REVISITING LOCAL CONTEXT FOR LONG-HORIZON STREAMING 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming 3D reconstruction from extremely long videos requires online estimation of camera motion and scene geometry with bounded model memory and per-frame computation. Early streaming models use finite context buffers or compact recurrent states, yet their estimates often deteriorate as sequences grow. Recent methods improve long-horizon stability through persistent or multi-level long-range memory. We pursue a different route: local geometric prediction with reliable sequential composition. We present *LC3R*, a simple streaming model that caches KV features from only the preceding 11 frames. It predicts a point map in the current camera coordinate system and an adjacent-frame relative pose, keeping prediction targets local as the stream grows. Global poses and geometry are recovered by composing these local measurements. To ensure reliable long-horizon composition, a lightweight temporal rotation refiner improves relative rotations using recent visual and motion context, while a composition-aware pose loss directly supervises multi-step pose composition. On Oxford Spires, *LC3R* achieves an ATE of 4.35 m and an RPE-R of $0.12°$, reducing both errors by approximately 40% relative to the respective best prior streaming results. It runs at 24.45 FPS with 6.71 GB peak GPU memory, excluding input storage, on an NVIDIA H100 GPU.

1 INTRODUCTION

Recovering camera motion and dense 3D scene geometry from a video stream is a fundamental capability for robotics, autonomous driving, and embodied intelligence. Recent feed-forward reconstruction models (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026b) have shown that point maps and camera parameters can be inferred directly from unposed images using learned geometric priors. Most of these models, however, operate on a fixed collection of views through pairwise or global interactions. Ultra-long video streams change the nature of the problem: observations arrive causally, model memory and per-frame computation must remain bounded, and the current view can be separated from early reference frames by thousands of time steps and substantial camera displacement.

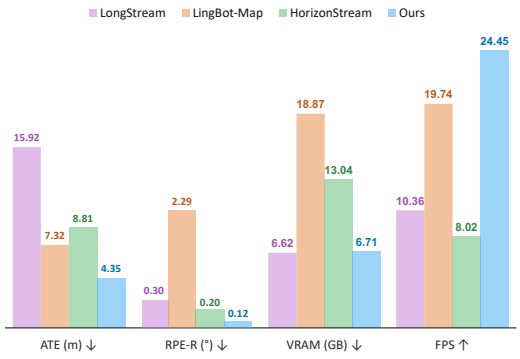


Figure 1: Pose accuracy on Oxford Spires (ATE and RPE_r) and streaming efficiency on KITTI-02 (FPS and GPU memory) excluding input storage, on an NVIDIA H100 GPU.

Existing streaming reconstruction models adapt learned geometric priors to online inference by propagating historical information through recurrent representations, cached features, or explicit spatial memories (Wang & Agapito, 2025; Wang et al., 2025b; Zhuo et al., 2026; Lan et al., 2026; Li et al., 2026c; Liu et al., 2025b; Chen et al., 2025b; Cheng et al., 2026a; Wang et al., 2026a; Chen et al., 2026b; Liu et al., 2026). To sustain reconstruction over increasingly long sequences, these methods employ various mechanisms to retain, compress, select, and update historical context. More recently, LingBot-Map (Chen et al., 2026a) and HorizonStream (Cheng et al., 2026b)

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.

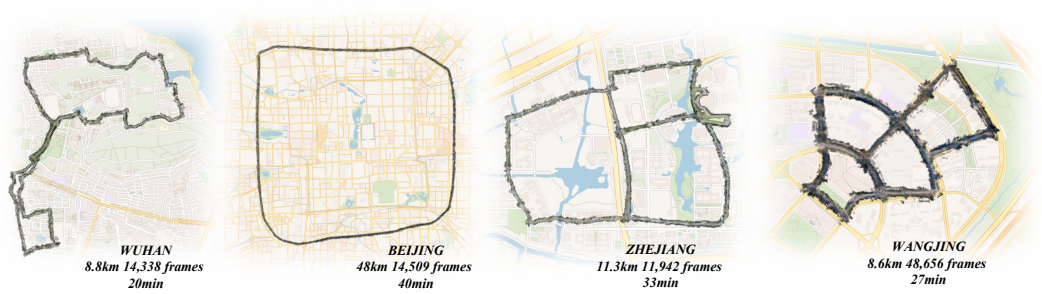


Figure 2: Long-range reconstruction results on real-world captured videos with loop closures. Our method enables stable inference on a **48,656**-frame sequence and a separate **48** km route.

have pushed streaming reconstruction beyond 10,000 frames by coordinating short-range observations with persistent long-range information. These advances show the effectiveness of long-range context, while making the management and fusion of historical state a central part of the architecture.

We focus on a different aspect of the problem: how geometric predictions are parameterized over time. A common design is to predict camera poses in a coordinate system anchored to the initial frames (Zhuo et al., 2026; Chen et al., 2026a). As the stream grows, this requires estimating motion over an increasingly long temporal span. We instead keep the learned temporal context strictly local and predict camera-local geometry and adjacent-frame motion. This design avoids elaborate long-range cache management and keeps the prediction horizon fixed as the stream grows. The model learns local measurements from short training sequences and applies the same prediction process throughout ultra-long streams, while global trajectories and geometry are assembled incrementally.

Based on this principle, we introduce *LC3R*, a streamlined streaming reconstruction model with a 12-frame causal attention window. For each incoming frame, the model attends only to KV features cached from the preceding 11 frames and maintains no persistent learned long-range state. *LC3R* predicts a point map P_i in the current camera coordinate system, a confidence map S_i, and an adjacent-frame transformation $T_{i-1\leftarrow i}$. These local measurements are predicted without a fixed global anchor. Global camera poses are recovered by chaining the adjacent transformations, and local point maps are transformed accordingly into a common coordinate system. This design keeps model memory and per-frame computation independent of total sequence length.

Accurate long-horizon reconstruction requires these local estimates to remain reliable under repeated composition: small errors in adjacent-frame poses, particularly rotations, can accumulate into substantial trajectory drift. We address this challenge at both the architecture and supervision levels. A lightweight temporal rotation refiner combines recent motion evolution with dense visual evidence to improve relative rotations before composition. A composition-aware pose objective supervises composed transformations across multiple temporal spans, optimizing local predictions according to their accumulated effect. Together, these components improve long-horizon trajectory accuracy while retaining bounded local temporal state.

Experiments on KITTI (Geiger et al., 2012), Oxford Spires (Tao et al., 2026b), and VBR (Brizi et al., 2024) demonstrate that *LC3R* generalizes from short training clips to ultra-long streams. On Oxford Spires, it reduces both ATE and RPE-R by approximately 40% relative to the respective best prior streaming results and achieves the best dense reconstruction scores among the compared methods. These accuracy gains come with efficient streaming inference at 24.45 FPS and 6.71 GB peak GPU memory on an NVIDIA H100 GPU (Fig. 1). The model also supports integration with a loop-closure backend without retraining, further reducing Oxford Spires ATE to 4.02 m. Beyond these benchmarks, we demonstrate large-scale outdoor reconstruction with loop closure on real-world videos, including a 48,656-frame sequence and a 48 km route (Fig. 2).

Our contributions are summarized as follows:

- We introduce *LC3R*, a local-context streaming reconstruction model that predicts camera-local geometry and adjacent-frame poses using KV features from only the preceding 11 frames, and recovers global poses and geometry through sequential composition.

- We develop a lightweight temporal rotation refiner and composition-aware pose supervision to improve the reliability of local predictions under repeated composition while maintaining bounded temporal state.
- We demonstrate accurate and efficient long-sequence reconstruction, reducing both ATE and RPE-R on Oxford Spires by approximately 40% relative to the respective best prior streaming results and achieving 24.45 FPS on KITTI-02.

2 Related Work

2.1 Feed-Forward Visual Geometry Reconstruction

Classical SfM and visual SLAM recover camera motion and scene structure through correspondence estimation and geometric optimization (Schönberger & Frahm, 2016; Campos et al., 2021). Feed-forward methods replace parts of these pipelines with learned geometric priors: DUSt3R (Wang et al., 2024) regresses point maps from unposed image pairs, while MASt3R (Leroy et al., 2024) adds dense features for matching. MV-DUSt3R+, Fast3R, DepthAnything v3, and VGGT extend this paradigm to joint multi-view prediction (Tang et al., 2025; Yang et al., 2025; Lin et al., 2025; Wang et al., 2025a), while π^3 (Wang et al., 2026b) introduces permutation equivariance across views. FLARE (Zhang et al., 2025b) jointly predicts geometry and appearance, PLANA3R (Liu et al., 2025a) represents scenes with metric planar primitives, and MonST3R (Zhang et al., 2025a) extends reconstruction to dynamic videos. These methods establish strong priors for pairwise and multi-view reconstruction; extending them to growing video streams requires causal inference with bounded temporal memory and per-frame computation.

2.2 Streaming Reconstruction and Temporal Memory

Spann3R (Wang & Agapito, 2025) and CUT3R (Wang et al., 2025b) maintain temporal context via external spatial memory and compact recurrent states, respectively; SLAM3R (Liu et al., 2025b) aligns overlapping local point maps. StreamVGGT (Zhuo et al., 2026) uses causal KV caching; STream3R (Lan et al., 2026), decoder-only causal prediction; WinT3R (Li et al., 2026c), sliding windows with global camera tokens; and Point3R (Wu et al., 2025), spatial pointer memory. For longer sequences, LONG3R (Chen et al., 2025b) combines gated 3D spatio-temporal memory with length-increasing curricula, while LongStream (Cheng et al., 2026a) pairs keyframe-relative prediction with periodic cache refresh. STAC (Wang et al., 2026a) compresses cached features, TTT3R (Chen et al., 2026b) uses confidence-guided recurrent updates, and Mem3R (Liu et al., 2026) separates tracking and mapping via fast-weight memory and a fixed-size token state. Beyond 10,000 frames, LingBot-Map (Chen et al., 2026a) uses anchor context, a pose-reference window, and trajectory memory, while HorizonStream (Cheng et al., 2026b) combines local and persistent linear attention. Our predictor instead caches KV features from only the preceding $K-1$ frames, without separate persistent long-range memory. It predicts current-frame point maps and adjacent-frame poses, keeping targets local as streams grow and reconstructing globally through pose composition.

