CLASP: COMPLEMENTARY CHUNK AND LONG PASSES FOR FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

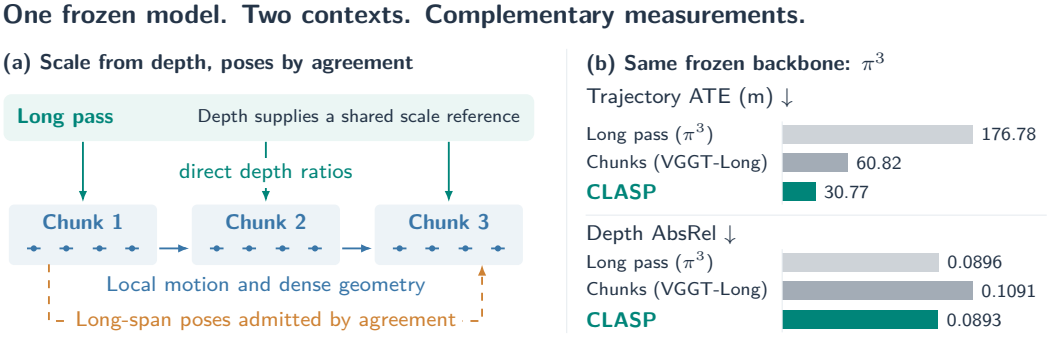

Figure 1: **CLASP combines two complementary inference contexts of one frozen model: a chunk-wise pass and a long full-sequence pass.** The chunk pass supplies local motion and geometry, while the long pass supplies a depth-scale reference and agreeing long-span poses. With the same frozen π^3 backbone, CLASP reduces ATE by 49% and depth AbsRel by 18% relative to the chunk-only VGGT-Long, matching the long pass in depth.

ABSTRACT

Feed-forward 3D reconstruction faces two distinct failures as sequences grow: camera predictions can deteriorate in a full-sequence pass while overlapping chunks accumulate pose and scale errors through their alignment chain. We present CLASP (Chunk-Linked, Aligned-Scale Pose graph), a training-free framework that combines both inference contexts of the same frozen model in one pose graph. Chunks supply local relative motion and output geometry. The long pass supplies a common scale reference through dense depth ratios between predictions of the same frame, independently of camera motion and overlap alignment. It also contributes long-range pose constraints where they agree with a chunk-derived trajectory. Separating these roles retains useful scale information when long-pass pose constraints are rejected, allowing one formulation to handle long and short captures without a dataset-specific switch. Experiments with three backbones on six datasets show that CLASP lowers trajectory error on long driving sequences and keeps pose accuracy close to or above full-sequence pass on shorter captures.

1 INTRODUCTION

Feed-forward geometric foundation models reconstruct 3D scenes directly from uncalibrated images. Following pairwise models such as DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024), multi-view models including Fast3R (Yang et al., 2025), MUSt3R (Cabon et al., 2025), VGGT (Wang et al., 2025a), π^3 (Wang et al., 2025c), VGGT-Ω (Wang et al., 2026a), MapAnything (Keetha et al., 2026) and Depth Anything 3 (Lin et al., 2026b) jointly predict camera poses and dense geometry. Long captures require these predictions to remain accurate locally and consistent across the full trajectory. Yet adding frames to a single forward pass can degrade camera predictions even when the sequence fits in memory (Appendix A.7). Streaming methods handle longer sequences with new architectures, such as learned memory and test-time-adapted context (Zhang

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.

et al., 2026b; Xie et al., 2026). Here, we study how two inference contexts of one frozen model complement each other: a chunk-wise pass over overlapping chunks and a long full-sequence pass.

Overlapping chunks preserve accurate local motion and geometry, but assembling them introduces a different failure mode. With sequential alignment, each chunk's coordinate frame and scale depend on all preceding overlap alignments. A single unreliable overlap can therefore corrupt every subsequent chunk's placement, including the scale of its depth predictions. VGGT-Long corrects drift with loop closure and global similarity optimization (Deng et al., 2026), while VGGT-SLAM aligns submaps with projective transformations (Maggio et al., 2025), LASER with layer-wise depth scales (Ding et al., 2026) and TALO with spatially varying warps (Zhang et al., 2026a). Their scale measurements, however, still come from overlapping or loop-closing chunks. So accurate geometry within each chunk can be placed at an incorrect scale in the assembled reconstruction.

A long full-sequence pass offers a complementary source of information. Its depth predictions are computed independently of the overlap chain and can remain useful as a common scale reference even when its camera trajectory is unreliable. On shorter captures, its relative poses can also preserve global structure that chunking loses. We build on this complementarity in CLASP (Fig. 1), a training-free framework that combines the two inference contexts of one frozen backbone: the chunk-wise pass and the long full-sequence pass. The two passes contribute different geometric quantities as measurements in one pose graph (Strasdat et al., 2010; Grisetti et al., 2010; Kümmerle et al., 2011), with robust reweighting reducing the influence of inconsistent measurements.

The chunk pass supplies local relative-pose links and the output point maps. The long pass provides a depth-scale reference through dense ratios between its depth and the chunk depth of the same frame, which constrain per-chunk scale without requiring camera motion. We separately test long-pass relative poses against a trajectory recovered from the chunk links and admit them over temporal spans where the two agree, countering drift accumulated along local links. Rejected pose constraints leave the depth-scale reference available, so the same formulation handles long and short captures without a dataset-specific switch. The output geometry reuses the chunk point maps, changing only their per-chunk scale and camera poses.

Experiments with π^3, VGGT and VGGT-Ω on six datasets show the benefits of both long-pass roles. On long driving sequences, the depth-scale reference complements chunk links and loop closures, and CLASP consistently reduces trajectory error relative to VGGT-Long with the same backbone. On shorter synthetic and indoor captures, where chunking can fall below the full-sequence pass, admitted long-span constraints keep pose accuracy comparable to or better than that pass. Correcting the scale and poses of the chunk point maps also improves the reconstructed geometry, bringing depth accuracy close to the full-sequence pass. In this paper, we make the following contributions:

- **Two complementary inference contexts in one pose graph.** We combine a chunk-wise pass and a long full-sequence pass of the same frozen model as measurements in a shared reconstruction, using the complementary information available at the two context lengths.
- **Each pass contributes what it measures reliably.** Chunks supply local motion and geometry. The long pass supplies an overlap-independent depth-scale reference and, where the passes agree, long-span poses, so its scale remains usable when its poses are rejected.
- **Consistent gains over chunk-based alignment.** Across six datasets, CLASP mostly achieves the best pose accuracy among training-free methods with improved geometric consistency and remains competitive with trained streaming models.

2 Related Work

Multi-view 3D Reconstruction. Classical pipelines recover geometry through structure-from-motion (Schönberger & Frahm, 2016; Pan et al., 2024) and multi-view stereo (Schönberger et al., 2016). Neural methods jointly optimize poses and scene representations (Meuleman et al., 2023) or directly predict Gaussian primitives (Charatan et al., 2024; Chen et al., 2024). DUSt3R (Wang et al., 2024) predicts point maps from image pairs with a pre-trained transformer architecture and then MASt3R (Leroy et al., 2024) adds learned matching prior. Fast3R (Yang et al., 2025), MUSt3R (Cabon et al., 2025) and VGGT (Wang et al., 2025a) jointly predict cameras and geometry by extending reconstruction to multiple views. Later on, π^3 (Wang et al., 2025c) introduces permutation-equivariant reasoning and VGGT-Ω (Wang et al., 2026a) improves accuracy

and efficiency through unsupervised video distillation. Further developments incorporate geometric priors (Keetha et al., 2026), depth-and-ray prediction (Lin et al., 2026b) and efficient aggregation (Elflein et al., 2025; Shen et al., 2026).

Long-sequence 3D Reconstruction. Long-sequence methods combine temporal context with geometric consistency. Earlier systems use recurrent volumetric fusion (Sun et al., 2021) and differentiable bundle adjustment (Teed & Deng, 2021). CUT3R (Wang et al., 2025b) maintains recurrent state, while STream3R (Lan et al., 2026) uses causal attention. Camera-token pools (Li et al., 2026) and spatial memory (Wu et al., 2025) retain historical context, while state adaptation (Chen et al., 2026a; Zheng et al., 2026) and cache pruning (Yuan et al., 2026) extend streaming without retraining. LoGeR (Zhang et al., 2026b) combines sliding-window attention with test-time-training memory, and Scal3R (Xie et al., 2026) learns adaptive global context. MASt3R-SLAM (Murai et al., 2025) integrates reconstruction priors with a geometric backend, VGGT-SLAM (Maggio et al., 2025) aligns submaps and MERG3R (Cheng et al., 2026) merges overlapping subsets. VGGT-Long (Deng et al., 2026) optimizes chunk similarities with loop constraints, LASER (Ding et al., 2026) aligns depth-layer scales and TALO (Zhang et al., 2026a) corrects spatially varying discrepancies. The relative-pose formulation of Scal3R (Lin et al., 2026a) learns pose queries on a frozen backbone for online graph optimization. Unlike these streaming and chunk-based methods, CLASP runs one frozen backbone on both overlapping chunks and the full sequence and fuses them in one pose graph. Appendix A.1 covers further chunk-alignment, streaming and acceleration methods.

3 METHOD

CLASP runs one frozen backbone in two inference contexts and combines their predictions in a pose graph (Fig. 2). Its design follows one principle: each pass contributes only the quantities it predicts reliably. The chunk pass supplies local relative motion and the output geometry (Sec. 3.1). The long pass supplies an overlap-independent depth-scale reference for every chunk (Sec. 3.2) and long-span pose constraints, which are admitted only over temporal spans where the two passes agree (Sec. 3.3). A robust pose graph then recovers one trajectory, on which the rescaled chunk geometry is placed (Sec. 3.4). All measurements and consistency tests use only the predicted geometry and the input images, without ground truth.

3.1 CHUNK AND LONG PASSES

Given an ordered image sequence $\{I_i\}_{i=0}^{N-1}$, we estimate camera-to-world poses $T_i = (R_i, p_i)$ and place the predicted camera-local point maps in one world frame, up to a global similarity. The backbone processes 60-frame chunks $\mathcal{C}_c$ with 30-frame overlap (the *chunk pass*) and the complete sequence in a single evaluation (the *long pass*). Each evaluation predicts camera poses, local point maps $X_i(u) \in \mathbb{R}^3$ over pixels u and confidences. Superscripts c and L mark predictions of chunk c and of the long pass, as in (R_i^c, p_i^c) and (R_i^L, p_i^L).

Pose graph. As in graph-based SLAM (Lu & Milios, 1997; Grisetti et al., 2010; Kümmerle et al., 2011), we fuse the predictions in a pose graph whose nodes are the unknown poses T_i. Each edge $(i, j) \in \mathcal{E}$ carries a relative-pose measurement $(\widehat{R}_{ij}, \widehat{t}_{ij})$ mapping camera j into camera i, so consistent poses satisfy $R_j = R_i \widehat{R}_{ij}$ and $p_j - p_i = R_i \widehat{t}_{ij}$, and the poses are estimated by weighted least squares. Unlike chunk- or submap-level graphs (Deng et al., 2026; Maggio et al., 2025; Xiong et al., 2026), CLASP uses one rigid pose per frame and solves chunk scales separately (Sec. 3.2).

Chunk links. Each chunk predicts poses in its own frame and units. Following VGGT-Long (Deng et al., 2026), we compose confidence-weighted Umeyama similarities (Umeyama, 1991) between overlapping chunk point maps into an initial scale s_c per chunk, in the first chunk's units. Only these scales are taken from the alignment, and the pose graph recovers rotations and positions. For pairs $j = i + d$ within chunk c, with gaps $d \in \mathcal{D}_C = \{1, 2, 3, 5, 10\}$, the chunk links are

$$\widehat{R}_{ij}^C = (R_i^c)^\top R_j^c, \qquad \widehat{t}_{ij}^C = s_c (R_i^c)^\top (p_j^c - p_i^c). \tag{1}$$

Each frame i keeps the point map of the later chunk covering it, denoted $c(i)$, as its output geometry. Because s_c accumulates along the chain of overlap alignments, a single unreliable overlap mis-scales both the links and the geometry of every later chunk.

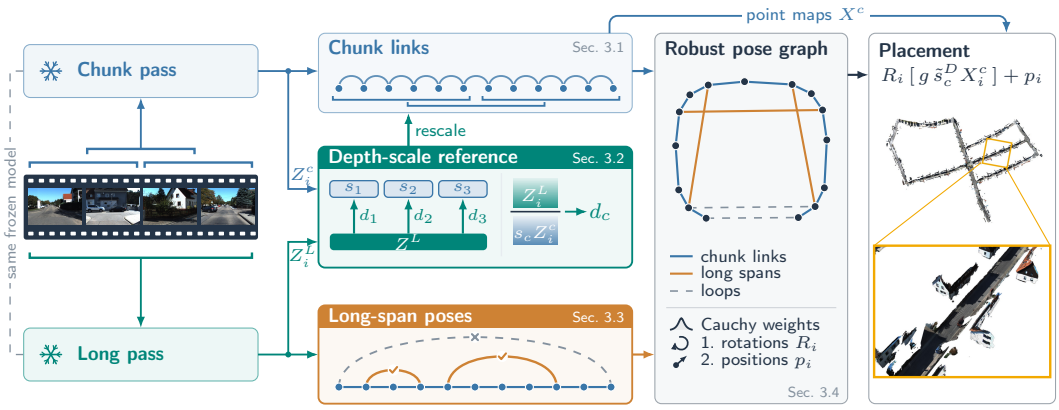

Figure 2: **Overview of CLASP.** One frozen backbone runs on overlapping chunks and on the full sequence (long pass). Chunks supply local links and the output point maps. The long pass supplies two measurements. Same-frame depth ratios give every chunk a direct scale measurement d_c against one depth reference, independent of overlaps, and long-pass poses. Long-span poses are admitted (✓) or rejected (×) by their agreement with the chunk trajectory. A robust pose graph recovers one trajectory, on which the rescaled chunk point maps are placed.

