.024 | 0.005 | 0.646 | 0.751 |
| **StreamCacheVGGT** | 300 | **0.035** | **0.022** | **0.014** | **0.003** | **0.643** | **0.743** |

Figure 5 presents a qualitative comparison on 7-Scenes and NRGBD. Compared with StreamVGGT, Evict3R, and the enlarged-context InfiniteVGGT, our reconstructions are more complete and geometrically coherent, consistent with the Completeness gains in Tables 1 and 2.




Figure 5: Qualitative reconstruction comparison on 7-Scenes and NRGBD. StreamCacheVGGT preserves more complete surfaces and thin structures in the highlighted regions.

Table 3 reports outdoor reconstruction on ETH3D. At 200K, StreamCacheVGGT improves Completeness and Normal Consistency but does not improve Accuracy over every baseline; raising the budget to 400K yields the lowest Accuracy and median Completeness.

Table 3: Outdoor reconstruction on ETH3D. Superscripts denote token budget in thousands.

| Method | Acc ↓ | | Comp ↓ | | NC ↑ | |
|---|---|---|---|---|---|---|
| | Mean | Med. | Mean | Med. | Mean | Med. |
| CUT3R (Wang et al., 2025b) | 0.9400 | 0.6070 | 0.7090 | 0.3740 | 0.7180 | 0.8120 |
| TTT3R (Chen et al., 2025a) | 0.5980 | 0.3740 | 0.5850 | 0.2230 | 0.7280 | 0.8260 |
| StreamVGGT (Zhuo et al., 2025) | 0.6010 | 0.3690 | 0.4420 | 0.1690 | 0.7910 | 0.9330 |
| Evict3R† (Deng et al., 2025) | 0.6050 | 0.3750 | 0.4420 | 0.1630 | 0.7920 | 0.9340 |
| InfiniteVGGT (Yuan et al., 2026) | 0.6030 | 0.3710 | 0.4440 | 0.1690 | 0.7920 | 0.9330 |
| **StreamCacheVGGT**$^{200}$ | 0.6074 | 0.3742 | **0.3916** | 0.1205 | **0.7994** | **0.9351** |
| **StreamCacheVGGT**$^{400}$ | **0.5347** | **0.3228** | 0.3952 | **0.1046** | 0.7951 | 0.9342 |

### 4.3 Video Depth

Table 4: Video depth on Bonn and KITTI (Abs Rel ↓).

| Method | Bonn | | | KITTI | | |
|---|---|---|---|---|---|---|
| | 100F | 300F | 500F | 100F | 300F | 500F |
| StreamVGGT (Zhuo et al., 2025) | **0.055** | *OOM* | *OOM* | 0.166 | *OOM* | *OOM* |
| Evict3R† (Deng et al., 2025) | 0.063 | 0.072 | 0.072 | 0.192 | 0.213 | 0.198 |
| InfiniteVGGT (Yuan et al., 2026) | 0.056 | 0.073 | **0.070** | 0.165 | 0.249 | 0.257 |
| **StreamCacheVGGT** | **0.055** | **0.070** | 0.072 | **0.125** | **0.127** | **0.123** |

Table 4 compares video depth estimation on Bonn and KITTI at 100–500 frames. On KITTI, StreamCacheVGGT improves Abs Rel by 24–52% over the streaming baselines (0.123–0.127 vs. 0.165–0.257), and the margin grows with sequence length as Evict3R and InfiniteVGGT degrade while our bounded cache remains stable. On Bonn, our method matches the best baseline at 100 frames and is best at 300 frames; StreamVGGT exhausts memory beyond 100 frames. A controlled same-backbone ablation across Bonn, continuous KITTI, and Sintel is given in Appendix E.

### 4.4 Pose and Efficiency

The controlled pose results (Table 5) show that CLCES and UGDC preserve pose quality within overlapping confidence intervals; UGDC gives the best point estimate on Sintel. The two-frame verification incurs a measured throughput cost, so we report it as an accuracy–throughput trade-off.

