Ω-SLAM: OMEGA-NATIVE LOOP CLOSURE FOR UNCALIBRATED DYNAMIC SLAM

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Ω-SLAM, a training-free dense RGB SLAM system built on VGGT-Ω for uncalibrated monocular videos. Extending feed-forward reconstruction to long sequences requires loop constraints that remain reliable under perceptual aliasing and scene motion. We address this challenge by using VGGT-Ω's own representations for both retrieval and geometric verification, eliminating a separate place-recognition network. Pre-attention patch embeddings retrieve candidate frames, while compact scene registers provide submap context for reranking. Jointly predicted cameras and depths then support a depth-conditioned rigid reprojection test on mutual patch correspondences, checking geometric consistency beyond epipolar agreement. Verified loops are integrated as quality-weighted robust factors in an SL(4) graph, with multi-anchor calibration bridges stabilizing submap alignment and geometry-based observation weights reducing the influence of moving content. Retrieval state uses 96 KiB for 16 new frames, while a fixed set of active submaps bounds resident dense-map memory. Experiments reduce mean absolute trajectory error relative to VGGT-SLAM 2.0 by 36.5% on 7-Scenes and 47.2% across Bonn and TUM dynamic sequences, demonstrating improved trajectory accuracy without additional training or an external retrieval network.

1 INTRODUCTION

Simultaneous localization and mapping (SLAM) estimates camera motion and reconstructs a 3D scene from an image sequence (Campos et al., 2021; Hartley & Zisserman, 2004). Traditional reconstruction pipelines combine local features, geometric verification, and bundle adjustment (Schönberger & Frahm, 2016; Lowe, 2004; DeTone et al., 2018; Fischler & Bolles, 1981; Triggs et al., 1999). Feed-forward geometric foundation models now predict much of this geometry directly. DUSt3R (Wang et al., 2024) reconstructs dense point maps from image pairs, MASt3R (Leroy et al., 2024) strengthens their correspondences, and VGGT (Wang et al., 2025a) jointly predicts cameras, depths, and point clouds from multiple views. VGGT-Ω (Wang et al., 2026a) improves reconstruction accuracy and efficiency through register attention and self-supervised training. Related models extend feed-forward reconstruction to broader settings and continuous perception (Lin et al., 2026; Wang et al., 2026b; Zhang et al., 2025; Wang et al., 2025b).

Building long-horizon feed-forward SLAM from these models requires incremental processing (Chen et al., 2026; Li et al., 2026). VGGT-Ω can process over 1,000 frames on an 80 GB A100 with FlashAttention v2 at 512×336 resolution, while its remaining global attention layers retain quadratic computation in sequence length (Wang et al., 2026a). VGGT-SLAM (Maggio et al., 2025) addresses this cost by reconstructing bounded frame windows and aligning the resulting submaps through factor graph optimization on SL(4), which accounts for the projective ambiguity of uncalibrated cameras. VGGT-SLAM 2.0 (Maggio & Carlone, 2026) further reduces drift with shared calibration across overlapping frames and keyframe-level graph nodes.

Loop closure in these systems relies on an external place-recognition network to retrieve revisits (Izquierdo & Civera, 2024). Recent work also extends reconstruction to long sequences through chunked alignment, loop correction, and retrieval of historical context (Deng et al., 2026; Zou et al., 2026). Internal features and attention also provide signals for view relevance and match verification (Han et al., 2026; Zou et al., 2026; Maggio & Carlone, 2026). These signals suggest that reconstruction representations could support both retrieval and verification.

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.

For feed-forward SLAM, the challenge is to turn retrieved revisits into reliable constraints between separately reconstructed submaps. A view that improves reconstruction context does not necessarily support a valid loop constraint. Cross-frame attention depends on the reconstruction window, while perceptual aliases and moving foregrounds can produce matches that violate rigid scene geometry. Geometry-aware denoising and dynamic reconstruction address related inconsistencies (Kim et al., 2026; Li et al., 2025), yet an incorrect loop can propagate errors through global optimization. We therefore use the model's own geometric predictions to assess whether a retrieved pair supports a loop constraint and how strongly that constraint should influence global map correction.

We present Ω-SLAM, a training-free feed-forward SLAM system for dense monocular mapping that unifies retrieval and geometric verification using VGGT-Ω's outputs. Scene registers provide compact long-term memory, while pre-attention patch embeddings index frames independently of temporal context. Together, they require 96 KiB for 16 new frames. *Depth-conditioned rigid reprojection* checks candidate patch matches against predicted cameras and depths, helping reject moving foreground matches that satisfy epipolar geometry but violate rigid 3-D consistency. Verified loops enter the SL(4) graph as quality-weighted factors, with multi-anchor bridges stabilizing submap alignment and geometry-based weights suppressing inconsistent dynamic observations. Main trajectory results use this pairwise gate. An optional reliability mode tests multi-view support and context stability by rechecking the central pair without temporal neighbours.

Our contributions are as follows:

- We introduce an internal loop-closure interface for VGGT-Ω that combines compact scene memory and context-free frame retrieval with depth-conditioned geometric verification.
- We develop Ω-SLAM, a training-free SLAM system that integrates verified loops, multi-anchor alignment, and dynamic observation weighting in an SL(4) graph without an external retrieval network.
- Experiments reduce mean ATE relative to VGGT-SLAM 2.0 by 36.5% on 7-Scenes and 47.2% across Bonn and TUM dynamic sequences, with ablations examining the contribution of each component.

2 REVIEW: VGGT-Ω

VGGT-Ω (Wang et al., 2026a) is a feed-forward transformer that maps N input images $\{I_1, \dots, I_N\}$ to per-frame cameras $\mathbf{g}_i = (\mathbf{q}_i, \mathbf{t}_i, \mathbf{f}_i) \in \mathbb{R}^9$ and depth maps $D_i \in \mathbb{R}^{H \times W}$. Each image is tokenized by a DINOv3 backbone into patch tokens $z_i^F \in \mathbb{R}^{H'W' \times C}$, augmented with one camera token and sixteen *scene registers* $z_i^{\text{scene}} \in \mathbb{R}^{16 \times C}$ (Oquab et al., 2024; Siméoni et al., 2025). An alternating-attention backbone interleaves frame-wise self-attention with global cross-frame attention, after which lightweight heads decode cameras and depths using dense-prediction heads in the style of DPT (Ranftl et al., 2021).

Two properties are central to our design. First, the scene registers aggregate global information across frames. They carry useful downstream features in the pretrained model (Wang et al., 2026a). Second, the patch embeddings computed *before* cross-frame attention provide a per-image descriptor independent of the temporal context window because inter-frame information has not yet been mixed. The remaining global attention layers couple each frame to the reconstruction window and incur quadratic computation in N (Wang et al., 2026a; Vaswani et al., 2017). In the reported FlashAttention v2 setup, inference memory grows approximately linearly (Wang et al., 2026a). Bounded submaps control the cost of each forward pass as the image stream grows.

3 Ω-SLAM

Ω-SLAM connects bounded VGGT-Ω reconstructions through internal loop retrieval, geometric verification, and a keyframe factor graph (Fig. 1). We first describe submaps and their compact state (Sec. 3.1), then loop retrieval and verification (Sec. 3.2), submap alignment (Sec. 3.3), and global optimization on SL(4) (Sec. 3.4). Algorithm 1 summarizes the online pipeline, and App. E gives the supporting derivations and assumptions.

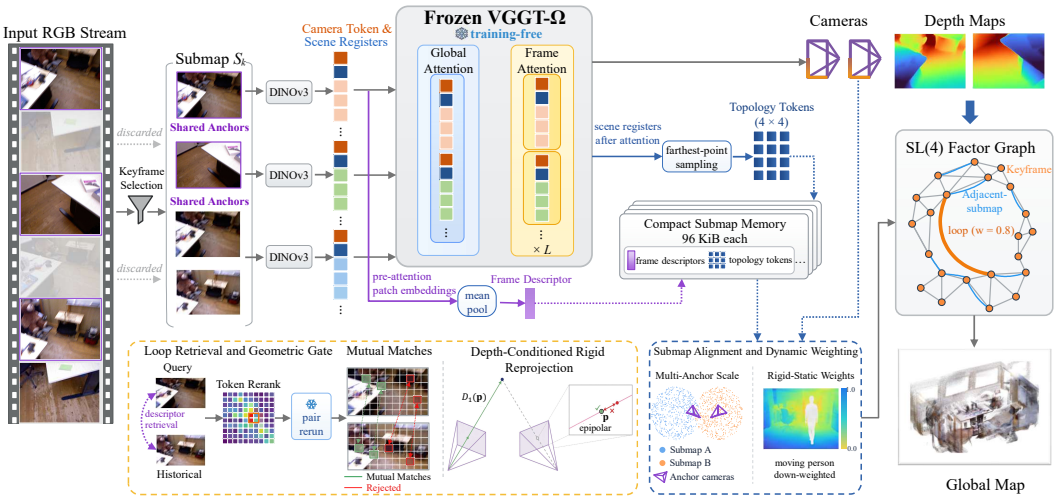

Figure 1: Overview of Ω-SLAM.

3.1 Submap Construction and Compact State

Keyframe selection. We select a keyframe when its mean sparse optical-flow disparity from the previous keyframe exceeds τ_{kf}. Removing near-duplicate frames before quadratic attention reduces redundant computation, an efficiency principle also explored by S-VGGT (Li et al., 2026). Once n keyframes have accumulated, VGGT-Ω reconstructs submap $\mathcal{S}_k$. The disparity uses the current frame and previous keyframe, so selection operates on a live stream without future frames or access to the global map.

Compact state. VGGT-Ω emits 16 scene registers per frame, so a submap of N_k frames carries $N_k \times 16$ registers of dimension $d_{\text{r}} = 2048$ in the released 1B model. After $\mathcal{S}_k$ enters the factor graph, we retain three items. The first is a set of M topology tokens $\mathbf{U}_k \in \mathbb{R}^{M \times d_{\text{r}}}$, obtained by deterministic farthest-point sampling under the cosine metric. Each step selects the register farthest from its nearest selected token, preserving angular diversity for partial-overlap comparison at fixed size. This retains distinct register responses that averaging would merge, so a revisit can be retrieved through the subset associated with the overlapping part of a scene. The second is one descriptor $\mathbf{d}_{k,i} \in \mathbb{R}^{d_{\text{f}}}$ per keyframe, formed by mean-pooling the image-local patch embeddings *before* cross-frame attention and normalizing. Pooling at this point reuses the encoder output and gives each image a descriptor independent of its temporal window. We also retain camera graph nodes, intrinsics, frame paths, and a strided set of depth samples used only for relative scale estimation. Together, they support retrieval, alignment, and dense-map regeneration.

With fp16 storage the retrieval payload of a submap occupies $2(Md_{\text{r}} + N_k d_{\text{f}})$ bytes, which for $M = N_k = 16$ and $d_{\text{f}} = 1024$ is 96 KiB, roughly $500\times$ smaller than the 48 MiB dense point map of the same window at 512^2 resolution. Including A shared anchor keys adds $2A$ KiB to the 16-frame payload. The latest K submaps form the active dense set, bounding resident dense-map memory independently of trajectory length. The compact archive and factor graph still grow with the trajectory. Source keyframes remain on disk, and dense geometry is regenerated on demand by rerunning VGGT-Ω on the stored images. Two-frame loop submaps are archived after their factors enter the graph, keeping their dense output from accumulating with the number of loops.

3.2 Loop Retrieval and Geometric Verification

Retrieval. For the current submap q and a historical submap j, excluding recent neighbours which overlap by construction, we first score the strongest agreement between their frame descriptors,

$$s_{\text{c}}(q, j) = \max_{a,b} \mathbf{d}_{q,a}^{\top} \mathbf{d}_{j,b}. \tag{1}$$

The maximizing pair $(a^\star, b^\star)$ supplies the frames for geometric verification. We rerank the top L submaps using their topology tokens. To accommodate partial overlap, we compare $\mathbf{C} = \mathbf{U}_q \mathbf{U}_j^\top$ through its best-matching subset. Writing $T_\rho(\mathbf{v})$ for the mean of the largest ρ fraction of the entries of $\mathbf{v}$, the fine score symmetrizes the row-wise and column-wise assignment quality,

$$s_{\mathrm{f}}(q, j) = \tfrac{1}{2}\Big[T_\rho\big(\max_n \mathbf{C}_{m,n}\big) + T_\rho\big(\max_m \mathbf{C}_{m,n}\big)\Big]. \tag{2}$$

Trimming accommodates non-overlapping regions, while symmetry balances coverage in both directions. Thus frame similarity locates a candidate pair and register agreement supplies its submap context. Both scores reuse features retained during reconstruction. Candidate selection therefore requires neither another image encoder nor an external place-recognition network.

