THINK LOCALLY, REFINE GLOBALLY FOR MEMORY-EFFICIENT 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

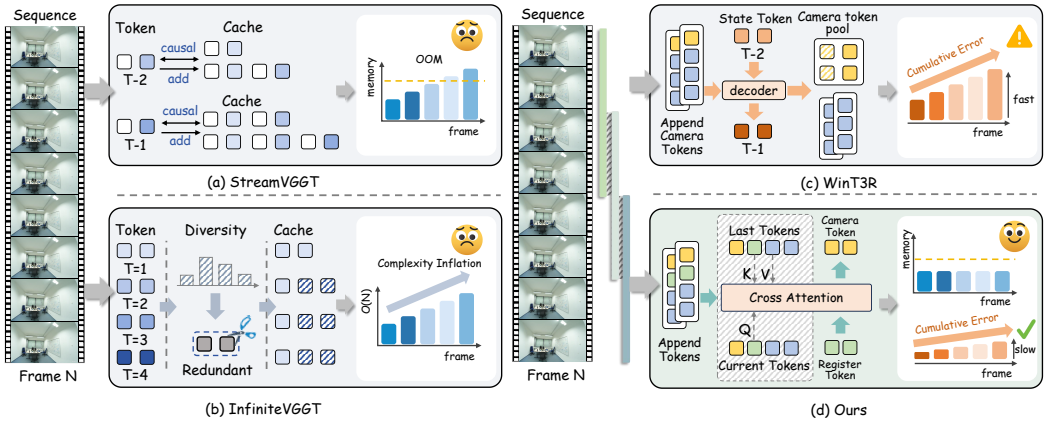


Figure 1: Comparison with previous baselines on long-horizon reconstruction. Our method achieves lower accumulated error while maintaining bounded memory consumption on long sequences.

ABSTRACT

We propose LoG-VGGT, a memory-efficient framework for long-sequence 3D reconstruction that balances local temporal modeling with global camera consistency. Instead of relying on full global attention, our method introduces cross-window attention at a small subset of transformer blocks, enabling effective information propagation across adjacent temporal windows while keeping memory usage bounded. To mitigate long-term pose drift, we further design a global camera consistency refinement module, where camera tokens interact with compact register tokens via cross-attention to enforce scene-level constraints across the entire sequence. This design enables joint optimization of camera representations and significantly improves long-horizon pose stability without incurring the high cost of sequence-wide attention. Extensive experiments demonstrate that LoG-VGGT achieves improved depth accuracy and robust camera pose estimation across multiple long-sequence benchmarks, while delivering competitive streaming reconstruction performance.

1 INTRODUCTION

Visual Geometry Grounded Transformer (VGGT) Wang et al. (2025a), a recently proposed end-to-end reconstruction model, has demonstrated strong performance in jointly estimating depth, camera pose, and other 3D attributes in offline settings. This image-based paradigm is becoming a foundation for applications such as SLAM Maggio et al. (2025); Deng et al. (2025); He et al. (2026), 3D object editing Sun et al. (2026); Huang et al. (2026); Zhou et al. (2026); Yang et al. (2026a), and embodied AI Ge et al. (2025); Vuong et al. (2025); Chen et al. (2025a); Zhu et al. (2026); Yang et al. (2026b). It offers an alternative to traditional reconstruction pipelines based on Structure-from-Motion (SfM) Schonberger & Frahm (2016) and Multi-View Stereo (MVS) Furukawa & Hernández (2015); Gu et al. (2020). Despite the growing impact across downstream tasks, the computational cost and memory footprint of VGGT grow prohibitively with sequence length, severely limiting their scalability to long-horizon 3D reconstruction scenarios.

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.

This challenge arises from the quadratic complexity of global attention over spatio-temporal tokens, which becomes increasingly impractical as the sequence expands. Recent efforts to scale VGGT-style models have diverged into two distinct paradigms. Offline batch methods Deng et al. (2025); Wang et al. (2025a); Shen et al. (2025) process entire sequences jointly to maximize global consistency, but incur substantial computation and memory overhead. In contrast, memory-efficient approaches Lan et al. (2025); Wang et al. (2025b); Zhuo et al. (2025) aim to support incremental inference by retaining all historical tokens or compressing past information into a global state. The former leads to unacceptable memory consumption, while the latter often suffers from catastrophic forgetting over long sequences, resulting in degraded geometric consistency.

Among existing long-sequence VGGT variants shown in Fig. 1, StreamVGGT Zhuo et al. (2025) caches all historical key value tokens in a causal transformer, enabling incremental inference but suffering from severe memory growth and out-of-memory failures on long sequences. In Fig. 1(c), WinT3R Li et al. (2025) introduces a sliding-window strategy together with a global camera token pool, allowing dense information exchange among frames within each window. Although this design successfully bounds memory consumption, it relies on maintaining a persistent global hidden state across windows. As the sequence length increases, small local estimation errors accumulate and propagate through the global state, resulting in unstable hidden representations and significant long-term pose drift. More recently, InfiniteVGGT Yuan et al. (2026) (Fig. 1(b)) tackles the memory explosion issue in StreamVGGT by introducing a rolling memory mechanism with a bounded but adaptive key value cache. However, it still relies on token-level eviction heuristics that operate independently at each layer and does not explicitly enforce cross-window geometric consistency. As a result, global camera alignment is only preserved through rolling memory, which may still lead to gradual pose drift under challenging trajectories.

These limitations reveal a fundamental tension in VGGT models: *can we retain sufficient temporal context for geometric consistency while preventing unbounded memory growth and long-term error drift?* To address this challenge, we propose LoG-VGGT, a memory-efficient framework that *thinks locally and refines globally* for scalable long-horizon 3D reconstruction. Instead of caching all historical tokens or maintaining a fragile global state, our approach propagates information through lightweight cross-window attention at key layers, preserving local temporal continuity with bounded memory. Furthermore, we introduce a global camera consistency refinement module, where camera tokens interact with compact register tokens, enabling global pose correction and long-term geometric consistency without reintroducing expensive full-sequence attention.

Overall, our approach balances efficiency and long-sequence modeling capacity. Our main contributions are as follows: (1) We introduce LoG-VGGT, a scalable feed-forward method that overcomes VGGT's efficiency limitations and enables online reconstruction of sequences containing thousands of frames while preserving geometric consistency. (2) We design a streaming backbone with Cross-Window Attention (CWA) for effective inter-window information propagation under strict streaming constraints. (3) We present a lightweight *global camera consistency refinement* (GCCR) module that uses register tokens to jointly refine camera tokens and mitigate long-sequence pose drift. (4) Experiments on several benchmarks demonstrate state-of-the-art long-sequence reconstruction performance with lower computational overhead than existing online reconstruction methods.