3.2 An overlap-independent scale reference

VGGT-Long and LASER measure scale between overlapping or loop-closing windows (Deng et al., 2026; Ding et al., 2026), so an earlier error in the chain persists. We instead compare each chunk directly with the long-pass depth of the same frames, predicted independently of the overlap chain.

One scale per chunk. Two diagnostic observations support one scale variable per chunk, measured from dense depth ratios. On four π^3 captures, the log-scale standard deviation relative to ground truth is 0.002–0.008 within chunks but 0.12–0.19 between chunks, so scale errors arise mainly between chunks. Moreover, the pixelwise log-depth ratios between the two passes have a median absolute deviation of only 0.005–0.02: their depth maps agree up to one scale factor even where their camera predictions differ. Ground truth is used only for this diagnosis. Appendix A.7 names these captures and shows that, on driving sequences where the long-pass poses fail, the long-pass depth measures chunk scales more accurately than both long-pass trajectory and chunk chain.

Depth-ratio measurements. Let $Z_i^c(u)$ and $Z_i^L(u)$ be the positive depths of frame i predicted by chunk c and by the long pass, each in its own camera frame. For each frame and chunk, we compute

$$\rho_i = \mathrm{median}_{u\in\mathcal{U}_i} \log \frac{Z_i^L(u)}{s_{c(i)} Z_i^{c(i)}(u)}, \qquad d_c = \mathrm{median}_{i:\, c(i)=c}\, \rho_i, \tag{2}$$

where $\mathcal{U}_i$ contains pixels with finite, positive depths in both predictions. We introduce a log-scale correction x_c per chunk, giving the corrected scale $\widetilde{s}_c = s_c \exp(x_c)$, and an auxiliary variable x_L for the arbitrary global scale of the long pass. Each chunk then yields the constraint $x_c - x_L = d_c$, which requires neither camera motion nor image matching. The pixelwise spread of the log ratios measures how closely the two depth maps agree up to scale and down-weights less consistent chunks.

Scale solve. A second measurement takes the same linear form (Table 5 in Appendix A.2). An overlap-trajectory row $x_b - x_a = \log(E_a/E_b)$ compares the extents E_a, E_b of the camera path shared by adjacent chunks a and b. These rows couple adjacent chunks through camera motion, independently of the point-map fit that sets s_c, and also constrain chunks without a valid depth row. Writing every row as $x_b - x_a = v_{ab}$, we solve

$$\min_{x:\, x_0=0} \sum_{(a,b)} w_{ab}(x_b - x_a - v_{ab})^2 + \lambda_s \|x\|_2^2 \tag{3}$$

by iteratively reweighted least squares, whose Cauchy weights w_{ab} suppress inconsistent rows. The constraint $x_0 = 0$ fixes the scale gauge, and the small ridge λ_s keeps weakly constrained chunks near their initialization.

Link and geometry scales. We solve Eq. (3) once. Its solution x rescales both the chunk-link translations, by $\exp(x_c)$, and the output point maps (Sec. 3.4), so the links and the geometry share the full dense depth reference. Neither row type uses the long-pass camera predictions, so the scale reference remains available when long-pass poses are rejected.

3.3 Long-span pose constraints by agreement

Long-pass relative poses connect distant frames directly and can correct drift accumulated along chunk links, but they become less reliable as sequences grow. We therefore admit them only where they agree with a reference trajectory $(\bar{R}_i, \bar{p}_i)$, which the solver of Sec. 3.4 recovers from the scale-corrected chunk links alone, with a negligible rotation-prior weight of 10^{-6}. A similarity fit from the long-pass camera centers to this reference supplies a translation scale s_L. At spans $d \in \mathcal{D}_L = \{15, 30, 60, 120, 240, 480\}$, the candidate measurements are

$$\widehat{R}^L_{ij} = (R^L_i)^\top R^L_j, \qquad \widehat{t}^L_{ij} = s_L (R^L_i)^\top (p^L_j - p^L_i), \quad j = i + d, \tag{4}$$

sampled at stride $\max(1, \lfloor d/20 \rfloor)$ and excluding pairs already covered by verified loops.

Span-level admission. The candidates at span d form one class L_d. For each candidate, we compare its rotation and translation with the reference:

$$a_{ij} = \operatorname{angle}\Big(\bar{R}_i^\top \bar{R}_j, \widehat{R}^L_{ij}\Big), \qquad q_{ij} = \frac{\|(\bar{p}_j - \bar{p}_i) - \bar{R}_i \widehat{t}^L_{ij}\|_2}{\max(\|\widehat{t}^L_{ij}\|_2, \epsilon)}, \tag{5}$$

where $\epsilon > 0$ prevents division by zero. A class is admitted as a whole if $\operatorname{median} a_{ij} < \tau$ and $\operatorname{median} q_{ij} < \tau_t$, with $\tau = 1^\circ$ and $\tau_t = 0.3$, and rejected otherwise. Testing a whole span, rather than individual edges, avoids selecting isolated edges that happen to match the reference's own drift. Admission is decided once, before the final solve. If no span is admitted, the graph relies on chunk links and loops, and the scale reference of Sec. 3.2 stays active. The same formulation thus handles long captures, where long-pass poses are often rejected, and short captures, where admitted spans preserve the global structure of the long pass, without a dataset-specific switch.

Noise calibration. Admitted edges constrain both rotations and translations. Their disagreement with the reference contains prediction noise as well as the accumulated chunk drift they are meant to correct, so estimating noise separately at each span would down-weight precisely the drift-correcting constraints. We therefore measure rotation and translation noise at the shortest admitted span, where accumulated drift is smallest, and use it as the class noise level ν_k in Eq. (6) for all admitted classes, while each class keeps its own adaptive outlier scale κ_k.

3.4 Pose-graph recovery and geometry placement

Graph construction. The final edge set $\mathcal{E}$ contains chunk links, admitted long-pass edges and verified loop closures. Loop candidates at least 100 frames apart are retrieved with GeM-pooled features from a frozen π^3 encoder and verified by SIFT matching, fundamental-matrix RANSAC and a trimmed rigid fit on the scale-corrected chunk points. Edges are grouped into robust classes by link gap, long-pass span and loops.

Robust weights. For an edge residual of magnitude r_e in class k, each solve uses the weight

$$w_e = \left(\frac{\sigma_1}{\nu_k}\right)^2 \frac{1}{1 + \left(r_e / (2.3849 \kappa_k)\right)^2}, \tag{6}$$

where the reference noise σ_1 is the disagreement between overlapping chunks on the same adjacent motion, ν_k sets the class influence and κ_k scales the Cauchy outlier weight with its standard tuning constant. Rotation and translation use separate noise levels and residuals. For the other classes, ν_k and κ_k are the residual median absolute deviation scaled by 1.4826 and floored at σ_1, with the exceptions listed in Appendix A.2.

Two-stage solve. As in chordal pose-graph initialization (Martinec & Pajdla, 2007; Carlone et al., 2015), we first recover rotations by relaxing $Y_i = R_i^\top$ to an unconstrained matrix and solving, with weights fixed,

$$\min_{\{Y_i\}} \sum_{(i,j)\in\mathcal{E}} w_{ij}^R \|Y_j - \widehat{R}_{ij}^\top Y_i\|_F^2 + \sum_i \alpha_i \|Y_i - (R_i^L)^\top\|_F^2, \tag{7}$$

where the weak, robustly weighted prior α_i toward the long-pass rotations sets the rotation reference. After each of four sparse linear solves, we project the matrices onto $\mathrm{SO}(3)$ by singular value decomposition and update the weights from angular residuals. With rotations fixed and without bundle adjustment, camera positions solve the reweighted sparse least-squares problem

$$\min_{\{p_i\}:\, p_0=0} \sum_{(i,j)\in\mathcal{E}} w_{ij}^p \|p_j - p_i - R_i \widehat{t}_{ij}\|_2^2. \tag{8}$$

Geometry placement. The corrected scales $\widetilde{s}_c = s_c \exp(x_c)$ and the recovered trajectory can still differ by one global factor. We estimate this factor g as the displacement-weighted median ratio between recovered and geometry-scaled chunk displacements over in-chunk frame pairs, rather than from the overall trajectory extent, which also changes with drift. The final world point map is

$$X_i^{\mathrm{world}}(u) = R_i \left[g\, \widetilde{s}_{c(i)} X_i^{c(i)}(u) \right] + p_i. \tag{9}$$

Each output map thus equals the chunk prediction up to a per-chunk scale and a rigid transform. Appendices A.2 and A.11 give the remaining solver details and discuss limitations.

4 Results

4.1 Evaluation setup

Datasets and metrics. We evaluate camera poses on six datasets: KITTI Odometry (Geiger et al., 2012), KITTI-360 (Liao et al., 2022), 4Seasons (Wenzel et al., 2020), TartanAir (Wang et al., 2020), Virtual KITTI (Cabon et al., 2020) and 7-Scenes (Shotton et al., 2013). Depth and point clouds require dense ground-truth depth, so we evaluate them on the three datasets that provide it: TartanAir, Virtual KITTI and 7-Scenes. Following the relative-pose evaluation used by VGGT (Wang et al., 2025a), we utilize AUC@30 to measure the accuracy curve up to 30°. RRA and RTA are the mean pairwise angular errors in degrees and ATE is camera-center RMSE in metres after Sim(3) alignment. For geometry, we report Chamfer distance (CD), F1, depth AbsRel and δ_1, the fraction of depth predictions within a ratio of 1.25 of ground truth.

Baselines. We compare with the streaming methods CUT3R (Wang et al., 2025b), TTT3R (Chen et al., 2026a), LoGeR (Zhang et al., 2026b) and Scal3R (Xie et al., 2026). Within each frozen-backbone block, we compare the base model with FastVGGT (Shen et al., 2026), VGGT-Long (Deng et al., 2026), MERG3R (Cheng et al., 2026), TALO (Zhang et al., 2026a) and CLASP. All results are produced with the public code and default hyperparameters. The main-text tables report these training-free methods with VGGT-Ω, and Appendix A.6 reports all three backbones. Base-model rows are shown in gray and excluded from within-block ranking.

4.2 Camera pose estimation

Table 1 compares camera poses on the three driving datasets. Over these long trajectories, the single pass loses global consistency, while chunk-based methods recover local accuracy but accumulate drift along their alignment chains. CLASP gives the most accurate trajectories among training-free methods on KITTI-360 and 4Seasons, and the lowest ATE on KITTI Odometry. Among the streaming methods, the recurrent CUT3R and TTT3R drift substantially over the long driving sequences. Scal3R is the strongest of them and outperforms CLASP on KITTI Odometry, but falls behind it on KITTI-360 and 4Seasons. Appendix A.6 shows similar trends with VGGT and π^3.

Table 2 covers synthetic and indoor sequences, where the single long pass already captures most of the global structure. Here chunking can hurt: VGGT-Long, MERG3R and TALO each fall below the long pass in AUC@30 on at least one of these datasets. CLASP instead admits long-pass pose

Table 1: Camera pose estimation on driving sequences. Best results are **bold** and second-best results are underlined among the training-free methods, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	KITTI Odometry				KITTI-360				4Seasons			
	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓
Streaming methods												
CUT3R	0.058	67.84	72.61	207.29	0.020	74.73	77.62	184.35	0.064	56.14	80.59	432.81
TTT3R	0.392	28.92	24.77	83.39	0.109	49.21	46.42	109.86	0.299	39.40	35.82	155.87
LoGeR	0.628	10.57	8.80	39.49	0.448	17.75	14.43	52.88	0.465	15.94	17.06	166.17
Scal3R	0.883	3.08	3.15	14.35	0.421	25.21	16.41	61.88	0.735	6.70	7.53	53.68
VGGT-Ω *(Base model)*	0.474	5.87	35.91	138.96	0.206	10.07	51.28	159.43	0.368	11.28	43.26	357.30
FastVGGT	—	—	—	—	—	—	—	—	—	—	—	—
VGGT-Long	**0.832**	4.67	**4.81**	23.47	0.614	10.10	9.83	38.18	0.795	4.79	5.71	31.60
MERG3R	0.317	51.66	50.99	189.85	0.049	88.19	73.03	202.97	0.108	81.65	63.99	440.73
TALO	0.748	5.81	7.25	24.90	0.601	9.09	11.74	35.73	0.764	5.24	6.48	37.13
CLASP (Ours)	0.820	**4.25**	4.96	**20.97**	**0.686**	**8.02**	**7.61**	**20.26**	**0.808**	**4.55**	**5.01**	**24.98**