2.3 Local-to-Global Pose and Trajectory Recovery

Local-to-global recovery is central to visual odometry and SLAM. DROID-SLAM (Teed & Deng, 2021) and ORB-SLAM3 (Campos et al., 2021) recover trajectories through geometric optimization. MASt3R-SLAM (Murai et al., 2025) uses pointmap matching for tracking and global optimization, while SLAM3R (Liu et al., 2025b), VGGT-Long (Deng et al., 2025), and TALO (Zhang et al., 2026a) construct and align local submaps. Anchor3R (Tao et al., 2026a) integrates current-centered relative poses from transient windows through a motion graph. R3 (Xu et al., 2026) predicts relative poses while retaining a first-frame reference, whereas LoGeR (Zhang et al., 2026c) combines chunk-wise inference with test-time global-consistency optimization. Our model learns local measurements for direct trajectory composition. Its predictor estimates adjacent-frame poses, while multi-gap supervision and local motion–visual rotation refinement improve their composition over long horizons. An optional, training-independent loop-closure backend provides additional global correction without changing the learned predictor.

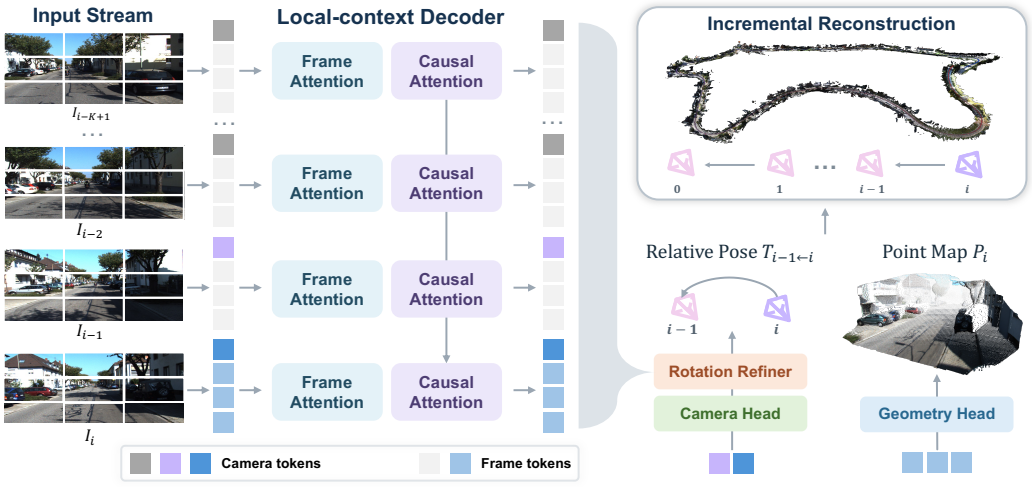


Figure 3: **Overview of our method.** For each incoming frame, a local-context decoder uses a fixed K-frame causal window to predict a point map in the current camera coordinate system and an adjacent-frame relative pose. A lightweight motion–visual refiner corrects the relative rotation before pose composition, enabling online trajectory recovery and incremental reconstruction.

3 METHOD

We formulate long-horizon streaming reconstruction as local geometric estimation, thereby enabling a simple and scalable architecture, as shown in Figure 3. Section 3.1 describes the local prediction formulation and windowed causal attention. Section 3.2 introduces the lightweight temporal refiner that improves relative rotations using recent motion and visual context. Section 3.3 presents a composition-aware objective that supervises pose chains over multiple temporal spans.

3.1 LOCAL GEOMETRY ESTIMATION

For each incoming frame I_i, an image encoder and a causal Transformer decoder produce dense frame tokens G_i and camera tokens C_i:

$$(G_i, C_i, \mathcal{M}_i) = \mathrm{Decoder}([\mathrm{Encoder}(I_i), C], \mathcal{M}_{i-1}), \tag{1}$$

where C contains L learnable camera tokens and $C_i = \{c_i^\ell\}_{\ell=1}^L$ denotes their decoded representations. Separate heads map G_i to a point map P_i in the current camera coordinate system and a confidence map S_i. The decoder uses windowed causal attention over the current frame and the KV states of the preceding $K-1$ frames, stored in $\mathcal{M}_{i-1}$. Older states are discarded, giving $O(K)$ temporal memory and $O(NK)$ temporal-attention computation for N frames.

We aggregate camera tokens with a MLP ϕ_{desc}, then predict adjacent-frame motion as:

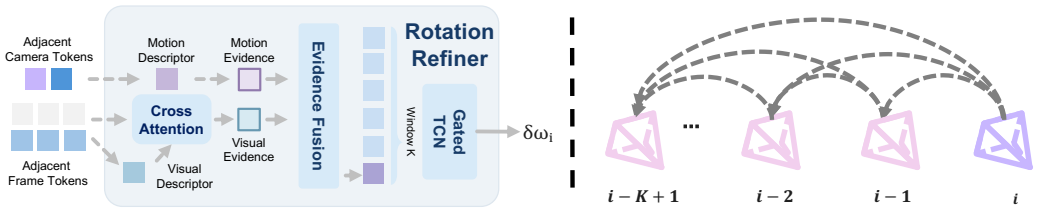
$$z_i = \frac{1}{L}\sum_{\ell=1}^{L} \phi_{\mathrm{desc}}(c_i^\ell), \quad q_i = \mathcal{R}(z_{i-1}, z_i), \quad T_{i-1\leftarrow i} = \mathrm{Head}_{\mathrm{pose}}(q_i). \tag{2}$$

Here, $\mathcal{R}(x, y) = [x, y, y-x, x \odot y]$ constructs a pairwise descriptor (Mou et al., 2016), and $T_{i-1\leftarrow i}$ maps coordinates from frame i to frame $i-1$.

Longer-range poses are recovered by composing adjacent transformations:

$$T_{i\leftarrow j} = \prod_{k=i+1}^{j} T_{k-1\leftarrow k}, \quad i < j. \tag{3}$$

Setting $i = 0$ yields global camera poses, which transform local point maps into a common coordinate system. In the full model, rotations are refined before composition, as described next.



(a) Motion-Visual Contextualized Rotation Refiner (b) Composed Relative Pose Loss

Figure 4: Motion–visual contextualized rotation refiner with composition-aware pose supervision. **(a)** Motion and visual evidence are fused and aggregated by a gated causal TCN to predict residual corrections to adjacent-frame rotations. **(b)** Refined relative poses are composed across multiple temporal gaps and supervised against ground truth to constrain accumulated pose errors.

3.2 Rotation Refiner

Although adjacent-frame pose prediction keeps the estimation target local, small rotation errors can accumulate into substantial trajectory drift. We introduce a lightweight motion–visual rotation refiner that uses recent motion and visual context to improve relative rotations before composition, while leaving the predicted translations unchanged.

As shown in Figure 4(a), we construct a fused representation for each adjacent-frame pair. The pairwise descriptor q_i provides compact motion evidence $f_i^m = \phi_m(q_i)$ through an MLP ϕ_m. To extract complementary visual evidence f_i^v, we spatially average-pool G_{i-1} and G_i into $\bar{G}_{i-1}$ and $\bar{G}_i$, construct their pairwise descriptor, and map it to a query over the original dense tokens:

$$f_i^v = \mathrm{CrossAttn}\big(\phi_v(\mathcal{R}(\bar{G}_{i-1}, \bar{G}_i)), [G_{i-1}; G_i]\big)\,, \tag{4}$$

where ϕ_v is an MLP. The pooled descriptor $\bar{G}_i$ and $\bar{G}_{i-1}$ summarize inter-frame visual changes, while cross-attention (Vaswani et al., 2017) retrieves spatially resolved evidence from the dense tokens. f_i^v and f_i^m are then fused to yield a compact $f_i = \phi_{\mathrm{fuse}}([f_i^m; f_i^v])$ by an MLP ϕ_{fuse}

We aggregate the recent fused features $\mathcal{W}_i = [f_{i-K+1}, \ldots, f_i]$ using a gated causal TCN (Lea et al., 2017). Its feature branch $\mathcal{T}_h$ models temporal motion–visual patterns, while its gating branch $\mathcal{T}_g$ adaptively modulates the resulting features. An output MLP ϕ_o predicts an axis–angle residual:

$$\delta\omega_i = \phi_o(\mathcal{T}_h(\mathcal{W}_i) \odot \sigma(\mathcal{T}_g(\mathcal{W}_i))) \in \mathbb{R}^3, \tag{5}$$

where σ is the sigmoid function.

Let $\widetilde{R}_{i-1\leftarrow i}$ denote the rotation component of the adjacent pose predicted in Section 3.1. We map the residual to $\mathrm{SO}(3)$ through the exponential map and apply it to the initial rotation:

$$\widehat{R}_{i-1\leftarrow i} = \widetilde{R}_{i-1\leftarrow i}\,\mathrm{Exp}([\delta\omega_i]_\times). \tag{6}$$

The refined rotations and unchanged translations form the adjacent poses used for trajectory composition and the composition-aware supervision introduced next.