2 Related Work

2.1 Incremental 3D Reconstruction in SLAM

Incremental 3D reconstruction has long been a core research topic in the field of simultaneous localization and mapping (SLAM). Traditional systems such as ORB-SLAM Mur-Artal et al. (2015) and DSO Wang et al. (2017) rely on keyframe-based optimization and sliding-window bundle adjustment to incrementally refine the estimated geometry. These methods achieve accurate results, but suffer from complex optimization pipelines and limited scalability when dealing with long-sequence inputs. Recent advances in neural implicit representations Zhu et al. (2022); Wang et al. (2023) combine dense neural field optimization with real-time tracking to enable online scene reconstruction. However, these methods rely heavily on explicit geometric constraints and iterative optimization procedures, which introduce substantial computational overhead and limit scalability and speed. Moreover, they struggle to handle sparse viewpoints effectively. These challenges motivate

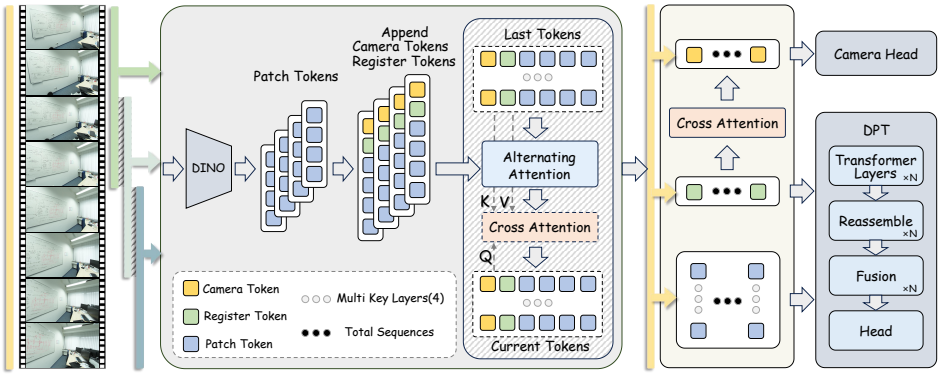


Figure 2: **Overview of LoG-VGGT.** Given a sequence, a frozen DINO visual encoder first converts the input frames into patch tokens. Camera tokens and learnable register tokens are appended to form the transformer input. To model temporal dependencies efficiently under streaming constraints, the model uses overlapping sliding windows in which current tokens attend to a limited set of cached tokens from previous frames. Finally, a lightweight global refinement module aligns camera and register tokens to mitigate long-term pose drift.

the need for end-to-end architectures capable of learning incremental 3D reconstruction from image sequences, reducing reliance on handcrafted optimization pipelines.

2.2 Feed-Forward 3D Reconstruction

Feed-forward methods Zhang et al. (2025b;a) have recently emerged as a compelling alternative to optimization-based pipelines, offering superior speed and generalization. Pioneering approaches such as PixelNeRF Yu et al. (2021) and MVSNeRF Chen et al. (2021) learn to infer neural radiance fields directly from multi-view images through end-to-end training. Subsequent works Liu et al. (2024); Jiang et al. (2025) extend this idea toward explicit and efficient 3D representations Kerbl et al. (2023). More recently, Visual Geometry Grounded Transformer (VGGT) Wang et al. (2025a) introduced an alternating-attention design that unifies camera pose estimation, depth prediction, and 3D point-cloud reconstruction within a single forward pass. However, its quadratic attention complexity severely restricts scalability to long or streaming input sequences. WinT3R Li et al. (2025) extends DUSt3R Wang et al. (2024) with an incremental transformer architecture for frame-by-frame reconstruction. StreamVGGT Zhuo et al. (2025) adapts VGGT with an implicit memory cache, while InfiniteVGGT Yuan et al. (2026) uses rolling memory with token pruning. Nevertheless, these approaches still struggle to maintain global coherence and suffer from information drift because they emphasize local temporal continuity without fully exploiting global context over extended sequences. Consequently, achieving globally consistent and temporally stable feed-forward reconstruction over long sequences remains an open challenge.

3 Motivation and Analysis

Keeping computation and memory bounded as a sequence grows is a fundamental challenge for VGGT-style 3D reconstruction models. Existing solutions typically fall into two extremes. Full-sequence global attention maximizes consistency but has quadratic complexity in the sequence length, making it impractical for long-horizon or streaming settings. Conversely, approaches that compress all historical information into a persistent memory state achieve bounded memory but often lose fine-grained temporal structure, degrading geometric consistency over time. We therefore seek efficient local propagation with selective global refinement.

One key insight is that effective long-sequence modeling does not require unrestricted global interaction at every layer. Although attention is localized within each window, the windows should not be processed in isolation, and temporal and geometric continuity must still be preserved across window boundaries. Our analysis shows that full cross-window interaction is unnecessary; instead, cross-window attention introduced at a small number of chosen transformer blocks is sufficient.

Concretely, cross-window attention is applied over overlapping regions between adjacent windows, explicitly aligning shared frames to enforce local temporal and geometric consistency. Constrained propagation also carries information across the sequence. By avoiding complete dense-token history, the backbone maintains bounded per-window memory while preserving useful temporal context.

However, while localized propagation effectively maintains mid-range geometric consistency, camera pose estimation presents a distinct challenge. Over long sequences, even minor pose errors can accumulate, resulting in significant drift when only local constraints are applied. To address this issue without reverting to expensive full-sequence attention, we introduce a lightweight global camera consistency refinement mechanism. By aggregating information from all time steps into a compact set of register tokens and performing cross-attention between these registers and camera tokens, the model can jointly refine camera poses across the sequence. This global refinement is independent of sliding-window attention, preserving efficiency while improving long-term pose stability.

Together, these observations motivate a design that *thinks locally and refines globally*: local windowed attention ensures efficiency and short-range consistency, while global camera pose refinement corrects long-term drift. This analysis forms the foundation of our LoG-VGGT framework.

4 Proposed Method

As illustrated in Fig. 2, LoG-VGGT extends VGGT with a bounded-memory streaming backbone and a separate camera-refinement module. We define the problem before detailing both components.

4.1 Preliminary and Problem Definition

VGGT Wang et al. (2025a) is a unified feed-forward framework for geometric reasoning, capable of jointly predicting per-frame depth, camera poses, and dense 3D point representations. The process begins by extracting patch-level tokens from each frame using a shared vision encoder. These tokens are then processed by a stack of transformer blocks that alternate between frame attention and global attention to propagate spatial information across all frames and tokens in sequence. Through this alternating attention scheme, VGGT predicts depth maps $\hat{D}_t$, camera poses $\hat{P}_t$, and global 3D point clouds $\hat{X}_t$ in one forward pass, with strong offline reconstruction performance.

Despite its effectiveness, VGGT relies on full global attention over all spatio-temporal tokens, resulting in quadratic computational and memory complexity $\mathcal{O}(N^2)$ with respect to the sequence length N. As the temporal horizon increases, the cost of global attention rapidly becomes prohibitive, severely limiting the scalability of VGGT for long-sequence geometric reconstruction. In this work, we address memory-efficient long-sequence 3D reconstruction under bounded computational resources. Given a potentially long image sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$, the objective is to estimate depth $\hat{D}_t$ and camera pose $\hat{P}_t$ for each frame while preserving long-range geometric consistency across the sequence. A key challenge is to maintain sufficient cross-frame context for stable geometric reasoning without incurring unbounded memory growth or expensive sequence-wide attention.

4.2 Memory-Efficient Streaming Backbone with Cross-Window Attention

To enable scalable inference on image streams, we reformulate VGGT using an overlapping sliding-window backbone. Given an input sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$, we partition it into a sequence of overlapping windows:

$$\mathcal{W}_i = \{I_i, I_{i+1}, \ldots, I_{i+K-1}\}, \tag{1}$$

where K denotes the window size and adjacent windows overlap with a fixed stride.

Within each window, VGGT blocks alternate frame attention with window-level global attention over patch, camera, and register tokens. Crucially, global attention is restricted to the current window, so the backbone's attention cost and memory scale with $\mathcal{O}(K^2)$ per window, independent of the total sequence length T.