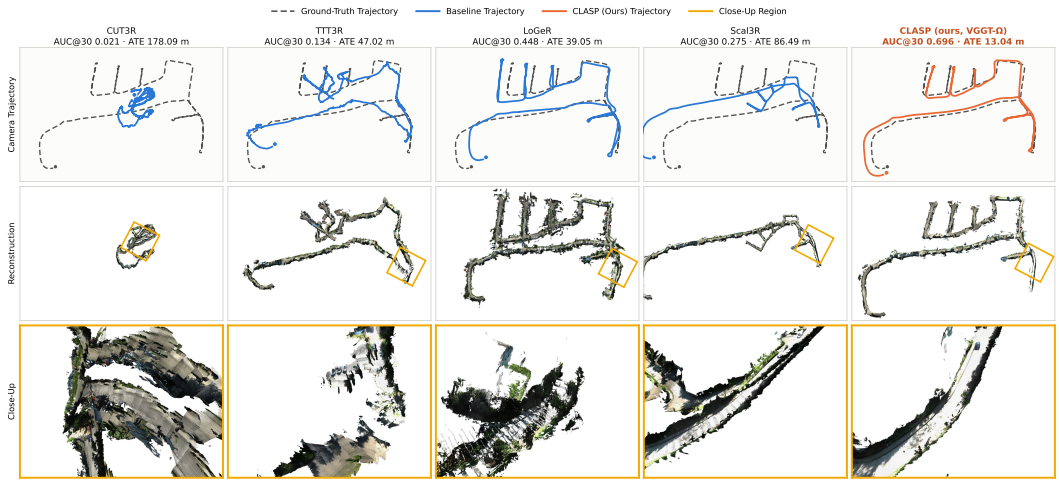

Figure 3: Comparison with streaming methods on KITTI-360 (0002/w001). CLASP using VGGT-Ω follows the road layout more closely and significantly reduces ATE.

constraints where they agree with the chunk trajectory. It improves on the single pass in AUC@30 on all three datasets and gives the best training-free result in every metric. Among the streaming methods, Scal3R is comparable to CLASP on TartanAir and Virtual KITTI, whereas CLASP is clearly more accurate on 7-Scenes. The trends are similar with other two backbones. These results support using long-pass information selectively: chunking corrects drift on long trajectories, while admitted long-pass constraints preserve global structure when the base model already captures it.

4.3 Depth and Reconstructed Geometry

Table 3 examines whether the recovered trajectories also yield consistent geometry. Among training-free methods, CLASP is best or second best in every depth and point-cloud metric, and it improves on VGGT-Long in CD, F1 and AbsRel on all three datasets. This advantage over VGGT-Long largely holds with the other two backbones (Appendix A.6). Since CLASP only rescales and places the chunk point maps (Eq. (9)), these gains come from more consistent scale and camera placement across chunks. Because per-frame depth still comes from the chunk predictions, the single pass retains slightly lower AbsRel on Virtual KITTI and 7-Scenes, where it is already accurate. Some streaming methods lead on individual datasets, such as Scal3R on Virtual KITTI, whereas CLASP is more accurate than all of them on 7-Scenes.

Table 2: Camera pose estimation on synthetic and indoor sequences. Best results are **bold** and second-best results are underlined among the training-free methods, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	TartanAir				Virtual KITTI				7-Scenes			
	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓
Streaming methods												
CUT3R	0.262	16.49	46.85	5.37	0.312	9.50	57.41	56.88	0.112	27.70	55.70	0.46
TTT3R	0.789	4.14	6.44	0.57	0.726	4.09	13.25	4.17	0.546	5.96	18.31	0.16
LoGeR	0.928	1.17	2.46	0.26	0.827	2.24	8.85	2.35	0.758	4.13	8.23	0.07
Scal3R	0.969	0.70	1.32	0.09	0.904	0.49	5.28	0.63	0.680	4.48	11.37	0.09
VGGT-Ω *(Base model)*	0.918	1.23	3.04	0.21	0.891	0.47	6.38	1.32	0.854	2.52	4.94	0.03
FastVGGT	—	—	—	—	0.884	0.50	6.94	0.66	—	—	—	—
VGGT-Long	0.949	1.04	1.82	0.16	0.889	0.48	10.29	0.71	0.764	2.83	8.78	0.06
MERG3R	0.863	1.35	6.20	0.70	0.862	0.84	6.59	3.76	0.819	2.91	6.17	0.04
TALO	0.865	0.98	5.81	1.43	0.887	**0.42**	7.26	6.01	0.789	2.73	7.18	0.08
CLASP (Ours)	**0.973**	**0.79**	**1.05**	**0.12**	**0.904**	**0.42**	**5.30**	**0.61**	**0.863**	**2.50**	**4.59**	**0.03**

Table 3: Point-cloud and depth accuracy on synthetic and indoor sequences. Best results are **bold** and second-best results are underlined among the training-free methods, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	TartanAir				Virtual KITTI				7-Scenes			
	CD↓	F1↑	AbsRel↓	δ_1↑	CD↓	F1↑	AbsRel↓	δ_1↑	CD↓	F1↑	AbsRel↓	δ_1↑
Streaming methods												
CUT3R	12.6357	0.0867	0.1534	0.879	6.0426	0.1636	0.0629	0.950	0.6805	0.1419	0.1072	0.894
TTT3R	6.6985	0.1980	0.1517	0.883	2.0429	0.3111	0.0611	0.952	0.2043	0.3535	0.0850	0.934
LoGeR	5.9594	0.2780	0.1092	0.919	3.1766	0.1628	0.0557	0.964	0.0786	0.5676	0.0706	0.950
Scal3R	7.3661	0.3159	0.1015	0.918	0.9710	0.6445	0.0369	0.972	0.1264	0.5162	0.0744	0.932
VGGT-Ω *(Base model)*	6.2027	0.3142	0.1605	0.908	2.1502	0.5261	0.0535	0.973	0.0748	0.7932	0.0484	0.960
FastVGGT	6.2969	0.3038	0.1585	0.906	2.2179	0.5235	**0.0540**	0.972	—	—	—	—
VGGT-Long	6.1443	0.3176	0.1584	**0.910**	2.0741	0.5727	0.0567	**0.973**	0.0904	0.6650	0.0553	**0.960**
MERG3R	**5.9655**	**0.3434**	0.1621	0.907	2.4400	0.4389	0.0585	0.972	0.0782	0.7620	**0.0484**	**0.960**
TALO	9.2296	0.1506	0.4086	0.515	3.0525	0.4423	0.1445	0.838	0.1094	0.5600	0.0804	0.953
CLASP (Ours)	6.1235	0.3391	**0.1570**	**0.910**	**2.0579**	**0.5867**	0.0555	**0.973**	**0.0666**	**0.7627**	0.0511	**0.960**

Table 4: Ablation of the two long-pass roles with π^3. **Bold** marks full CLASP.

	7-Scenes				4Seasons			
	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓
Long pass	0.785	3.56	7.57	0.058	0.310	14.76	48.39	409.19
Chunk graph	0.735	3.72	9.41	0.081	0.620	9.43	12.26	242.15
+ Long-pass scale	0.734	3.72	9.37	0.081	0.636	9.43	10.81	93.51
Full CLASP (+ long-range transl.)	**0.787**	**3.61**	**7.42**	**0.058**	**0.636**	**9.43**	**10.81**	**93.51**

Figures 3 and 4 show how trajectory errors propagate into the reconstruction. In the KITTI-360 example, CLASP follows the road layout more closely than the streaming methods. On KITTI Odometry sequence 05, it reduces the distortion of the chunk-based reconstructions and keeps a twice-traversed street aligned. The close-ups show that drift and scale errors displace otherwise detailed local predictions. The Appendix adds other backbones, trajectories and failure cases.

4.4 Ablation of key components

Table 4 isolates the two long-pass roles with π^3 by adding the scale reference and then the admitted translations to a chunk graph without long-pass measurements. On the long 4Seasons trajectories, where long-pass poses are unreliable, the scale reference more than halves the chunk graph's ATE,

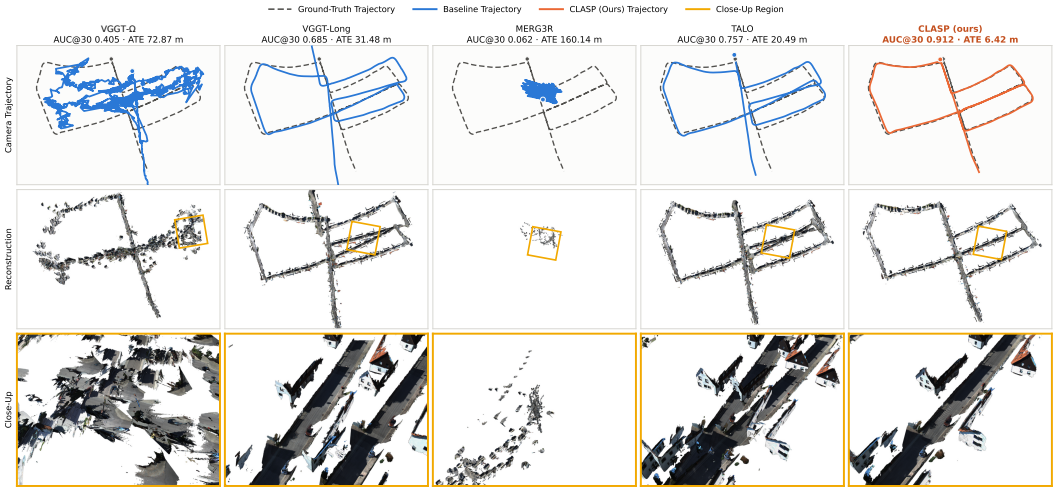

Figure 4: Training-free comparison with VGGT-Ω on KITTI Odometry sequence 05. CLASP reduces trajectory distortion and keeps the twice-traversed street in the close-up aligned, where VGGT-Long and TALO duplicate it.

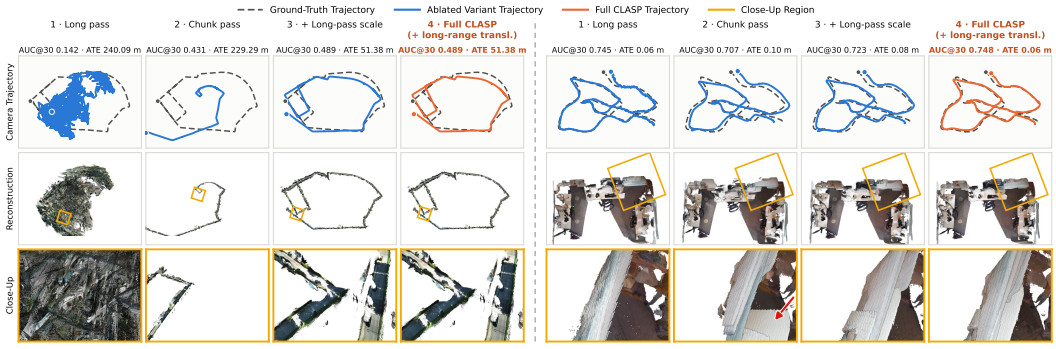

Figure 5: Ablation of the two long-pass roles on KITTI-360 (0009/w000, left) and 7-Scenes (office/seq-04, right). Scale correction accounts for the driving improvement, and admitted translations further improve the indoor trajectory. The red arrow marks a local misalignment.

and admitted translations leave the result unchanged. On the shorter 7-Scenes captures, scale alone leaves the chunk graph behind the long pass, whereas admitted translations recover its accuracy. The two roles are thus complementary, as Figure 5 also shows. Appendices A.8 and A.9 ablate the chunk-scale solve and the fusion admission thresholds.

5 Conclusion

We present CLASP, a training-free framework that combines chunk and full-sequence predictions of one frozen model in a single pose graph. Chunks supply local motion and geometry, while the long pass adds an overlap-independent scale reference and, where the passes agree, long-span pose constraints. Across three backbones and six datasets, CLASP improves trajectory accuracy on long sequences, maintains pose accuracy on shorter captures and yields more consistent geometry.

AI USE STATEMENT

We used large language model tools (Codex and Claude Code) for writing and for code. For writing, they helped edit, polish and format the manuscript text. For code, they helped implement and debug parts of the method and evaluation code and helped write the scripts that turn recorded evaluation outputs into our tables and figures. The research ideas, method design and experimental design are the authors' own. Language models did not generate or change any experimental results. All reported numbers come from executed runs, and we checked every table entry against the recorded outputs. The authors reviewed all AI-assisted text and tested all AI-assisted code. They take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Section 3 formulates every step of CLASP. Appendix A.2 lists the thresholds, iteration counts and noise settings used in these steps. Appendix A.5 describes the chunk layout, hardware, baseline configurations and ablation variants. It also notes that the reported results re-solve the scale and pose graphs offline from cached predictions and previously verified loop closures. CLASP needs no training: it uses the frozen weights of VGGT, π^3 and VGGT-Ω. All baselines are run with their public code (Section 4.1). For evaluation, Appendix A.3 describes the sequences and how long recordings are split, and Appendix A.4 defines the metrics and scoring details. Appendix A.10 reports runtime and memory. The codes will be released upon acceptance for full reproducibility.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work involves no human subjects and collects no new data. All experiments use public benchmarks (KITTI Odometry, KITTI-360, 4Seasons, TartanAir, Virtual KITTI and 7-Scenes) in accordance with their licenses. The real-world driving recordings were captured on public roads and may show pedestrians, vehicles and license plates. We use these images only as input for reconstruction and pose evaluation, and we make no attempt to identify individuals. Since CLASP adds no trained parameters, it inherits any biases of its frozen backbones. Better 3D reconstruction benefits robotics, autonomous driving, mapping and augmented reality. It could also make it easier to reconstruct private spaces from footage recorded without consent, although CLASP only improves the consistency of existing models rather than adding new capabilities. Finally, CLASP can fail, for example under false loop closures or inaccurate chunk predictions (Appendix A.12). Its outputs should therefore be validated independently before use in safety-critical navigation.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, 2022.