Table 5: Camera pose estimation (ATE RMSE ↓, bootstrap 95% CI) on TUM-dynamics and Sintel.

| | TUM-dynamics | | Sintel | |
|---|---|---|---|---|
| Configuration | ATE | 95% CI | ATE | 95% CI |
| Evict-only | 0.0722 | [0.042, 0.108] | 0.2586 | [0.112, 0.444] |
| + CLCES | 0.0736 | [0.046, 0.103] | 0.2535 | [0.110, 0.436] |
| + CLCES + UGDC | 0.0747 | [0.046, 0.105] | **0.2490** | [0.110, 0.427] |

Figure 6 and Table 6 report the scaling behavior directly. Full-cache methods (StreamVGGT, Evict3R) exhaust GPU memory before 200 frames, and InfiniteVGGT survives only with a $\sim 6\times$ larger token budget and steadily growing memory. Our bounded cache keeps throughput nearly constant (4.46–4.51 FPS from 50 to 1000 frames) at the lowest peak memory (8.4→10.6 GB), so sequences an order of magnitude longer than the baselines' failure point remain processable.

Table 6: Streaming FPS (↑) and peak VRAM.

| Method | 50F | 100F | 200F | 500F | 1000F | Peak VRAM (GB) |
|---|---|---|---|---|---|---|
| StreamVGGT (Zhuo et al., 2025) | 4.55 | 3.24 | *OOM* | – | – | 12.5→18.2→*OOM* |
| Evict3R (Deng et al., 2025) | 1.68 | 0.98 | *OOM* | – | – | 13.8→20.6→*OOM* |
| InfiniteVGGT (Yuan et al., 2026) | **5.04** | 4.05 | 3.74 | 3.54 | 3.43 | 14.9→17.4 |
| **StreamCacheVGGT** | 4.51 | **4.51** | **4.48** | **4.48** | **4.46** | **8.4→10.6** |

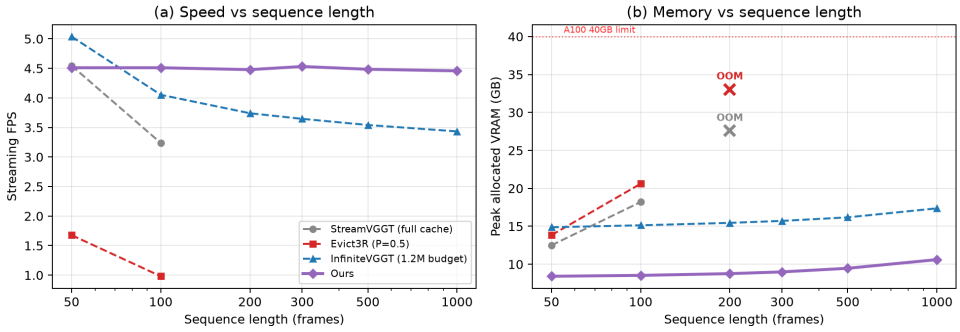



Figure 6: Streaming FPS and peak VRAM versus sequence length.

### 4.5 Two-Module Ablation

Table 7: Controlled two-module ablation at a common 200K-token budget.

| Configuration | 7-Scenes Acc ↓ | Long3D Acc ↓ / Comp ↓ | KITTI-cont (500F) Abs Rel ↓ |
|---|---|---|---|
| Evict-only | 0.0259 | 0.108 / 15.90 | 0.1822 |
| + CLCES | 0.0306 | 0.090 / 15.05 | **0.1745** |
| + CLCES + UGDC | **0.0232** | **0.087** / **14.53** | 0.1748 |

Table 7 details the module-wise ablation of our system under a strict 200K-token limit. CLCES provides strong geometric filtering, improving accuracy on long-sequence reconstruction (Long3D) and continuous video depth (KITTI) by mitigating single-layer activation noise. Integrating UGDC further enhances robustness under severe cache pressure, yielding the best overall Accuracy and Completeness on Long3D (0.087 and 14.53, respectively). While the immediate merge branch may introduce a marginal metric shift in highly dynamic environments (e.g., KITTI Abs Rel moves slightly from 0.1745 to 0.1748), it successfully preserves the critical global geometric context vital for static and long-horizon stability.

### 4.6 Visualization and More Results

A visualization of the token fates assigned by UGDC's immediate branch (Figure 7), the sensitivity study, and the long-sequence and depth results are provided in Appendices D–G.

## 5 Conclusion

We present a memory efficient streaming visual geometry transformer equipped with two modules requiring no additional training. Cross Layer Consistency Enhanced Scoring evaluates rank stability to extract robust token importance scores and explicit uncertainty estimates. Uncertainty Gated Deferred Compaction then routes tokens through an immediate retain, merge, and evict pipeline or a short verification branch kept strictly within the fixed budget. Across diverse benchmarks, StreamCacheVGGT significantly improves geometric reconstruction quality and continuous video depth accuracy over existing baselines. By sustaining constant throughput at minimal peak memory, our system processes massive video sequences far beyond the failure points of traditional methods while maintaining a strict $\mathcal{O}(1)$ persistent memory footprint.