A direct consequence of windowed processing is the potential loss of temporal continuity across window boundaries. Naively caching all historical tokens to address this issue is memory prohibitive. Instead, we introduce *Cross-Window Attention* (CWA) at a small subset of transformer blocks, referred

to as key layers, which enables effective information propagation across windows under bounded memory. Importantly, the key layers are not randomly selected. Following the original VGGT design, we use the same fixed set of four intermediate transformer layers that is employed for geometry prediction. These layers have been empirically shown to encode geometry-relevant representations and serve as the primary feature sources for depth and pose estimation in VGGT. By reusing this fixed layer set, our design remains fully consistent with the original VGGT architecture.

Let $\mathbf{H}_i^{(l)}$ denote the token representations at layer l for window $\mathcal{W}_i$. After processing window $\mathcal{W}_{i-1}$, we extract and cache the key-layer tokens $\tilde{\mathbf{H}}_{i-1}^{(l)}$ for all $l \in \mathcal{L}_{\text{key}}$. During inference of the current window $\mathcal{W}_i$, we inject these cached tokens into the corresponding layers via cross-attention. Formally, for each key layer $l \in \mathcal{L}_{\text{key}}$, the attention operation is defined as:

$$\mathbf{H}_i^{(l+1)} = \mathcal{F}\big(\mathbf{Q}(\mathbf{H}_i^{(l)}),\ \mathbf{K}([\mathbf{H}_i^{(l)}, \tilde{\mathbf{H}}_{i-1}^{(l)}]),\ \mathbf{V}([\mathbf{H}_i^{(l)}, \tilde{\mathbf{H}}_{i-1}^{(l)}])\big), \tag{2}$$

where $\mathcal{F}$ is the attention operation and $[\cdot,\cdot]$ denotes token concatenation. The cross-window interaction explicitly aligns overlapping frames across adjacent windows, thereby enforcing local temporal continuity. By restricting cross-window attention to a fixed and small set of geometry-critical layers, our method preserves the modeling capacity of VGGT while avoiding the memory overhead associated with storing the full token history.

4.3 Global Camera Consistency Refinement

After streaming inference over the entire sequence, we adopt task-specific post-processing strategies for different geometric predictions. For depth estimation, which primarily relies on local temporal continuity and progressive consistency propagation, the key-layer representations aggregated through sliding-window inference are sufficient to produce stable and coherent depth predictions. In contrast, camera pose estimation is more susceptible to error accumulation over long temporal horizons, often resulting in noticeable pose drift.

To mitigate this issue, we introduce a *Global Camera Consistency Refinement* (GCCR) module that operates at the sequence level. Let $\mathbf{c}_t$ denote the camera token associated with frame t. Then, we collect register tokens from all windows and time steps, forming a global register set:

$$\mathbf{R} = \{\mathbf{r}_t^j \mid t = 1, \ldots, S,\ j = 1, \ldots, R\}, \tag{3}$$

where S is the number of processed windows and R is the number of register tokens per window. These register tokens preserve scene-level geometric constraints in a compact global representation. We refine each camera token by performing cross-attention with the register tokens:

$$\mathbf{c}_t' = \mathrm{Attn}\big(\mathbf{Q}(\mathbf{c}_t),\ \mathbf{K}(\mathbf{R}),\ \mathbf{V}(\mathbf{R})\big). \tag{4}$$

This token-level refinement jointly aligns camera representations across the sequence, improving long-horizon pose stability without full-sequence dense attention.

4.4 Training Objective

Our LoG-VGGT is trained end-to-end in a multi-task learning framework. The overall objective combines camera pose and dense depth supervision:

$$\mathcal{L} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}}, \tag{5}$$

where the two loss terms exhibit comparable magnitudes and do not require additional re-weighting.

The camera loss $\mathcal{L}_{\text{camera}}$ supervises predicted camera parameters $\hat{g}_i$, which encode translation, rotation, and intrinsics in a unified pose representation. The corresponding ground-truth parameters g_i are transformed into the same encoding space. The camera loss is defined as:

$$\mathcal{L}_{\text{camera}} = \frac{1}{S}\sum_{s=1}^{S} \gamma^{S-s} \sum_{i=1}^{N} \left\| \hat{g}_i^{(s)} - g_i \right\|_1, \tag{6}$$

where S denotes the number of prediction stages and $\gamma \in (0, 1]$ is a temporal decay factor that emphasizes later-stage predictions. Following VGGT Wang et al. (2025a), the pose error is computed as a weighted sum of the corresponding ℓ_1 distances:

$$\|\hat{g}_i - g_i\|_1 = \lambda_t \left\|\hat{\boldsymbol{t}}_i - \boldsymbol{t}_i\right\|_1 + \lambda_R \left\|\hat{\boldsymbol{q}}_i - \boldsymbol{q}_i\right\|_1 + \lambda_F \left\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}_i\right\|_1. \tag{7}$$

Table 1: **Quantitative 3D reconstruction results on 7-Scenes and NRGBD datasets.**

Method	Input	7-Scenes						NRGBD					
		Acc.↓		Comp.↓		NC↑		Acc.↓		Comp.↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R	200	0.132	0.088	0.065	0.028	0.547	0.569	0.221	0.125	0.068	0.010	0.583	0.638
Point3R		0.045	0.026	0.025	0.011	0.559	0.599	0.072	0.048	**0.017**	0.005	0.619	0.694
TTT3R		0.040	0.025	0.024	0.005	0.567	0.601	0.102	0.044	0.025	0.005	0.612	0.675
WinT3R		0.043	0.021	0.025	0.005	0.551	0.576	0.037	0.032	0.019	0.004	0.621	0.697
InfiniteVGGT		0.038	0.013	0.025	0.005	0.579	0.612	0.049	0.031	0.021	0.005	0.652	0.761
Ours		**0.036**	**0.013**	**0.024**	**0.005**	**0.609**	**0.617**	**0.034**	**0.028**	0.023	**0.002**	**0.671**	**0.770**
CUT3R	500	0.183	0.130	0.091	0.033	0.530	0.543	0.326	0.243	0.132	0.042	0.556	0.582
Point3R		0.063	0.026	0.031	0.015	0.555	0.583	0.113	**0.048**	0.037	0.005	0.613	0.684
TTT3R		0.062	0.036	0.029	0.005	0.552	0.577	0.165	0.084	0.095	0.015	0.594	0.648
WinT3R		0.078	0.040	0.041	0.015	0.537	0.553	0.106	0.065	0.036	0.006	0.579	0.621
InfiniteVGGT		0.043	0.018	0.025	**0.005**	0.561	0.593	**0.080**	0.054	0.037	0.008	0.643	0.746
Ours		**0.037**	**0.016**	**0.025**	0.006	**0.612**	**0.624**	0.088	0.057	**0.032**	**0.004**	**0.668**	**0.759**
CUT3R	1000	0.219	0.152	0.101	0.036	0.501	0.517	0.375	0.267	0.152	0.053	0.543	0.565
Point3R		0.082	0.054	0.049	0.038	0.545	0.557	0.150	0.130	0.151	0.079	0.588	0.671
TTT3R		0.101	0.077	0.357	0.096	0.514	0.533	0.188	0.095	0.151	0.053	0.517	0.565
WinT3R		0.134	0.076	0.055	0.028	0.522	0.531	0.197	0.125	0.124	0.049	0.548	0.585
InfiniteVGGT		0.065	0.038	**0.039**	0.027	0.512	0.535	0.097	0.089	0.056	0.028	0.616	0.698
Ours		**0.060**	**0.034**	0.040	**0.021**	**0.551**	**0.569**	**0.091**	**0.083**	**0.051**	**0.020**	**0.647**	**0.752**

The depth loss $\mathcal{L}_{\text{depth}}$ follows DUSt3R Wang et al. (2024) and models aleatoric uncertainty. For each frame i, the model predicts a depth map $\hat{D}_i$ and uncertainty Σ_i^D formulated as:

$$\mathcal{L}_{\text{depth}} = \sum_{i=1}^{N} \Big(\big\| \Sigma_i^D \odot (\hat{D}_i - D_i) \big\|_2 + \big\| \Sigma_i^D \odot (\nabla \hat{D}_i - \nabla D_i) \big\|_1 - \alpha \log \Sigma_i^D \Big), \tag{8}$$

where $\odot$ denotes channel-broadcast element-wise multiplication and α controls the regularization strength imposed on the predicted uncertainty. Although no explicit point-map loss is used, the reconstructed 3D points are fully determined by the predicted depth and camera parameters, and are implicitly regularized by the above supervision.