Geometric verification gate. Geometric verification resolves perceptual aliases among visually similar places before graph insertion. For the top candidate we rerun VGGT-Ω on $(I_{a^\star}, I_{b^\star})$ and request final patch tokens together with the jointly predicted cameras $(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)$ and depths $\mathbf{D}_i$. Mutual nearest-neighbour patch matches above a cosine threshold form a candidate set, which we evaluate with two complementary geometric tests. Patch mutuality supplies local correspondence support (DeTone et al., 2018; Sun et al., 2021). The first is the epipolar constraint induced by the predicted cameras, measured by the Sampson error of the fundamental matrix $\mathbf{F} = \mathbf{K}_2^{-\top}[\mathbf{t}_{21}]_\times \mathbf{R}_{21}\mathbf{K}_1^{-1}$ (Hartley & Zisserman, 2004). Pure rotation and planar views create degeneracies, so we also fit a RANSAC homography and retain the better supported model (Fischler & Bolles, 1981).

The second test checks depth-conditioned rigid consistency in VGGT-Ω's joint prediction, drawing on the cross-view consistency principle of learned multi-view stereo (Yao et al., 2018; 2019). We use world-to-camera extrinsics $\mathbf{X}_i = \mathbf{R}_i \mathbf{X}_w + \mathbf{t}_i$, so that $\mathbf{R}_{21} = \mathbf{R}_2 \mathbf{R}_1^\top$ and $\mathbf{t}_{21} = \mathbf{t}_2 - \mathbf{R}_{21}\mathbf{t}_1$. Unprojecting a match $\mathbf{p}$ at depth $\mathbf{D}_1(\mathbf{p})$ and reprojecting it into the second camera gives

$$\hat{\mathbf{p}}_{1\to 2} = \pi\Big(\mathbf{K}_2\big[\mathbf{R}_{21}\mathbf{D}_1(\mathbf{p})\mathbf{K}_1^{-1}\tilde{\mathbf{p}} + \mathbf{t}_{21}\big]\Big). \tag{3}$$

The residual is the RMS of forward and backward pixel errors, using the same predicted intrinsics, motion, and depth convention in both directions. The epipolar constraint fixes a line, whereas depth-conditioned reprojection fixes the position along it. A coherently moving foreground can satisfy a fundamental matrix while violating rigid reprojection, so the joint inlier set must exceed both an absolute count and a fraction of all matches. The count requires enough correspondences to support the bridge, while the fraction measures their consistency across the candidate set. The accepted two-frame cameras and depths then provide the same bridge as an adjacent submap, inserting the verified loop through the ordinary odometry path. The test operates in the predicted scale and requires translation for depth sensitivity (App. E.2).

Multi-view consensus certificate. The main benchmarks use the two-frame gate. An optional strict reliability mode gathers a clipped temporal neighbourhood around each retrieved frame and runs joint VGGT-Ω inference. The central pair is always checked. The remaining cross-neighbour pairs are ranked by the context-free frame descriptors and checked separately with the same patch, epipolar, and depth-rigid tests. A candidate is certified only if at least m pairs pass and those pairs cover at least c distinct frames on both sides. With $m = c = 2$, the decision is

$$V_{\mathrm{mv}} = V_{a^\star b^\star}\mathbf{1}\left[\sum_{a,b} V_{ab} \geq m, \quad |\{a : \exists b,\ V_{ab} = 1\}| \geq c, \quad |\{b : \exists a,\ V_{ab} = 1\}| \geq c\right]. \tag{4}$$

The central pair defines the loop bridge. Additional multi-view support scales its quality weight, weakening factors with marginal evidence while keeping persistent retrieval memory compact. Requiring distinct frames on both sides prevents several matches to one view from satisfying the coverage test. Retrieval therefore identifies where to spend inference, and local geometric agreement determines whether that candidate can constrain the graph.

Context-intervention stability. Global attention can make a candidate's support depend on its neighbours. To test this dependence, strict mode reruns only $(I_{a^\star}, I_{b^\star})$ and compares the two predictions. Let $\mathcal{M}_{\mathrm{ctx}}$ and $\mathcal{M}_{\mathrm{pair}}$ be their mutual patch matches. The context-stability score is

$$\kappa = \frac{|\mathcal{M}_{\mathrm{ctx}} \cap \mathcal{M}_{\mathrm{pair}}|}{\min(|\mathcal{M}_{\mathrm{ctx}}|, |\mathcal{M}_{\mathrm{pair}}|)}, \qquad \Delta_r = |r_{\mathrm{ctx}} - r_{\mathrm{pair}}|, \tag{5}$$

where r is the rigid inlier ratio. We also record the change in median rigid reprojection error. Empty match sets fail the test. We accept the central pair when κ is above its validation-set operating point and both residual changes are below theirs. The pair-only prediction supplies the loop submap, keeping the inserted measurement in the camera gauge that passed verification.

Calibration protocol for strict acceptance. For a prospective calibrated operating point, the scoring rule and proposal policy must first be frozen, then evaluated on a sequence-disjoint calibration set of labelled hard negatives produced by that policy. Let a be the nonconformity score (large means less trustworthy) formed from retrieval and patch margins, geometric and rigid inlier deficits, reprojection residual, match support, and multi-view support. For n calibration negatives, strict acceptance uses

$$k_\alpha = \lfloor (n+1)\alpha \rfloor, \qquad \tau_\alpha = \begin{cases} a_{(k_\alpha)}, & k_\alpha \geq 1, \\ -\infty, & k_\alpha = 0, \end{cases} \qquad V_{\text{cal}} = \mathbf{1}[a < \tau_\alpha]. \tag{6}$$

Here $0 < \alpha < 1$ and $a_{(k)}$ is the kth smallest calibration score. If the calibration scores and a future false-proposal score are exchangeable, the rank argument underlying split conformal calibration gives $\Pr(a_{\text{false}} < \tau_\alpha) \leq \alpha$ (Angelopoulos & Bates, 2021). The strict inequality handles ties conservatively. The bound concerns per-proposal false acceptance (App. E.3). This optional calibration augments the geometric gates when a held-out split is available. The reported strict-mode results use fixed multi-view and context gates.

3.3 Submap alignment

Adjacent submaps share $A \geq 1$ anchor frames, formed by prepending the last A keyframes of $\mathcal{S}_{k-1}$ to $\mathcal{S}_k$. Shared images provide a common camera-frame reference for alignment. Expressing 3D points in that reference makes their two estimates directly comparable without an unrestricted submap-wide homography (Maggio & Carlone, 2026). Let $\mathbf{H}_j^i$ map camera-j coordinates into camera-i coordinates, and let s multiply the source reconstruction's depth to match the target. For duplicate frames, the bridge compensates for predicted calibration differences and estimates this scale. Within a window, it uses the predicted rigid relation:

$$\mathbf{H}_j^i = \mathcal{N}\left(\begin{bmatrix} \mathbf{K}_i^{-1}\mathbf{K}_j & \mathbf{0} \\ \mathbf{0}^\top & s^{-1} \end{bmatrix} \right), \qquad \mathbf{H}_j^i\big|_{\text{intra}} = \mathbf{T}_i^{-1}\mathbf{T}_j, \tag{7}$$

where $\mathcal{N}(\mathbf{H}) = \mathbf{H}/(\det \mathbf{H})^{1/4}$ for $\det \mathbf{H} > 0$, and $\mathbf{T}_i$ is a camera-to-window pose, the inverse of the world-to-camera extrinsic matrix. The s^{-1} entry ensures that dehomogenization multiplies 3-D coordinates by s. Jointly predicted poses supply the rigid intra-submap model, and bridges enforce relative agreement between duplicate estimates under the remaining global gauge (App. E.1).

Multi-anchor scale. A single anchor ties alignment to the depth quality of one frame, the failure mode noted by Maggio & Carlone (2026). We estimate a log-scale offset within each anchor and aggregate across anchors with a nested robust median,

$$\ell^\star = \operatorname*{med}_{a \in \mathcal{A}} \Big[\operatorname*{med}^{w}_{\mathbf{p} \in \mathcal{P}_a} \big(\log \|\mathbf{X}_a^{k-1}(\mathbf{p})\| - \log \|\mathbf{X}_a^{k}(\mathbf{p})\| \big) \Big], \quad s = \exp\big(\ell^\star\big), \tag{8}$$

where the inner median is weighted by the static weights defined below and taken after MAD rejection of non-finite, near-zero, and outlying samples. The outer median gives each viewpoint equal weight. The spread of per-anchor estimates ℓ_a around $\ell^\star$ provides uncertainty σ_ℓ, reflecting disagreement among depth predictions. Each anchor contributes one scale observation after its own robust filtering. An anchor with many valid pixels consequently cannot dominate the common scale solely through its sample count. The norm ratio is an exact depth-scale ratio when the calibration rays agree. Otherwise, calibration-ray differences contribute an estimation error. Robustness requires a majority of reliable anchors, so two anchors cannot tolerate one arbitrarily corrupted anchor (App. E.4). This consensus view is related to similarity synchronization with learned depth, but our median operates on shared-frame depth ratios rather than solving a separate global synchronization problem (Yu & Yang, 2024). Every usable duplicate frame contributes a factor carrying the common scale. The A estimates originate from the same two VGGT-Ω windows and are therefore correlated, so we inflate their standard deviations by $\sqrt{A}$. The single-anchor configuration supports comparison with Maggio & Carlone (2026).

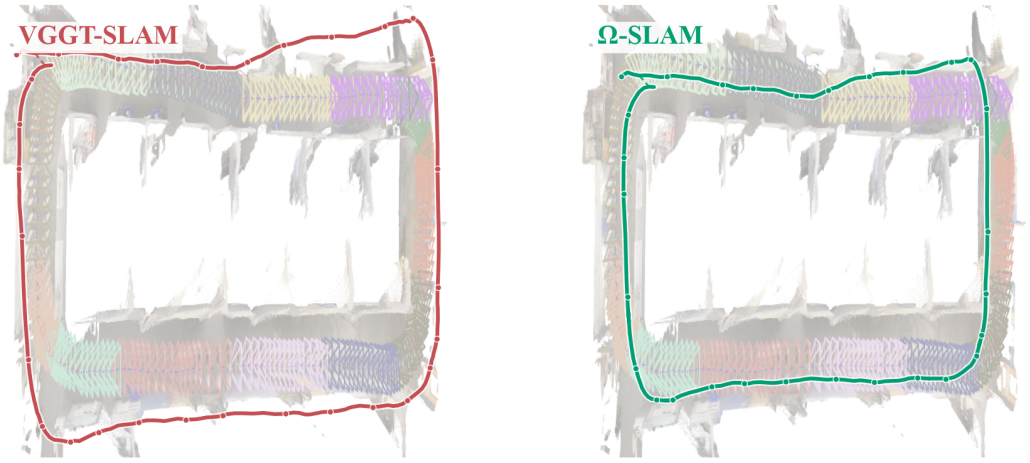

Figure 2: Office-loop reconstructions in a shared display gauge. Colored paths show estimated poses. A thin bridge marks the accepted revisit.

Rigid-static observation weights. We weight observations by their agreement with the local rigid geometry. Within each bounded window, every valid depth sample of a frame is transferred to its temporal neighbours and compared against the depth observed there. From the relative discrepancy r we form a Cauchy weight $w = 1/(1 + (r/\tau_s)^2)$ and average it over the neighbours that carry evidence. The relative form makes w invariant to the unknown scale of the submap. A sample that falls outside the common field of view, lies near a depth discontinuity, or is occluded by a nearer surface yields no evidence and keeps unit weight. Visible, inconsistent samples receive lower weights. These weights enter the inner median of Eq. (8), remove low-weight pixels from the active and exported dense maps, and, through a low mean support $\bar{w}$, inflate the noise of the adjacent-submap factor so that the backend can discount a window dominated by moving content.