3.3 Composition-aware Loss Functions

To constrain accumulated pose errors, we supervise refined pose compositions over the frame pairs

$$\mathcal{P} = \{(i, j) \mid 0 \le i < j < N,\ j - i \le K - 1\}, \tag{7}$$

where N is the training clip length and K is the local window size. Each composition follows Eq. 3 and is compared with ground truth using the translation and rotation losses of π^3 (Wang et al., 2026b) (Figure 4(b)):

$$\mathcal{L}_{\mathrm{pose}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(\lambda_{\mathrm{trans}}\ell_{\mathrm{trans}}^{(i,j)} + \alpha_{ij}\lambda_{\mathrm{rot}}\ell_{\mathrm{rot}}^{(i,j)}\right). \tag{8}$$

The rotational weights emphasize longer composition chains and are normalized to have mean one:

$$\alpha_{ij} = \frac{|\mathcal{P}|(j-i)^\gamma}{\sum_{(m,n)\in\mathcal{P}}(n-m)^\gamma}, \qquad 0 < \gamma < 1. \tag{9}$$

We regularize the magnitude and temporal variation of rotation residuals $\delta\omega_i$ through $\mathcal{L}_{\text{smooth}}$, and retain the point-map, surface-normal, and confidence losses of π^3. The overall objective is

$$\mathcal{L} = \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{pts}}\mathcal{L}_{\text{pts}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}, \tag{10}$$

where the λ coefficients balance the loss terms.

4 Experiments

We evaluate the ability of *LC3R* to maintain accuracy on streams extending far beyond its training horizon. We first assess long-horizon camera pose accuracy, 3D reconstruction quality, and streaming efficiency. We then conduct ablation studies to analyze the contributions of composition-aware supervision, long-sequence training, and rotation refinement. Additional training details, evaluation protocols, and more per-sequence pose results are provided in Appendix B and Appendix C.

4.1 Experimental Setup

We evaluate camera pose estimation on KITTI (Geiger et al., 2012), Oxford Spires (Tao et al., 2026b), and VBR (Brizi et al., 2024), covering sequences from 271 to 18,846 frames. We report Sim(3)-aligned absolute trajectory error (ATE), rotational relative pose error (RPE_r), and translational relative pose error (RPE_t). For dense 3D reconstruction, we evaluate on 7Scenes (Shotton et al., 2013), TUM-Dynamic (Sturm et al., 2012), and Oxford Spires, using Chamfer distance (CD) and F1 score. We compare against both optimization-based and learning-based streaming methods, including recent long-horizon approaches such as LongStream (Cheng et al., 2026a), LingBot-Map (Chen et al., 2026a), and HorizonStream (Cheng et al., 2026b). We evaluate baselines using public implementations and checkpoints where available under a unified protocol. Methods marked with * require camera intrinsics, whereas our model takes RGB images only. Appendix B.3 provides the full baseline list and evaluation details.

4.2 Implementation Details

We train on a mixture of 30 synthetic and real-world datasets spanning indoor, driving, handheld, aerial, and unordered multiview captures. Training is performed in two stages. Stage I initializes from the released π^3 weights (Wang et al., 2026b) and trains on 32-frame clips using our adjacent-frame relative pose formulation. Stage II extends the training clips to 128 frames while keeping the local prediction window fixed at 12 frames, and introduces the motion–visual contextualized

Table 1: Camera pose estimation on KITTI. Lower is better. Within streaming methods, best and second-best results are marked in blue and underlined. LC denotes loop closure.

		Per-sequence ATE (m) ↓											Sequence averages		
		00	01	02	03	04	05	06	07	08	09	10	Avg.	Avg.	Avg.
		4541f	1101f	4661f	801f	271f	2761f	1101f	1101f	4071f	1591f	1201f	ATE (m)↓	RPE_r (°)↓	RPE_t (m)↓
Group	Method	3.70 km	2.50 km	5.10 km	0.60 km	0.40 km	2.20 km	1.20 km	0.70 km	3.20 km	1.70 km	0.90 km			
Opt.-based	VGGT-SLAM (Maggio et al., 2025)	27.99	214.23	260.88	53.50	26.02	99.61	23.62	64.52	243.87	169.10	80.13	114.86	1.57	2.73
	MASt3R-SLAM (Murai et al., 2025)	190.12	507.15	232.17	169.82	88.98	126.37	133.33	62.92	199.70	178.03	151.69	185.48	0.98	2.58
	VGGT-Long (Deng et al., 2025)	8.63	58.37	52.79	8.83	4.23	9.83	4.65	2.67	73.12	31.84	27.67	25.69	0.26	0.42
	SLAMFormer-∞ (Fang et al., 2026)	15.39	96.58	28.81	4.97	4.16	11.36	13.54	8.53	37.00	18.85	13.93	23.01	–	–
	Droid-SLAM* (Teed & Deng, 2021)	131.25	267.79	217.63	3.40	1.47	80.41	64.20	18.19	167.80	90.05	22.74	96.81	0.23	0.60
	DPV-SLAM* (Lipson et al., 2024)	113.19	16.76	113.06	2.44	0.99	59.42	53.32	18.90	110.36	74.60	13.74	52.43	0.05	0.27
	DPVO* (Teed et al., 2023)	113.17	16.56	113.37	2.45	0.98	59.59	55.63	19.40	110.64	74.44	13.80	52.73	0.05	0.27
	Droid-W* (Li et al., 2026b)	48.26	207.57	52.80	1.76	4.65	25.16	5.34	3.35	29.05	35.56	8.01	38.32	0.05	0.36
Streaming	InfiniteVGGT (Yuan et al., 2026)	186.48	624.69	289.03	168.42	73.88	143.81	116.65	85.53	225.14	214.28	156.87	207.71	9.44	19.21
	OVGGT (Lu et al., 2026)	185.65	716.66	304.04	155.78	65.01	155.48	108.51	85.92	246.97	215.64	178.67	219.85	11.02	18.02
	Stream3R-w (Lan et al., 2026)	188.66	695.02	301.28	161.68	98.99	159.85	122.82	87.19	264.61	219.90	194.32	226.76	7.03	11.52
	CUT3R (Wang et al., 2025b)	185.21	654.37	300.51	157.95	23.72	153.77	133.09	68.57	230.95	206.95	184.34	209.04	2.94	2.79
	CUT3R w/ reset (Wang et al., 2025b)	190.70	91.56	269.58	20.95	5.97	104.88	72.01	27.44	164.69	93.11	41.53	98.40	0.55	0.65
	TTT3R (Chen et al., 2026b)	165.12	524.16	272.38	105.06	12.32	146.24	131.02	61.75	262.88	187.14	124.13	181.11	3.85	3.49
	TTT3R w/ reset (Chen et al., 2026b)	118.19	102.98	242.05	14.37	3.52	31.62	37.66	11.99	81.52	79.10	29.81	68.44	0.49	0.60
	LongStream (Cheng et al., 2026a)	90.02	63.46	236.42	5.04	2.57	80.62	12.76	13.84	89.61	95.07	25.67	65.01	0.35	0.33
	LingBot-Map (Chen et al., 2026a)	33.57	151.48	66.02	**2.72**	2.09	12.79	6.06	4.82	**13.36**	19.50	**8.01**	29.13	0.61	3.04
	HorizonStream (Cheng et al., 2026b)	29.67	24.15	97.24	5.53	**0.72**	13.94	**5.46**	6.12	22.63	29.79	12.35	22.51	0.17	0.11
	HorizonStream w/ LC (Cheng et al., 2026b)	15.45	24.15	71.01	5.53	**0.72**	7.84	7.54	3.14	22.63	27.45	12.35	17.98	0.30	**0.09**
	Ours	21.86	**21.55**	69.87	4.38	2.93	13.54	12.04	2.02	22.30	14.11	16.17	18.25	**0.10**	0.10
	Ours w/ LC	**11.70**	**21.55**	**39.00**	4.38	2.93	**6.00**	8.70	**1.65**	22.30	**13.99**	16.17	**13.49**	**0.10**	0.10

rotation refiner. We then add the confidence branch and optimize it in a short calibration stage while keeping the reconstruction and pose networks fixed. All inputs are resized to 504×280. Appendix A provides details on data preprocessing, while Appendix B.2 describes the detailed training schedule.

4.3 Main Results

Camera pose estimation. Tables 1 and 2 report the three long-sequence benchmarks. Among streaming methods without loop closure, *LC3R* obtains the best Oxford Spires ATE and the lowest rotational error on both KITTI and Oxford. On KITTI, it improves average ATE from 22.51m for HorizonStream to 18.25m while reducing RPE_r from $0.17°$ to $0.10°$. VBR is more challenging: *LC3R* is competitive before global optimization with the lowest local pose errors. The optional backend reduces VBR ATE from 30.14m to 9.99m and achieves the best ATE among streaming methods on all three benchmarks. Unlike several optimization-based baselines in the full tables, *LC3R* does not require test-time intrinsics. Figure 5 shows qualitative trajectory comparisons.

Table 2: Average camera-pose accuracy of all streaming methods on Oxford Spires and VBR.

Method	Oxford Spires			VBR		
	ATE	RPE_r	RPE_t	ATE	RPE_r	RPE_t
InfiniteVGGT	31.76	5.30	2.70	102.90	8.27	4.84
OVGGT	33.19	7.83	3.61	104.08	11.43	6.07
Stream3R-w	34.11	9.05	2.67	103.87	7.81	2.72
CUT3R	32.79	1.78	0.25	103.79	1.50	0.37
CUT3R w/ reset	14.74	0.66	0.05	70.91	0.75	0.11
TTT3R	25.96	3.19	0.31	100.23	3.58	0.50
TTT3R w/ reset	9.40	0.29	0.04	66.61	0.81	0.10
LongStream	15.92	0.30	0.06	92.93	0.71	0.13
LingBot-Map	7.32	2.29	0.53	29.45	4.42	2.70
HorizonStream	8.81	0.20	0.03	29.02	**0.60**	**0.04**
HorizonStream + LC	7.93	0.20	0.03	17.64	0.61	**0.04**
LC3R	4.35	**0.12**	**0.02**	30.14	**0.60**	**0.04**
LC3R + LC	**4.02**	**0.12**	**0.02**	**9.99**	**0.60**	**0.04**

Long-Horizon Stability. Long sequences pose two distinct challenges for streaming pose estimation: local predictions may deteriorate during extended operation, while small errors can accumulate when motion is composed across increasingly distant frames. We evaluate both effects in Figure 6. On the complete 4,661-frame KITTI-02 sequence, *LC3R* maintains stable adjacent-frame rotational and translational RPE throughout the stream, showing no degradation in local pose quality as processing continues. When RPE is evaluated over increasing temporal gaps and averaged across all KITTI sequences, *LC3R* achieves the lowest error at every tested gap, demonstrating more reliable long-range composition. Together, these results establish stable online pose estimation at both local and extended temporal scales.