5 Experiments

5.1 Experimental Settings

Implementation Details. Our model follows the VGGT design, with 24 alternating frame- and global-attention layers. Global attention operates at the window level with a window size of $K = 4$ and a stride of 2. Training proceeds in two stages: the model is first trained with batches of 12 frames for 100 epochs and then fine-tuned with batches of 64 frames for an additional 10 epochs. All experiments were conducted on a computing cluster with 16 NVIDIA H20 GPUs. Appendix A provides further implementation details.

Training Dataset. Our method is trained on a curated multi-domain collection of 11 datasets: Co3Dv2 Reizenstein et al. (2021), BlendMVS Yao et al. (2020), ARKitScenes Dehghan et al. (2021), WildRGBD Xia et al. (2024), ScanNet Dai et al. (2017), MVS-Synth Huang et al. (2018), PointOdyssey Zheng et al. (2023), Virtual KITTI Cabon et al. (2020), Spring Mehl et al. (2023), HyperSim Roberts et al. (2021), and Replica Straub et al. (2019). The collection balances synthetic and real data across indoor and outdoor scenes, temporal scales, geometry, lighting, and viewpoints.

Comparison Baselines. We compare our method with representative streaming feed-forward 3D reconstruction baselines. Specifically, we include CUT3R Wang et al. (2025b), Point3R Wu et al. (2025), and TTT3R Chen et al. (2025b) as incremental reconstruction methods, as well as VGGT-based streaming approaches WinT3R Li et al. (2025) and InfiniteVGGT Yuan et al. (2026). All baselines are evaluated using their official implementations and standard configurations under comparable streaming settings. Performance is reported across 3D reconstruction, video depth estimation, and camera pose estimation tasks on multiple benchmarks.



Figure 3: **Qualitative 3D reconstruction results on 7-Scenes.** We compare WinT3R, TTT3R, InfiniteVGGT, and our method, with sequence length increasing from top to bottom.

Table 2: **Video depth estimation results on Bonn, Virtual KITTI, and Sintel.**

Method	Sintel (50 frames)		Bonn (110 frames)		Virtual KITTI (110 frames)	
	Abs Rel↓	$\delta < 1.25 \uparrow$	Abs Rel↓	$\delta < 1.25 \uparrow$	Abs Rel↓	$\delta < 1.25 \uparrow$
CUT3R	0.421	0.479	0.078	0.937	0.118	0.881
Point3R	0.452	0.489	0.060	0.960	0.136	0.842
TTT3R	0.404	0.500	0.068	0.954	0.113	0.904
WinT3R	0.404	**0.696**	0.058	0.929	0.130	0.876
InfiniteVGGT	0.323	0.657	0.059	**0.972**	0.173	0.721
Ours	**0.321**	0.663	**0.054**	0.968	**0.109**	**0.904**

5.2 3D Reconstruction

In line with prior work, we evaluate scene-level 3D reconstruction performance on the 7-Scenes Shotton et al. (2013) and Neural-RGBD (NRGBD) Azinović et al. (2022) datasets. We uniformly sample subsequences of 200, 500, and 1000 frames with a fixed stride to assess performance on each sequence. We report standard metrics, including accuracy (Acc), completeness (Comp), and normal consistency (NC). As shown in Tab. 1, our method achieves consistently strong performance across both 7-Scenes and NRGBD benchmarks under different input sequence lengths, demonstrating its robustness and effectiveness in streaming 3D reconstruction. For the 7-Scenes dataset, our approach yields the best overall accuracy and normal consistency across all tested sequence lengths, outperforming recent streaming approaches such as InfiniteVGGT and WinT3R. This advantage remains stable as the input sequence length increases to 500 and 1000 frames, indicating that our model effectively mitigates long-term drift while preserving geometric details. Fig. 3 further provides qualitative comparisons on the 7-Scenes dataset, where our reconstructions exhibit cleaner geometry, fewer artifacts, and more consistent surface normals under varying input sequence lengths. Overall, these results confirm that our method scales favorably to long sequences without sacrificing reconstruction fidelity.

5.3 Video Depth Estimation

We evaluate video depth estimation on three benchmarks: Virtual KITTI, Sintel Butler et al. (2012), and Bonn Palazzolo et al. (2019), using multi-frame sequences. These datasets collectively cover a diverse range of dynamic and static scenes across both indoor and outdoor environments. Evaluation sequences are disjoint from training sequences; Virtual KITTI is present in both splits, so those results do not measure cross-dataset generalization. Following DUSt3R Wang et al. (2024), we report Absolute Relative Error (Abs Rel) and the percentage of depths within a factor of 1.25 of ground truth ($\delta_{1.25}$). As shown in Tab. 2, our method achieves performance comparable to InfiniteVGGT while outperforming representative streaming-based methods across all evaluated datasets. This suggests that our design effectively leverages multi-frame information without relying on full global attention, enabling stable depth estimation.

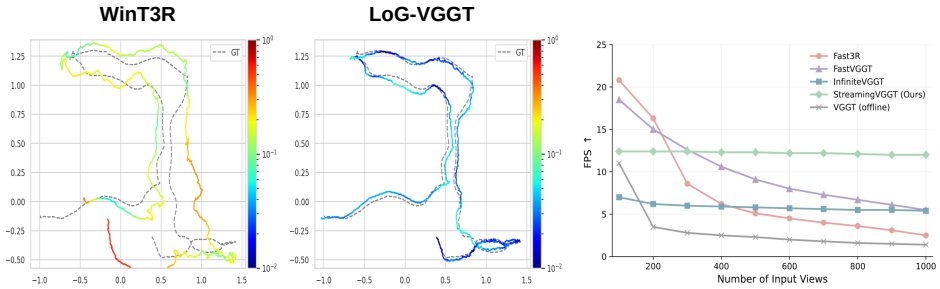


Figure 4: **Left:** Comparison of pose estimation performance between LoG-VGGT and WinT3R. **Right:** Inference speed comparison across different methods as the sequence length increases.