3.4 Backend: factor graph optimization on SL(4)

To correct rotation and translation drift throughout each window, we place an SL(4) node $\mathbf{H}_i$ on every keyframe within the projective graph formulation (Maggio et al., 2025; Maggio & Carlone, 2026). Duplicate observations in different submaps have separate nodes linked by bridges. Three kinds of edge connect the graph, all expressed through Eq. (7). Intra-submap edges carry the rigid relations predicted within one VGGT-Ω window, adjacent-submap edges carry the calibration and scale bridge on shared anchors, and loop edges carry the same bridge on verified two-frame submaps. Given the edge set $\mathcal{E}$, the global estimate is the maximum a posteriori solution

$$\hat{\mathcal{H}} = \underset{\{\mathbf{H}_i \in \mathrm{SL}(4)\},\, \mathbf{H}_0 = \mathbf{I}_4}{\arg\min} \sum_{(i,j)\in\mathcal{E}} \rho_{ij}\left(\left\| \mathrm{Log}\big((\mathbf{H}_j^i)^{-1} \mathbf{H}_i^{-1} \mathbf{H}_j\big) \right\|_{\mathbf{\Sigma}_{ij}^{-1}} \right), \tag{9}$$

where $\mathrm{Log}(\cdot)$ maps a group element to the 15-dimensional tangent coordinates of $\mathfrak{sl}(4)$ on a local logarithm chart, $\mathbf{\Sigma}_{ij} = \sigma_{ij}^2 \mathbf{I}_{15}$, and ρ_{ij} is quadratic within windows and Huber on adjacent and loop factors. Fixing $\mathbf{H}_0$ removes the left-action gauge for a connected graph. SE(3) embeds directly in SL(4), and positive-scale Sim(3) embeds after determinant normalization, giving all edge types a common residual (Dellaert & Kaess, 2017; Triggs et al., 1999). Structured camera and scale measurements constrain alignment while the nodes retain projective freedom. We solve Eq. (9) with Levenberg-Marquardt in GTSAM (Dellaert, 2012). Figure 2 illustrates the resulting office-loop reconstruction and the bridge connecting an accepted revisit.

Uncertainty of learned measurements. The gate of Sec. 3.2 is binary, while the supporting evidence varies. Writing $m(x;\tau) = \mathrm{clip}((x-\tau)/(1-\tau), 0, 1)$ for the normalized margin for a score $x \in [0,1]$ with $\tau < 1$, we combine the register score, the patch similarity, the inlier ratio of the selected geometric model, the rigid-reprojection support, and the match count into a scalar quality $q \in [0,1]$, weighting the two geometric terms most heavily. Match counts are first mapped to a bounded support score. The loop factor deviation interpolates as

$$\sigma_{\mathrm{loop}}(q) = \sigma_{\max} - q(\sigma_{\max} - \sigma_{\min}), \qquad 0 < \sigma_{\min} < \sigma_{\max}. \tag{10}$$

Broadly supported loops receive greater weight. The binary gate decides whether a loop enters the graph, while its quality determines how strongly that loop influences the trajectory relative to other measurements. Adjacent-submap factors inherit their deviation from the estimation quantities of Sec. 3.3, growing with the anchor disagreement σ_ℓ and with the scarcity of static support $1 - \bar{w}$, and inflated by $\sqrt{A}$ as discussed above. These uncertainties enter before optimization and are shared by all graph iterations. Both families use the Huber model of Eq. (9) to bound the score in whitened residual space. Its effect on graph updates also depends on factor Jacobians and graph conditioning (App. E.5).

Algorithm 1 Ω-SLAM

Require: $\{I_t\}$; $n, A, K, L, \tau_{\mathrm{kf}}$; frozen gates; strict-mode flag $\eta \in \{0, 1\}$
Ensure: Graph $\mathcal{G}$, compact archive $\mathcal{D}$, dense set $\mathcal{R}$; gauge $\mathbf{H}_0 = \mathbf{I}_4$

$\mathcal{G}, \mathcal{D}, \mathcal{R}, \mathcal{B} \leftarrow \varnothing$; $k \leftarrow 0$
for I_t: first frame or $\Delta(I_t, I_{\text{last kf}}) > \tau_{\mathrm{kf}}$ **do**
  $\mathcal{B} \leftarrow \mathcal{B} \parallel I_t$; **if** $|\mathcal{B}| < n$ **then continue**
  $(\mathcal{S}_k, \mathbf{U}_k, \{\mathbf{d}_{k,i}\}) \leftarrow \mathcal{F}_\Omega(\mathcal{B})$; $\mathbf{w}_k \leftarrow \mathrm{StaticWeights}(\mathcal{S}_k)$
  $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{V}_k \cup \mathcal{E}_{\mathrm{intra}}(\mathcal{S}_k)$ — Eq. (7)
  if $k > 0$: $s \leftarrow e^{\ell^\star}$; $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{E}_{\mathrm{bridge}}(s, \mathbf{w}_k)$ — Eq. (8)
  $\mathcal{C} \leftarrow \mathrm{TopL}_{j \in \mathcal{D} \setminus \mathcal{D}_{\mathrm{recent}}} s_{\mathrm{c}}(k, j)$ — Eq. (1)
  if $\mathcal{C} \neq \varnothing$ **then**
    $j^\star \leftarrow \arg\max_{j \in \mathcal{C}} s_{\mathrm{f}}(k, j)$; $(a^\star, b^\star) \leftarrow \arg\max_{a,b} \mathbf{d}_{k,a}^\top \mathbf{d}_{j^\star,b}$ — Eq. (2)
    $\mathcal{S}_{\mathrm{pair}} \leftarrow \mathcal{F}_\Omega(I_{a^\star}, I_{b^\star})$; $(V, q) \leftarrow \mathrm{Verify}(\mathcal{S}_{\mathrm{pair}})$ — Eq. (3)
    if $\eta V = 1$ **then**
      $V \leftarrow V\, V_{\mathrm{mv}}\, V_{\mathrm{context}}\, V_{\mathrm{cal}}$ — Eq. (4), Eq. (5), Eq. (6)
    end if
    if $V = 1$ **then**
      $\sigma_{\mathrm{loop}} \leftarrow \sigma_{\max} - q(\sigma_{\max} - \sigma_{\min})$ — Eq. (10)
      $\mathcal{G} \leftarrow \mathcal{G} \cup \mathrm{LoopFactors}(\mathcal{S}_{\mathrm{pair}}, \sigma_{\mathrm{loop}})$ — Eq. (7)
    end if
  end if
  $\widehat{\mathcal{H}} \leftarrow \mathrm{LM}(\mathcal{G}; \mathbf{H}_0 = \mathbf{I}_4)$; $\mathcal{D} \leftarrow \mathcal{D} \cup \mathrm{Compact}(\mathcal{S}_k)$ — Eq. (9)
  $\mathcal{R} \leftarrow \mathrm{Latest}_K(\mathcal{R} \cup \{\mathcal{S}_k\})$; $\mathcal{B} \leftarrow \mathrm{Tail}_A(\mathcal{B})$; $k \leftarrow k + 1$
end for

$\mathcal{F}_\Omega$ denotes inference and compact feature extraction, and Δ is mean flow disparity. Verify returns pairwise acceptance and quality. Edge insertion includes nodes and noise weights. LoopFactors adds the pair's intra-edge and endpoint bridges. Strict mode evaluates neighbouring views and the pair-only context intervention, with $V_{\mathrm{cal}} = 1$ unless held-out calibration is available. Pair-only dense geometry is released after verification and factor insertion.

4 Experiments

4.1 Experimental setup

Datasets. We evaluate all 18 7-Scenes (Shotton et al., 2013), five Bonn Dynamic (Palazzolo et al., 2019), and five TUM-Dynamic (Sturm et al., 2012) test sequences. Each system processes the ordered RGB stream at stride two with its released frontend. We report ATE, RPE, pose coverage, and throughput.

Baseline taxonomy. The pose tables compare 13 configurations: ORB-SLAM3 (Campos et al., 2021), DeepV2D (Teed & Deng, 2020), DeepFactors (Czarnowski et al., 2020), DPV-SLAM and DPV-SLAM++ (Lipson et al., 2024), GO-SLAM (Zhang et al., 2023), DROID-SLAM (Teed & Deng, 2021) with known intrinsics or RGB-only GeoCalib (Veicht et al., 2024), calibrated and native-uncalibrated MASt3R-SLAM (Murai et al., 2025), VGGT-SLAM (Maggio et al., 2025), VGGT-SLAM 2.0 (Maggio & Carlone, 2026), and Ω-SLAM. Each configuration is locally rerun under the same protocol.

Camera protocol. We fit one Umeyama Sim(3) per sequence, retaining the last pose for repeated frame IDs. ATE is the aligned camera-center RMSE. RPE measures relative translation and rotation between consecutive exported poses.

Table 1: RMSE ATE (m) on 18 7-Scenes sequences, with scene and overall macro means. Bold and underline mark the best overall and uncalibrated values.

	Method	Venue	N	Chess	Fire	Heads	Office	Pumpkin	Red Kit.	Stairs	Avg.
Calib.	ORB-SLAM3	T-RO'21	18	0.0377	0.0344	0.0778	0.1290	0.0971	0.0658	0.0690	0.0777
	DeepV2D	ICLR'20	18	0.2287	0.1801	0.1460	0.2134	0.2080	0.2021	0.1139	0.1929
	DeepFactors	RA-L'20	18	0.5086	0.2024	0.0862	0.2362	0.3737	0.3448	0.1112	0.2859
	DPV-SLAM	ECCV'24	18	0.0354	0.0296	0.0297	0.1286	0.3375	0.0699	0.0483	0.0997
	DPV-SLAM++	ECCV'24	18	0.0358	0.0293	0.0297	0.1243	0.2993	0.0694	0.0554	0.0952
	GO-SLAM	ICCV'23	18	0.0341	0.0308	0.0259	0.0936	0.1073	0.0868	0.0587	0.0720
	DROID-SLAM	NeurIPS'21	18	0.0346	0.0266	0.0224	0.1048	0.1046	0.0636	0.0689	0.0683
	MASt3R-SLAM	CVPR'25	18	0.0428	**0.0241**	0.0149	0.1109	0.0795	0.0717	**0.0134**	0.0631
Uncalib.	DROID-SLAM	NeurIPS'21	18	0.0476	0.0452	0.0586	0.1229	0.1206	0.1291	0.0720	0.0981
	MASt3R-SLAM	CVPR'25	18	0.0623	0.0335	0.0294	0.1219	0.1014	0.0854	0.0453	0.0794
	VGGT-SLAM	NeurIPS'25	18	0.5205	0.5152	0.2929	0.7383	0.7031	0.7072	0.0531	0.5759
	VGGT-SLAM 2.0	RSS'26	18	0.0459	0.0305	0.0228	0.1138	0.1049	0.0682	0.0530	0.0715
	Ω-SLAM (Ours)	Ours	18	**0.0283**	0.0254	**0.0131**	**0.0491**	**0.0547**	**0.0451**	0.0826	**0.0454**

Table 2: RMSE ATE (m) on five Bonn and five TUM-Dynamic sequences. "Avg. ± Std." is the macro mean and population standard deviation. Bold and underline mark the best overall and uncalibrated values.

	Method	Bonn Dynamic					TUM-Dynamic					Avg. ±Std.
		Ball.	Crowd	Kid.	Obs.	Person	Sitting XYZ	Half	RPY	Static	XYZ	
Calib.	ORB-SLAM3	0.0443	0.0392	0.0201	**0.0363**	0.2972	0.0096	0.0205	0.0704	0.0149	0.0123	0.0565±0.0822
	DeepV2D	0.1902	0.0745	0.0952	0.1761	0.1783	0.0437	0.3526	0.1398	0.0111	0.1729	0.1434±0.0915
	DeepFactors	0.3838	0.0762	0.4494	0.1383	0.6003	0.0788	0.2742	0.1451	0.0132	0.2743	0.2434±0.1789
	DPV-SLAM	0.0417	0.0327	0.0264	0.1696	0.0371	0.0109	0.0226	0.0439	0.0195	0.0190	0.0423±0.0436
	DPV-SLAM++	0.2733	0.0295	0.0263	0.1707	0.0369	0.0109	0.0240	0.0552	0.0099	0.0210	0.0658±0.0823
	GO-SLAM	0.0688	0.0803	0.0258	0.1654	0.2052	0.0115	0.0218	0.0451	0.0208	0.0154	0.0660±0.0641
	DROID-SLAM	0.0512	0.0473	0.0185	0.1948	0.1493	0.0097	0.0180	0.0385	0.0107	0.0140	0.0552±0.0610
	MASt3R-SLAM	0.0504	0.0216	**0.0176**	0.0565	**0.0216**	0.0061	0.0370	0.0349	0.0044	0.0161	0.0266±0.0167
Uncalib.	DROID-SLAM	0.0518	0.0461	0.0209	0.2473	0.2206	0.0252	0.0465	0.0497	0.0112	0.0343	0.0753±0.0805
	MASt3R-SLAM	0.1580	0.0688	0.0441	0.0856	0.0935	0.0492	0.1782	0.1198	0.0099	0.1568	0.0964±0.0529
	VGGT-SLAM	0.4520	0.0469	0.4085	0.2878	0.5518	0.0090	0.2128	0.0771	**0.0014**	0.0260	0.2073±0.1956
	VGGT-SLAM 2.0	0.0580	0.0221	0.0260	0.1011	0.0902	0.0097	0.0208	0.0390	0.0017	0.0146	0.0383±0.0323
	Ω-SLAM (Ours)	**0.0355**	**0.0128**	0.0273	0.0594	0.0232	**0.0058**	**0.0107**	**0.0178**	0.0018	**0.0079**	**0.0202±0.0164**

Unless noted otherwise, Ω-SLAM uses $w = 16$, two shared anchors, one active dense submap, bfloat16 inference, seed zero, and one RTX 3090. The nine Freiburg1 TUM RGB-D sequences use $w = 32$ following VGGT-SLAM 2.0 (Maggio & Carlone, 2026). Appendix C gives camera, keyframe, and resizing settings.