### AI USE STATEMENT

In this work, generative AI tools were used only to polish the language of the manuscript. The authors reviewed all language edits and take responsibility for the final content of this submission.

## REFERENCES


Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pp. 690–708. Springer, 2022.

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*, pp. 6290–6301, 2022.

George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *Ieee Potentials*, 13(4): 27–31, 2002.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. 2022.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012.

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

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.

Joonmyung Choi, Sanghyeok Lee, Byungoh Ko, Eunseo Kim, Jihyung Kil, and Hyunwoo J Kim. Representation shift: Unifying token compression with flashattention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20456–20466, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. volume 35, pp. 16344–16359, 2022.

Jinhui Deng, Zhili Li, Yijin Ma, Xin Yang, and Pengfei Wan. Evict3R: eci. *arXiv preprint arXiv:2507.14890*, 2025.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *arXiv preprint arXiv:2407.11550*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2821–2830, 2018.

Deyi Ji, Wenwei Jin, Hongtao Lu, and Feng Zhao. Pptformer: Pseudo multi-perspective transformer for uav segmentation. *International Joint Conference on Artificial Intelligence*, pp. 893–901, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. pp. 71–91, 2024.

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.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. 2023.

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)*, pp. 7855–7862. IEEE, 2019.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

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*, 2025.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. pp. 78–89, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. 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*, pp. 20697–20709, 2024.

Yifan Wang et al. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c.

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

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. pp. 21924–21935, 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.

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.

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

## A LIMITATIONS

First, while our training-free modules effectively mitigate KV-cache growth without retraining, the method introduces spatial and temporal hyperparameters (e.g., consistency weight $\lambda$ and recurrence threshold $\tau_{recur}$); hence, achieving absolute peak performance may require slight dataset-specific tuning. Second, our token merging and deferred verification mechanisms predominantly assume static environments. Consequently, the immediate merge branch can occasionally oversmooth fine details in highly dynamic or rapid-motion scenes.

## B METHOD DERIVATION DETAILS

**Rank extraction via permutation operators.** For each layer $l$, the permutation $\pi^{(l)} \in \mathfrak{S}_N$ sorts the importance scores in ascending order, $\pi^{(l)} = \arg\min_{\pi \in \mathfrak{S}_N} \sum_{j=1}^{N} j \cdot s^{(l)}_{\pi(j)}$, where $\mathfrak{S}_N$ denotes the symmetric group on $N$ elements. The rank of token $i$ is the inverse permutation $R^{(l)}_i = (\pi^{(l)})^{-1}(i) - 1 \in \{0, 1, \ldots, N-1\}$, computed in practice via a double `argsort` in $\mathcal{O}(N \log N)$ time. Ranks are normalized to the unit interval, $\tilde{R}^{(l)}_i = R^{(l)}_i/(N-1)$, for scale-invariance across layers with different token counts, and collected over the sliding window $\mathcal{W}_l = \{l - W + 1, \ldots, l\}$ into the rank tensor $\boldsymbol{\mathcal{R}} \in [0,1]^{N \times W}$ with $\boldsymbol{\mathcal{R}}_{i,k} = \tilde{R}^{(k)}_i$.

**Order-statistical calibration.** For each token $i$, the windowed statistics are $\mu_i = \frac{1}{W} \sum_{k \in \mathcal{W}_l} \tilde{R}^{(k)}_i$ and $\sigma^2_i = \frac{1}{W-1} \sum_{k \in \mathcal{W}_l} (\tilde{R}^{(k)}_i - \mu_i)^2$. Under the null hypothesis that a token's rank is uniformly distributed across layers (importance purely random), the normalized ranks are i.i.d. $\mathrm{Uniform}(0,1)$ samples with variance $\frac{1}{12}$, giving the reference deviation $\sigma_{\mathrm{ref}} = 1/\sqrt{12}$ used in Eq. 3. The uniform reference is the maximum-entropy baseline when the rank of a token among $N$ peers is equally likely to take any value.