Table 3: **Camera pose estimation on ScanNet, TUM-D, and CO3Dv2.**

Method	ScanNet			TUM-D			CO3Dv2		
	RRA@30↑	RTA@30↑	AUC@30↑	RRA@30↑	RTA@30↑	AUC@30↑	RRA@30↑	RTA@30↑	AUC@30↑
CUT3R	58.46	57.13	54.91	77.62	73.51	75.59	76.33	72.67	75.94
Point3R	66.73	62.55	61.06	78.33	74.12	76.35	75.51	71.21	67.99
TTT3R	75.32	75.25	71.36	88.59	84.92	86.75	78.80	81.17	83.31
WinT3R	61.73	58.25	56.34	65.60	68.21	64.66	73.67	74.52	75.68
InfiniteVGGT	81.15	78.25	78.46	92.47	92.78	90.03	82.89	85.55	88.02
Ours	**82.10**	**78.93**	**78.56**	**92.51**	**93.46**	**91.17**	**83.52**	**86.61**	**88.24**

5.4 Camera Pose Estimation

For camera pose estimation, we evaluate our method on ScanNet, TUM-D Sturm et al. (2012), and CO3Dv2. For each scene, we randomly sample 500 frames to construct an evaluation sequence. All evaluated models are either trained on the same datasets or on none of them, avoiding potential bias from additional data. Together, these datasets cover object- and scene-level settings, including both static and dynamic scenes. We report Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) at specified thresholds, such as RRA@30 at 30 degrees. AUC is the area under the accuracy–threshold curve of the minimum of RRA and RTA. As shown in Tab. 3, our method achieves state-of-the-art performance among streaming approaches. Furthermore, to illustrate the accumulation of pose drift commonly observed in hidden-state-based methods such as WinT3R, and to highlight the benefit of our Global Camera Consistency Refinement, Fig. 4 visualizes the estimated camera trajectories produced by our method and WinT3R. Our approach effectively suppresses cumulative drift over long sequences, demonstrating its robustness and practical applicability for real-world deployment.

5.5 Ablation Study

We conduct ablation studies to evaluate the contributions of the CWA mechanism and GCCR module in LoG-VGGT. Experiments are performed on the 7-Scenes and NRGBD datasets, covering both 3D reconstruction and camera pose estimation. Reconstruction quality is assessed using Chamfer Distance (CD) and Normal Consistency (NC), while pose estimation is evaluated by comparing pose accuracy across ablated variants. All ablations are conducted under short-sequence (10 frames) and long-sequence (200 frames) settings to assess robustness to sequence length.

As shown in Tab. 4, CWA is essential for extending VGGT to the streaming setting, enabling bounded-memory processing of arbitrarily long sequences. Nevertheless, using only local cross-window interactions causes noticeable degradation on long sequences, mainly due to accumulated pose errors. GCCR reduces this drift and improves both camera pose and reconstruction quality. We also evaluate periodic GCCR on CO3Dv2 using only past frames. As shown in Tab. 5, it improves on the streaming baseline without future frames, while full-sequence GCCR achieves the best accuracy. This trades immediate predictions for global alignment.

Appendix C.1 (Tab. 8) provides a detailed runtime breakdown: the streaming CWA backbone accounts for 99.84% of the total runtime, whereas full-sequence GCCR contributes only 0.16% because it performs a single lightweight cross-attention refinement over compact tokens. Appendix C.2 (Tab. 9)

Table 4: **Ablation studies on 3D reconstruction and camera pose estimation.**

Input	Method	7-Scenes		NRGBD		CO3Dv2		
		CD↓	NC↑	CD↓	NC↑	RRA@30↑	RTA@30↑	AUC@30↑
10	w/o CWA	1.846	0.512	2.470	0.495	67.53	68.54	42.35
10	w/o GCCR	0.027	0.634	0.029	0.669	74.64	79.27	62.63
10	Full model	**0.022**	**0.688**	**0.021**	**0.714**	**97.12**	**98.66**	**91.61**
200	w/o CWA	OOM	OOM	OOM	OOM	OOM	OOM	OOM
200	w/o GCCR	0.045	0.503	0.061	0.525	60.34	62.28	44.27
200	Full model	**0.032**	**0.609**	**0.026**	**0.671**	**83.52**	**86.61**	**88.24**

evaluates a fully online GCCR variant, which supports frame-by-frame processing with only a minor FPS drop while retaining a clear efficiency advantage over InfiniteVGGT.

Table 5: **Ablation study of GCCR settings on CO3Dv2.**

GCCR Setting	Uses Future Frames	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
No GCCR	×	60.34	62.28	44.27
Periodic GCCR (every 50 frames)	×	72.10	74.36	63.85
Periodic GCCR (every 100 frames)	×	77.84	81.02	74.91
Final full-sequence GCCR	✓	**83.52**	**86.61**	**88.24**

5.6 Discussion

Efficiency Analysis. To address the scalability limitations of VGGT on long sequences, InfiniteVGGT introduces rolling memory with token pruning to bound memory usage and enables infinite-horizon inference. However, this design requires maintaining and dynamically updating an explicit cache at each time step, including per-layer and per-head token selection, which incurs substantial computational overhead and leads to a degraded inference speed as the sequence length increases. In contrast, our LoG-VGGT enforces locality at the architectural level by restricting attention to fixed sliding windows. This design avoids costly cache maintenance and ensures that the computational cost per frame remains constant with respect to the sequence length. As a result, LoG-VGGT achieves significantly higher and more stable inference speed across varying input lengths, as shown in Fig. 4.

Online Operation and Limitations. Full-sequence GCCR refines camera tokens after observing all frames, so the main pose results describe postprocessed rather than strictly causal predictions. The periodic variant in Tab. 5 uses only past frames and trades pose accuracy for responsiveness. Appendix C.2 evaluates a fully online variant: at 1000 frames it runs at 8.2 FPS, versus 11.9 FPS for full-sequence GCCR and 5.4 FPS for InfiniteVGGT. The faster full-sequence result includes delayed refinement and should not be interpreted as immediate output. The sliding-window backbone bounds dense-attention memory, but full-sequence GCCR retains compact camera and register tokens as the sequence grows. With fixed stride, $S = \mathcal{O}(T)$ windows require $\mathcal{O}(SR)$ register storage and $\mathcal{O}(TSR)$ token-level attention work to refine all camera tokens; the complete method therefore lacks a constant total-memory guarantee. Depth performance varies by metric: on Sintel and Bonn, the method leads in Abs Rel but not in $\delta < 1.25$ (Tab. 2). Since Virtual KITTI appears in the training collection, its held-out sequences measure within-dataset generalization, unlike the unseen Sintel and Bonn datasets. Evaluation extends to 1000 frames, leaving behavior at longer horizons or on memory-limited devices untested. These limitations motivate causal global correction and compression of refinement tokens.

6 Conclusion

In this paper, we present LoG-VGGT, a memory-efficient extension of VGGT that enables scalable 3D reconstruction from streaming image sequences. To explicitly preserve local temporal continuity across adjacent windows, we introduce cross-window attention at a small set of geometry-critical layers while keeping memory usage bounded. Then, we propose a lightweight global camera consistency refinement module that jointly optimizes camera tokens through cross-attention with compact register tokens. This module effectively corrects long-term pose drift without incurring the cost of expensive sequence-wide attention. Our method achieves improved depth accuracy and more stable camera pose estimation compared to existing baselines, while significantly reducing memory and computational overhead.

AI USE STATEMENT

Generative AI assisted with English-language editing, manuscript organization, and LaTeX formatting, including flagging potential inconsistencies for author review. It did not generate experimental data or conduct experiments. The authors retain responsibility for the claims, citations, and submitted text.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6290–6301, 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)*, pp. 611–625. Springer-Verlag, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14124–14133, 2021.

Shuhang Chen, Xianliang Huang, Zhizhou Zhong, Juhong Guan, and Shuigeng Zhou. A focused human body model for accurate anthropometric measurements extraction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22658–22667. IEEE, 2025a.

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

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.

Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *NeurIPS Datasets and Benchmarks*, 2 (6):16, 2021.

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.

Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and trends in computer graphics and vision*, 9(1-2):1–148, 2015.

Shijia Ge, Yinxin Zhang, Shuzhao Xie, Weixiang Zhang, Mingcai Zhou, and Zhi Wang. Vggt-dp: Generalizable robot control via vision foundation models. *arXiv preprint arXiv:2509.18778*, 2025.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2020.