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

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *CVPR*, 2025.

Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert. Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization. In *ICRA*, 2015.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, 2024.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *ICLR*, 2026a.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *ECCV*, 2024.

Yutian Chen, Yuheng Qiu, Ruogu Li, Jay Patrikar, and Sebastian Scherer. Co-Me: Confidence-guided token merging for visual geometric transformers. In *CVPR*, 2026b.

Leo Kaixuan Cheng, Abdus Shaikh, Ruofan Liang, Zhijie Wu, Yushi Guan, and Nandita Vijaykumar. MERG3R: A divide-and-conquer approach to large-scale neural visual geometry. In *CVPR*, 2026.

Weichen Dai, Wenhan Su, Da Kong, Yuhang Ming, and Wanzeng Kong. Keyframe-based feed-forward visual odometry. *arXiv preprint arXiv:2601.16020*, 2026.

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. In *ICRA*, 2026.

Tianye Ding, Yiming Xie, Yiqing Liang, Moitreya Chatterjee, Pedro Miraldo, and Huaizu Jiang. LASER: Layer-wise scale alignment for training-free streaming 4D reconstruction. In *CVPR*, pp. 36433–36443, 2026.

Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3R-SfM: Towards feed-forward structure-from-motion. In *CVPR*, 2025.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. In *CVPR*, 2026.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2010.

Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T. Barron, Noah Snavely, and Aleksander Holynski. ZipMap: Linear-time stateful 3D reconstruction via test-time training. *CVPR*, 2026.

Seonghyun Jin and Jong Chul Ye. FILT3R: Latent state adaptive Kalman filter for streaming 3D reconstruction. In *ECCV*, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *3DV*, 2026.

Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g2o: A general framework for graph optimization. In *ICRA*, 2011.

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 *ICLR*, 2026.

Jungho Lee, Minhyeok Lee, Sunghun Yang, Minseok Kang, and Sangyoun Lee. SwiftVGGT: A scalable visual geometry grounded transformer for large-scale scenes. *CVPR Findings*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, pp. 71–91, 2024.

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. *ICLR*, 2026.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *TPAMI*, 2022.

Chin-Yang Lin, Yang-Che Sun, Cheng Sun, Fu-En Yang, Min-Hung Chen, Yen-Yu Lin, Wei-Chen Chiu, and Yu-Lun Liu. Scal3R: Learning efficient multi-relative pose query for scalable online 3D reconstruction. In *ECCV*, 2026a.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026b.

Feng Lu and Evangelos Milios. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, 1997.

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. In *RSS*, 2026.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. In *NeurIPS*, 2025.

Daniel Martinec and Tomáš Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, 2007.

Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *CVPR*, 2025.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited. In *ECCV*, 2024.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *ICCV*, 2023.

Haosong Peng, Hao Li, Jiaqi Chen, Yuhao Pan, Runmao Yao, Yalun Dai, Fushuo Huo, Fangzhou Hong, Zhaoxi Chen, Haozhao Wang, Dingwen Zhang, Ziwei Liu, and Wenchao Xu. SpatialBench: Is your spatial foundation model an all-round player? *arXiv preprint arXiv:2605.27367*, 2026.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pp. 4104–4113, 2016.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Training-free acceleration of visual geometry transformer. In *ICLR*, 2026.

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 *CVPR*, 2013.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, and Xiao-Xiao Long. LiteVGGT: Boosting vanilla VGGT via geometry-aware cached token merging. In *CVPR*, 2026.

Hauke Strasdat, J. M. M. Montiel, and Andrew J. Davison. Scale drift-aware large scale monocular SLAM. In *RSS*, 2010.

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *CVPR*, 2021.

Xianbing Sun, Zhikai Zhu, Zhengyu Lou, Bo Yang, Jinyang Tang, Liqing Zhang, He Wang, and Jianfu Zhang. AVGGT: Rethinking global attention for accelerating VGGT. In *CVPR*, 2026.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *NeurIPS*, 2021.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 1991.

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

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025b.

Runze Wang, Yuxuan Song, Youcheng Cai, and Ligang Liu. STAC: Plug-and-play spatio-temporal aware cache compression for streaming 3D reconstruction. In *CVPR*, 2026b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, pp. 20697–20709, 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 *IROS*, 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. *arXiv preprint arXiv:2507.13347*, 2025c.

Zipeng Wang and Dan Xu. FlashVGGT: Efficient and scalable visual geometry transformers with compressed descriptor attention. In *CVPR*, 2026.

Patrick Wenzel, Rui Wang, Nan Yang, Qing Cheng, Qadeer Khan, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. 4Seasons: A cross-season dataset for multi-weather SLAM in autonomous driving. *arXiv preprint arXiv:2009.06364*, 2020.

Yuchen Wu, Jiahe Li, Xiaohan Yu, Lina Yu, Jin Zheng, and Xiao Bai. SCE-SLAM: Scale-consistent monocular SLAM via scene coordinate embeddings. *CVPR*, 2026.

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

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3D reconstruction. In *CVPR*, 2026.

Zhuang Xiong, Chen Zhang, Qingshan Xu, and Wenbing Tao. VGGT-Motion: Motion-aware calibration-free monocular SLAM for long-range consistency. *ICML*, 2026.

Yangfan Xu, Lilian Zhang, Xiaofeng He, Pengdong Wu, Wenqi Wu, and Jun Mao. GPA-VGGT: Adapting VGGT to large scale localization by self-supervised learning with geometry and physics aware loss. *arXiv preprint arXiv:2601.16885*, 2026.

Zhisong Xu and Takeshi Oishi. FrameVGGT: Coherence-preserving memory for bounded streaming geometry. *arXiv preprint arXiv:2603.07690*, 2026.

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 *CVPR*, 2025.

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

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 *CVPR*, 2026a.

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*, 2026b.

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

A Appendix

This appendix extends the discussion of related work and completes the method details. It then describes the evaluation protocol and implementation configurations, reports comparisons with all three frozen backbones, collects diagnostics of the two inference contexts and ablates the chunk-scale solve and the fusion admission thresholds. It closes with resource measurements, scope and limitations, failure cases and additional visual comparisons.

A.1 Extended related work

This section extends the main-text related work with further methods that target long image sequences, directly or indirectly, grouped into chunk and submap alignment, long-context and streaming models, and single-pass acceleration.

Chunk and submap alignment. Several systems build on the chunk-and-align design of VGGT-Long (Deng et al., 2026) and the submap alignment of VGGT-SLAM (Maggio et al., 2025). SwiftVGGT (Lee et al., 2026) keeps the training-free VGGT-Long pipeline, but aligns neighboring chunks with a single SVD-based Sim(3) fit on sampled points instead of iteratively reweighted least squares and detects loops from VGGT's own encoder features instead of a separate place-recognition model. VGGT-Motion (Xiong et al., 2026) partitions the sequence into submaps guided by optical flow, registers them with anchor-driven dense Sim(3) alignment and enforces global consistency with a submap-level pose graph. VGGT-SLAM 2.0 (Maggio & Carlone, 2026) redesigns the factor graph of VGGT-SLAM to remove its 15-degree-of-freedom drift and planar degeneracy, and uses one of VGGT's attention layers to verify loop-closure candidates without additional training. Among the methods in our comparisons, TALO (Zhang et al., 2026a) refines submap alignment with thin-plate-spline warps, LASER (Ding et al., 2026) aligns depth-layer scales between consecutive windows and MERG3R (Cheng et al., 2026) merges overlapping subsets through global alignment and confidence-weighted bundle adjustment. In all of these systems, the scale of a submap follows from its alignment to overlapping or loop-closing submaps. CLASP differs in using a second inference context of the same model. Each chunk's scale is measured against the depth of the full-sequence pass, which a preceding overlap error cannot alter, and the long-pass relative poses enter the graph over temporal spans where they agree with the chunk trajectory.

Long-context and streaming models. A second group changes how the model holds context, reducing the growth of memory and computation with sequence length. VGG-T^3 (Elflein et al., 2026) and ZipMap (Jin et al., 2026) use test-time-training layers to compress the scene into a fixed-size state, giving linear-time reconstruction that still aggregates all views. Streaming models instead bound their key-value cache or recurrent state. InfiniteVGGT (Yuan et al., 2026) prunes the cache of a causal VGGT with a training-free rule, FrameVGGT (Xu & Oishi, 2026) keeps a fixed number of frame-level key-value segments together with sparse long-range anchors and STAC (Wang et al., 2026b) retains tokens with high decayed cumulative attention while compressing spatially redundant tokens into voxel-aligned representations. FILT3R (Jin & Ye, 2026) treats the recurrent state update as Kalman-style filtering in token space, balancing the retention of history against new observations without training. Other methods train components for odometry and SLAM. SCE-SLAM (Wu et al., 2026) maintains a consistent scale with learned scene coordinate embeddings and a scene-coordinate bundle adjustment. GPA-VGGT (Xu et al., 2026) fine-tunes VGGT on unlabeled sequences with photometric and geometric self-supervision for large-scale localization, and Keyframe FF-VO (Dai et al., 2026) learns a keyframe-selection policy for feed-forward visual odometry with reinforcement learning. These methods change the architecture, memory, state update or weights of the model, or train an additional module. CLASP leaves the frozen full-attention backbone unchanged and combines its predictions from two inference contexts.

Table 5: Scale measurements in the chunk-scale solve.

Measurement	Constraint	Role
Overlap trajectory	$x_b - x_a = \log(E_a/E_b)$	Couple adjacent chunks through their shared camera extents, independently of the point-map fit.
Long-pass depth	$x_c - x_L = d_c$	Read scale densely from the same frames in both contexts.

Single-pass acceleration. A third group addresses long sequences indirectly, by reducing the time and memory of one full-attention pass so that it can cover more frames. FastVGGT (Shen et al., 2026), which appears in our comparisons, merges tokens in the global-attention layers without training. LiteVGGT (Shu et al., 2026) selects merge anchors by geometric importance and reuses merge indices across layers, with optional fine-tuning and FP8 quantization. Co-Me (Chen et al., 2026b) merges low-confidence tokens ranked by a distilled confidence predictor, without retraining the base model. FlashVGGT (Wang & Xu, 2026) replaces dense global attention with cross-attention from all image tokens to a compact set of per-frame descriptor tokens, and reuses cached descriptors across chunks for online inference. AVGGT (Sun et al., 2026) converts early global-attention layers to frame attention and subsamples keys and values in the remaining ones. With these changes, FastVGGT and LiteVGGT process 1,000-image scenes in one pass, and FlashVGGT scales beyond 3,000 images. These methods are orthogonal to CLASP and could reduce the cost of its long pass, which sets its peak memory (Appendix A.10).

A.2 Method details

This section gives the thresholds, the overlap-trajectory measurement and solver details of the main-text Method.

Chunk pass. For VGGT and VGGT-Ω, local points are obtained by unprojecting predicted depth with predicted intrinsics. π^3 predicts them directly. The initial overlap alignment fits the Umeyama similarity to corresponding pixels of overlapping world point maps, sampled at spatial stride four, and performs two refits, each retaining the 80% smallest residuals. If several chunks cover a link pair, we use the chunk whose center is closest to the pair's midpoint. The initial alignment sets only the translation units, and the graph estimates rotations and positions from the relative measurements. Overlapping chunks predict the same adjacent motion twice. For rotation and translation separately, the median disagreement between these predictions, divided by $\sqrt{2}$ and bounded by a numerical floor, gives the reference noise σ_1 used in the graph weights.

Overlap-trajectory measurement. Table 5 summarizes the two row types of the chunk-scale solve. An overlap-trajectory row $x_b - x_a = \log(E_a/E_b)$ compares the camera motion shared by adjacent chunks, where E_a and E_b are the root-mean-square radii of the shared camera centers in the initially aligned units. We discard the row if either extent is below three median chunk step lengths or a similarity fit between the overlap trajectories leaves a normalized residual above 0.5.

Scale solve. A depth row requires at least 100 valid pixels per frame and three valid frames per chunk. The scale solve uses eight Cauchy reweighting iterations, a nominal log-scale uncertainty of 0.04 and ridge $\lambda_s = 10^{-4}$. Depending on the row type, the uncertainty is multiplied by the normalized overlap-fit residual relative to 0.05 or the depth log-ratio spread relative to 0.05. These factors are floored at one, reducing the influence of less consistent measurements. The auxiliary long-pass variable x_L is discarded after solving. Besides the link translations and the output geometry, the correction $\exp(x_c)$ also rescales the point maps used for loop verification.

Loop closure. The retrieval step proposes five neighbors per frame at least 100 frames away. A loop requires at least 25 geometric inliers, a fit residual at most ten times the median adjacent-fit residual, rotation disagreement below 20° with the reference and translation magnitude below four times the median positive depth. When the rotation from an essential-matrix estimate agrees with the fitted rotation within 3°, we use its translation direction and retain the fitted magnitude. Verification

reuses the existing point maps, and loops form a separate robust class. When a long-pass span of at least 60 frames is admitted and at least 20 loops exist, the disagreement between the loops and the long pass sets the loop translation noise.