**Bayesian interpretation of score-proportional fusion.** The score-proportional weighting in Eq. 7 can be interpreted through the lens of optimal Bayesian estimation. If each token's key vector is modeled as a noisy observation of an underlying geometric feature, with the importance score inversely proportional to the observation noise variance ($\hat{s}_i \propto 1/\sigma^2_i$), then the weighted average in Eq. 7 corresponds precisely to the minimum-variance unbiased estimator (MVUE) of the underlying feature. This provides a principled justification for using importance scores as fusion weights, beyond the intuitive argument that more important tokens should contribute more.

## C IMPLEMENTATION DETAILS

All experiments are conducted on a single NVIDIA A100-40GB GPU in BF16. We use a CLCES window of $W = 5$ and consistency weight $\lambda = 0.5$. UGDC uses a two-frame verification horizon $H = 2$ and descriptor recurrence threshold $\tau_{\mathrm{recur}} = 0.9$. The deferred region $D_l$ is reserved inside each layer's persistent budget, and an expired slot returns to MAIN immediately. The immediate branch uses per-layer retain/merge/evict thresholds and score-weighted cosine-nearest fusion. Unless stated otherwise, the persistent cache budget is 200K tokens.

## D VERY-LONG SEQUENCE RECONSTRUCTION

Table 8 reports the external Long3D comparison. This benchmark stresses cache management because the sequences extend to 9545 frames. StreamCacheVGGT is competitive with the enlarged-context baseline and improves Accuracy and Chamfer distance on the longest Lecture hall sequence while keeping a fixed persistent budget.

The corrected protocol provides the clearest evidence for the deferred branch under severe cache pressure. CLCES alone improves mean Accuracy, while the complete two-module configuration gives the best measured mean Accuracy and Completeness. Unreported cells were not transplanted from another protocol.

Table 8: Long3D results under the active-row protocol.

| Method | Scene/Input | Acc. ↓ | | Comp. ↓ | | NC ↑ | | CD ↓ |
|---|---|---|---|---|---|---|---|---|
| | | Mean | Med. | Mean | Med. | Mean | Med. | |
| CUT3R | | 0.496 | 0.374 | 0.085 | 0.036 | 0.520 | 0.525 | 0.291 |
| TT3R | *Classroom* | 0.396 | 0.319 | 0.081 | 0.035 | 0.530 | 0.540 | 0.239 |
| InfiniteVGGT | 2128 | **0.357** | **0.298** | **0.057** | **0.033** | 0.576 | 0.612 | **0.207** |
| StreamCacheVGGT | | 0.369 | 0.342 | 0.479 | 0.438 | **0.621** | **0.697** | 0.424 |
| CUT3R | | 1.800 | 1.372 | 0.404 | 0.090 | 0.501 | 0.495 | 1.102 |
| TT3R | *Dormitory* | 1.965 | 1.749 | **0.329** | 0.100 | 0.515 | 0.509 | 1.147 |
| InfiniteVGGT | 4208 | 1.438 | 1.159 | 0.575 | 0.089 | **0.526** | **0.538** | 1.007 |
| StreamCacheVGGT | | **1.132** | **0.835** | 0.863 | **0.080** | 0.511 | 0.512 | **0.997** |
| CUT3R | | 1.907 | 1.437 | **0.193** | 0.079 | 0.504 | 0.507 | 1.050 |
| TT3R | *Library* | 2.175 | 1.484 | 0.430 | 0.095 | 0.494 | 0.481 | 1.303 |
| InfiniteVGGT | 4726 | **1.121** | **0.821** | 0.571 | **0.077** | **0.508** | **0.514** | **0.846** |
| StreamCacheVGGT | | 1.549 | 1.176 | 1.087 | 0.125 | 0.493 | 0.464 | 1.318 |
| CUT3R | | 2.489 | 2.432 | 5.802 | 5.071 | 0.495 | 0.483 | 4.146 |
| TT3R | *Badminton Court* | 2.791 | 2.392 | 3.160 | 2.673 | 0.509 | 0.502 | 2.975 |
| InfiniteVGGT | 6067 | **1.843** | **1.555** | **1.854** | **0.816** | 0.510 | 0.509 | **1.848** |
| StreamCacheVGGT | | 1.981 | 1.834 | 2.359 | 2.029 | **0.549** | **0.571** | 2.170 |
| CUT3R | | 8.062 | 5.650 | **0.673** | **0.251** | 0.496 | 0.491 | 4.638 |
| TT3R | *Lecture hall* | 7.710 | 5.793 | 6.192 | 5.159 | **0.513** | **0.519** | 6.951 |
| InfiniteVGGT | 9545 | 5.733 | 4.603 | 1.206 | **0.251** | 0.495 | 0.490 | 3.470 |
| StreamCacheVGGT | | **4.980** | **3.798** | 1.812 | 0.335 | 0.480 | 0.455 | **3.396** |