Zebin He, Mingxin Yang, Shuhui Yang, Hanxiao Sun, Xintong Han, Chunchao Guo, and Wenhan Luo. Tango3d: Towards alignment for global and local 2d-3d correspondence. *arXiv preprint arXiv:2605.19727*, 2026.

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.

Xianliang Huang, Chen Xiao, Yuanxiang Ni, Guanming Liu, Mingkai Liu, Dikai Fan, Xiao Liu, and Hao Zhang. Semantic-guided progressive object removal with gaussian splatting. *arXiv preprint arXiv:2607.04144*, 2026.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13229–13239, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

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.

Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2024.

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

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4991, 2023.

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.

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.

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.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

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

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.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Hanxiao Sun, Mingxin Yang, Shuhui Yang, Zebin He, Xintong Han, Hongbo Fu, Chunchao Guo, and Wenhan Luo. Roar-3d: Routing arbitrary views for high-fidelity 3d generation. *arXiv preprint arXiv:2605.21121*, 2026.

An Dinh Vuong, Minh Nhat Vu, and Ian Reid. Improving robotic manipulation with efficient geometry-aware vision encoder. *arXiv preprint arXiv:2509.15880*, 2025.

Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13293–13302, 2023.

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*, pp. 5294–5306, 2025a.

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*, pp. 10510–10522, 2025b.

Rui Wang, Martin Schworer, and Daniel Cremers. Stereo dso: Large-scale direct sparse visual odometry with stereo cameras. In *Proceedings of the IEEE international conference on computer vision*, pp. 3903–3911, 2017.

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.

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

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22378–22389, 2024.

Zhenjie Yang, Xingyu Jiao, Guopeng Zhong, Shuzhe Yang, Shi Che, Chao Wu, Chenyu Jiang, Dongjie Zhang, Yideng Zhang, Zheng Zhang, et al. Handedit: A unified benchmark for egocentric human-to-robot dexterous hand image editing. *arXiv preprint arXiv:2608.12122*, 2026a.

Zhenjie Yang, Yideng Zhang, Dongjie Zhang, Chenyu Jiang, Xianshuai Liu, Yufeng Li, Zuhao Ge, Xingyu Jiao, Zheng Zhang, Kaiyu He, et al. Bench2dex: Benchmarking visuo-tactile bimanual dexterous manipulation across dexterous hands. *arXiv preprint arXiv:2609.15726*, 2026b.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

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.

Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiao-Xiao Long, Hanxue Liang, Zexiang Xu, Hao Su, et al. Advances in feed-forward 3d reconstruction and view synthesis: A survey. *arXiv preprint arXiv:2507.14501*, 2025a.

Wei Zhang, Yihang Wu, Songhua Li, Wenjie Ma, Xin Ma, Qiang Li, and Qi Wang. Review of feed-forward 3d reconstruction: From dust3r to vggt. *arXiv preprint arXiv:2507.08448*, 2025b.

Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19855–19865, 2023.

Jingke Zhou, Xianliang Huang, Yixin Ren, Dikai Fan, Shengyu Gu, Bin Tian, Xiao Liu, and Tianjia Shao. Ov3dseg-vggt: Open-vocabulary 3d segmentation with visual geometry-grounded transformers. *Visual Informatics*, pp. 100311, 2026.

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12786–12796, 2022.

Zijie Zhu, Weiren Cai, Yizhou Wang, Zhenjie Yang, Yide Liu, Jiahao Chen, and Guanqi He. Mint: A unified model for world-space camera and hand motion estimation from scalable egocentric pipeline supervision. *arXiv preprint arXiv:2609.04958*, 2026.

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 Implementation Details

Our model follows the VGGT design, with $L = 24$ alternating frame- and global-attention layers. Global attention is applied at the window level with a window size of 4 and a stride of 2, following WinT3R. For cross-window attention, we select representations of overlapping frames at layers block4DPT_idx $= [4, 11, 17, 23]$ in the previous window and propagate them as keys (K) and values (V) for the next window. We select these layers because VGGT's DPTHead uses their tokens for prediction.

Training proceeds in two stages. We first train the model with batches of 12 frames for 100 epochs to stabilize local geometric reasoning. We then fine-tune it with batches of 64 frames for 10 additional epochs to improve long-range temporal consistency. During each iteration, frames are randomly sampled from diverse training sequences. Following Point3R Wu et al. (2025), input images with varying aspect ratios are resized so that the longer edge is 518 pixels while preserving the original aspect ratio.

All experiments were conducted on a computing cluster equipped with 16 NVIDIA H20 GPUs. This setup enabled efficient processing of long sequences and extensive evaluation across multiple benchmarks.

B Additional Results

B.1 3D Reconstruction

Figures 5 and 6 present additional qualitative comparisons of scene-level 3D reconstruction. On 7-Scenes, we sample each sequence with a stride of 2, yielding approximately 500 frames per evaluation sequence. On NRGBD, we use the first 100 frames of each sequence, following common practice for qualitative visualization. Across scenes, our method produces more complete reconstructions with better geometric consistency, particularly in regions affected by occlusion or limited viewpoints.

To further evaluate the robustness of our method across diverse data regimes, Fig. 8 shows qualitative results on object-centric scenes from WildRGBD, dynamic scenes from Dynamic Replica Karaev et al. (2023), and outdoor environments from Virtual KITTI. In all settings, our method produces more complete and geometrically consistent reconstructions.

Figure 9 shows qualitative reconstructions and estimated camera trajectories for large-scale, long-sequence outdoor scenes. Despite the larger scene scale and longer temporal horizon, our method maintains coherent geometry and stable camera estimates. These results demonstrate its robustness in challenging real-world scenarios and further validate its effectiveness for long-horizon streaming 3D reconstruction.



Figure 5: Additional qualitative 3D reconstruction results on 7-Scenes.

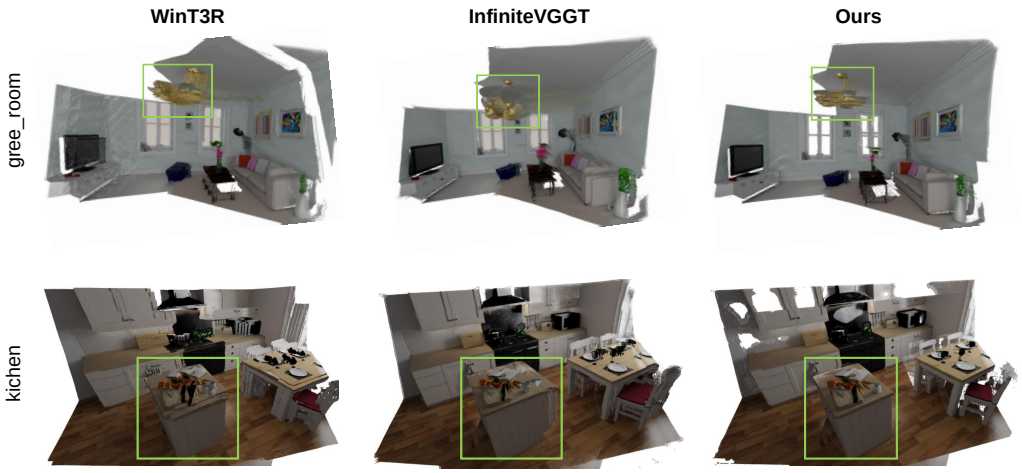


Figure 6: Qualitative 3D reconstruction results on NRGBD.

Quantitative evaluations on 7-Scenes and NRGBD (Table 6) compare our method with XStreamVGGT using accuracy (Acc), completeness (Comp; lower is better), and normal consistency (NC; higher is better) for sequences of 200 and 1000 frames. With 200 frames, our method improves Acc and NC on 7-Scenes and all three metrics on NRGBD. With 1000 frames, it improves Acc and Comp on 7-Scenes and Comp on NRGBD. These results are consistent with the qualitative comparisons.