Robust pose recovery. In the class weights, the adjacent-link class uses σ_1 directly, and classes with fewer than 20 edges use the pooled spread. The noise calibration of the main text sets ν_k for admitted long-pass edges, and the loop-noise rule above sets it for loop translations when applicable. The rotation-prior weight α_i has base value 0.002, multiplied by a Cauchy factor with angular scale 5°. The position solve uses five reweighting iterations, each a sparse graph-Laplacian system.

Geometry placement. For in-chunk pairs at gaps $\{5, 10, 20, 40\}$, let $\ell_{ij} = \widetilde{s}_c \|p_j^c - p_i^c\|_2$ be the chunk displacement in corrected-scale units. The global geometry factor is

$$g = \exp\left[\mathrm{wmedian}_{c,(i,j)\subset\mathcal{C}_c}\left(\log \frac{\|p_j - p_i\|_2}{\ell_{ij}}; \ell_{ij}\right)\right], \tag{10}$$

where wmedian denotes the weighted median. We use positive displacements and weight each log ratio by ℓ_{ij}, reducing the influence of nearly stationary pairs. If no pair is valid, we set $g = 1$.

A.3 Evaluation sets

The comparison tables use 51 sequences from six datasets. KITTI Odometry, Virtual KITTI and 7-Scenes are taken from the evaluation data released with SpatialBench (Peng et al., 2026), whereas KITTI-360, 4Seasons and TartanAir are downloaded from their official releases. Camera poses are evaluated on all six datasets, and depth and point clouds on TartanAir, Virtual KITTI and 7-Scenes, which provide dense ground-truth depth. Path lengths below sum the distances between consecutive ground-truth camera centers.

KITTI Odometry (Geiger et al., 2012) is a real-world benchmark of urban, residential and highway driving recorded in Karlsruhe, with ground-truth poses from a GPS/IMU system. We evaluate all 11 sequences with public ground truth (00–10).

KITTI-360 (Liao et al., 2022) records 73.7 km of suburban driving in Karlsruhe with geo-localized vehicle poses. We use its left perspective camera and split each drive into consecutive sequences of at most 1,000 frames, subsampled to about every third camera frame. We evaluate 12 sequences, where each sequence contains 630–1,000 frames and spans 1.67–2.96 km.

4Seasons (Wenzel et al., 2020) is a cross-season driving dataset with more than 350 km of stereo recordings in nine environments, from a multi-level parking garage and urban streets to countryside roads and highways, captured under different seasons, weather and illumination. As for KITTI-360, recordings are split into consecutive sequences of at most 1,000 frames. We evaluate 10 sequences from seven recordings in six environments: city loop, countryside, neighborhood, office loop, old town and parking garage. These recordings cover spring and winter, sunny, cloudy and snowy weather, and morning and afternoon capture times. Sequences contain 232–1,000 frames and span 0.54–3.52 km.

TartanAir (Wang et al., 2020) is a synthetic visual SLAM benchmark rendered with AirSim in photo-realistic environments that vary in lighting, weather and scene content, and its trajectories include diverse camera motion. It provides ground-truth camera poses and dense depth. We evaluate six trajectories from an abandoned factory, a Japanese alley, an office, a seaside town and a forest rendered in two seasons. They contain 301–612 frames at 640×480 and span 22–72 m.

Virtual KITTI (Cabon et al., 2020) recreates five KITTI driving sequences in the Unity game engine and renders each under variations in weather, time of day and camera orientation, with ground-truth depth and camera poses. We use the left camera and one variant per scene: the clone variant, which reproduces the original conditions, for Scenes 01 and 02, fog for Scene 06, rain for Scene 18 and sunset for Scene 20. The sequences contain 233–837 frames and span 52–711 m.

7-Scenes (Shotton et al., 2013) contains handheld RGB-D sequences of seven indoor scenes (chess, fire, heads, office, pumpkin, red kitchen and stairs), recorded with a Kinect at 640×480 and registered with KinectFusion to obtain ground-truth poses. We evaluate one sequence per scene. Each contains 1,000 frames, except stairs with 500, and the camera travels 4.4–11.9 m. Our single-pass evaluation harness limits a forward pass to 900000 patch tokens, which admits 868 frames at

518×392. On the six 1000-frame sequences, it therefore keeps one frame from each of 868 equal consecutive temporal bins, choosing the frame most similar to the rest of its bin, and every comparison method receives these frames. Stairs is used whole. The ablations instead score CLASP on all frames. For the long pass and full CLASP, the resulting 7-Scenes means agree at the displayed precision except RTA, which differs by 0.01°, and ATE differs by less than 1 mm.

A.4 Metric definitions and scoring details

Camera poses. Relative rotation and translation-direction errors are evaluated over all frame pairs in a sequence. A pair contributes to the accuracy curve at threshold τ when both angular errors are below τ. AUC@30 averages this curve in one-degree bins up to 30°. RRA and RTA average the two angular errors separately. ATE is the root-mean-square camera-center error after a single Umeyama Sim(3) alignment (Umeyama, 1991). These metrics measure different aspects of the trajectory: the pairwise angular metrics are scale invariant, whereas ATE measures camera-position accuracy after one global similarity alignment and thus also reflects scale drift.

Depth. Valid pixels have finite ground-truth depth in $[10^{-3}, 100]$ metres and satisfy the dataset's depth, sky and available aliasing masks. Ground truth is transformed by each model's input resize and crop before scoring. For example, a 640×480 image yields 518×392 on the VGGT grid, 592×448 on the VGGT-Ω grid, 512×384 for CUT3R/TTT3R and 518×378 for Scal3R. Cross-backbone depth comparisons therefore include differences in the pixels seen by each model.

One median scale s is fitted per sequence. On valid pixels, AbsRel averages $|s\widehat{Z} - Z|/Z$, and δ_1 is the fraction satisfying $\max(s\widehat{Z}/Z, Z/(s\widehat{Z})) < 1.25$. A single sequence scale preserves errors in relative depth scale across frames and an independent fit to every frame would remove part of the inconsistency that CLASP is designed to correct.

Point clouds. Reference clouds are formed from ground-truth depth, intrinsics and poses on the evaluated frames. Predicted clouds are transformed with the same trajectory-based Sim(3) fit used for ATE. After voxel downsampling, CD is the mean of the two directional nearest-neighbor distances, in metres. Precision and recall count points within a distance threshold of the other cloud, and F1 is their harmonic mean. The voxel size and F1 radius are $(0.1, 0.25)$ m for TartanAir, $(0.2, 0.5)$ m for Virtual KITTI and $(0.0078125, 0.05)$ m for 7-Scenes. CD should be compared within each dataset because the scene scale and voxel resolution differ. The reference covers the observed surfaces, rather than a complete scene mesh.

A.5 Implementation and comparison configurations

All methods are evaluated on an NVIDIA H100 GPU with 80 GB of memory. CLASP uses the same frozen weights for its chunk and long passes, with 60-frame chunks overlapping by 30 frames. The reported results re-solve the scale and pose graphs offline from cached inputs, namely the predictions of both passes and the loop closures verified in an earlier run of the full pipeline. The loop closures are therefore held fixed rather than re-verified after the scale solve. Chunk links and output geometry are rescaled with the scale solution defined in the Method section. On 4Seasons with π^3, this re-solve and a fresh end-to-end run differ by less than 0.001 in AUC@30 and 0.03 m in ATE. The long pass ran on the complete sequence for every evaluation sequence and backbone, up to 1591 frames, so no reported result uses the chunk-and-loop fallback of Appendix A.11. For VGGT and VGGT-Ω, it keeps the aggregator layers read by the prediction heads in a memory-mapped host cache and runs the heads over chunks of frames, which reproduces the standard forward pass within 80 GB. The base-model rows of the pose tables are this long pass. The geometry tables instead take their base-model rows from the standard single pass of the benchmark harness, in bf16, which with VGGT-Ω covers 95% of the frames on one of the 18 geometry sequences.

The TTT3R comparison resets its recurrent state every 100 frames, and the accuracy tables use LoGeR with 64-frame windows. Scal3R refers to the test-time-training model (Xie et al., 2026). TALO uses its released defaults: 12-frame submaps, 50% overlap, pose-graph recovery followed by thin-plate-spline refinement. LASER uses its released evaluation defaults: 20-frame windows overlapping by 5 frames, layer-wise depth-scale alignment and no loop closure. Its release supports only π^3 so for VGGT and VGGT-Ω, we run the same pipeline on each backbone's depth maps, back-

Table 6: Camera pose estimation on driving sequences for all three backbones. Best results are **bold** and second-best results are underlined within each backbone block, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	KITTI Odometry				KITTI-360				4Seasons			
	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓
Streaming methods												
CUT3R	0.058	67.84	72.61	207.29	0.020	74.73	77.62	184.35	0.064	56.14	80.59	432.81
TTT3R	0.392	28.92	24.77	83.39	0.109	49.21	46.42	109.86	0.299	39.40	35.82	155.87
LoGeR	0.628	10.57	8.80	39.49	0.448	17.75	14.43	52.88	0.465	15.94	17.06	166.17
Scal3R	0.883	3.08	3.15	14.35	0.421	25.21	16.41	61.88	0.735	6.70	7.53	53.68
VGGT *(Base model)*	0.319	10.73	53.24	188.34	0.115	13.75	73.48	183.63	0.158	22.91	81.22	480.92
FastVGGT	0.327	9.11	52.52	185.13	0.116	**12.92**	72.29	181.68	0.160	21.15	79.29	461.83
VGGT-Long	**0.703**	10.22	**7.19**	37.22	0.351	29.99	22.20	118.28	0.541	18.10	15.44	93.66
LASER	0.580	10.63	18.52	126.00	0.328	19.96	27.45	141.38	0.451	27.79	17.14	246.43
MERG3R	0.212	57.07	64.42	211.12	0.031	91.69	83.14	205.80	0.098	69.07	78.63	512.27
TALO	0.640	10.50	9.24	40.48	0.393	20.12	17.12	64.07	0.561	15.26	14.58	**51.08**
CLASP (Ours)	0.682	**8.90**	8.75	**35.07**	**0.462**	15.05	**15.21**	**56.08**	**0.674**	**13.64**	**13.25**	52.94
π^3 *(Base model)*	0.439	6.60	39.44	152.72	0.168	13.70	55.17	164.63	0.310	14.76	48.39	409.19
FastVGGT	0.433	7.17	41.44	154.40	0.140	15.33	59.48	169.61	0.315	14.10	49.33	416.07
VGGT-Long	0.841	4.14	4.21	26.48	0.617	15.90	11.00	51.42	0.595	17.21	13.74	174.11
LASER	0.781	4.23	7.38	45.76	0.588	10.49	11.77	48.96	0.534	13.16	11.72	143.72
MERG3R	0.323	40.85	49.64	174.15	0.078	50.59	66.71	191.78	0.190	33.60	63.02	463.19
TALO	0.790	4.42	6.22	30.41	0.642	8.86	9.81	33.11	0.562	11.07	11.19	135.73
CLASP (Ours)	**0.860**	**3.36**	**4.04**	**18.20**	**0.721**	**7.61**	**6.69**	**17.13**	**0.636**	**9.43**	**10.81**	**93.51**
VGGT-Ω *(Base model)*	0.474	5.87	35.91	138.96	0.206	10.07	51.28	159.43	0.368	11.28	43.26	357.30
FastVGGT	—	—	—	—	—	—	—	—	—	—	—	—
VGGT-Long	**0.832**	4.67	**4.81**	23.47	0.614	10.10	9.83	38.18	0.795	4.79	5.71	31.60
LASER	0.713	6.03	9.77	51.06	0.583	9.14	13.33	64.70	0.775	5.87	5.20	71.08
MERG3R	0.317	51.66	50.99	189.85	0.049	88.19	73.03	202.97	0.108	81.65	63.99	440.73
TALO	0.748	5.81	7.25	24.90	0.601	9.09	11.74	35.73	0.764	5.24	6.48	37.13
CLASP (Ours)	0.820	**4.25**	4.96	**20.97**	**0.686**	**8.02**	**7.61**	**20.26**	**0.808**	**4.55**	**5.01**	**24.98**

projected with its predicted intrinsics. All VGGT-Long results, including the teaser, the qualitative figures and the runtime comparison, use the released implementation at its default configuration, with its own chunk layout, loop closure and global chunk-graph optimization.

Ablation configuration. The pass ablation uses fixed π^3 predictions. Its four variants are 1) the long pass alone, 2) the chunk graph with long-pass scale measurements and admitted long-range translations removed, 3) the graph with scale measurements restored and 4) full CLASP with admitted translations restored. On sequences with no admitted long-pass span, the last two variants coincide. The quantitative pass and scale-row ablations average all seven 7-Scenes and all 10 4Seasons sequences.

A.6 Comparisons across frozen backbones

Tables 6, 7 and 8 extend the main-text comparisons from VGGT-Ω to VGGT and π^3, using the same sequences, metrics and streaming-method rows, and add LASER (Ding et al., 2026) to each backbone block.