3D reconstruction results. As shown in Table 3, *LC3R* remains competitive on 7Scenes and TUM-Dynamic across both CD and F1, despite relying only on limited local temporal context. These benchmarks are dominated by relatively compact indoor environments, where substantial spatial overlap allows persistent scene information to remain useful over long sequences. In contrast, Ox-

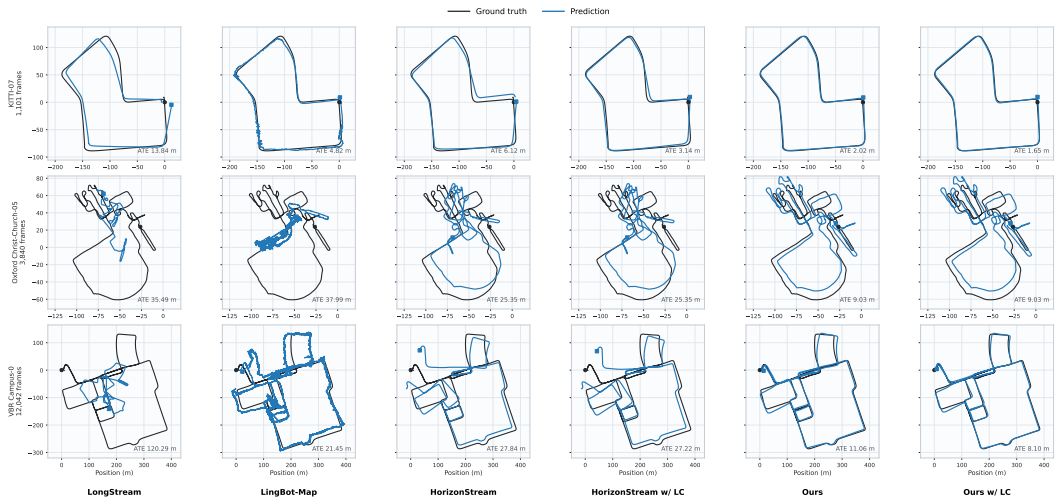
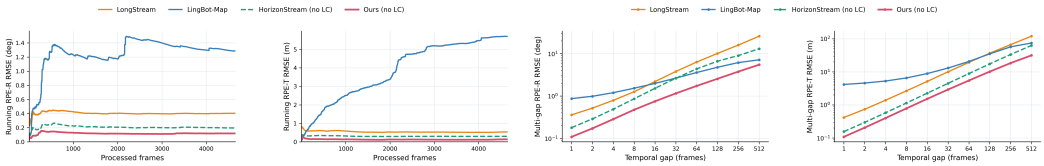


Figure 5: Qualitative camera-pose comparison on representative sequences. Predicted trajectories (blue) are aligned to ground truth (black). *LC3R* preserves global trajectory geometry across scene scales and motion patterns; loop closure further reduces drift when revisits are available.

Table 3: Streaming efficiency and dense reconstruction. Input storage is excluded; CD is in meters (↓) and F1 in percent (↑).

Method	Streaming efficiency (KITTI-02)		Dense reconstruction					
	FPS ↑	Mem. (GB) ↓	7Scenes		TUM-Dyn.		Oxford Spires	
			CD ↓	F1 ↑	CD ↓	F1 ↑	CD ↓	F1 ↑
InfiniteVGGT (Yuan et al., 2026)	7.30	15.95	0.08	91.96	0.10	92.94	7.38	47.83
OVGGT (Lu et al., 2026)	11.87	8.05	0.07	92.91	0.09	93.91	6.84	55.38
Stream3R-w (Lan et al., 2026)	12.76	5.26	0.12	87.57	0.20	87.40	6.44	47.75
CUT3R (Wang et al., 2025b)	29.62	3.16	0.18	75.18	0.11	90.68	7.26	41.22
TTT3R (Chen et al., 2026b)	25.53	4.65	0.11	86.98	**0.06**	94.45	7.77	40.42
LongStream (Cheng et al., 2026a)	10.36	6.62	0.06	94.60	0.14	88.29	5.69	53.84
LingBot-Map (Chen et al., 2026a)	19.74	18.87	**0.05**	96.01	0.08	**95.97**	1.68	90.58
HorizonStream (Cheng et al., 2026b)	8.02	13.04	0.23	**98.22**	0.12	90.97	2.00	84.69
LC3R	24.45	6.71	0.06	94.88	0.11	92.19	**1.37**	**91.81**



(a) Running adjacent-frame error on KITTI-02.

(b) Multi-gap error averaged over KITTI.

Figure 6: Long-horizon pose stability on KITTI. Left: running RMSE of adjacent-frame RPE_r and RPE_t over the complete KITTI-02 sequence. Right: multi-gap RPE_r and RPE_t averaged over all KITTI sequences. All methods are evaluated causally without loop closure.

ford Spires covers a substantially larger spatial extent. On this benchmark, *LC3R* achieves the best CD and F1 scores among the compared methods, indicating that local geometric prediction remains effective even as the reconstruction extends beyond the model's learned temporal context.

Figure 7 provides complementary qualitative comparisons across scenes with different lengths and spatial scales. On KITTI-02, Oxford Obs.-Q1, and 7Scenes Stairs-01, the compared methods recover the overall scene structure, while *LC3R* maintains coherent geometry across compact indoor, long driving, and landmark-scale sequences. The zoomed regions on KITTI and Oxford further show that local structures remain well aligned within the reconstructed scene.

Streaming efficiency. *LC3R* combines efficient streaming inference with accurate long-horizon pose estimation, running at 24.45 FPS with 6.71 GB of GPU memory on an NVIDIA H100 (Table 3). Although CUT3R and TTT3R are faster and use less memory, *LC3R* achieves substantially lower long-sequence pose drift.

4.4 Ablation Studies

To assess the contribution of each component, we progressively augment the adjacent-pose baseline with composition-aware supervision, longer training sequences, and a lightweight rotation refiner. All variants share the same causal 12-frame inference window and evaluation protocol (Table 4). Each addition improves performance on all three benchmarks; together, they reduce ATE by 67.8% on KITTI, 57.2% on Oxford, and 43.0% on VBR relative to the baseline.

Table 4: Ablation study on KITTI, Oxford, and VBR.

Variant	Training configuration				KITTI			Oxford			VBR		
	Frames	Model size	Refiner	$\mathcal{L}_{\mathrm{comp}}$	ATE	RPE_r	RPE_t	ATE	RPE_r	RPE_t	ATE	RPE_r	RPE_t
Baseline	32	996M	–	–	56.60	0.19	0.17	10.16	0.19	0.03	52.90	0.83	0.06
+ Comp. loss	32	996M	–	✓	27.66	0.14	0.13	9.40	0.15	0.03	44.72	0.63	0.06
+ Long-seq. training	128	996M	–	✓	22.31	0.11	0.12	5.95	0.13	**0.02**	36.77	0.62	0.06
+ Rot. Refiner	128	1B	✓	✓	**18.25**	**0.10**	**0.10**	**4.35**	**0.12**	**0.02**	**30.14**	**0.60**	**0.04**

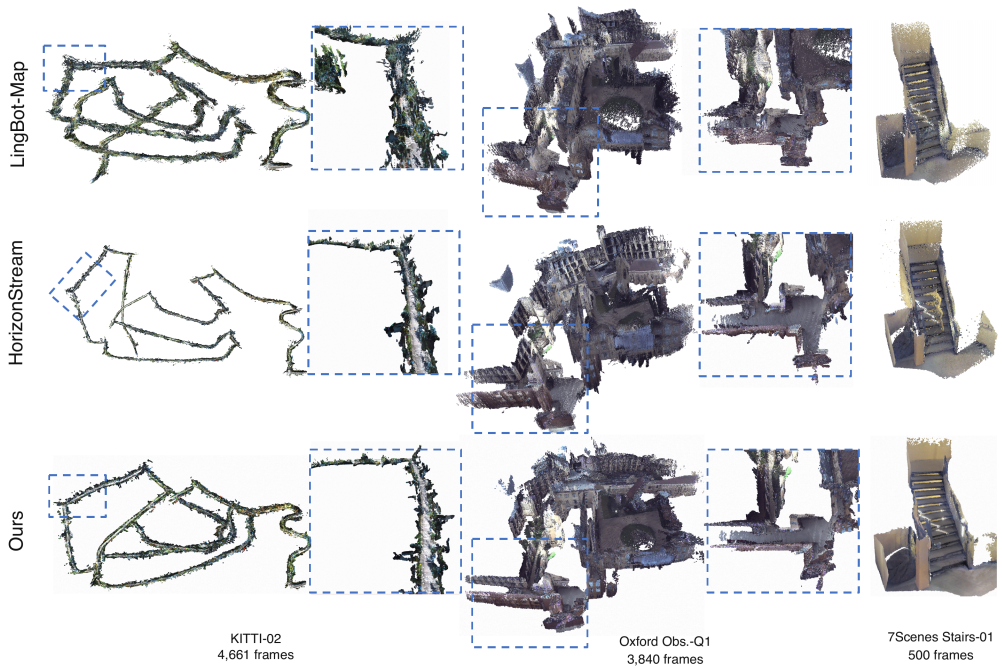


Figure 7: Qualitative point-cloud reconstruction on KITTI-02, Oxford Obs.-Q1, and 7Scenes Stairs-01. Across sequence lengths and scene scales, *LC3R* preserves coherent global structure and well-aligned local geometry using only local temporal context.