Loop reliability is measured before factor insertion with frozen thresholds. The original cross-traversal audit labels positives within 0.25 m and 45° viewing-direction difference, with cross-scene and repeated-place hard negatives. Table 3 pairs candidate-quality ablations with complete trajectories. Main benchmarks use the pairwise gate.

4.2 Camera pose estimation

Ω-SLAM reaches 0.0454 m ATE on 7-Scenes (Table 1), improving over VGGT-SLAM 2.0 by 36.5% and VGGT-SLAM by 92.1%. It leads uncalibrated methods on six scene groups. Calibrated MASt3R-SLAM has lower ATE on Fire and Stairs.

On ten dynamic sequences (Table 2), ATE is 0.0202 m, 47.2% below VGGT-SLAM 2.0 and 90.2% below VGGT-SLAM, with the best uncalibrated result on eight sequences. Mean RPE is 0.1161 m in translation and 6.254° in rotation. Table 3 relates loop precision and recall to trajectory accuracy.



Figure 3: Recovered world-frame reconstructions on static Office and dynamic TUM streams.

Within the same SLAM scaffold, the complete mechanism set reduces Bonn ATE by 19.6% with VGGT and 42.7% with VGGT-Ω (Appendix C.5). Adaptive loop weights lower VGGT-Ω ATE by 3.8% to 9.3% relative to uniform weights across four benchmarks.

Runtime. With BF16 reconstruction weights and common 448×560 inputs, Ω-SLAM uses 3.39 GiB and runs at 3.88 to 6.03 keyframes/s across three benchmarks. Appendix C.6 reports the accuracy, runtime, and memory comparisons.

4.3 Reconstruction quality and loop reliability

Figures 2 and 3 show aligned submaps, an accepted revisit, and motion-induced scatter in static Office and dynamic TUM scenes. Table 3 evaluates loop verification with the backbone, keyframes, alignment, and backend fixed, using the candidate audits and complete trajectory protocols in Appendix B. Depth rigidity lowers ATE by 18.1% on Bonn and 11.1% on TUM-Dynamic at unchanged recall. Multi-view support further improves dynamic-scene ATE, while context stability raises precision with a small increase in ATE. On 7-Scenes, the pairwise gates make identical audit decisions and give the lowest ATE, showing that higher candidate precision need not improve the trajectory. For retrieval, frame keys with topology reranking yield lower ATE than SALAD on all three benchmarks (Table 4). R@1 rises from 68.8% to 75.0% on Bonn and from 62.5% to 75.0% on TUM-Dynamic, while 7-Scenes rankings are unchanged.

Table 3: Loop verification precision (P, %), recall (R, %), and trajectory ATE (m).

	7-Scenes (static)			Bonn (dynamic)			TUM-Dynamic		
Verification	P ↑	R ↑	ATE ↓	P ↑	R ↑	ATE ↓	P ↑	R ↑	ATE ↓
No loop closure	–	–	0.0498	–	–	0.0520	–	–	0.0108
Patch + 2D	22.2	20.8	**0.0454**	27.3	37.5	0.0387	33.3	50.0	0.0099
+ Depth-rigid (main)	22.2	20.8	**0.0454**	46.2	37.5	0.0317	57.1	50.0	0.0088
+ Multi-view	40.0	16.7	0.0466	55.6	31.2	**0.0306**	60.0	37.5	**0.0084**
+ Context stability	46.2	12.5	0.0479	57.1	25.0	0.0314	75.0	37.5	0.0086

5 Conclusion

We presented Ω-SLAM, a training-free system for uncalibrated dynamic SLAM using VGGT-Ω's internal representations for retrieval and rigid reprojection verification. Quality-weighted SL(4) loop factors, multi-anchor alignment, and rigid-static weighting yield 0.0454 m ATE on 18 7-Scenes sequences and 0.0202 m on ten dynamic sequences, with a 96 KiB retrieval payload for 16 new frames. The verification ablations show different precision and trajectory trade-offs across static and dynamic scenes.

AI USE STATEMENT

We used AI tools for writing and language polishing, literature discovery, code refactoring, and experiment organization. We carefully reviewed AI-assisted content, checked claims and citations against original sources for factual errors and fabricated references, and tested generated code. We take responsibility for the final manuscript and results.

ETHICS STATEMENT

This work uses public RGB-D benchmarks and contains no human intervention, additional personal data collection, or deployment in safety-critical control.

REPRODUCIBILITY STATEMENT

The project artifacts document the fixed manifests, thresholds, raw trajectories, rendering commands, and coordinate conventions used for the reported results. The paper package contains the final source, PDF, and publication figures.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

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. URL https://doi.org/10.1109/TRO.2021.3075644.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aMs6FtNaY5.

Jakub Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J. Davison. DeepFactors: Real-time probabilistic dense monocular SLAM. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020.

Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical Report GT-RIM-CP&R-2012-002, Georgia Institute of Technology, 2012.

Frank Dellaert and Michael Kaess. Factor graphs for robot perception. *Foundations and Trends in Robotics*, 6(1–2):1–139, 2017. doi: 10.1561/2300000043.

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 *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2507.16443.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 427–437, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_Emergent_Outlier_View_Rejection_in_Visual_Geometry_Grounded_Transformers_CVPR_2026_paper.html.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2 edition, 2004. ISBN 978-0-521-54051-3. doi: 10.1017/CBO9780511811685. URL https://doi.org/10.1017/CBO9780511811685.

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, 2024. doi: 10.1109/CVPR52733.2024.01672. URL https://openaccess.thecvf.com/content/CVPR2024/html/Izquierdo_Optimal_Transport_Aggregation_for_Visual_Place_Recognition_CVPR_2024_paper.html.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. SplaTAM: Splat, track and map 3D gaussians for dense RGB-D SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21357–21366, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Keetha_SplaTAM_Splat_Track__Map_3D_Gaussians_for_Dense_RGB-D_CVPR_2024_paper.html.

Jin Hyeon Kim, Jaeeun Lee, Claire Kim, Kyoungjin Oh, Paul Hyunbin Cho, Jaewon Min, Yeji Choi, Jihye Park, Hyunhee Park, Minkyu Park, and Seungryong Kim. Geometry-aware representation denoising for robust multi-view 3D reconstruction. *arXiv preprint arXiv:2605.26230*, 2026. URL https://arxiv.org/abs/2605.26230.

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

Xinze Li, Pengxu Chen, Yiyuan Wang, Weifeng Su, and Wentao Cheng. S-VGGT: Structure-aware subscene decomposition for scalable 3D foundation models. In *IEEE International Conference on Multimedia and Expo*, 2026. URL https://arxiv.org/abs/2603.17625.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_MegaSaM_Accurate_Fast_and_Robust_Structure_and_Motion_from_Casual_CVPR_2025_paper.html.

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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yirunib8l8.

Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual SLAM. In *European Conference on Computer Vision*, pp. 424–440, 2024. URL https://arxiv.org/abs/2408.01654.

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

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. In *Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2601.19887.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the $\mathrm{SL}(4)$ manifold. In *Advances in Neural Information Processing Systems*, volume 38, pp. 143976–144004, 2025. doi: 10.52202/085713-4324. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bc65ab11abfbad890171109686233f4e-Abstract-Conference.html.

Yu A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

Avilasha Mandal, Rajesh Kumar, Sudarshan Sunil Harithas, and Chetan Arora. VGGT-SLAM++. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1008–1018, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/VOCVALC/html/Mandal_VGGT-SLAM_CVPRW_2026_paper.html.

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, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Murai_MASt3R-SLAM_Real-Time_Dense_SLAM_with_3D_Reconstruction_Priors_CVPR_2025_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3d reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IROS*, 2019.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021. doi: 10.1109/ICCV48922.2021.01196.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html.

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.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Joan Solà, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018.

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

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1812.04605.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment—a modern synthesis. In *International Workshop on Vision Algorithms*, pp. 298–372. Springer, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

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, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21499, 2026a. URL https://arxiv.org/abs/2605.15195.

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, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

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. URL https://openreview.net/forum?id=DTQIjngDta.

Thomas Whelan, Renato F. Salas-Moreno, Ben Glocker, Andrew J. Davison, and Stefan Leutenegger. Elasticfusion: Real-time dense slam and light source estimation, 2016.

Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19595–19604, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, pp. 767–783, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5525–5534, 2019.

Xihang Yu and Heng Yang. Sim-sync: From certifiably optimal synchronization over the 3d similarity group to scene reconstruction with learned depth. *IEEE Robotics and Automation Letters*, 2024.

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*, 2025. URL https://openreview.net/forum?id=lJpqxFgWCM.

Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. GO-SLAM: Global optimization for consistent 3D instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3727–3737, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_GO-SLAM_Global_Optimization_for_Consistent_3D_Instant_Reconstruction_ICCV_2023_paper.html.

Zihan Zhu, Songyou Peng, Viktor Larsson, Wei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. doi: 10.48550/arXiv.2507.11539. URL https://arxiv.org/abs/2507.11539.

Zichen Zou, Xiaosong Jia, Zuxuan Wu, and Yu-Gang Jiang. Attention itself could retrieve. RetrieveVGGT: Training-free long context streaming 3D reconstruction via query-key similarity retrieval. *arXiv preprint arXiv:2605.09644*, 2026. doi: 10.48550/arXiv.2605.09644. URL https://arxiv.org/abs/2605.09644.

A Related work

A.1 Visual SLAM

Visual SLAM maintains a map as images arrive and uses revisits to correct accumulated drift. ORB-SLAM3 (Campos et al., 2021) combines feature tracking and bundle adjustment with place recognition and map merging. Learned systems retain geometric optimization while changing the measurements or map representation. DeepFactors (Czarnowski et al., 2020) optimizes compact learned depth codes in a probabilistic framework, and DROID-SLAM (Teed & Deng, 2021) recurrently updates camera poses and dense depths through differentiable bundle adjustment. Deep Patch Visual SLAM (Lipson et al., 2024) extends patch-based odometry with loop closure while keeping memory use low. Dense mapping also explores neural implicit scene representations (Zhu et al., 2022) and 3D Gaussians (Keetha et al., 2024). Reconstruction priors provide another route to online mapping. MASt3R-SLAM (Murai et al., 2025) integrates two-view geometry and matching into tracking, local fusion, and global optimization. VGGT-SLAM (Maggio et al., 2025) aligns jointly reconstructed submaps on SL(4), and VGGT-SLAM 2.0 (Maggio & Carlone, 2026) refines the graph with shared calibration and keyframe-level variables. VGGT-Long (Deng et al., 2026) extends reconstruction to long sequences through overlapping chunks, Sim(3) alignment, and loop optimization. VGGT-SLAM++ (Mandal et al., 2026) uses digital elevation maps for DINOv2 and AnyLoc retrieval, supplying spatial neighbours for local Sim(3) correction. We connect retrieval with geometric constraints at the submap level. Frame keys retrieve revisits, scene registers score submap context, and predicted cameras and depths verify the resulting loop pairs.