Driving trajectories. The trends of the main text hold with the other two backbones. Each single pass loses global consistency over the long driving sequences. CLASP gives the highest AUC@30 among training-free methods on KITTI-360 and 4Seasons with every backbone, and the lowest ATE in all dataset–backbone combinations except 4Seasons with VGGT, where TALO is slightly lower. On KITTI Odometry, VGGT-Long has a slightly higher AUC@30 with VGGT and VGGT-Ω, while CLASP still gives the lowest ATE. Among the streaming methods, Scal3R remains the strongest on KITTI Odometry and also has a higher AUC@30 than CLASP on 4Seasons with VGGT and π^3.

Table 7: Camera pose estimation on synthetic and indoor sequences for all three backbones. Best results are **bold** and second-best results are underlined within each backbone block, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	TartanAir				Virtual KITTI				7-Scenes			
	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓	AUC@30↑	RRA°↓	RTA°↓	ATE (m)↓
Streaming methods												
CUT3R	0.262	16.49	46.85	5.37	0.312	9.50	57.41	56.88	0.112	27.70	55.70	0.46
TTT3R	0.789	4.14	6.44	0.57	0.726	4.09	13.25	4.17	0.546	5.96	18.31	0.16
LoGeR	0.928	1.17	2.46	0.26	0.827	2.24	8.85	2.35	0.758	4.13	8.23	0.07
Scal3R	0.969	0.70	1.32	0.09	0.904	0.49	5.28	0.63	0.680	4.48	11.37	0.09
VGGT *(Base model)*	0.821	2.50	6.74	0.61	0.821	1.56	9.68	7.78	0.696	9.34	13.87	0.11
FastVGGT	0.820	2.27	6.30	0.57	0.796	1.64	10.53	6.74	0.717	5.36	11.88	0.10
VGGT-Long	0.950	0.88	1.89	0.16	0.887	0.79	7.17	1.37	0.667	4.69	12.07	0.10
LASER	0.960	0.90	1.45	0.53	0.899	**0.75**	5.28	4.32	0.736	4.27	9.19	0.08
MERG3R	0.802	4.23	9.13	1.00	0.793	1.83	12.19	14.94	**0.780**	**3.55**	**8.05**	**0.07**
TALO	0.859	**0.84**	5.85	1.56	**0.905**	**0.75**	4.81	3.46	0.709	4.27	10.39	0.10
CLASP (Ours)	**0.969**	0.94	**1.20**	**0.11**	0.904	0.95	**4.25**	**1.26**	0.693	7.68	12.34	0.11
π^3 *(Base model)*	0.951	0.83	1.96	0.15	0.850	1.08	9.12	6.01	0.785	3.56	7.56	0.06
FastVGGT	0.930	1.03	2.58	0.18	0.832	1.35	10.84	7.14	**0.790**	**3.32**	7.47	**0.06**
VGGT-Long	0.954	0.72	1.82	0.16	0.827	1.49	10.19	5.18	0.680	4.30	12.31	0.09
LASER	0.943	0.88	2.10	0.65	0.823	1.85	12.73	7.82	0.722	4.62	9.50	0.08
MERG3R	0.844	5.20	9.01	1.15	0.772	1.72	16.46	17.21	0.780	3.62	7.86	**0.06**
TALO	0.892	0.81	4.25	1.07	0.842	1.63	11.61	3.48	0.727	4.62	9.21	0.08
CLASP (Ours)	**0.978**	**0.60**	**0.97**	**0.08**	**0.874**	**0.71**	**9.43**	**2.53**	0.787	3.61	**7.41**	**0.06**
VGGT-Ω *(Base model)*	0.918	1.23	3.04	0.21	0.891	0.47	6.38	1.32	0.854	2.52	4.94	0.03
FastVGGT	—	—	—	—	0.884	0.50	6.94	0.66	—	—	—	—
VGGT-Long	0.949	1.04	1.82	0.16	0.889	0.48	10.29	0.71	0.764	2.83	8.78	0.06
LASER	0.961	1.03	1.42	0.25	0.894	**0.41**	7.02	1.69	0.811	2.74	6.33	0.05
MERG3R	0.863	1.35	6.20	0.70	0.862	0.84	6.59	3.76	0.819	2.91	6.17	0.04
TALO	0.865	0.98	5.81	1.43	0.887	0.42	7.26	6.01	0.789	2.73	7.18	0.08
CLASP (Ours)	**0.973**	**0.79**	**1.05**	**0.12**	**0.904**	0.42	**5.30**	**0.61**	**0.863**	**2.50**	**4.59**	**0.03**

Table 8: Point-cloud and depth accuracy on synthetic and indoor sequences for all three backbones. Best results are **bold** and second-best results are underlined within each backbone block, excluding the base model. Entries marked — are unavailable due to the GPU memory limit.

Method	TartanAir				Virtual KITTI				7-Scenes			
	CD↓	F1↑	AbsRel↓	δ_1↑	CD↓	F1↑	AbsRel↓	δ_1↑	CD↓	F1↑	AbsRel↓	δ_1↑
Streaming methods												
CUT3R	12.6357	0.0867	0.1534	0.879	6.0426	0.1636	0.0629	0.950	0.6805	0.1419	0.1072	0.894
TTT3R	6.6985	0.1980	0.1517	0.883	2.0429	0.3111	0.0611	0.952	0.2043	0.3535	0.0850	0.934
LoGeR	5.9594	0.2780	0.1092	0.919	3.1766	0.1628	0.0557	0.964	0.0786	0.5676	0.0706	0.950
Scal3R	7.3661	0.3159	0.1015	0.918	0.9710	0.6445	0.0369	0.972	0.1264	0.5162	0.0744	0.932
VGGT *(Base model)*	5.9016	0.2634	0.0998	0.916	1.7136	0.4232	0.0502	0.966	0.1434	0.5192	0.0686	0.934
FastVGGT	5.4528	0.2486	0.1010	0.915	1.6794	0.3708	0.0497	0.967	0.1160	0.5407	**0.0666**	**0.936**
VGGT-Long	5.6195	0.2979	0.0942	0.920	1.6790	0.4002	0.0496	0.971	0.1353	0.4344	0.0783	0.931
LASER	5.5850	0.2897	0.1411	0.816	1.9163	0.3914	0.0790	0.922	0.1098	0.4865	0.0826	0.931
MERG3R	**5.0065**	0.3300	0.1011	0.916	1.9988	0.3742	0.1013	0.792	**0.0952**	**0.5934**	0.0676	0.934
TALO	8.3898	0.1318	0.3907	0.460	1.8146	0.3593	0.0777	0.954	0.1387	0.4624	0.1038	0.898
CLASP (Ours)	5.5046	**0.3632**	**0.0941**	**0.921**	**1.3804**	**0.4620**	**0.0413**	**0.972**	0.1420	0.4868	0.0694	0.933
π^3 *(Base model)*	5.4525	0.3852	0.1011	0.898	1.1570	0.4924	0.0758	0.942	0.0657	0.6412	0.0580	0.950
FastVGGT	5.2510	0.3771	0.1029	0.896	1.4811	0.3896	**0.0811**	**0.937**	0.0641	0.6607	**0.0568**	**0.950**
VGGT-Long	**5.1594**	0.3778	**0.0949**	**0.910**	1.4216	0.3981	0.1261	0.847	0.0794	0.5739	0.0662	**0.950**
LASER	5.9958	0.2613	0.1342	0.839	2.1547	0.3310	0.1563	0.773	0.0866	0.5010	0.0814	0.933
MERG3R	5.6481	0.3560	0.1021	0.898	1.6487	0.4102	0.0996	0.897	**0.0576**	**0.6739**	0.0582	**0.950**
TALO	8.1898	0.1890	0.2905	0.550	1.8680	0.3567	0.1050	0.903	0.0914	0.5193	0.0723	0.949
CLASP (Ours)	5.1950	**0.3999**	0.0959	0.906	**1.3153**	**0.4391**	0.0813	0.928	0.0632	0.6229	0.0601	**0.950**
VGGT-Ω *(Base model)*	6.2027	0.3142	0.1605	0.908	2.1502	0.5261	0.0535	0.973	0.0748	0.7932	0.0484	0.960
FastVGGT	6.2969	0.3038	0.1585	0.906	2.2179	0.5235	**0.0540**	0.972	—	—	—	—
VGGT-Long	6.1443	0.3176	0.1584	**0.910**	2.0741	0.5727	0.0567	**0.973**	0.0904	0.6650	0.0553	**0.960**
LASER	6.1480	0.2942	0.1722	0.908	2.1973	0.5571	0.0677	0.972	0.0855	0.6235	0.0604	0.953
MERG3R	**5.9655**	**0.3434**	0.1621	0.907	2.4400	0.4389	0.0585	0.972	0.0782	0.7620	**0.0484**	**0.960**
TALO	9.2296	0.1506	0.4086	0.515	3.0525	0.4423	0.1445	0.838	0.1094	0.5600	0.0804	0.953
CLASP (Ours)	6.1235	0.3391	**0.1570**	**0.910**	**2.0579**	**0.5867**	0.0555	**0.973**	**0.0666**	**0.7627**	0.0511	**0.960**

Table 9: Camera predictions of π^3 versus context length on KITTI Odometry sequence 00. Adding frames to the pass degrades the predictions of the same 75 frames, from an AUC@30 of 0.929 to at most 0.245 beyond 300 frames. Each pass holds the first N frames, and the first-75 rows score the same frames in every pass. Direction error is the angular error of adjacent-frame translation directions.

Frames in the pass, N	75	150	300	500	750	1000	1250	1514
Path length (m)	156	316	638	1088	1690	2297	2922	3722
First 75 frames								
AUC@30↑	0.929	0.632	0.527	0.096	0.245	0.115	0.095	0.151
Direction error° ↓	4.7	24.1	73.9	94.1	58.0	81.3	85.2	89.3
All N frames								
AUC@30↑	0.929	0.787	0.191	0.115	0.118	0.114	0.121	0.145
ATE (m)↓	1.10	8.50	110.95	105.30	161.86	178.35	169.62	174.70

Synthetic and indoor trajectories. With every backbone, CLASP improves on the single pass on TartanAir and Virtual KITTI and stays close to it on 7-Scenes, where the single pass already captures most of the global structure. It gives the best training-free AUC@30 and ATE on TartanAir with all three backbones. The weakest case is 7-Scenes with VGGT: CLASP remains at the single-pass accuracy, whereas MERG3R is clearly more accurate. Elsewhere, CLASP is the best or a close second training-free method in AUC@30.

Depth and point clouds. Compared with VGGT-Long on the same backbone, CLASP improves F1 in every dataset–backbone combination and AbsRel in all but TartanAir with π^3. CD is less uniform: MERG3R reaches a lower CD in several combinations, and VGGT-Long in two. As in the main text, the single pass sometimes retains slightly lower AbsRel, because CLASP does not refine per-pixel depth. Consistent with the pose results, MERG3R also gives more accurate point clouds on 7-Scenes with VGGT.

A.7 Diagnostics of the two inference contexts

This section gives the π^3 diagnostics behind the premise of CLASP.

Context length. In Table 9, the same first 75 frames of KITTI Odometry sequence 00 reach AUC@30 of 0.929 when processed alone, but at most 0.245 in passes of 500 or more frames. On 15 further 75-frame windows from sequences 00, 02 and 07, AUC@30 drops from 0.80–0.97 alone to 0.03–0.23 inside the full pass, and a higher input resolution changes neither range. The loss is local: in the full passes, frames at most ten apart already have translation-direction errors of 56–82° on 8 of the 11 sequences. This is consistent with π^3's global attention, which repeats one 2D positional grid for every frame and spreads almost uniformly over frames. So the share of attention on co-visible frames falls from 23.4% at 75 frames to 2.6% at 1000.

Scale within and between chunks. The four captures of Sec. 3.2 are Virtual KITTI scene06 (fog), 7-Scenes pumpkin, ADT (Pan et al., 2023) meal (sequence 135) and NRGB-D (Azinović et al., 2022) grey_white_room. The statistics are log standard deviations of the per-frame ground-truth depth scale within chunks and of the per-chunk median scales.

Long-pass depth as a scale reference. Table 10 shows that the long-pass depth measures chunk scales accurately even where the long-pass poses fail. It compares three per-chunk scale estimates with ground truth: the chunk chain s_c, a similarity fit to the long-pass cameras and the depth ratio d_c of Eq. (2). On the three driving sequences, the long-pass poses are poor (AUC@30 of 0.095–0.291), and scales taken from them have errors of 0.242–0.631. Scales from the long-pass depth have errors of only 0.028–0.072, lower than those of both the long-pass trajectory and the chunk chain (0.067–0.298). On 7-Scenes, where the long-pass poses are accurate, the two long-pass estimates are similar. CLASP therefore takes the chunk scales from the long-pass depth rather than its poses, and keeps this reference when long-pass pose edges are rejected (Sec. 3.3).

Table 10: Per-chunk scale references with π^3. Where the long-pass poses fail, the long-pass depth still gives the most accurate chunk scales. The log-ratio spread between long-pass and chunk depth is the median pixelwise MAD per frame and the median standard deviation of ρ_i within a chunk. Scale error is the per-chunk log-scale MAD against ground truth. **Bold** marks the lowest.