Composition-aware supervision. Adding the composition loss lowers ATE from 56.60 to 27.66 m on KITTI, with consistent gains on Oxford and VBR. The especially large KITTI improvement shows that supervising adjacent motions alone is insufficient: errors that appear small locally can grow rapidly after recursive trajectory composition.

Long-sequence training. Increasing the training horizon from 32 to 128 frames further reduces ATE by 19.3%, 36.7%, and 17.8% on KITTI, Oxford, and VBR, respectively. These results show that longer training clips improve trajectory accuracy while retaining the same local attention window at inference.

Rotation refinement. The 4.75M-parameter refiner (0.48% of the full model) provides another 18.2%, 26.9%, and 18.0% reduction. Refining relative rotations with recent visual and motion context improves the accuracy of the composed trajectories.

5 Conclusion

We presented *LC3R*, a simple streaming 3D reconstruction model that uses a fixed 12-frame causal attention window. Instead of storing or propagating persistent long-range state, *LC3R* predicts geometry in the current camera frame and motion between adjacent frames, keeping the estimation targets local as the sequence grows. Global camera trajectories and scene geometry are recovered by composing these local measurements over time. To make this composition reliable, a lightweight motion-visual rotation refiner improves relative rotations, while composition-aware pose supervision directly constrains their accumulated effect over multiple steps. Experiments on long driving, hand-held, and landmark-scale sequences show that *LC3R* delivers accurate camera tracking and dense reconstruction with bounded model memory and per-frame computation. These results demonstrate that persistent learned long-range memory is not a prerequisite for ultra-long streaming reconstruction: a simple short-context model can scale effectively when its local predictions are formulated to remain accurate and reliably composable.

AI USE STATEMENT

In this work, we used generative AI tools to assist to preprocess partial datasets, check experimental code, and edit manuscript for readability and language. All AI-assisted outputs were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide comprehensive implementation and experimental details in the appendix. In Appendix A, we describe the sampling strategies and hyperparameter settings for all datasets, as well as the data filtering procedures used for selected datasets. We also provide detailed data augmentation configurations. In Appendix B, we provide detailed training configurations and inference procedures, together with all relevant details of our efficiency evaluation, including the evaluation settings and protocols. These details are provided to enable the reproduction of our experimental results and efficiency measurements.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

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

Armen Avetisyan et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, 2024.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. Uasol, a large-scale high-resolution outdoor stereo dataset. *Scientific data*, 6(1):162, 2019.

Leonardo Brizi, Emanuele Giacomini, Luca Di Giammarino, Simone Ferrari, Omar Salem, Lorenzo De Rebotti, and Giorgio Grisetti. VBR: A vision benchmark in rome. In *IEEE International Conference on Robotics and Automation*, 2024. doi: 10.1109/ICRA57147.2024.10611395.

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

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. doi: 10.1109/TRO.2021.3075644.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a. URL https://arxiv.org/abs/2604.14141.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9158–9168. IEEE, 2025a.

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

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, October 2025b.

Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyang Guo, and Hao Wang. LongStream: Long-sequence streaming autoregressive visual geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 272–283, June 2026a.

Chong Cheng, Peilin Tao, Nanjie Yao, Guanzhi Ding, Xianda Chen, Yuansen Du, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, Zhengqing Chen, and Hao Wang. HorizonStream: Long-horizon attention for streaming 3d reconstruction. *arXiv preprint arXiv:2605.23889*, 2026b. URL https://arxiv.org/abs/2605.23889.

Xiaoya Cheng, Rouwan Wu, Xinyi Liu, Zeyu Cui, Yan Liu, Na Zhao, Yu Liu, Maojun Zhang, and Shen Yan. Airzoo: A unified large-scale dataset for grounding aerial geometric 3d vision, 2026c. URL https://arxiv.org/abs/2604.26567.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

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. URL https://arxiv.org/abs/2507.16443.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Zhijian Fang, Weicheng Zheng, Yijun Yuan, Weibang Wang, Zhuoguang Chen, Chang Sun, Junhao Huang, Kenan Li, Minghui Qin, and Hang Zhao. Slamformer-∞: Infinite slam transformer for unbounded frontend and backend processing, 2026. URL https://arxiv.org/abs/2608.03429.

Michaël Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 553–562, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9546, 2026a.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 414–424, 2026b.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction, 2026. URL https://arxiv.org/abs/2601.04090.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830, 2018.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17658–17668. IEEE, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

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.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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. In *International Conference on Learning Representations*, 2026.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1003–1012. IEEE, 2017. URL https://arxiv.org/pdf/1611.05267.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91. Springer, 2024. doi: 10.1007/978-3-031-73220-1_5.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 132324–132345, 2026a.

Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. Droid-slam in the wild, 2026b. URL https://arxiv.org/abs/2603.19076.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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. In *International Conference on Learning Representations*, 2026c.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023. doi: 10.1109/TPAMI.2022.3179507.

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

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *European Conference on Computer Vision*, 2024.

Changkun Liu, Bin Tan, Zeran Ke, Shangzhan Zhang, Jiachen Liu, Ming Qian, Nan Xue, Yujun Shen, and Tristan Braud. PLANA3R: Zero-shot metric planar 3d reconstruction via feed-forward planar splatting. In *Advances in Neural Information Processing Systems*, 2025a.

Changkun Liu, Jiezhi Yang, Zeman Li, Yuan Deng, Jiancong Guo, and Luca Ballan. Mem3R: Streaming 3d reconstruction with hybrid memory via test-time training. *arXiv preprint arXiv:2604.07279*, 2026. doi: 10.48550/arXiv.2604.07279. URL https://arxiv.org/abs/2604.07279.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16651–16662, June 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

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.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. In *Advances in Neural Information Processing Systems*, 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.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural language inference by tree-based convolution and heuristic matching. In *Annual Meeting of the Association for Computational Linguistics*, pp. 130–136, 2016.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3d reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16695–16705, June 2025. doi: 10.1109/CVPR52734.2025.01556.

Jing Ou, Zidong Cao, Yinrui Ren, Zhuoxiao Li, Jinjing Zhu, Tongyan Hua, Shuai Zhang, Hui Xiong, and Wufan Zhao. Holo360d: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3d reconstruction and beyond. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2604.22482.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025. doi: 10.1109/IROS60139.2025.11246002.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Ming Qian, Tianjian Ouyang, Mingchao Sun, Zijian Wang, Jincheng Xiong, Jiarong Han, Yongchang Zhang, Jiawei Zhang, Xu Wang, Yu Liu, Luyang Tang, Fei Yu, Zengye Ge, Mengmeng Du, Yuan Liu, Nianfei Fan, Song Wang, Yingliang Peng, Chunxue Jia, Yang Liu, Shiying Zeng, Haozhe Shi, Junnan Lai, Hongyu Pan, Zheng Wu, Ning Guo, Mu Xu, and Hang Zhang. Abot-earth 0.5: Generative 3d earth model, 2026a. URL https://arxiv.org/abs/2606.09967.

Ming Qian, Zijian Wang, Minchao Sun, Jincheng Xiong, Hang Zhang, Mu Xu, Chi Wang, and Baoquan Chen. Gs-voxel: Fitting-free structured latents for large-scale 3dgs generation, 2026b. URL https://arxiv.org/abs/2608.17988.

Ming Qian, Zimin Xia, Changkun Liu, Shuailei Ma, Wen Wang, Zeran Ke, Bin Tan, Hang Zhang, and Gui-Song Xia. Sat3DGen: Comprehensive street-level 3d scene generation from single satellite image. In *International Conference on Learning Representations*, 2026c.

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.

Katharina Schmid, Nicolas von Lützow, Jozef Hladký, Angela Dai, and Matthias Nießner. GenRecon: Bridging generative priors for multi-view 3d scene reconstruction, 2026. URL https://arxiv.org/abs/2605.23888.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, June 2016. doi: 10.1109/CVPR.2016.445.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013.

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*, pp. 573–580. IEEE, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. Vggt-world: Transforming vggt into an autoregressive geometry world model. *arXiv preprint arXiv:2603.12655*, 2026.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander G. Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5283–5293, June 2025.

Peilin Tao, Chong Cheng, Yuansen Du, Caiwei Song, Zhengqing Chen, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, Hainan Cui, and Shuhan Shen. Anchor3R: Streaming 3d reconstruction with transient anchors for long-horizon visual mapping. *arXiv preprint arXiv:2606.05035*, 2026a. URL https://arxiv.org/abs/2606.05035.

Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The oxford spires dataset: Benchmarking large-scale LiDAR-visual localisation, reconstruction and radiance field methods. *The International Journal of Robotics Research*, 45 (6):839–857, 2026b. doi: 10.1177/02783649251369905.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, pp. 16558–16569, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems*, 2023.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8942–8952, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Hengyi Wang and Lourdes Agapito. Amb3r: Accurate feed-forward metric-scale 3d reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14612–14625, 2026.

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

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

Runze Wang, Yuxuan Song, Youcheng Cai, and Ligang Liu. STAC: Plug-and-play spatio-temporal aware cache compression for streaming 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7567–7576, June 2026a.

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, June 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4909–4916. IEEE, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

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

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.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, June 2025.

Congrong Xu, Huachen Gao, Xingyu Chen, Yuliang Xiu, Jun Gao, and Anpei Chen. R^3: 3d reconstruction via relative regression. *arXiv preprint arXiv:2605.26519*, 2026. doi: 10.48550/arXiv.2605.26519. URL https://arxiv.org/abs/2605.26519.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21924–21935, June 2025.

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.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Bin Yu, Shijie Lian, Xiaopeng Lin, Zhaolong Shen, Yuliang Wei, Haishan Liu, Changti Wu, Hang Yuan, Bailing Wang, Cong Huang, et al. 3d-mix for vla: A plug-and-play module for integrating vggt-based 3d information into vision-language-action models. *arXiv preprint arXiv:2603.24393*, 2026.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams, 2026.