B.2 Depth Estimation

To further evaluate depth estimation performance under extended temporal settings, we conduct additional experiments on the Virtual KITTI and Bonn datasets using longer video sequences. Following the protocol in the main paper, we randomly sample input sequences of 200 and 500 frames from each dataset to examine the behavior of different methods as the temporal horizon increases.

Table 6: Quantitative 3D reconstruction results on 7-Scenes and NRGBD datasets with XStreamVGGT.

Method	Input	7-Scenes			NRGBD		
		Acc (Mean/Med.) ↓	Comp (Mean/Med.) ↓	NC (Mean/Med.) ↑	Acc (Mean/Med.) ↓	Comp (Mean/Med.) ↓	NC (Mean/Med.) ↑
XStreamVGGT	200	0.040 / 0.022	0.017 / 0.003	0.569 / 0.606	0.108 / 0.095	0.052 / 0.023	0.619 / 0.701
Ours	200	0.036 / 0.013	0.024 / 0.005	0.609 / 0.617	0.034 / 0.028	0.023 / 0.002	0.671 / 0.770
XStreamVGGT	1000	0.142 / 0.068	0.125 / 0.048	0.734 / 0.848	0.085 / 0.049	0.075 / 0.038	0.850 / 0.986
Ours	1000	0.060 / 0.034	0.040 / 0.021	0.551 / 0.569	0.091 / 0.083	0.051 / 0.020	0.647 / 0.752

Table 7: **Video depth estimation results on Bonn and Virtual KITTI.**

Method	Input	Bonn		Virtual KITTI	
		Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
CUT3R	200	0.072	0.947	0.125	0.850
Point3R	200	0.069	0.954	0.191	0.728
TTT3R	200	0.068	0.953	0.110	0.893
InfiniteVGGT	200	**0.063**	**0.964**	0.173	0.720
Ours	200	0.065	0.957	**0.110**	**0.898**
CUT3R	500	0.084	0.939	0.152	0.809
Point3R	500	0.081	0.946	0.210	0.694
TTT3R	500	0.076	0.953	0.132	0.868
InfiniteVGGT	500	0.069	**0.960**	0.179	0.714
Ours	500	**0.065**	0.952	**0.126**	**0.852**

Evaluation sequences are excluded from training. Virtual KITTI also contributes training data, so its test results measure generalization to held-out sequences rather than to an unseen dataset. We report the same evaluation metrics as in the main paper, including Absolute Relative Error (Abs Rel) and the percentage of depth predictions within a factor of 1.25 of the ground-truth depth ($\delta < 1.25$).

Quantitative results are summarized in Tab. 7. As shown in the table, our method achieves competitive performance across both datasets and sequence lengths and consistently compares favorably with streaming baselines. In particular, our approach maintains stable accuracy as the input sequence length increases, indicating its robustness to extended temporal inputs.

B.3 Camera Pose Estimation

We further evaluate camera pose estimation over extended temporal horizons. Figure 7 reports Absolute Trajectory Error (ATE) as a function of the number of input views, comparing our method with representative streaming baselines—TTT3R, CUT3R, Point3R, and WinT3R—and the offline VGGT model. Compared with cache- or hidden-state-based streaming methods, our method accumulates error substantially more slowly as the number of views grows. This behavior confirms that GCCR mitigates long-term pose drift without full-sequence attention. Overall, LoG-VGGT balances long-horizon pose accuracy, memory efficiency, and inference stability, making it well suited to large-scale, long-duration streaming reconstruction.

C Additional Efficiency Analysis

C.1 Runtime Breakdown

Table 8 provides a detailed runtime breakdown on a 1000-frame sequence. The streaming CWA backbone accounts for 99.84% of the total runtime, while full-sequence GCCR contributes only 0.16%. This negligible overhead follows from GCCR's lightweight design: rather than operating on dense 3D representations or repeatedly reconstructing global geometry, it performs a single cross-attention refinement over compact camera and register tokens. GCCR therefore improves global consistency while preserving the efficiency required for long-sequence inference.

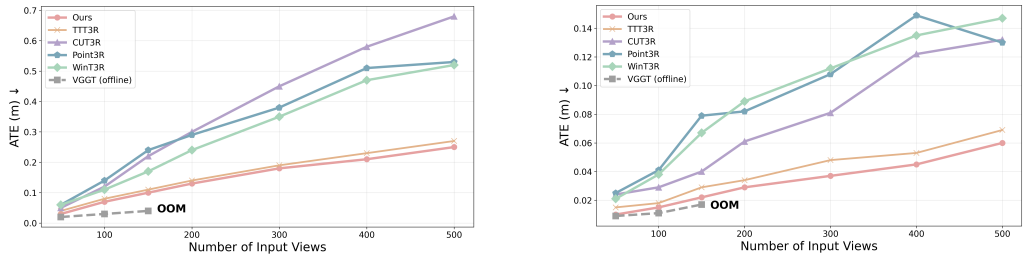


Figure 7: Camera pose evaluation on ScanNet (left) and TUM-D (right).

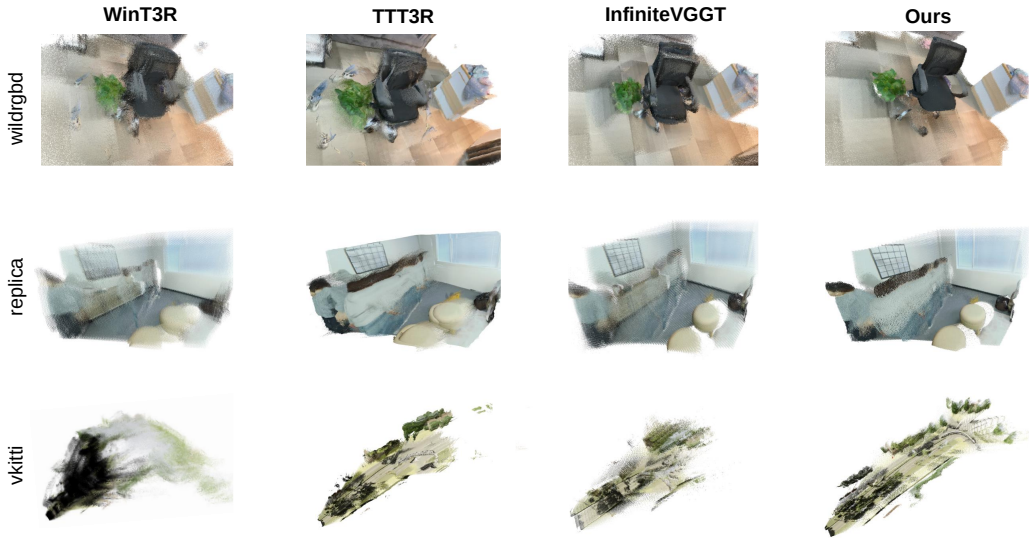


Figure 8: Qualitative 3D reconstruction results on additional datasets.

C.2 Fully Online GCCR Variant

To support strict online scenarios, we additionally introduce a fully online GCCR variant. Unlike full-sequence GCCR, which applies a one-time global refinement after processing the entire sequence, online GCCR performs incremental cross-attention updates after each CWA window inference. This enables frame-by-frame processing without delayed batch refinement, making the method applicable to latency-sensitive streaming settings.