Table 9: Corrected Long3D protocol: controlled mean results at a common 200K budget.

| Configuration | Mean Acc ↓ | Mean Comp ↓ | Mean NC ↑ |
|---|---|---|---|
| Evict-only | 0.108 | 15.90 | 0.471 |
| + CLCES | 0.090 | – | – |
| + CLCES + UGDC | **0.087** | **14.53** | **0.504** |

## E VIDEO DEPTH

The main-text Table 4 is a system-level external comparison; the controlled component evidence below is limited to the explicitly measured cells under a common backbone and budget.

On continuous KITTI, the two-module row is best among the measured bounded-cache configurations at 300 and 500 frames. On dynamic Sintel, no row improves over the evict-only baseline, which is consistent with the method's static-geometry assumption.

## F TOKEN FATE VISUALIZATION

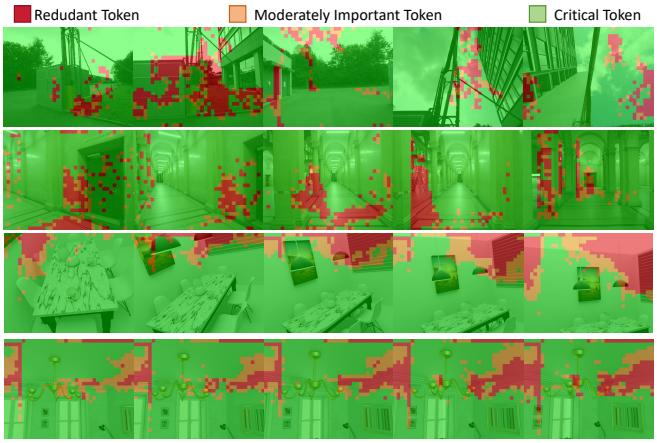



Figure 7: Visualization results of UGDC's immediate-branch token fates.

Table 10: Extended depth ablation under a common backbone and 200K budget (Abs Rel ↓).

| | Bonn | | | KITTI-cont | | Sintel |
|---|---|---|---|---|---|---|
| Configuration | 100F | 300F | 500F | 300F | 500F | Abs Rel ($\delta$) |
| Evict-only | 0.0549 | **0.0711** | **0.0734** | 0.1830 | 0.1822 | **0.3631** (**0.547**) |
| + CLCES | 0.0567 | 0.0721 | 0.0743 | 0.1737 | 0.1745 | 0.3836 (0.540) |
| + CLCES + UGDC | **0.0572** | 0.0722 | 0.0745 | **0.1740** | **0.1748** | 0.3791 (0.540) |

Figure 7 visualizes the token fates assigned by UGDC's immediate branch. Merge candidates concentrate around continuous surfaces and transition regions, where redundant observations can be fused without losing geometry, while evicted candidates are more often located in low-information areas.

## G CLCES SENSITIVITY

Table 11: CLCES sensitivity on 7-Scenes.

| $\lambda$ | $W$ | Acc ↓ | Comp ↓ | NC ↑ | CD ↓ |
|---|---|---|---|---|---|
| 0.0 | – | 0.026 | 0.020 | 0.571 | 0.033 |
| 0.25 | 5 | 0.024 | 0.019 | 0.574 | 0.031 |
| 0.5 | 5 | **0.021** | **0.016** | **0.582** | **0.027** |
| 0.75 | 5 | 0.022 | 0.017 | 0.580 | 0.028 |
| 0.5 | 3 | 0.023 | 0.018 | 0.577 | 0.029 |
| 0.5 | 7 | 0.022 | 0.017 | 0.580 | 0.028 |

Table 11 reports the CLCES sensitivity study on 7-Scenes. Every non-trivial fusion weight improves over the $\lambda$=0 baseline, and $\lambda$=0.5 with window $W$=5 gives the best Accuracy and Completeness; the ranking is stable across $W \in \{3, 5, 7\}$, so the module is not sensitive to the window width.