Fengyi Zhang, Tianjun Zhang, Kasra Khosoussi, Zheng Zhang, Zi Huang, and Yadan Luo. TALO: Pushing 3d vision foundation models towards globally consistent online reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21870–21879, June 2026a.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv preprint arXiv:2603.27437*, 2026b.

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

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. Loger: Long-context geometric reconstruction with hybrid memory. *arXiv preprint arXiv:2603.03269*, 2026c.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21936–21947, June 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.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. In *International Conference on Learning Representations*, 2026.

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

A Data Processing

A.1 Temporal Sampling

Video datasets We construct every training sample as an ordered sequence. For video datasets, the interval $[a, b]$ denotes the range of frame-index steps between consecutive inputs. We use three temporal sampling strategies according to the structure and length of the source trajectory.

1. **Forward fixed-stride sampling.** We uniformly sample one feasible interval from $[a, b]$ and use it throughout the clip. Frames are always traversed in chronological order. This strategy is used for long trajectories that can provide the requested number of frames without reversing direction.
2. **Adaptive forward sampling.** We sample only from the forward portion of a trajectory. When the remaining sequence is too short for the configured maximum interval, we reduce the upper bound to fit the complete clip. Depending on the dataset, the resulting clip uses either one fixed stride or independently sampled per-step strides.
3. **Foldback sampling.** We traverse the source sequence with a stride sampled from $[a, b]$. Upon reaching a sequence boundary, the traversal reverses direction and a new stride is sampled for the next segment. Foldback sampling allows a short source video to provide a long training clip while preserving local continuity and avoiding jumps between unrelated frames.

Multi-view datasets For unordered multi-view data, we construct fixed-length sequences using a camera pose graph. Let θ_{ij} denote the relative rotation between views i and j, and let $d_{ij} = \|\mathbf{t}_i - \mathbf{t}_j\|_2/s$ denote their translation normalized by the median nearest-neighbor camera distance s. For each view, we consider its K nearest neighbors and retain at most K_g edges satisfying $\theta_{ij} < \theta_{\max}$ and $d_{ij} < \tau_t$. Beam search then expands valid paths and keeps the best B candidates according to their motion and turning costs. Repeated views are permitted only through valid graph edges, subject to a minimum unique-view ratio $\rho_{\min}$.

Table 5: Pose-graph sequence sampling parameters.

Dataset	$\theta_{\max}$	τ_t	K	K_g	B	$\rho_{\min}$
HyperSim	20°	5	48	24	192	0.50
US4K	20°	5	64	24	128	1.00
BlendedMVS	40°	5	24	14	96	0.75

US4K requires all views to be unique, while HyperSim and BlendedMVS allow constrained revisiting when necessary.

Dataset percentages The temporal sampling strategy or multi-view construction rule used for each dataset is summarized in Table 6.

A.2 Data Filtering

We find that public datasets contain occasional annotation errors and noisy samples. By inspecting sampled training examples and analyzing the samples associated with loss spikes during training, we identify problematic data and apply dataset-specific filtering. Removing these erroneous samples substantially improves model performance.

BlendedMVS We observed that some BlendedMVS scenes contain sideways or upside-down images, as shown in Fig. 8. We manually identified and excluded 38 such scenes.

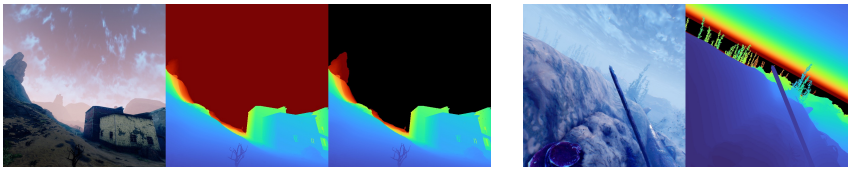


Figure 8: Examples of sideways and upside-down images in BlendedMVS.

Table 6: Per-dataset sequence-construction and interval policies used for training, resolved from the training configuration and dataset-loader defaults. Interval bounds are measured in source frames. For pose-graph data, b_{med} denotes the scene's median nearest-neighbor camera baseline.

Dataset	Type	Sampling strategy	Interval / constraint
TartanAir-v2	Synth.	Foldback	Hard $[1, 8]$, Easy $[1, 16]$
TartanAir	Synth.	Adaptive forward	$[1, 20]$
OmniWorld-Game	Synth.	Foldback	$[4, 16]$
TartanGround	Synth.	Foldback	$[1, 8]$
Virtual KITTI 2	Synth.	Adaptive forward	$[1, 5]$
HyperSim	Synth.	Pose-graph path	$\Delta R < 20^\circ, \Delta t < 5b_{\text{med}}$
PointOdyssey	Synth.	Adaptive forward	$[1, 4]$
Unreal4K	Synth.	Pose-graph path	$\Delta R < 20^\circ, \Delta t < 5b_{\text{med}}$
Spring	Synth.	Foldback	$[1, 4]$
MVS-Synth	Synth.	Foldback	$[1, 4]$
Dynamic Replica	Synth.	Adaptive forward	$[1, 16]$
ASE	Synth.	Foldback	$[1, 2]$
MidAir	Synth.	Foldback	$[1, 8]$
MatrixCity	Synth.	Forward fixed-stride	1
AirZoo	Synth.	Foldback	$[1, 8]$
BlendedMVS	Synth.	Pose-graph path	$\Delta R < 40^\circ, \Delta t < 5b_{\text{med}}$
Internal Synthetic	Synth.	Forward fixed-stride	$[2, 4]$
DL3DV	Real	Adaptive forward	$[1, 20]$
Waymo	Real	Adaptive forward	$[1, 8]$
ARKitScenes	Real	Foldback	1
ScanNet++	Real	Foldback	1
WildRGBD	Real	Foldback	$[1, 4]$
ScanNet	Real	Adaptive forward	$[1, 30]$
MapFree	Real	Adaptive forward	$[1, 30]$
ARKitScenes HR	Real	Foldback	1
UASOL	Real	Adaptive forward	$[1, 40]$
DDAD	Real	Forward fixed-stride	1
KITTI-360	Real	Forward fixed-stride	$[2, 6]$
Internal Real	Real	Forward fixed-stride	$[1, 2]$
Holo360D	Real	Foldback	$[1, 2]$

TartanAir & TartanAir v2 For the TartanAir datasets, we found that some scenes contained erroneous depth values in sky regions, as shown in Figure 9(a). We invalidated the corresponding depth values using semantic masks. We also observed erroneous water-surface depths in the Ocean scene, as shown in Figure 9(b), and therefore excluded this scene from both TartanAir and TartanAir-v2.



(a) Incorrect Depth (b) Ocean

Figure 9: Examples of incorrect scenes in TartanAir and TartanAir-v2.

OmniWorld-Game We observed erroneous geometric annotations in the dataset, particularly anomalous focal lengths in the camera intrinsics, as shown in Figure 10. We therefore remove unreliable scenes by validating cross-frame correspondences. Specifically, we sample multiple image pairs at different temporal positions in each sequence and obtain reliable 2D correspondences using SIFT (Lowe, 2004) feature matching and fundamental-matrix RANSAC. The matched pixels are then back-projected into 3D using the annotated depths and camera intrinsics. We estimate the relative rigid motion between the two frames and compare it with the annotated camera pose. Rotation error, translation direction and scale errors, and depth-normalized geometric residuals are jointly used to identify inconsistent frame pairs. A sequence is removed when severe inconsistencies account for at least 15% of its valid frame pairs and form a cluster of at least three pairs.

DL3DV Since DL3DV is constructed from captured videos using structure-from-motion (SfM) and multi-view stereo (MVS), some scenes contain erroneous pose reconstructions. We assume



Figure 10: Examples of incorrect scenes in OmniWorld-Game.

that scenes with discontinuous camera trajectories are likely to suffer from reconstruction failures. Therefore, we remove any scene in which the rotation between two consecutive frames exceeds $70°$, or the inter-frame translation exceeds $15\times$ median inter-frame translation of that scene. Moreover, because the DL3DV depth maps contain noise, we first discard the farthest 2% of depth values, apply a morphological opening with a 5×5 kernel, and then perform erosion with a 3×3 kernel to remove isolated depth outliers.

HyperSim We found that all sampled RGB frames from `ai_003_001/cam_00` and `ai_004_009/cam_01` were completely black; therefore, we excluded both sequences from training.

ScanNet++ We observed floating artifacts within the reconstructed meshes of some scenes, which occluded large portions of otherwise valid depth. We therefore excluded the affected scenes from training.

ARKitScenes We found that the low-resolution depth maps in ARKitScenes, captured using the iPhone LiDAR sensor, are not sufficiently accurate. Therefore, we do not use these depth maps for supervision.

A.3 Spatial Preprocessing and Augmentation

With probability 0.9, we preserve the source field of view by resizing the image isotropically to 504 pixels in width and producing a 504×280 canvas through vertical center cropping or symmetric mean-color padding. The padding value is $(0.485, 0.456, 0.406)$ in normalized RGB space. Geometric annotations and camera intrinsics are transformed consistently with the image. Photometric transformations are applied only to valid image content; synthetic padding is restored to its mean value afterward. The photometric pipeline comprises color jitter (brightness, contrast, saturation, hue, and gamma), JPEG degradation, and image blur. For 20% of training sequences, a single set of augmentation parameters—including the JPEG and blur decisions—is sampled and shared by all frames. For the remaining 80%, these parameters are sampled independently for each frame. This mixed strategy retains temporal coherence in part of the training data while exposing the model to frame-wise photometric variation.

B Experimental Details

This section supplies implementation and evaluation details omitted from the main paper for space.