A.2 Feed-forward 3D reconstruction

Feed-forward reconstruction learns geometric priors that allow camera and scene estimation from uncalibrated images. DUSt3R (Wang et al., 2024) predicts dense point maps for an image pair and aligns pairwise predictions for larger collections. MASt3R (Leroy et al., 2024) couples point-map regression with local descriptors, grounding image matching in the predicted geometry. VGGT (Wang et al., 2025a) jointly processes multiple views to infer cameras, depths, point maps, and tracks, bringing camera estimation and dense reconstruction into a shared network. VGGT-Ω (Wang et al., 2026a) develops this design with register-based attention and self-supervised refinement, improving reconstruction accuracy and efficiency. Temporal extensions maintain history in different forms. CUT3R (Wang et al., 2025b) updates a persistent recurrent state as observations arrive. StreamVGGT (Zhuo et al., 2025) distills VGGT into a causal transformer that caches historical

keys and values for incremental reconstruction. RetrieveVGGT (Zou et al., 2026) selects historical context using similarity between queries and keys in the first global attention layer. For videos containing independently moving objects, MonST3R (Zhang et al., 2025) adapts point-map prediction to scene motion. We use the geometry and internal features of VGGT-Ω to connect submaps. Pre-attention frame keys retain an image-specific retrieval representation, post-attention registers summarize the submap, and predicted cameras and depths supply the geometric evidence for loop verification and alignment.

B Loop verification and retrieval ablations

Table 3 evaluates loop verification and its effect on trajectory accuracy. The 2D and depth-rigid gates share two-frame predictions. Multi-view support and context stability are added cumulatively. The trajectory runs fix the backbone, keyframes, timestamps, static weighting, anchor aggregation, and backend. Accepted edges use the same pairwise quality weight q, with the additional strict-mode multipliers disabled. ATE follows the per-sequence Sim(3) protocol in Sec. 4.1.

The additional candidate audits fix the retriever, historical archive, and ordered top-three candidates. Ground-truth poses and sensor depth label background overlap and relative-motion accuracy using thresholds frozen on disjoint development data. Each query contributes its first accepted candidate. Precision and recall are $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and TP/Q_+, reported as percentages. Here Q_+ counts queries with a positive in the eligible history after excluding shared frames and temporal neighbours. For 7-Scenes, P and R for Patch + 2D, Depth-rigid, and Context stability come from the original pose-proximity audit described below. All ATE values come from complete trajectory runs.

Depth rigidity improves dynamic-scene verification without lowering recall. Precision rises from 27.3% to 46.2% on Bonn and from 33.3% to 57.1% on TUM-Dynamic. ATE falls from 0.0387 to 0.0317 m and from 0.0099 to 0.0088 m, respectively. Multi-view support further lowers dynamic-scene ATE to 0.0306 and 0.0084 m despite reduced recall. Adding context stability raises precision but slightly increases these errors to 0.0314 and 0.0086 m. On 7-Scenes, both pairwise gates give the same audit decisions and the lowest ATE of 0.0454 m, while stricter checks raise ATE. The trajectory benefit of additional verification therefore varies across the static and dynamic benchmarks.

Original 7-Scenes audit. The cross-traversal audit contains 66 queries, including 48 positives under the 0.25 m and 45° pose-proximity criterion. Register retrieval alone accepts all 66 queries (11 TP, 55 FP). Both pairwise gates accept the same 83 of 198 candidates, yielding 45 accepted queries (10 TP, 35 FP). Strict verification accepts 13 queries (6 TP, 7 FP). False accepts per 1,000 queries, computed as $1000\,\mathrm{FP}/Q$, fall from 833.3 without verification to 530.3 with either pairwise gate and 106.1 with strict checks. The pairwise gates have median match count, rigid residual, and rigid inlier ratio of 325, 25.73 px, and 0.477. Strict verification gives 352, 23.68 px, and 0.506, with median support from three view pairs, 0.527 context overlap, and 1.55 px residual change across contexts. Residuals use inference-image coordinates. This audit uses fixed gates without conformal calibration.

B.1 Compact retrieval memory

In Table 4, we vary the retrieval module while keeping the backbone, pairwise verifier, bridges, and backend fixed for trajectory evaluation. SALAD uses the same frame-to-submap aggregation as frame keys. The additional audits compress registers symmetrically with centroid-initialized farthest-point sampling. Retrieval-score normalization in q is frozen on development data. R@1 is reported as a percentage over queries with a historical positive. MRR averages the reciprocal rank of the first positive in the eligible gallery. Topology scores rerank the three leading frame-key candidates and retain the remaining order, leaving R@3 unchanged by construction.

Frame keys provide the main retrieval signal, while topology reranking adds useful context on the dynamic datasets. R@1 rises from 68.8% to 75.0% on Bonn and from 62.5% to 75.0% on TUM-Dynamic, with MRR also increasing. The corresponding ATE reductions are 2.8% and 3.3%. On 7-Scenes, the original audit gives identical ranking metrics with and without topology. The full trajectories also differ little in ATE, at 0.0457 and 0.0454 m. Combining frame keys with topology gives the lowest measured ATE on all three datasets, including lower errors than using SALAD for retrieval.

Table 4: Retrieval quality and trajectory ATE (m).

	7-Scenes (static)			Bonn (dynamic)			TUM-Dynamic		
Retrieval	R@1 ↑	MRR ↑	ATE ↓	R@1 ↑	MRR ↑	ATE ↓	R@1 ↑	MRR ↑	ATE ↓
SALAD	72.9	0.821	0.0463	62.5	0.741	0.0358	50.0	0.635	0.0096
Registers only	8.3	0.180	0.0514	12.5	0.162	0.0531	12.5	0.179	0.0108
Frame key only	75.0	0.845	0.0457	68.8	0.791	0.0326	62.5	0.729	0.0091
Frame key + topology	75.0	0.845	**0.0454**	75.0	0.834	**0.0317**	75.0	0.802	**0.0088**

The fp16 descriptor payload is $2N$ KiB for N frame keys and $4M$ KiB for M topology tokens. With 16 new frames and $M = 16$, the per-submap archive entry uses 96 KiB excluding shared anchors. Including A shared anchors adds $2A$ KiB in frame keys, while the topology payload stays fixed. For 16 or 32 new frames with two anchors, the totals are 100 or 132 KiB. These values describe descriptor storage, with model weights, dense maps, and graph state contributing additional memory. Internal retrieval reuses network outputs and adds pooling and sampling, whereas SALAD requires another encoder pass.

Original retrieval audit. The 48-query 7-Scenes audit excludes the query traversal, leaving 61 to 64 candidates per query. Query registers are compressed to 16 tokens by sampling from token zero, while candidate registers remain full. Scoring uses top-quarter partial cosine. This audit supplies the 7-Scenes R@1 and MRR for Registers only, Frame key only, and Frame key + topology in Table 4. R@3 is 16.7% for registers and 93.8% for both frame keys and hybrid retrieval. Median hybrid lookup time is 2.46 ms on a 65-submap CPU bank. The compact payload figures above correspond to symmetric storage. A dense fp32 XYZ map for 16 frames at 512^2 resolution occupies 48.0 MiB.

B.2 Internal retrieval diagnostics

Figure 4 shows a query and its three highest-ranked frame-key candidates. The first candidate is a labelled cross-traversal positive. The other two are visually similar hard negatives under the fixed pose-overlap criterion. The lower panels compare frame-key similarity and register-topology scores, linking the retrieved images to the evidence used for reranking. The reported cache uses deterministic exact cosine ranking. Approximate-nearest-neighbour indices such as FAISS and HNSW offer alternatives for larger galleries (Johnson et al., 2021; Malkov & Yashunin, 2020).

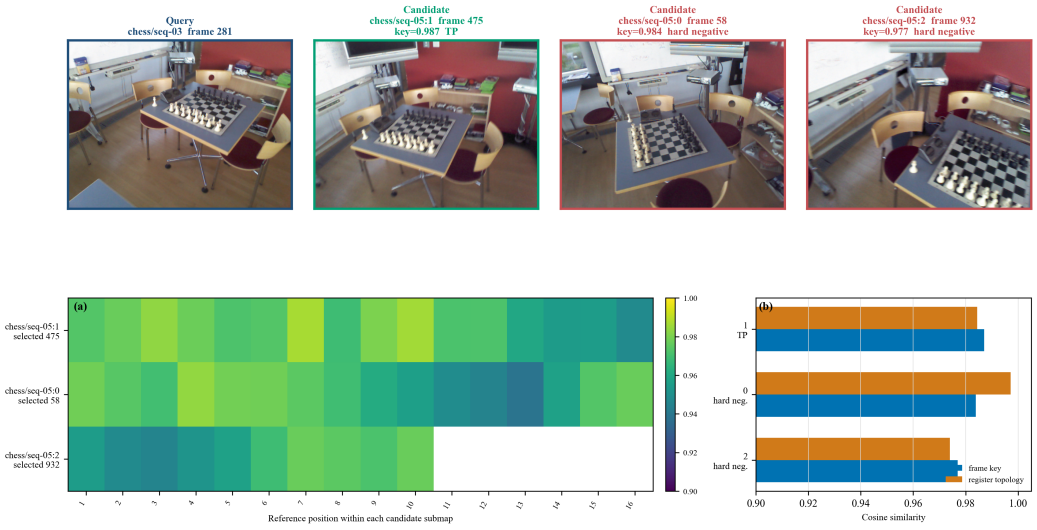

Figure 4: Retrieved views and internal similarity scores on 7-Scenes.

B.3 Gallery size and retrieval recall

Figure 5 and Table 5 use the earlier 48 queries throughout. Each pool contains the first S cached submaps, with the query traversal excluded before ranking. Queries without a positive in the prefix count as misses. Here both query and candidate registers are compressed to 16 tokens using farthest-point sampling from token zero. Recall grows as matching traversals enter the gallery. At $S = 65$, frame-key and hybrid retrieval reach $75.0\%/93.8\%$ Recall@1/@3, compared with $16.7\%/20.8\%$ for registers alone. Hybrid and frame-key retrieval have equal recall at every tested prefix. Frame keys therefore attain the measured recall with 32 KiB per submap. Topology tokens add 64 KiB for context scoring.

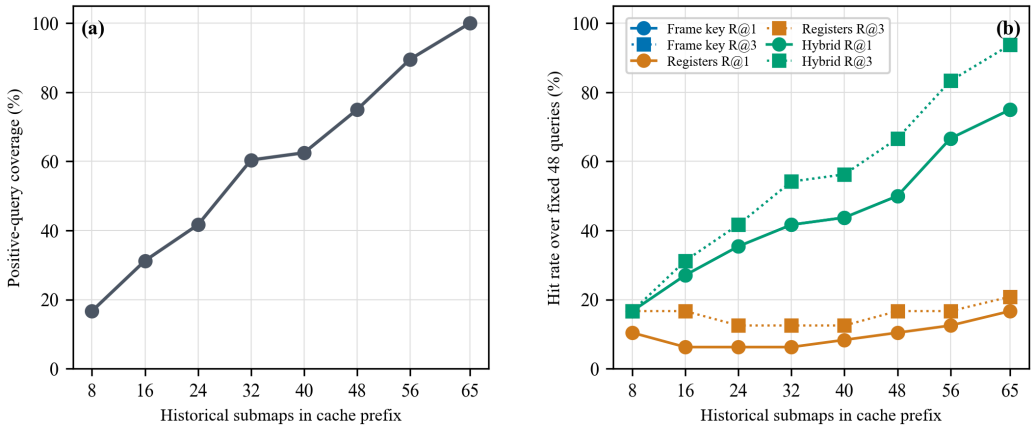

Figure 5: Positive-query coverage and retrieval recall as the gallery grows.

Table 5: Recall@1 / Recall@3 (%) across gallery prefixes for 48 queries.

Prefix size S	Positive queries in prefix	Frame key	Registers	Hybrid
8	8	16.7 / 16.7	10.4 / 16.7	16.7 / 16.7
16	15	27.1 / 31.2	6.2 / 16.7	27.1 / 31.2
24	20	35.4 / 41.7	6.2 / 12.5	35.4 / 41.7
32	29	41.7 / 54.2	6.2 / 12.5	41.7 / 54.2
40	30	43.8 / 56.2	8.3 / 12.5	43.8 / 56.2
48	36	50.0 / 66.7	10.4 / 16.7	50.0 / 66.7
56	43	66.7 / 83.3	12.5 / 16.7	66.7 / 83.3
65	48	75.0 / 93.8	16.7 / 20.8	75.0 / 93.8

C Additional trajectory and ablation results

C.1 Per-sequence trajectory results

Tables 7 and 8 give per-sequence ATE. C/U denote calibrated/uncalibrated inputs, and G denotes GeoCalib. Numbers under the 7-Scenes scene names identify test sequences.