			Long-pass poses		**Log-ratio spread**		**Scale error↓**		
Sequence	Frames	Chunks	AUC@30↑	ATE (m)↓	Pixels	Frames in chunk	Chunk chain	Long-pass trajectory	Long-pass depth
KITTI Odometry 07	1101	36	0.291	54.67	0.030	0.021	0.067	0.242	**0.028**
KITTI-360 0009, window 0	1000	33	0.142	240.09	0.044	0.033	0.068	0.631	**0.037**
4Seasons parking garage	582	19	0.095	19.77	0.054	0.044	0.298	0.334	**0.072**
7-Scenes office	1000	33	0.745	0.06	0.007	0.003	0.050	**0.028**	0.031

Table 11: Depth AbsRel with π^3 for one scale per sequence and one scale per frame. On Virtual KITTI and 7-Scenes, one scale per sequence costs the long pass and CLASP little, unlike VGGT-Long. So the long-pass depth keeps a common scale that CLASP transfers to the chunks. The increase from the per-frame to the per-sequence fit measures departure from a common depth scale.

	TartanAir		**Virtual KITTI**		**7-Scenes**	
	Per sequence	Per frame	Per sequence	Per frame	Per sequence	Per frame
Long pass (π^3)	0.1011	0.1242	0.0758	0.0740	0.0580	0.0563
VGGT-Long	0.0949	0.1146	0.1261	0.0855	0.0662	0.0579
CLASP	0.0959	0.1176	0.0813	0.0749	0.0601	0.0578

Table 12: Ablation of the two scale measurements in the chunk-scale solve with π^3. Scale error is the per-chunk log-scale MAD against ground truth. **Bold** marks the default solve.

	7-Scenes					**4Seasons**				
	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓	Scale err.↓	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓	Scale err.↓
All rows	**0.787**	**3.61**	**7.42**	**0.058**	**0.055**	**0.636**	**9.43**	**10.81**	**93.51**	**0.154**
w/o overlap trajectory	0.787	3.61	7.45	0.059	0.068	0.628	9.43	11.12	98.68	0.157
w/o long-pass depth	0.786	3.61	7.50	0.060	0.061	0.620	9.43	12.26	242.15	0.726

Depth scale over complete sequences. Table 11 shows that the long-pass depth keeps one scale across a whole sequence and that CLASP transfers this scale to the chunks. If all frames share a common depth scale, fitting one scale per sequence scores almost as well as fitting one per frame. For the long pass, the per-sequence fit is only 0.0018 and 0.0017 worse on Virtual KITTI and 7-Scenes, although its Virtual KITTI ATE is 2.4 times that of CLASP. VGGT-Long, which places its chunks by overlap and loop alignments, is 0.0406 and 0.0083 worse, whereas the rescaled chunks of CLASP are 0.0064 and 0.0023 worse. Taking the chunk scales from the long-pass trajectory instead of its depth raises the AbsRel of CLASP from 0.0813 to 0.1013 on Virtual KITTI and from 0.0601 to 0.0792 on 7-Scenes, consistent with Table 10. TartanAir is uninformative here because per-frame fits are unstable on its seasonsforest scene.

A.8 Terms of the chunk-scale solve

Table 12 removes each of the two measurement types from the chunk-scale solve. The long-pass depth rows account for most of the scale recovery. On 4Seasons, removing them raises ATE from 93.51 to 242.15 m and the scale error from 0.154 to 0.726, whereas removing the overlap-trajectory rows raises ATE only to 98.68 m. AUC@30 falls by at most 0.016 in either case, because the pairwise angular metrics are scale invariant and do not reflect scale recovery. On the shorter 7-Scenes captures, all variants stay within 0.002 m in ATE. Removing either term slightly increases the scale error.

Table 13: Sensitivity to the fusion admission thresholds with π^3. **Bold** marks the adopted settings $\tau = 1^\circ$ and $\tau_t = 0.3$.

Threshold	7-Scenes				4Seasons			
	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓	AUC@30↑	RRA° ↓	RTA° ↓	ATE (m)↓
Rotation threshold τ (°)								
0.5	0.771	3.64	8.10	0.066	0.636	9.43	10.81	93.51
1 (adopted)	**0.787**	**3.61**	**7.42**	**0.058**	**0.636**	**9.43**	**10.81**	**93.51**
2	0.789	3.60	7.38	0.058	0.636	9.43	10.81	93.51
4	0.789	3.60	7.38	0.057	0.635	9.43	10.81	93.51
8	0.788	3.60	7.39	0.057	0.635	9.43	10.81	93.51
16	0.788	3.60	7.39	0.057	0.635	9.43	10.81	93.51
32	0.788	3.60	7.39	0.057	0.635	9.43	10.81	93.51
90	0.788	3.60	7.39	0.057	0.635	9.43	10.81	93.51
Relative-translation threshold τ_t								
0.15	0.779	3.61	7.83	0.062	0.636	9.43	10.81	93.51
0.3 (adopted)	**0.787**	**3.61**	**7.42**	**0.058**	**0.636**	**9.43**	**10.81**	**93.51**
0.6	0.787	3.61	7.42	0.058	0.636	9.43	10.81	93.51
1.2	0.787	3.61	7.42	0.058	0.636	9.43	10.81	93.51
2.4	0.787	3.61	7.42	0.058	0.636	9.43	10.81	93.51
4.8	0.787	3.61	7.42	0.058	0.636	9.43	10.81	93.51

A.9 Fusion admission robustness

Table 13 varies the rotation threshold τ and relative-translation threshold τ_t used to admit long-pass spans. Each sweep changes one threshold from the adopted pair $(1^\circ, 0.3)$, holding the remaining configuration fixed. Both sweeps use the complete seven-sequence 7-Scenes and 10-sequence 4Seasons evaluation sets.

The results are stable over a wide range of both thresholds. On 7-Scenes, only the strictest setting of each threshold reduces accuracy, since admission then excludes useful long-pass constraints, and all looser settings perform on par with the adopted pair. On 4Seasons, the thresholds have almost no effect, consistent with the main-text ablation, where admitted translations leave the 4Seasons results unchanged.

A.10 Runtime and memory

Figure 6 reports one 964-frame KITTI-360 sequence on an NVIDIA H100 with 80 GB memory. Methods ran sequentially on the reserved GPU, with image decoding reported separately. The method times exclude decoding, ground-truth loading and scoring. Peak VRAM includes the process's allocator pool and CUDA context.

With π^3, CLASP takes 122.0 s and peaks at 24.82 GiB GPU memory, compared with 34.8 s and 7.49 GiB for VGGT-Long. The additional full-sequence inference therefore has a substantial cost. MERG3R and TALO take 120.0 and 154.2 s in this run. The VGGT-Ω single-pass bar uses the released inference, which fits only 616 of 964 frames in GPU memory, marked by the hatched time bar, and its time and memory are not measurements of the full sequence. CLASP's own VGGT-Ω long pass processes all frames with the memory-mapped cache of Appendix A.5.

The long pass is the main bottleneck for scaling CLASP to longer sequences. It sets the peak GPU memory, which is close to that of the π^3 single pass, and its global attention grows quadratically with the number of tokens, whereas the cost of the chunk passes grows linearly with sequence length. Accelerating the long pass, for example with training-free token merging as in FastVGGT (Shen et al., 2026), is an interesting direction for future work.

A.11 Scope and limitations

CLASP is an offline procedure whose full-sequence pass retains the backbone's long-context resource requirements. If that pass cannot run, we omit its measurements and use the chunk-and-loop

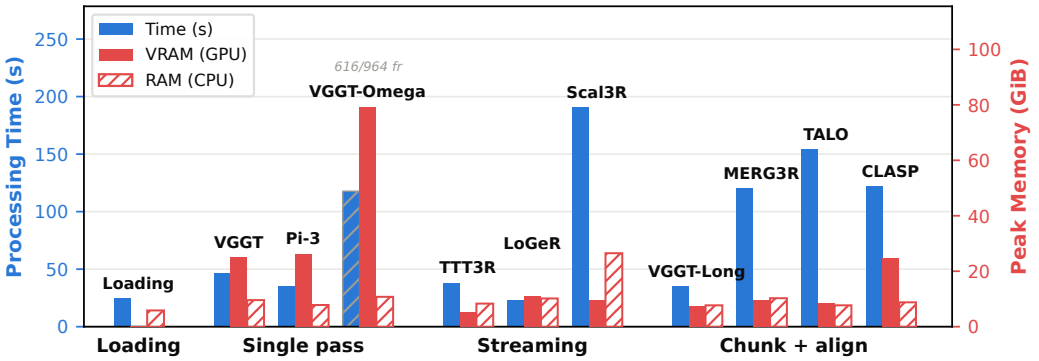

Figure 6: Processing time and peak GPU/CPU memory on a 964-frame KITTI-360 sequence (0007/w000) with one NVIDIA H100 80 GB. Chunk-and-align methods use π^3, and loading is shown separately. The hatched VGGT-Ω time bar covers only 616 frames.

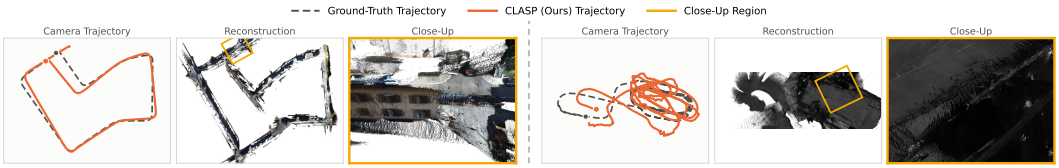

Figure 7: Failure cases with π^3. Left: on KITTI Odometry sequence 07, unreliable loop verification leaves a mismatch near the revisited segment. Right: on 4Seasons (parking_garage_1/w000), inaccurate chunk predictions leave the graph without reliable local motion.

graph, retaining relative-scale constraints from overlapping chunks and a negligible rotation reference. When long-pass pose edges are rejected, drift correction relies on the chunk links and verified loops. Repeated structures can still produce false loops. A single scale per chunk cannot capture spatially varying or within-chunk scale errors, and agreement between the passes cannot reveal biases they share.

A.12 Failure cases

Figure 7 shows two limits of the measurement separation. On KITTI Odometry sequence 07, unreliable loop verification may leave a mismatch near the revisited segment. On the 4Seasons parking-garage sequence, the chunk predictions themselves provide inaccurate local motion, and neither scale recovery nor the admitted-pose test supplies the missing constraints. A common scale reference cannot repair arbitrary local geometry or motion errors. These cases also illustrate why agreement between two predictions is a consistency test rather than an independent correctness guarantee.

A.13 Additional visual comparisons

Figures 8–11 extend the trajectory, reconstruction and close-up comparisons to additional sequences and all three frozen backbones. Gray dashed curves show ground truth, blue curves show comparison methods and orange curves show CLASP. The boxed regions identify the close-ups. These views illustrate how errors in trajectory and scale displace detailed local geometry, including duplicated structures around revisited streets. Figures 12 and 13 show additional trajectories from Virtual KITTI, 4Seasons, TartanAir and 7-Scenes. For the training-free comparison, each row uses the same backbone for all methods.



Figure 8: Additional comparisons with streaming methods, with CLASP using π^3. Top to bottom: KITTI-360 (0000/w001, 0009/w002) and KITTI Odometry sequence 08.



Figure 9: Additional training-free comparisons with π^3. Top to bottom: KITTI-360 (0009/w002, 0006/w002, 0000/w001).

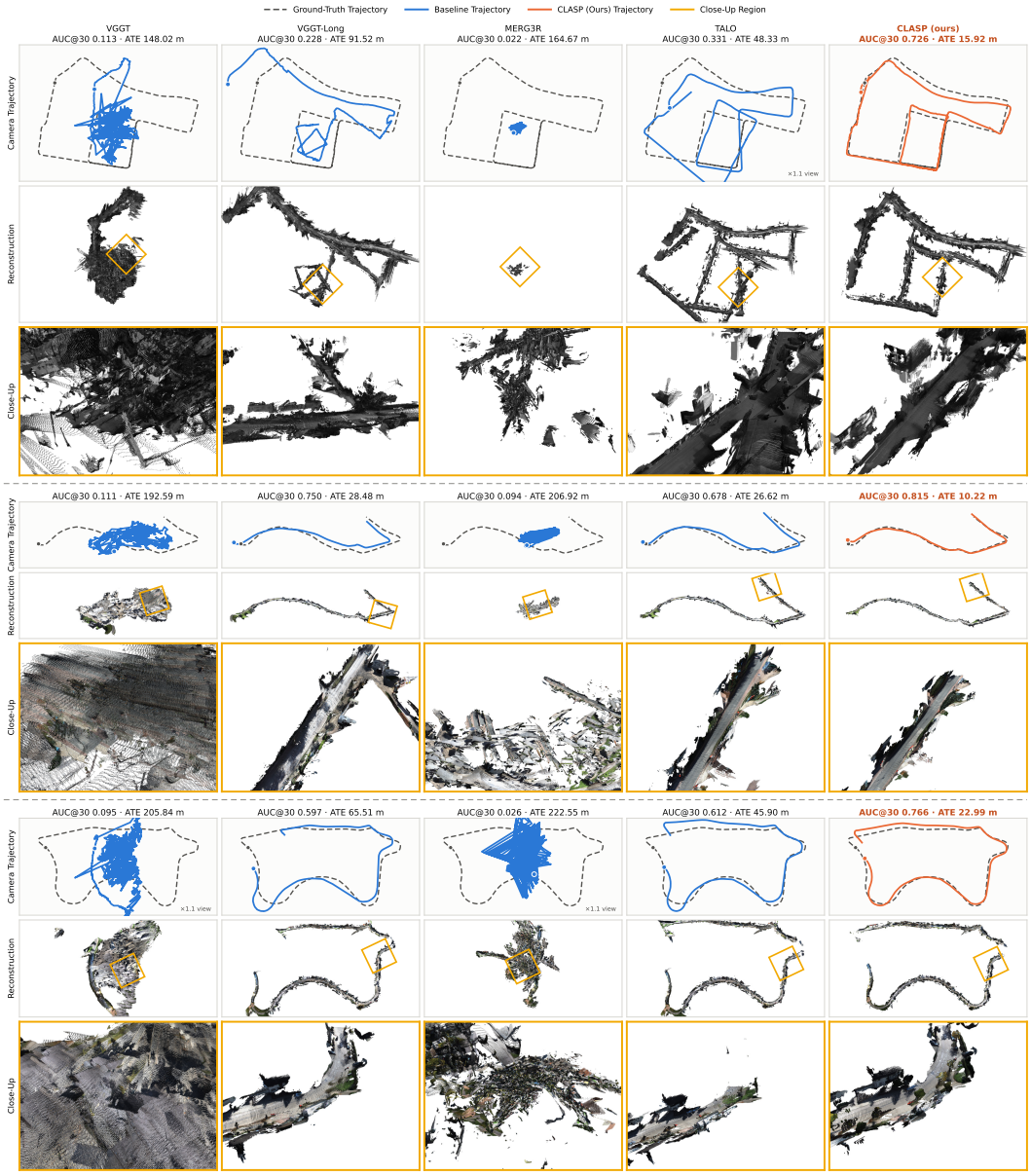

Figure 10: Additional training-free comparisons with VGGT. Top to bottom: 4Seasons (neighborhood_1/w000) and KITTI Odometry sequences 10 and 09.