B.1 Training Data

We train on a diverse mixture of 30 synthetic and real-world datasets covering indoor scenes, outdoor environments, autonomous driving, handheld capture, and aerial trajectories. As summarized in Table 7, synthetic and real-world data account for 62.05% and 37.95% of the sampling distribution, respectively. We assign larger sampling probabilities to datasets containing long, temporally coherent trajectories, while sampling from the remaining datasets to preserve diversity in scene appearance, geometry, and camera motion.

Table 7: Training-data composition. Sampling ratios are obtained by normalizing the dataset weights used by the training sampler. Internal data are divided into two categories: *Internal Synthetic* and *Internal Real*.

Synthetic data		Real-world data	
Dataset	Ratio (%)	Dataset	Ratio (%)
TartanAir-v2 Wang et al. (2020)	8.96	DL3DV Ling et al. (2024)	8.96
TartanAir Wang et al. (2020)	6.72	Waymo Sun et al. (2020)	5.38
OmniWorld-Game Zhou et al. (2026)	8.96	ARKitScenes Baruch et al. (2021)	1.34
TartanGround Patel et al. (2025)	6.72	ScanNet++ Yeshwanth et al. (2023)	3.81
VirtualKITTI2 Cabon et al. (2020)	6.27	WildRGBD Xia et al. (2024)	2.91
HyperSim Roberts et al. (2021)	6.27	ScanNet Dai et al. (2017)	2.46
PointOdyssey Zheng et al. (2023)	0.90	MapFree Arnold et al. (2022)	1.12
Unreal4K Tosi et al. (2021)	1.34	ARKitScenes HR Baruch et al. (2021)	2.46
Spring Mehl et al. (2023)	0.23	UASOL Bauer et al. (2019)	0.45
MVS-Synth Huang et al. (2018)	0.67	DDAD Guizilini et al. (2020)	1.34
Dynamic Replica Karaev et al. (2023)	0.45	KITTI-360 Liao et al. (2023)	2.02
ASE Avetisyan et al. (2024)	2.69	Holo360D Ou et al. (2026)	2.02
MidAir Fonder & Van Droogenbroeck (2019)	2.69	Internal Real	3.68
MatrixCity Li et al. (2023)	1.34		
AirZoo Cheng et al. (2026c)	1.34		
BlendedMVS Yao et al. (2020)	2.47		
Internal Synthetic	4.03		
Synthetic total	**62.05**	**Real-world total**	**37.95**

B.2 Training Configuration

We initialize Stage I from the public π^3 checkpoint (Wang et al., 2026b), replace its absolute-pose output with adjacent-frame relative poses, and adapt the network to causal inference. Stage II extends the training clips while keeping the local prediction window fixed at 12 frames and activates composition-aware supervision and the rotation refiner. In the final calibration phase, the geometry and pose networks are frozen and only the confidence branch is optimized, as in Table 8.

Table 8: Training schedule. Batch size is the global number of sequences.

Phase	Frames	Iterations	Batch	Hardware
Stage I	32	32K	48	48 NVIDIA H20
Stage II	128	38K	32	32 AMD MI308
Confidence calibration	128	4K	32	32 AMD MI308

We use AdamW (Loshchilov & Hutter, 2019). The Stage-I learning rate peaks at 1×10^{-4} and decays to 1×10^{-6}. In Stage II, the refiner and the remaining parameters use peak learning rates of 5×10^{-5} and 2×10^{-5}, respectively. Confidence calibration uses a peak rate of 2×10^{-5}. Training uses bfloat16, gradient clipping at 1.0, and an exponential moving average with decay 0.999 (Polyak & Juditsky, 1992).

B.3 Inference and Evaluation Protocol

Evaluation datasets We evaluate our method on two complementary tasks: camera pose estimation and dense 3D reconstruction. For camera pose estimation, we use KITTI Geiger et al. (2012), Oxford Spires Tao et al. (2026b), and VBR Brizi et al. (2024). These benchmarks cover kilometer-scale driving sequences as well as handheld and vehicle-mounted capture. For dense 3D reconstruction, we report results on 7Scenes Shotton et al. (2013) , TUM-Dynamic Sturm et al. (2012), and Oxford Spires Tao et al. (2026b), covering indoor and outdoor environments as well as static and dynamic scenes. All methods process every evaluated sequence once in temporal order with a stride of 1. For 7Scenes, we select `seq-01` from each of its seven scenes; for all other benchmarks, we use all sequences in the corresponding evaluation set. Dataset statistics are summarized in Table 9.

Inference Mode The model processes each stream once in temporal order without chunk-wise resets. Its learned state is the fixed-size local KV cache. Each frame predicts its local geometry and adjacent-frame motion; global geometry and camera poses are recovered incrementally by composition. For optional loop closure, we retrieve revisited frames by FAISS (Johnson et al., 2019) over DINOv2-SALAD descriptors (Izquierdo & Civera, 2024), estimate relative poses between lo-

Table 9: Evaluation datasets and statistics.

Dataset	Task	Frames/sequence	#Seq.	Inference stride
KITTI Geiger et al. (2012)	Camera pose	271–4,661	11	1
VBR Brizi et al. (2024)	Camera pose	8,815–18,846	7	1
Oxford Spires Tao et al. (2026b)	Pose / Reconstruction	3,821–3,840	10	1
7Scenes Shotton et al. (2013)	Reconstruction	500–1,000	7	1
TUM-Dynamic Sturm et al. (2012)	Reconstruction	707–1,261	8	1

cal windows around retrieved pairs, and optimize the resulting sparse pose graph. This backend does not alter the causal predictor.

Evaluation Protocol Camera trajectories are aligned to ground truth with Umeyama Sim(3). We report RMSE ATE, rotational RPE (RPE_r), and translational RPE (RPE_t). For reconstruction, point maps are resampled to the ground-truth grid, aligned by Umeyama registration, voxelized, and refined by point-to-point ICP following LingBot-Map (Chen et al., 2026a). Voxel sizes are $4/512$ m for 7Scenes and TUM-Dynamic and 0.05 m for Oxford Spires. F1 uses thresholds of 0.25 m indoors and 4 m on Oxford. Oxford inference uses every frame, while its point-cloud evaluation uses every tenth frame.

Baseline methods Streaming baselines are InfiniteVGGT (Yuan et al., 2026), OVGGT (Lu et al., 2026), Stream3R-w (Lan et al., 2026), CUT3R (Wang et al., 2025b), TTT3R (Chen et al., 2026b), LongStream (Cheng et al., 2026a), LingBot-Map (Chen et al., 2026a), and HorizonStream (Cheng et al., 2026b). Optimization-based pose baselines are VGGT-SLAM (Maggio et al., 2025), MASt3R-SLAM (Murai et al., 2025), VGGT-Long (Deng et al., 2025), SLAMFormer-∞ (Fang et al., 2026), Droid-SLAM (Teed & Deng, 2021), DPV-SLAM (Lipson et al., 2024), DPVO (Teed et al., 2023), and Droid-W (Li et al., 2026b). Except for the unavailable SLAMFormer-∞ KITTI result, all numbers use public implementations and checkpoints under our unified protocol.

B.4 Runtime and Memory Evaluation Protocol

We evaluate runtime and memory efficiency on the full 4,661-frame KITTI-02 sequence at stride 1 using a single NVIDIA H100 GPU. Each method uses its specified inference precision and input resolution and all are evaluated with frame-by-frame streaming inference mode. Before measurement, we perform an untimed 64-frame warm-up and then reset the streaming state. The complete 4,661-frame sequence is subsequently processed for timing. All input frames are preloaded onto the GPU, and image loading and preprocessing are excluded. We synchronize CUDA once immediately before and after the complete forward pass; thus, the reported FPS includes model inference and all streaming-state updates, without introducing per-frame synchronization overhead. Peak memory is reported as the maximum PyTorch-allocated GPU memory excluding the storage occupied by the preloaded input sequence. It includes model parameters, intermediate activations, cached tensors, and persistent streaming state. All methods are evaluated on the same sequence and hardware using the same timing and memory accounting protocol. The resulting throughput and memory measurements are reported in Table 3.

C More Quantitative Results

The main paper reports the complete per-sequence results on KITTI. Table 10 and Table 11 provide the corresponding results on the remaining pose benchmarks. Beyond these benchmarks, we further examine generalization on four long real-world videos collected from different environments. Since accurate ground-truth camera trajectories are unavailable for these sequences, we use them only for qualitative evaluation.

Large-scale real-world sequences We additionally evaluate on four long videos as shown in Figure 11. These sequences cover different scene scales and capture distributions, including extended urban driving, campus roads, repeated structures, intersections, and substantial variations in camera motion and visual appearance. They therefore complement the standard benchmarks with less curated and more heterogeneous long-video streams.

Table 10: Camera pose estimation on Oxford Spires. Within streaming methods, best and second-best results are marked in blue and underlined, respectively. * denotes methods that require camera intrinsics.