Benchmark configuration. Ω-SLAM uses the released 1B checkpoint with bfloat16 autocast on an RTX 3090, seed zero, nominal resolution 512, fixed resizing, and 25% confidence filtering. Keyframes require mean tracked displacement > 50 original-image pixels. Stride-two input supplies $w = 16$ new keyframes on 7-Scenes, Bonn, and TUM-Dynamic. Freiburg1 uses stride one and $w = 32$. Two shared anchors give 18/34-frame windows.

Tracking uses pyramidal Lucas-Kanade flow on up to 1,000 Shi-Tomasi corners, a 21×21 window, and four pyramid levels. The displacement mean uses successful tracks. Fewer than ten tracks trigger a new keyframe.

After optimization, Ω-SLAM retains $K = 1$ dense submap. Older submaps keep descriptors, camera/graph state, image paths, and sampled alignment geometry (Maggio et al., 2025; Maggio & Carlone, 2026). The fp16 retrieval archive stores 16 topology tokens and one 1024-dimensional key per frame. For $w = 16$, it uses 96 KiB excluding shared anchors. Adding A anchors costs $2A$ KiB, giving 100 KiB for $w = 16, A = 2$ and 132 KiB for $w = 32, A = 2$.

Calibrated baselines receive dataset intrinsics, and GeoCalib DROID-SLAM estimates them from RGB. Camera-to-world poses use (x, y, z, w) quaternions. ATE and RPE follow Sec. 4.1, with rotation in degrees.

Sequence-level variation. Ω-SLAM improves over VGGT-SLAM 2.0 on all four Office and five RedKitchen sequences. Office ATE ranges from 0.0350 to 0.0810 m, compared with 0.0978 to 0.1426 m. The largest regression occurs on Stairs-01, where ATE rises from 0.0798 to 0.1372 m. Dynamic-scene gains are largest on Bonn person and obstructing-box, with higher ATE on Bonn kidnap and TUM walking static.

Runtime and verification cost. Each sequence is run once on an RTX 3090 with PyTorch 2.5.1 and CUDA 11.8. Synchronized timing covers loading, keyframe selection, inference, loop closure, and graph updates, including first-call overhead but excluding initialization and export. Table 15 also uses bfloat16 reconstruction weights and common 448×560 bicubic crops that preserve aspect ratio, with fp32 geometry outputs. Submaps retain 16 new keyframes and one shared frame for VGGT-SLAM or two for Ω-SLAM, with native retrieval and graph policies.

Native inputs are 392×518 for VGGT-SLAM and 448×592 for Ω-SLAM, both yielding 1,036 patches per frame. Native weights are fp32 for VGGT-SLAM and Ω-SLAM, and bfloat16 for version 2.0. Ω-SLAM's fp32 parameters occupy 4.26 GiB, plus cast weights and activations during inference. On the first 18-frame Office-06 window, bfloat16 weights and mixed-precision heads reduce the peak from 7.25 to 3.46 GiB at native resolution.

Table 6 profiles the native pairwise gate across all 28 sequences. The extra forward takes about 152 ms, and full verification takes 229 to 241 ms per candidate including loading, transfers, and geometry. Verification uses 2.0% to 5.3% of online time and peaks at 6.09 GiB, below reconstruction's 7.25 GiB. Stage means pool the indicated units.

Table 6: Native runtime breakdown of Ω-SLAM with pairwise verification.

Stage / measurement	7-Scenes	Bonn Dynamic	TUM-Dynamic
Reconstruction (ms/submap)	1483.6	1483.4	1358.7
Retrieval (ms/query)	2.2	2.4	2.2
Extra forward (ms/candidate)	152.0	151.1	152.2
Geometry checks (ms/candidate)	45.7	47.0	46.2
Full verification (ms/candidate)	241.2	229.3	239.9
Verification share (%)	5.27	2.03	2.37

Table 7: Per-sequence RMSE ATE (m) on the 18-sequence 7-Scenes test split.

Method/config.	Chess		Fire		Heads	Office			
	03	05	03	04	01	02	06	07	09
ORB-SLAM3 (C)	0.0419	0.0334	0.0343	0.0346	0.0778	0.1159	0.1185	0.1131	0.1686
DeepV2D (C)	0.2607	0.1968	0.2039	0.1563	0.1460	0.1357	0.2315	0.2809	0.2054
DeepFactors (C)	0.5677	0.4495	0.2516	0.1531	0.0862	0.1654	0.2278	0.2681	0.2835
DPV-SLAM (C)	0.0392	0.0315	0.0306	0.0286	0.0297	0.1514	0.1178	0.1094	0.1358
DPV-SLAM++ (C)	0.0395	0.0320	0.0305	0.0281	0.0297	0.1336	0.1181	0.1095	0.1358
GO-SLAM (C)	0.0352	0.0329	0.0305	0.0311	0.0259	0.0758	0.1106	0.1027	0.0854
DROID-SLAM (C)	0.0376	0.0315	0.0291	0.0241	0.0224	0.0885	0.1030	0.1199	0.1079
DROID-SLAM (G)	0.0528	0.0425	0.0618	0.0287	0.0586	0.1431	0.1008	0.1522	0.0954
MASt3R-SLAM (C)	0.0411	0.0445	0.0279	0.0203	0.0149	0.0882	0.0992	0.1158	0.1404
MASt3R-SLAM (U)	0.0795	0.0451	0.0309	0.0361	0.0294	0.1176	0.1114	0.1404	0.1183
VGGT-SLAM	0.4727	0.5683	0.5825	0.4478	0.2929	0.7544	0.7706	0.7388	0.6895
VGGT-SLAM 2.0	0.0497	0.0422	0.0351	0.0260	0.0228	0.1068	0.0978	0.1426	0.1081
Ω-SLAM (Ours)	0.0310	0.0256	0.0331	0.0177	0.0131	0.0350	0.0371	0.0810	0.0432

Method/config.	Pumpkin		RedKitchen					Stairs	
	01	07	03	04	06	12	14	01	04
ORB-SLAM3 (C)	0.1175	0.0767	0.0630	0.0489	0.0798	0.0877	0.0495	0.1057	0.0323
DeepV2D (C)	0.2233	0.1928	0.1074	0.0934	0.2663	0.3803	0.1630	0.1419	0.0860
DeepFactors (C)	0.3362	0.4112	0.1177	0.1486	0.5417	0.4317	0.4843	0.1548	0.0675
DPV-SLAM (C)	0.4504	0.2246	0.0546	0.0520	0.0734	0.1131	0.0562	0.0695	0.0270
DPV-SLAM++ (C)	0.4492	0.1495	0.0549	0.0518	0.0735	0.1108	0.0560	0.0823	0.0286
GO-SLAM (C)	0.1436	0.0709	0.0662	0.1727	0.0652	0.0768	0.0531	0.0314	0.0860
DROID-SLAM (C)	0.1362	0.0730	0.0592	0.0534	0.0747	0.0763	0.0546	0.0235	0.1143
DROID-SLAM (G)	0.1469	0.0943	0.1918	0.0615	0.1108	0.1804	0.1007	0.0230	0.1209
MASt3R-SLAM (C)	0.0876	0.0715	0.0557	0.0450	0.0921	0.0871	0.0786	0.0107	0.0160
MASt3R-SLAM (U)	0.1121	0.0906	0.0503	0.0619	0.0937	0.1217	0.0991	0.0320	0.0586
VGGT-SLAM	0.8303	0.5759	0.5378	0.6279	0.7704	1.0288	0.5713	0.0804	0.0259
VGGT-SLAM 2.0	0.1353	0.0746	0.0473	0.0579	0.0886	0.0823	0.0648	0.0798	0.0262
Ω-SLAM (Ours)	0.0428	0.0666	0.0341	0.0425	0.0486	0.0628	0.0376	0.1372	0.0280

Table 8: Per-sequence RMSE ATE (m) on Bonn and TUM-Dynamic.

Method/config.	Bonn Dynamic					TUM-Dynamic				
	Balloon	Crowd	Kidnap	Obstruct	Person	Sitting XYZ	Walking Half	Walking RPY	Walking Static	Walking XYZ
ORB-SLAM3 (C)	0.0443	0.0392	0.0201	0.0363	0.2972	0.0096	0.0205	0.0704	0.0149	0.0123
DeepV2D (C)	0.1902	0.0745	0.0952	0.1761	0.1783	0.0437	0.3526	0.1398	0.0111	0.1729
DeepFactors (C)	0.3838	0.0762	0.4494	0.1383	0.6003	0.0788	0.2742	0.1451	0.0132	0.2743
DPV-SLAM (C)	0.0417	0.0327	0.0264	0.1696	0.0371	0.0109	0.0226	0.0439	0.0195	0.0190
DPV-SLAM++ (C)	0.2733	0.0295	0.0263	0.1707	0.0369	0.0109	0.0240	0.0552	0.0099	0.0210
GO-SLAM (C)	0.0688	0.0803	0.0258	0.1654	0.2052	0.0115	0.0218	0.0451	0.0208	0.0154
DROID-SLAM (C)	0.0512	0.0473	0.0185	0.1948	0.1493	0.0097	0.0180	0.0385	0.0107	0.0140
DROID-SLAM (G)	0.0518	0.0461	0.0209	0.2473	0.2206	0.0252	0.0465	0.0497	0.0112	0.0343
MASt3R-SLAM (C)	0.0504	0.0216	0.0176	0.0565	0.0216	0.0061	0.0370	0.0349	0.0044	0.0161
MASt3R-SLAM (U)	0.1580	0.0688	0.0441	0.0856	0.0935	0.0492	0.1782	0.1198	0.0099	0.1568
VGGT-SLAM	0.4520	0.0469	0.4085	0.2878	0.5518	0.0090	0.2128	0.0771	0.0014	0.0260
VGGT-SLAM 2.0	0.0580	0.0221	0.0260	0.1011	0.0902	0.0097	0.0208	0.0390	0.0017	0.0146
Ω-SLAM (Ours)	0.0355	0.0128	0.0273	0.0594	0.0232	0.0058	0.0107	0.0178	0.0018	0.0079

Table 9 evaluates all nine Freiburg1 TUM RGB-D sequences with stride-one RGB input and independent Sim(3) alignment. Calibration variants are identified in the row labels. Ω-SLAM reaches 0.0461 m mean ATE, compared with 0.0415 m for VGGT-SLAM 2.0, with lower errors on Floor, Room, RPY, and XYZ. DeepFactors uses a local photometric-only branch on Freiburg1 Floor and 7-Scenes Office-06.

Table 9: RMSE ATE (m) on the nine-sequence Freiburg1 TUM RGB-D benchmark.

Method/config.	360	Desk	Desk2	Floor	Plant	Room	RPY	Teddy	XYZ	Avg.
ORB-SLAM3 (calib.)	0.0150	0.0386	0.2835	0.0165	0.0347	0.4436	0.0356	0.3460	0.0089	0.1358
DeepV2D (calib.)	0.1172	0.1516	0.2165	0.3883	0.2027	0.2994	0.0445	0.2133	0.0585	0.1880
DeepFactors (calib.)	0.1010	0.1976	0.2413	0.1827	0.2941	0.3265	0.0285	0.5800	0.0227	0.2194
DPV-SLAM (calib.)	0.1479	0.0163	0.0260	0.0283	0.0175	0.3985	0.0228	0.0841	0.0104	0.0835
DPV-SLAM++ (calib.)	0.1534	0.0163	0.0250	0.0281	0.0173	0.1721	0.0402	0.0830	0.0103	0.0606
GO-SLAM (calib.)	0.1893	0.0182	0.0344	0.0232	0.0271	0.7029	0.0189	0.0619	0.0102	0.1207
DROID-SLAM (calib.)	0.0712	0.0174	0.7799	0.0223	0.0146	0.0405	0.0376	0.0407	0.0105	0.1150
DROID-SLAM (uncalib.)	0.2032	0.0468	0.7977	0.0558	0.0466	0.9031	0.0428	0.1620	0.0154	0.2526
MASt3R-SLAM (calib.)	0.0465	0.0165	0.0271	0.0245	0.0174	0.0640	0.0226	0.0425	0.0101	0.0301
MASt3R-SLAM (uncalib.)	0.0715	0.0380	0.0582	0.0504	0.0383	0.0991	0.0413	0.1175	0.0198	0.0593
VGGT-SLAM	0.1731	0.6388	0.8456	0.7505	0.6967	0.9367	0.0457	0.8711	0.0803	0.5598
VGGT-SLAM 2.0	0.0503	0.0263	0.0305	0.0990	0.0269	0.0648	0.0259	0.0365	0.0136	0.0415
Ω-SLAM (Ours)	0.0890	0.0520	0.0350	0.0668	0.0318	0.0611	0.0242	0.0420	0.0130	0.0461

C.2 Pose accuracy and coverage

Table 10 reports sequence macro means over 18 7-Scenes, five Bonn, and five TUM-Dynamic sequences. Pose coverage is the mean fraction of presented frames with an exported pose across all 28 sequences. Ω-SLAM and both VGGT-SLAM versions report 8.9% coverage, reflecting their keyframe outputs. We report coverage alongside ATE to account for differences in pose-export density. Table 15 reports throughput on these full streams.