Table 9 compares FPS at different sequence lengths. Replacing full-sequence GCCR with online GCCR causes only a minor FPS drop because refinement still operates on compact camera and register tokens rather than dense geometric representations. Under the strict online setting, our method retains a clear efficiency advantage over InfiniteVGGT, confirming the effectiveness of combining streaming CWA inference with lightweight global optimization.

D Limitations and Broader Impacts

While LoG-VGGT demonstrates strong performance and favorable scalability for long-sequence streaming 3D reconstruction, it still has several limitations. These limitations mainly arise from the central design principle of our method: introducing global reasoning only where it is most beneficial, while preserving local and streaming-friendly computation elsewhere. We discuss these limitations below, together with their broader implications and possible directions for future work.

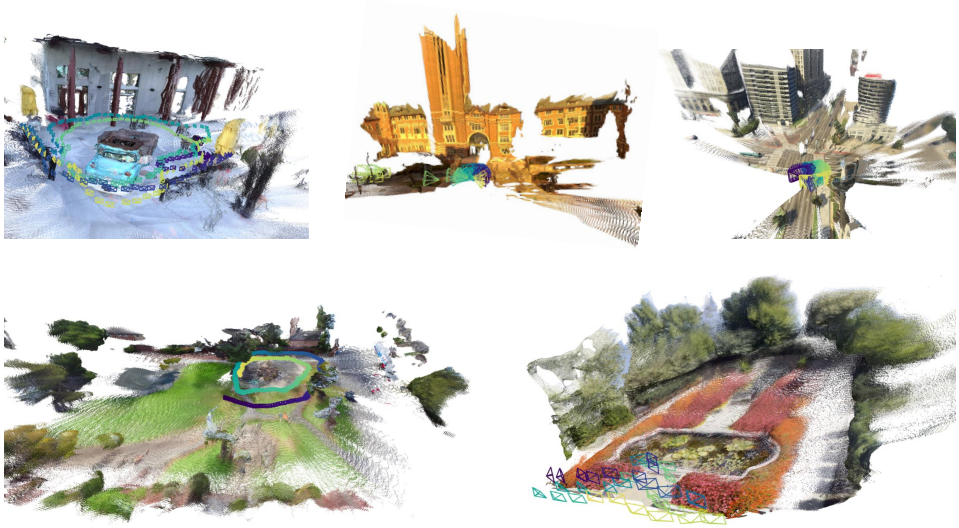


Figure 9: Qualitative in-the-wild 3D reconstruction results.

Table 8: Runtime breakdown of core components on a 1000-frame sequence.

Component	Runtime (ms)	Share
Streaming CWA backbone	83,900	99.84%
Full-sequence GCCR	134	0.16%
Total	**84,034**	**100.00%**

Table 9: FPS comparison under different sequence lengths.

Method	Sequence Length			
	100	**200**	**500**	**1000**
Fast3R	**20.8**	**16.3**	5.1	2.5
FastVGGT	18.5	15.0	9.1	5.5
InfiniteVGGT	7.0	6.2	5.8	5.4
Ours w/ full-sequence GCCR	12.4	12.3	**12.3**	**11.9**
Ours w/ online GCCR	11.5	10.8	9.4	8.2

D.1 Limitations

Depth Estimation under Local and Weak Global Constraints. A key design choice in LoG-VGGT is the decoupling of depth estimation from global sequence-level optimization. Motivated by the intrinsic differences between depth prediction and camera pose estimation, our method relies solely on local cross-window attention and weakly propagated temporal consistency for depth estimation, without introducing explicit global refinement across the entire sequence. This design preserves bounded memory usage and stable streaming inference, while avoiding the high cost of full-sequence attention.

Empirically, this strategy proves effective. As shown in Tab. 2 and Tab. 7, our method achieves performance comparable to the strongest existing streaming-based approaches, indicating that local context propagation combined with weak global consistency is sufficient for stable multi-frame depth estimation. These results support the effectiveness of our task-specific design choice.

However, our method does not consistently achieve the best performance across all depth metrics and datasets. This suggests an inherent limitation of relying exclusively on local temporal cues, particularly in scenarios involving large viewpoint changes, long-range occlusions, or globally ambiguous depth configurations. In such cases, stronger or more selective forms of global reasoning may further improve depth accuracy.

Efficiency Trade-off in Global Camera Refinement. The proposed *Global Camera Consistency Refinement* (GCCR) module mitigates long-term camera pose drift through compact, register-token-based global reasoning. As demonstrated by the ablation results in Table 4, removing GCCR substantially degrades camera pose accuracy and reconstruction quality on long sequences. This confirms that global camera-level refinement is important for maintaining long-term geometric

consistency. Nevertheless, GCCR also introduces a small additional computational cost. Although LoG-VGGT maintains significantly more stable inference speed than cache-based methods such as InfiniteVGGT, Fig. 4 shows a mild decline in throughput as the sequence length increases. This is caused by the cross-attention computation between camera tokens and accumulated register tokens at the sequence level. This overhead remains much smaller than that of full-sequence attention or dense rolling-memory mechanisms, since GCCR operates only on compact camera and register tokens rather than dense 3D representations. However, it still reflects a trade-off between long-term pose consistency and absolute inference speed. Future work may further reduce this overhead by compressing or sparsifying register tokens, performing GCCR asynchronously or at configurable temporal intervals, or designing incremental refinement schemes that reuse previously refined camera representations.

D.2 Broader Impacts

This work introduces LoG-VGGT, a memory-efficient framework for long-sequence 3D reconstruction under streaming constraints. By enabling scalable geometric perception with bounded memory usage, the method may benefit applications in robotics, AR/VR, online mapping, spatial computing, and large-scale scene understanding. Its reduced computational and memory requirements may also make 3D perception more accessible on resource-constrained platforms such as mobile devices, AR/VR headsets, and small robots, while lowering inference cost and energy consumption.

At the same time, efficient long-sequence 3D reconstruction raises important privacy and security considerations. Reconstructed scenes may reveal sensitive spatial layouts, personal belongings, private indoor environments, or location-specific information. Practical deployments should therefore obtain appropriate user consent, favor on-device processing when possible, apply secure data handling, and anonymize or redact sensitive regions before storage or sharing. The technology could also be misused for unauthorized mapping or surveillance, which highlights the need for transparent deployment policies, regulatory compliance, and clear restrictions on data collection and downstream use.

In addition, performance may degrade in challenging conditions such as low-light scenes, textureless regions, reflective surfaces, fast motion, or highly dynamic environments. Before deployment in safety-critical settings, systems based on this work should be evaluated across diverse real-world scenarios and combined with uncertainty estimation or failure detection mechanisms.

Overall, LoG-VGGT contributes to scalable and efficient streaming 3D reconstruction. Its societal benefits are best realized through responsible deployment practices that prioritize privacy, transparency, safety, and robust evaluation.

D.3 Future Directions

Overall, these limitations point toward a broader research direction: *task-adaptive global reasoning under strict streaming constraints*. Instead of applying costly global attention uniformly to all representations, future systems may selectively strengthen global reasoning for the sub-tasks that benefit most from it, such as long-term camera pose consistency, while maintaining local computation for dense predictions such as depth. Promising directions include:

- lightweight global depth refinement based on low-rank, hierarchical, or sparse temporal representations;
- adaptive selection, clustering, or compression of register tokens for more efficient GCCR;
- asynchronous or periodic global refinement to balance latency and consistency;
- incremental refinement schemes that reuse previously refined camera representations;
- confidence- or uncertainty-aware mechanisms that trigger global reasoning only when needed.

We believe such task-adaptive designs offer a promising path toward scalable, efficient, and reliable 3D perception systems for long-sequence streaming scenarios.