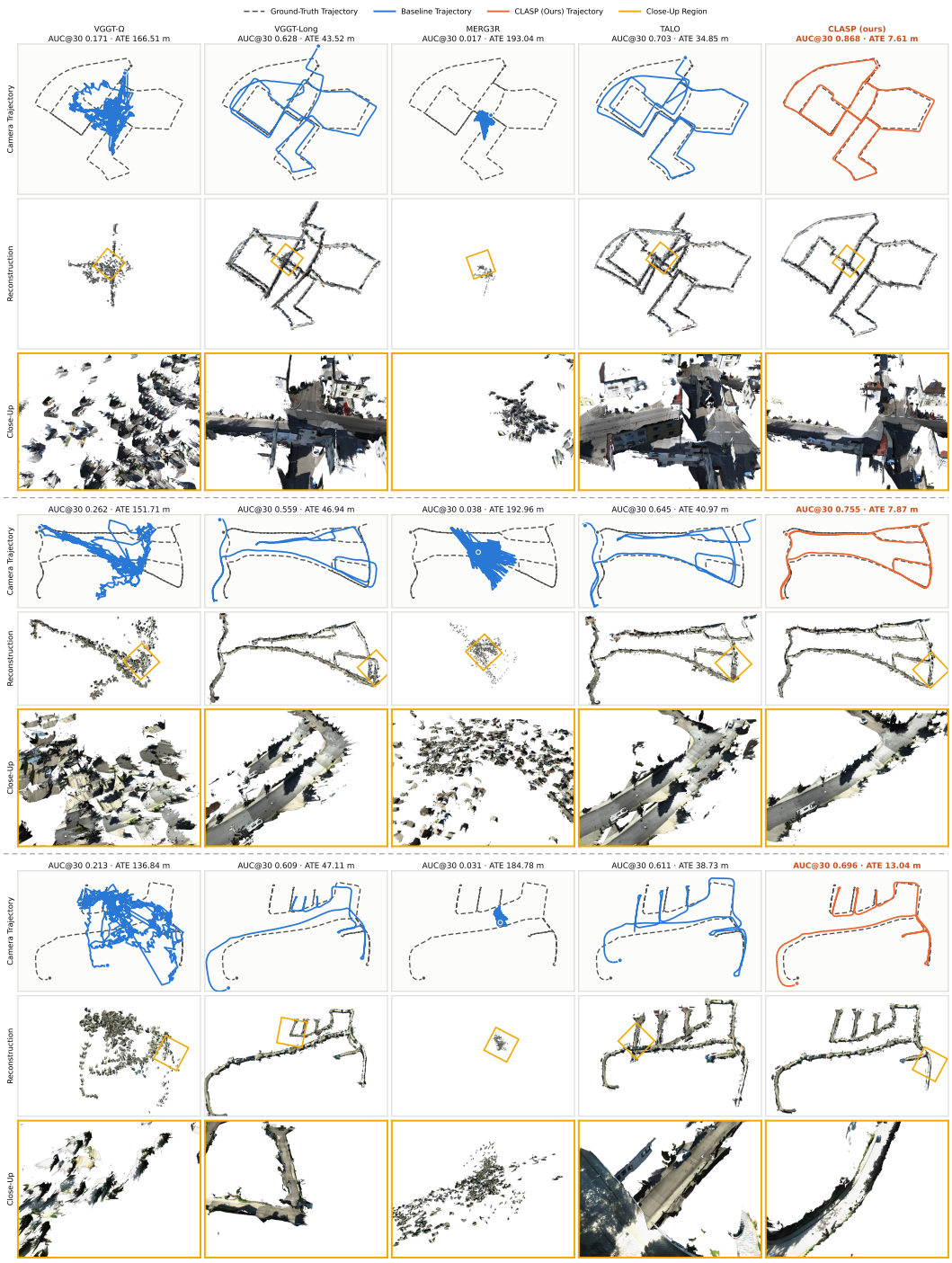

Figure 11: Additional training-free comparisons with VGGT-Ω. Top to bottom: KITTI Odometry sequence 00 and KITTI-360 (0006/w000, 0002/w001).

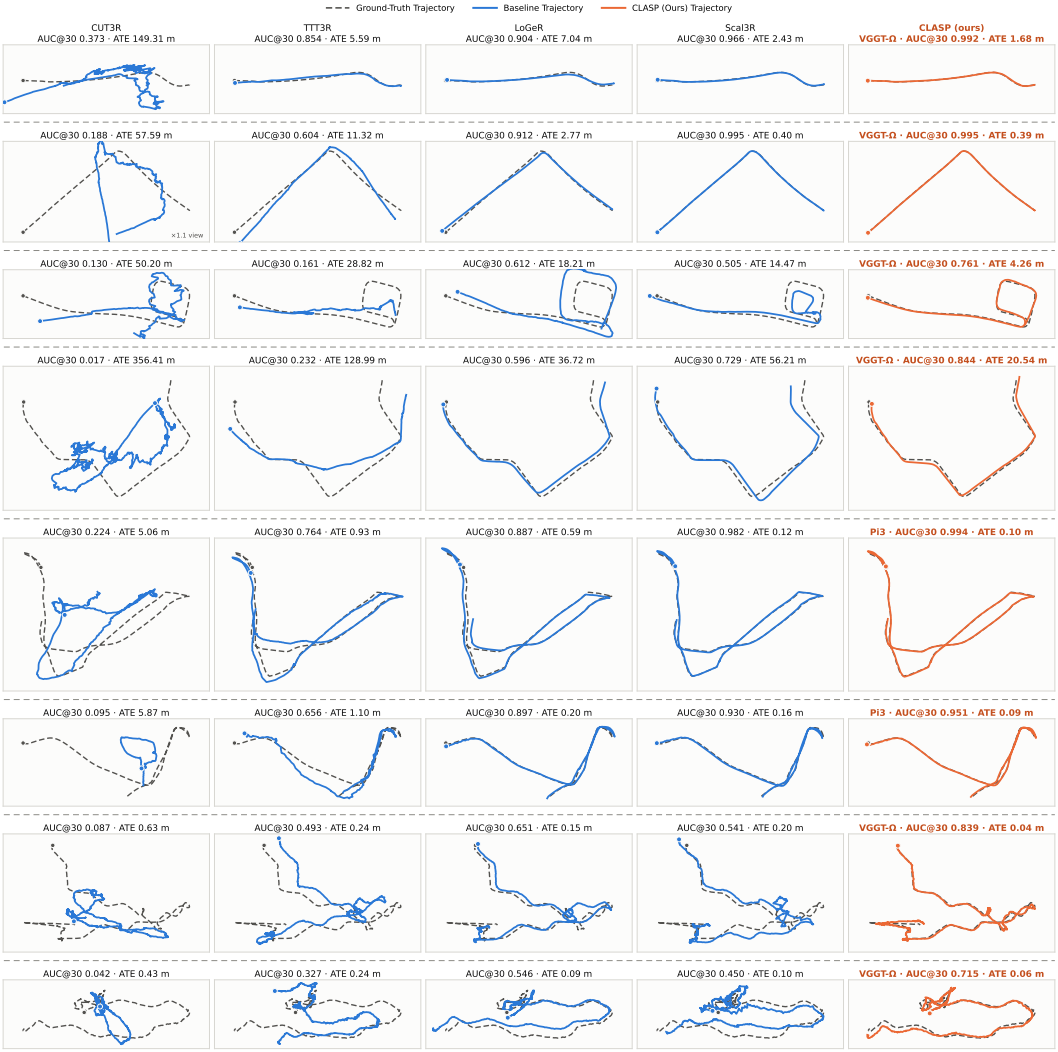

Figure 12: Trajectories compared with streaming methods. Row pairs: Virtual KITTI (scene20/sunset, scene01/clone), 4Seasons (countryside_1/w002, old_town_3/w001), TartanAir (japanesealley/P007, seasidetown/P003) and 7-Scenes (pumpkin/seq-01, redkitchen/seq-05).

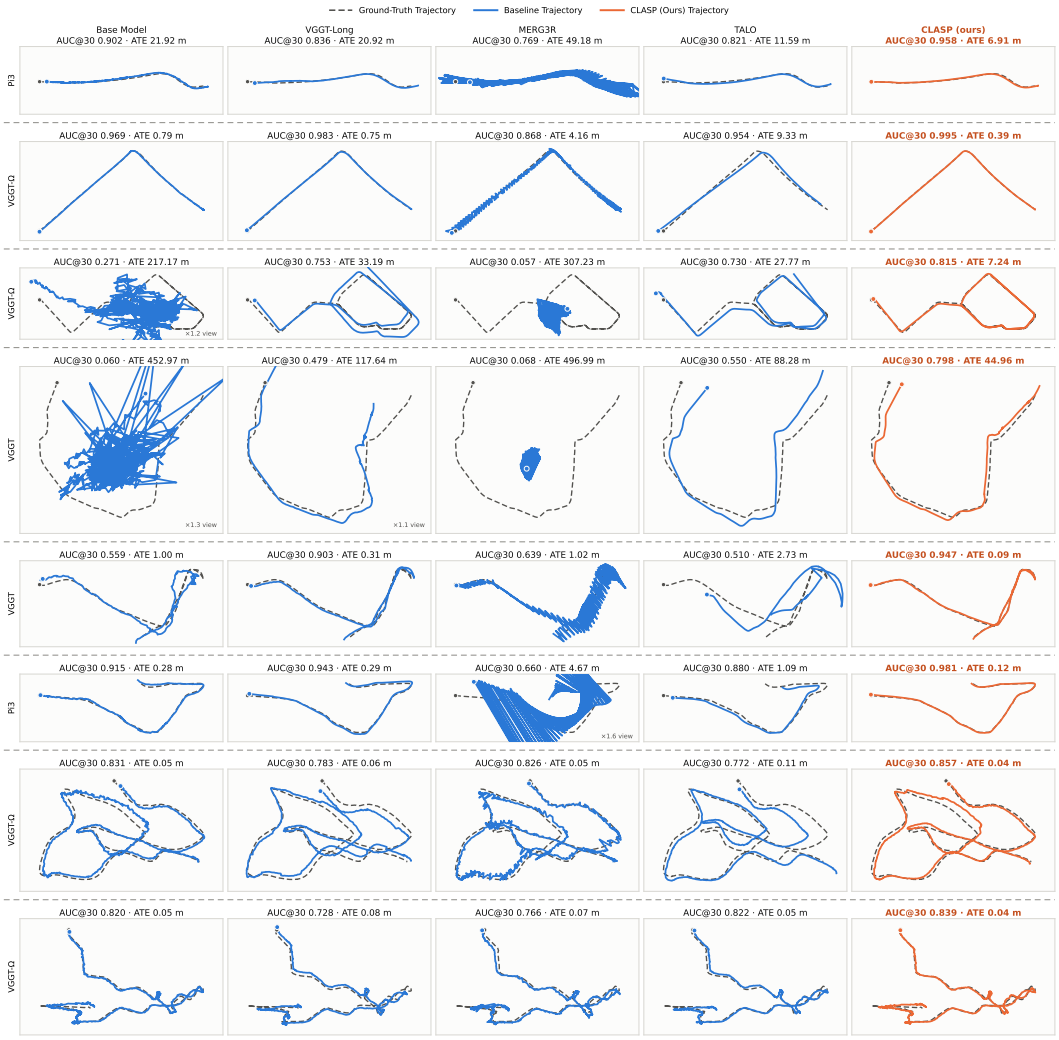

Figure 13: Trajectories compared with training-free methods, all using the backbone labeled at left. Row pairs: Virtual KITTI (scene20/sunset, scene01/clone), 4Seasons (office_loop_1/w000, old_town_3/w000), TartanAir (seasidetown/P003, office2/P003) and 7-Scenes (office/seq-04, pumpkin/seq-01).