Table 10: Continuous-stream ATE (m) and exported-pose coverage (%).

Input	Method/config.	7-Scenes (18)	Bonn Dyn. (5)	TUM Dyn. (5)	Pose cov. (%)
Calib.	ORB-SLAM3	0.0777	0.0874	0.0255	23.0
Calib.	DeepV2D	0.1929	0.1429	0.1440	37.2
Calib.	DeepFactors	0.2859	0.3296	0.1571	26.4
Calib.	DPV-SLAM	0.0997	0.0615	0.0232	100.0
Calib.	DPV-SLAM++	0.0952	0.1074	0.0242	100.0
Calib.	GO-SLAM	0.0720	0.1091	0.0229	100.0
Calib.	DROID-SLAM	0.0683	0.0922	0.0182	100.0
Uncalib.	DROID-SLAM	0.0981	0.1173	0.0334	100.0
Calib.	MASt3R-SLAM	0.0631	0.0335	0.0197	4.2
Uncalib.	MASt3R-SLAM	0.0794	0.0900	0.1028	4.4
Uncalib.	VGGT-SLAM	0.5759	0.3494	0.0653	8.9
Uncalib.	VGGT-SLAM 2.0	0.0715	0.0595	0.0172	8.9
Uncalib.	**Ω-SLAM (Ours)**	**0.0454**	**0.0317**	**0.0088**	8.9

C.3 Dense-map diagnostics

Table 11 measures median nearest-neighbour distance from each reconstructed cloud to the GT cloud. All methods use the same 100 sampled frames and camera-center Sim(3) alignment on four 7-Scenes sequences. Ω-SLAM has the lowest four-sequence mean, 1.93 cm, while VGGT-SLAM has lower distances on Chess-03 and Office-06.

Table 12 reports pose errors and throughput from the dense-map reruns used for the TUM walking visualizations. Each sequence uses 100 uniformly sampled RGB frames with ground-truth associations, all retained as keyframes at 384×512 resolution. One Sim(3) per method aligns both poses

Table 11: Median cloud-to-GT nearest-neighbour distance (cm) on 7-Scenes.

Method	Chess-03	Fire-03	Office-06	RedKitchen-04	Mean
VGGT	3.26	4.03	7.48	3.88	4.66
VGGT-Ω	2.66	1.98	3.58	4.12	3.08
VGGT-SLAM	**1.75**	2.52	**2.07**	2.84	2.29
Ω-SLAM (Ours)	1.79	**1.17**	2.78	**1.97**	**1.93**

and clouds to GT. Ω-SLAM lowers ATE and both RPE measures on all three walking sequences. Throughput is lower on walking_xyz and walking_halfsphere but higher on walking_rpy.

The reconstruction comparison evaluates RGB geometry. Surfel fusion (Whelan et al., 2016), neural implicit maps (Zhu et al., 2022), and Gaussian maps (Yan et al., 2024; Keetha et al., 2024) offer alternative map representations. We evaluate calibrated RGB-D SplaTAM separately from the RGB continuous-stream pose benchmarks.

Table 12: Pose accuracy and throughput for the TUM dense-map reruns.

Sequence	Method	ATE (m)	RPE-t (m)	RPE-R (°)	FPS
walking_xyz	VGGT-SLAM	0.01949	0.01552	0.383	8.42
walking_xyz	Ω-SLAM	0.00834	0.00769	0.325	6.28
walking_halfsphere	VGGT-SLAM	0.02528	0.01788	0.571	8.46
walking_halfsphere	Ω-SLAM	0.01315	0.01024	0.511	6.65
walking_rpy	VGGT-SLAM	0.02799	0.02386	0.654	6.50
walking_rpy	Ω-SLAM	0.01357	0.01223	0.500	10.85

C.4 Component ablation

Table 13 varies rigid-static weighting (S), multi-anchor alignment (A), and verified loop closure (L). Disabling A leaves one anchor. Loop closure produces the largest single-component ATE reduction on all three benchmarks. Static weighting rejects 3.82% of observations and modestly improves ATE on its own. When added to alignment and loop closure, it further reduces Bonn and TUM-Dynamic ATE by 8.9% and 6.4%. The full configuration has the lowest errors of 0.0491, 0.0317, and 0.0088 m on Office, Bonn, and TUM-Dynamic, matching the main benchmark results.

Table 13: ATE (m) and rejected observations (%) in the component ablation.

S	A	L	Office ATE ↓	Bonn ATE ↓	TUM-Dyn. ATE ↓	Reject. (%) ↑
			0.0530	0.0553	0.0124	0.00
✓			0.0527	0.0539	0.0121	3.82
	✓		0.0519	0.0536	0.0112	0.00
		✓	0.0508	0.0418	0.0101	0.00
✓	✓		0.0516	0.0520	0.0108	3.82
✓		✓	0.0503	0.0385	0.0096	3.82
	✓	✓	0.0498	0.0348	0.0094	0.00
✓	✓	✓	**0.0491**	**0.0317**	**0.0088**	**3.82**

C.5 Backbone and mechanism ablation

Table 14 compares VGGT and VGGT-Ω within the same SLAM scaffold, then examines the effect of the mechanism set and loop weighting. Each backbone retains its native preprocessing, while the scaffold and gate thresholds remain fixed. S, A, and L denote rigid-static weighting, multi-anchor alignment, and loop closure. Disabling A leaves one anchor. Uniform loop weights use $\sigma_{\text{loop}} = 0.115$, and adaptive weights follow Eq. (10). Tables 1 and 2 report results for the released VGGT-SLAM systems.

Table 14: Backbone and mechanism ablation. Mean per-sequence RMSE ATE (m), with sequence counts in parentheses.

Backbone	S	A	L	Loop weights	7-Scenes (18)	Bonn (5)	TUM-Dyn. (5)	Freiburg1 (9)
VGGT				–	0.0762	0.0698	0.0218	0.0647
VGGT	✓	✓	✓	Adaptive	0.0709	0.0561	0.0173	0.0548
VGGT-Ω				–	0.0516	0.0553	0.0124	0.0621
VGGT-Ω	✓	✓		–	0.0498	0.0520	0.0108	0.0594
VGGT-Ω	✓	✓	✓	Uniform	0.0472	0.0342	0.0097	0.0489
VGGT-Ω	✓	✓	✓	Adaptive	**0.0454**	**0.0317**	**0.0088**	**0.0461**

With S, A, and L disabled, VGGT-Ω yields lower ATE than VGGT on all four benchmarks, including 0.0516 versus 0.0762 m on 7-Scenes. Enabling the complete mechanism set further reduces ATE on VGGT-Ω by 12.0%, 42.7%, 29.0%, and 25.8% on 7-Scenes, Bonn, TUM-Dynamic, and Freiburg1, respectively. The same additions reduce ATE on VGGT by 7.0%, 19.6%, 20.6%, and 15.3%. Both backbones therefore benefit from the added mechanisms.

On VGGT-Ω, adding uniformly weighted loop closure to S and A reduces Bonn ATE from 0.0520 to 0.0342 m and TUM-Dynamic ATE from 0.0108 to 0.0097 m. Adaptive weighting then lowers these errors to 0.0317 and 0.0088 m, improving on uniform weighting by 7.3% and 9.3%. The corresponding reductions on 7-Scenes and Freiburg1 are 3.8% and 5.7%. Together with the individual component comparisons in Table 13, these results support combining verified loop closure with quality-dependent factor weights.

C.6 Runtime and memory comparison

Table 15 compares native settings with common BF16 reconstruction weights and 448×560 inputs across all 28 sequences. ATE and keyframe FPS are sequence means, and memory is the maximum CUDA allocation over each dataset. Under common settings, Ω-SLAM uses 3.39 GiB, about 48% less than VGGT-SLAM 2.0, with higher throughput and lower ATE than both VGGT-SLAM versions on all three benchmarks. Its Bonn ATE rises from 0.0317 to 0.0353 m relative to native settings.

Table 15: ATE (m), keyframe FPS, and peak GPU memory (GiB) on one RTX 3090. Bold marks the best value within each setting.

	7-Scenes			Bonn Dynamic			TUM-Dynamic		
Method	ATE ↓	FPS ↑	GiB ↓	ATE ↓	FPS ↑	GiB ↓	ATE ↓	FPS ↑	GiB ↓
Native settings									
VGGT-SLAM	0.5759	4.71	12.25	0.3494	5.39	11.96	0.0653	3.34	12.00
VGGT-SLAM 2.0	0.0715	**5.27**	**5.83**	0.0595	5.09	**5.75**	0.0172	3.50	**5.79**
Ω-SLAM (Ours)	**0.0454**	4.81	7.25	**0.0317**	**5.58**	7.25	**0.0088**	**3.63**	7.25
BF16 weights, common 448×560 *input*									
VGGT-SLAM	0.5735	4.56	6.69	0.3580	5.10	6.46	0.0660	3.42	6.51
VGGT-SLAM 2.0	0.0738	4.54	6.57	0.0610	5.09	6.46	0.0188	3.42	6.51
Ω-SLAM (Ours)	**0.0453**	**5.22**	**3.39**	**0.0353**	**6.03**	**3.39**	**0.0084**	**3.88**	**3.39**

D Qualitative reconstructions

Figure 6 shows VGGT, VGGT-Ω, VGGT-SLAM, and Ω-SLAM in a shared GT-derived display gauge. The GT cloud defines the gauge and is omitted from the panels. Table 11 provides the corresponding geometric-distance metric. Ω-SLAM reaches 2.78 cm on Office-06, compared with 2.07 cm for VGGT-SLAM. Camera-path accuracy and point-cloud accuracy capture different aspects of reconstruction.



Figure 6: Dense reconstructions of Office `seq-06` in a shared display gauge.

Figure 7 compares Fire `seq-03` point clouds under the same view, sampling, and pose convention. Camera pyramids carry their source RGB images, showing how the observations are placed within the reconstructed scene.



VGGT-SLAM Ω-SLAM (ours)

Figure 7: World-frame point clouds and RGB camera frusta on Fire `seq-03`.

Figure 8 compares Office `seq-06` camera paths in a common PCA display gauge and reports per-keyframe position errors after Sim(3) alignment. The error curves locate deviations along the sequence, complementing the trajectory's spatial view.

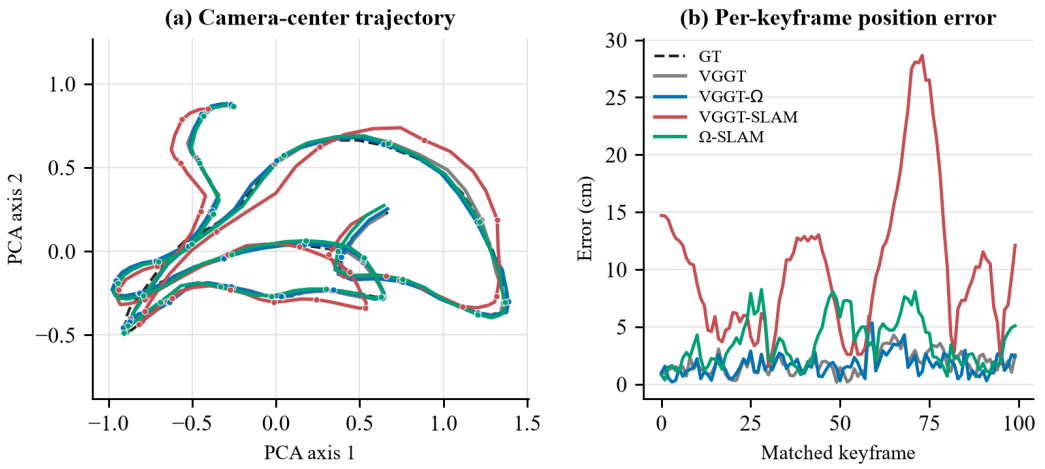

Figure 8: Camera trajectories and per-keyframe position errors on Office `seq-06`.

E THEORY

We derive the geometric and statistical basis of Algorithm 1. Projective geometry defines the alignment space, and depth reprojection tests candidate correspondences. Robust scale aggregation and quality-weighted factors control how accepted evidence enters the graph. We also analyze memory and retrieval costs. The multi-view and calibration analysis describes the optional reliability mode.