		Per-sequence ATE (m) ↓										Sequence averages		
		Keble-2	Keble-3	Keble-4	Keble-5	Observ.-1	Observ.-2	Christ-2	Christ-3	Christ-5	Bodleian-2			
		3840f	3840f	3840f	3840f	3840f	3839f	3821f	3840f	3840f	3840f			
Group	Method	0.20 km	0.18 km	0.23 km	0.21 km	0.27 km	0.28 km	0.25 km	0.18 km	0.80 km	0.55 km	Avg. ATE (m)↓	Avg. RPE_r (°)↓	Avg. RPE_t (m)↓
Opt.-based	VGGT-SLAM	2.64	3.78	3.89	1.86	5.74	3.39	7.41	1.02	12.94	18.83	6.15	0.39	0.05
	MASt3R-SLAM	7.06	18.03	11.27	14.64	10.26	10.30	2.90	6.95	19.85	19.99	12.12	0.36	0.12
	VGGT-Long	9.07	9.37	11.77	3.96	6.03	4.05	2.41	0.48	14.38	12.59	7.41	0.29	0.08
	Droid-SLAM*	16.79	15.99	29.40	21.00	0.87	1.08	1.31	1.15	7.58	1.82	9.70	0.12	0.03
	DPV-SLAM*	0.20	1.31	0.03	0.30	0.16	0.45	0.15	0.04	25.77	1.08	2.95	0.15	0.03
	DPVO*	0.12	0.84	0.05	0.29	0.16	0.51	0.15	0.35	12.04	1.08	1.56	0.16	0.01
	Droid-W*	2.45	0.99	2.64	4.27	1.07	1.06	1.63	0.75	1.56	6.28	2.27	0.08	0.01
Streaming	InfiniteVGGT	25.02	20.39	25.09	24.15	29.10	28.14	33.02	12.95	40.39	79.32	31.76	5.30	2.70
	OVGGT	32.97	23.01	28.53	23.23	28.51	27.75	30.01	13.09	42.35	82.41	33.19	7.83	3.61
	Stream3R-w	31.28	22.39	31.07	24.63	28.38	27.18	38.12	12.99	42.39	82.66	34.11	9.05	2.67
	CUT3R	31.68	21.67	22.76	25.07	26.77	29.41	33.93	13.06	43.66	79.93	32.79	1.78	0.25
	CUT3R w/ reset	13.10	11.55	15.76	8.98	13.58	15.60	12.36	5.71	31.31	19.42	14.74	0.66	0.05
	TTT3R	24.66	22.56	27.33	18.67	26.94	28.18	19.83	9.68	26.76	55.01	25.96	3.19	0.31
	TTT3R w/ reset	8.27	7.26	8.77	7.55	9.43	10.30	4.56	2.11	21.75	14.00	9.40	0.29	0.04
	LongStream	11.57	9.44	16.88	10.67	12.43	14.34	17.93	6.42	35.49	24.02	15.92	0.30	0.06
	LingBot-Map	**2.71**	3.04	2.04	**3.23**	5.38	4.43	**2.21**	0.74	37.99	11.43	7.32	2.29	0.53
	HorizonStream	5.79	6.47	7.36	6.27	6.33	5.30	10.69	3.11	25.35	11.41	8.81	0.20	0.03
	HorizonStream w/ LC	5.79	6.47	**1.25**	6.24	6.33	5.30	10.69	**0.47**	25.35	11.41	7.93	0.20	0.03
	Ours	4.24	**1.70**	5.08	4.44	**4.60**	**4.40**	3.07	1.15	**9.03**	**5.76**	4.35	**0.12**	**0.02**
	Ours w/ LC	4.24	**1.70**	2.04	4.44	**4.60**	**4.40**	3.07	0.85	**9.03**	**5.76**	**4.02**	**0.12**	**0.02**

Table 11: Camera pose estimation on VBR. Within streaming methods, best and second-best results are marked in blue and underlined, respectively. * denotes methods that require camera intrinsics.

		Per-sequence ATE (m) ↓							Sequence averages		
		Colosseo-0	Campus-0	Campus-1	Pincio-0	Spagna-0	Diag-0	Ciampino-1			
		8815f	12042f	11671f	11142f	14141f	10021f	18846f			
Group	Method	1.45 km	2.73 km	2.95 km	1.27 km	1.56 km	1.02 km	5.20 km	Avg. ATE (m)↓	Avg. RPE_r (°)↓	Avg. RPE_t (m)↓
Opt.-based	VGGT-SLAM	54.30	122.07	117.15	32.48	24.86	33.52	188.30	81.81	1.08	0.31
	MASt3R-SLAM	88.75	116.21	114.37	67.11	54.62	35.67	171.76	92.64	1.25	0.30
	VGGT-Long	44.03	132.67	115.62	64.29	59.04	33.66	179.43	89.82	0.82	0.37
	Droid-SLAM*	98.12	132.85	119.53	63.21	64.20	35.86	198.21	101.71	0.82	0.23
	DPV-SLAM*	86.34	107.85	70.29	51.56	55.60	33.87	105.91	73.06	0.70	0.16
	DPVO*	90.42	109.72	58.49	52.83	57.84	33.84	102.60	72.25	1.34	0.15
	Droid-W*	50.27	125.02	96.89	21.96	22.82	24.59	171.52	73.29	0.58	0.13
Streaming	InfiniteVGGT	89.89	136.76	116.63	77.76	63.35	33.84	202.09	102.90	8.27	4.84
	OVGGT	96.94	137.74	117.39	77.76	62.49	34.84	201.43	104.08	11.43	6.07
	Stream3R-w	92.51	138.38	116.38	77.43	64.57	35.64	202.14	103.87	7.81	2.72
	CUT3R	99.21	136.32	117.92	75.05	62.29	33.66	202.10	103.79	1.50	0.37
	CUT3R w/ reset	76.68	59.51	67.21	48.57	48.55	26.66	169.18	70.91	0.75	0.11
	TTT3R	91.11	129.47	116.59	74.31	60.96	34.39	194.75	100.23	3.58	0.50
	TTT3R w/ reset	73.32	56.51	52.76	37.39	35.33	34.97	175.98	66.61	0.81	0.10
	LongStream	83.01	120.29	106.92	71.39	60.81	30.61	177.51	92.93	0.71	0.13
	LingBot-Map	**11.11**	21.45	12.39	46.30	35.40	16.16	63.36	29.45	4.42	2.70
	HorizonStream	40.89	27.84	24.28	26.77	28.17	25.39	29.83	29.02	**0.60**	**0.04**
	HorizonStream w/ LC	24.66	27.22	8.72	16.29	21.33	**7.36**	17.92	17.64	0.61	**0.04**
	Ours	48.12	11.06	17.69	21.01	29.24	28.06	55.82	30.14	**0.60**	**0.04**
	Ours w/ LC	16.70	**8.10**	**4.30**	**5.16**	**17.96**	9.71	**8.01**	**9.99**	**0.60**	**0.04**

Across the four sequences, our method produces trajectories with more consistent large-scale structure than the compared streaming baselines. This behavior is consistent with the motivation of our design: the network always predicts an adjacent relative transform, so its regression target does not

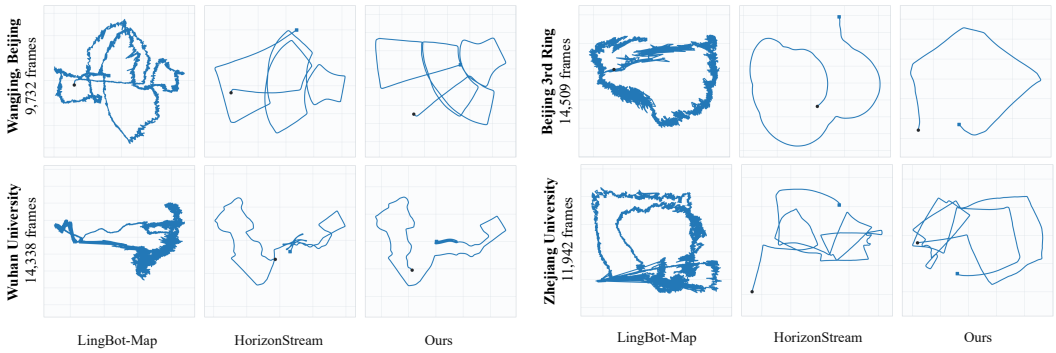


Figure 11: Qualitative no-loop trajectory comparisons on four large-scale real-world videos. For efficiency and fair comparison, all methods are evaluated with temporal stride 5. The results illustrate the behavior of streaming models under variations in scene scale, camera motion, and data distribution.

change as the explored area or elapsed sequence length increases. The temporal refiner improves relative rotations before they are composed into the global trajectory.

Taken together, these qualitative results extend the benchmark evaluation to camera viewpoints, motion patterns, and spatial scales. While accurate ground truth is unavailable for quantitative assessment, the observed trajectory consistency suggests that keeping pose prediction local provides a practical basis for generalization across heterogeneous long-video streams.

D Discussion

Our results show that the proposed local formulation is particularly effective for long, temporally continuous streams. The model maintains stable local pose estimation as the processing horizon grows and remains consistent under longer pose compositions, leading to strong trajectory accuracy on KITTI, VBR, and Oxford Spires. On compact indoor benchmarks, the gains are less pronounced and our reconstruction results are not uniformly the best. One possible reason is that, in spatially confined scenes with frequent revisits and high visual overlap, persistent geometric context can remain useful over long portions of the sequence. Our method does not explicitly retain such scene-level state, yet remains competitive in these settings, suggesting that the local formulation remains effective across both compact and substantially larger environments.

The optional loop-closure backend complements the local predictor when scene revisits are available. Because it is independent of model training, loop closure introduces sparse long-range constraints only at inference time while leaving the learned streaming formulation unchanged. This provides additional global correction on sequences with useful revisits, without requiring persistent learned long-range state inside the predictor. In this sense, local prediction and loop closure play complementary roles: the former provides a bounded, sequence-length-independent streaming backbone, while the latter offers optional global refinement when additional long-range constraints are available.

The resulting modular design also leaves room for further extensions. The local streaming backbone can incorporate advances in feed-forward reconstruction, caching, and inference acceleration without changing the underlying local-to-global formulation. Future work may explore stronger handling of dynamic scenes (Chen et al., 2025a; Hu et al., 2026b), external memory (Wang & Agapito, 2026; Zhang et al., 2026c), larger-scale pretraining (Lin et al., 2025), and selective long-range constraints for cases where local temporal continuity becomes ambiguous, as well as applications to long-horizon embodied perception (Agarwal et al., 2026; Li et al., 2026a; Kim et al., 2024; Yu et al., 2026), 3D-VLM (Hu et al., 2026a; Fan et al., 2026; Zhang et al., 2026b), 3D Generation (Qian et al., 2026b;a;c; Schmid et al., 2026; Xiang et al., 2025) and world modeling (Sun et al., 2026; Huang et al., 2026; Wu et al., 2026).