E.1 PROJECTIVE AMBIGUITY AND THE SL(4) PARAMETERIZATION

Ambient ambiguity and coordinate convention. For a static scene, exact correspondences, and a nondegenerate uncalibrated camera configuration, projective reconstruction determines structure only up to an invertible 3-D homography (Hartley & Zisserman, 2004). Thus two exact reconstructions can satisfy

$$\widetilde{\mathbf{X}}^{\,i} \sim \overline{\mathbf{H}}^{i}_{j}\widetilde{\mathbf{X}}^{\,j}, \qquad \overline{\mathbf{H}}^{i}_{j} \in \mathrm{GL}(4). \tag{11}$$

The symbol $\sim$ denotes equality up to homogeneous scale, leaving $16 - 1 = 15$ degrees of freedom. We use this space to align learned reconstructions, with prediction errors represented by graph residuals. Here $\mathbf{H}^{i}_{j}$ maps camera j into camera i, and $\mathbf{H}_i$ maps camera i into global coordinates. Positive-determinant transforms admit the normalization

$$\mathcal{N}(\overline{\mathbf{H}}) = \frac{\overline{\mathbf{H}}}{(\det \overline{\mathbf{H}})^{1/4}} \in \mathrm{SL}(4), \qquad \det \overline{\mathbf{H}} > 0. \tag{12}$$

Normalization preserves dehomogenized points and restricts the graph to the positive-determinant projective component. Structured measurements and continuous updates select consistent representatives of the equivalent maps $\mathbf{H}$ and $-\mathbf{H}$. SE(3) already has determinant one. Positive-scale similarities enter SL(4) through $\mathcal{N}$.

Structured measurements. Let $\mathbf{T}_i$ be the camera-to-window pose obtained by inverting the predicted world-to-camera extrinsics. Within a window the estimator adopts the predicted rigid-camera model, giving

$$\mathbf{H}^{i}_{j} = \mathbf{T}_i^{-1}\mathbf{T}_j. \tag{13}$$

For two reconstructions of the *same image*, use indices i and j for its target and source copies. At a shared pixel $\mathbf{p}$, $\mathbf{X}_i = d_i\mathbf{K}_i^{-1}\tilde{\mathbf{p}}$ and $\mathbf{X}_j = d_j\mathbf{K}_j^{-1}\tilde{\mathbf{p}}$. If $d_i = sd_j$ on the retained static support, then with $\mathbf{B}_{ij} = \mathbf{K}_i^{-1}\mathbf{K}_j$,

$$\mathbf{X}_i = s\mathbf{B}_{ij}\mathbf{X}_j, \qquad \mathbf{H}^{i}_{j} = \mathcal{N}\left(\begin{bmatrix}\mathbf{B}_{ij} & \mathbf{0}\\ \mathbf{0}^{\top} & s^{-1}\end{bmatrix}\right). \tag{14}$$

Eq. (14) gives the bridge in Eq. (7): the shared image aligns camera rays, and s aligns depth scale. The bridge follows VGGT-SLAM 2.0 (Maggio & Carlone, 2026) and replaces a general 15-DoF

edge fit with calibration compensation and one scale estimate. Spatially varying depth bias contributes to the bridge residual.

Global gauge and camera recovery. The graph is invariant to $\mathbf{H}_i \mapsto \mathbf{Q}\mathbf{H}_i$ for a common $\mathbf{Q} \in \mathrm{SL}(4)$. Fixing one node removes this redundancy in a connected graph. The global projection associated with a node is

$$\mathbf{P}_i = [\mathbf{K}_i \quad \mathbf{0}]\,\mathbf{H}_i^{-1}, \tag{15}$$

and global points follow by dehomogenizing $\mathbf{H}_i\widetilde{\mathbf{X}}_i$. A finite, nonsingular projection decomposes into calibration and pose in the chosen gauge. Metric interpretation relies on the learned camera prior, while absolute scale remains unobservable. We evaluate trajectories after a common Sim(3) alignment.

E.2 Depth-conditioned rigid consistency

From an epipolar line to a predicted pixel. For a matched pixel $\mathbf{p}_i$ with homogeneous coordinate $\tilde{\mathbf{p}}_i$, let $d_i = D_i(\mathbf{p}_i)$ and define the standard two-view camera relation (Hartley & Zisserman, 2004):

$$\mathbf{X}_1(d_1) = d_1\mathbf{K}_1^{-1}\tilde{\mathbf{p}}_1, \qquad \mathbf{X}_2(d_1) = \mathbf{R}_{21}\mathbf{X}_1(d_1) + \mathbf{t}_{21}, \tag{16}$$

$$\hat{\mathbf{p}}_{1\to2}(d_1) = \pi(\mathbf{K}_2\mathbf{X}_2(d_1))\,, \qquad \mathbf{F} = \mathbf{K}_2^{-\top}[\mathbf{t}_{21}]_\times\mathbf{R}_{21}\mathbf{K}_1^{-1}. \tag{17}$$

Here depths are positive optical-axis depths, $\pi([x,y,z]^\top) = [x/z, y/z]^\top$, and the reverse map $\hat{\mathbf{p}}_{2\to1}$ is defined by exchanging the two views. The epipolar line is $\boldsymbol{\ell}_2 = \mathbf{F}\tilde{\mathbf{p}}_1$, and the Sampson residual is

$$e_{\mathrm{epi}} = \frac{|\tilde{\mathbf{p}}_2^\top\mathbf{F}\tilde{\mathbf{p}}_1|}{\sqrt{(\mathbf{F}\tilde{\mathbf{p}}_1)_{1:2}^\top(\mathbf{F}\tilde{\mathbf{p}}_1)_{1:2} + (\mathbf{F}^\top\tilde{\mathbf{p}}_2)_{1:2}^\top(\mathbf{F}^\top\tilde{\mathbf{p}}_2)_{1:2}}}. \tag{18}$$

For a nondegenerate fundamental matrix, every point on $\boldsymbol{\ell}_2$ has zero numerator. Epipolar consistency therefore leaves one image degree of freedom, whereas the predicted depth selects a pixel. The gate uses the symmetric residual

$$\begin{aligned} r_{1\to2} &= \|\mathbf{p}_2 - \hat{\mathbf{p}}_{1\to2}(D_1(\mathbf{p}_1))\|_2\,, \qquad r_{2\to1} = \|\mathbf{p}_1 - \hat{\mathbf{p}}_{2\to1}(D_2(\mathbf{p}_2))\|_2\,, \\ r_{\mathrm{sym}} &= \sqrt{\tfrac{1}{2}(r_{1\to2}^2 + r_{2\to1}^2)}. \end{aligned} \tag{19}$$

Both transfers require valid positive projected depth. Exact cameras, depths, and static correspondences give $r_{\mathrm{sym}} = 0$, making this residual a direct test of agreement with the predicted rigid geometry.

Depth sensitivity and degeneracy. With $\mathbf{J}_\pi([x,y,z]^\top) = \begin{bmatrix} z^{-1} & 0 & -xz^{-2} \\ 0 & z^{-1} & -yz^{-2} \end{bmatrix}$, the depth sensitivity is

$$\frac{\partial\hat{\mathbf{p}}_{1\to2}}{\partial d_1} = \mathbf{J}_\pi(\mathbf{K}_2\mathbf{X}_2)\mathbf{K}_2\mathbf{R}_{21}\mathbf{K}_1^{-1}\tilde{\mathbf{p}}_1. \tag{20}$$

Writing $\mathbf{a} = \mathbf{K}_2\mathbf{R}_{21}\mathbf{K}_1^{-1}\tilde{\mathbf{p}}_1$ and $\mathbf{b} = \mathbf{K}_2\mathbf{t}_{21}$ gives, for component $u \in \{1,2\}$,

$$\frac{\partial\hat{p}_u}{\partial d_1} = \frac{a_u b_3 - b_u a_3}{(d_1 a_3 + b_3)^2}. \tag{21}$$

Translational parallax makes the predicted pixel sensitive to depth. Under pure rotation, $\mathbf{b} = \mathbf{0}$ and the derivative vanishes, so reprojection reduces to a rotation-induced homography. Zero-parallax rays have the same degeneracy, and weak parallax reduces sensitivity. The homography fallback verifies image alignment in these cases. The rigid reprojection gate still applies, but has no depth sensitivity at pure rotation. Bridge scale comes from repeated depth predictions of shared images (Eq. 25). Rotational image motion alone does not determine scale.

Scale invariance. Replacing $(D_1, D_2, \mathbf{t}_{21})$ by $(\gamma D_1, \gamma D_2, \gamma\mathbf{t}_{21})$, $\gamma > 0$, leaves both projections unchanged because $\pi(\gamma\mathbf{x}) = \pi(\mathbf{x})$. The gate therefore tests rigid consistency independently of the pair's common scale. It rejects moving matches that violate predicted motion. Coherent foreground motion and jointly biased camera and depth predictions remain sources of false acceptance. The optional checks below test additional views and prediction contexts.

E.3 MULTI-VIEW EVIDENCE AND FALSE-ACCEPT CALIBRATION

Distinct-view support. Represent the tested cross-neighbour pairs as a bipartite graph, with a frame on each vertex and a passing verification on each edge. Eq. (4) requires the central edge, at least m passing edges, and at least c incident vertices on each side. The required vertex coverage enforces support across distinct views. Shared images and joint inference correlate the edge decisions, so false-accept calibration operates on the combined candidate score.

Context intervention. Removing neighbouring frames while fixing the central images and patch indices isolates the effect of attention context. Eq. (5) compares match retention and rigid residuals across the two predictions. For nonempty match sets, $0 \le \kappa \le 1$, with $\kappa = 1$ whenever the smaller set is contained in the larger. Pairwise support thresholds and residual-change tests complement this overlap measure. The pair-only prediction supplies the graph measurement.

Finite-sample false-accept bound. Suppose the score function and candidate-selection policy are fixed before calibration and that $a_1, \dots, a_n, a_{\text{false}}$ are exchangeable scores of false proposals under that policy. With k_α and τ_α from Eq. (6),

$$\Pr(a_{\text{false}} < \tau_\alpha) \le \frac{k_\alpha}{n+1} \le \alpha. \tag{22}$$

Proof. Randomly order any tied scores. Exchangeability makes the rank of the future score uniform among $n+1$ positions. The event $a_{\text{false}} < a_{(k_\alpha)}$ implies that this rank is at most k_α, even with ties. For $k_\alpha = 0$, acceptance is impossible. The bound follows. Adding geometric or context gates only reduces the acceptance event. The proof follows the rank-calibration argument in conformal inference (Angelopoulos & Bates, 2021).

Eq. (22) bounds per-proposal false acceptance, marginally over calibration data and one future false proposal. Applying the bound under distribution shift requires exchangeable negatives across calibration and test data, including with sequence-disjoint splits. The reported strict-mode audit uses fixed gates, with optional held-out calibration described in Sec. 3.2.

E.4 ROBUST MULTI-ANCHOR SCALE ESTIMATION

Scale and calibration error. For anchor a, let $\mathcal{P}_a$ be the samples *retained after* validity checks and MAD filtering, and let $w_{a,p} > 0$ be their static weights. Define

$$x_{a,p} = \log \frac{\|\mathbf{X}_a^{k-1}(p)\|_2}{\|\mathbf{X}_a^{k}(p)\|_2}, \qquad \ell_a = \operatorname{wmed}_{p\in\mathcal{P}_a}(x_{a,p}; w_{a,p}), \tag{23}$$

$$\operatorname{wmed}(x_i; w_i) = \inf\left\{ x : \sum_i w_i \mathbf{1}[x_i \le x] \ge \tfrac{1}{2}\sum_i w_i \right\}. \tag{24}$$

The direction agrees with Eq. (14): s scales the current window into the preceding one. Writing $\mathbf{r}_a^k(p) = (\mathbf{K}_a^k)^{-1}\tilde{\mathbf{p}}$ gives

$$x_{a,p} = \log \frac{D_a^{k-1}(p)}{D_a^k(p)} + \log \frac{\|\mathbf{r}_a^{k-1}(p)\|_2}{\|\mathbf{r}_a^k(p)\|_2}. \tag{25}$$

Eq. (25) separates the depth-scale offset from calibration-ray bias. Matching calibration rays removes the second term. The tolerance below accounts for both sources of error.

Nested-median error bound. Proposition 1. Let $\ell_\star = \log s_\star$ be a common target log-scale. Suppose strictly more than half of the A usable anchors each have a subset $\mathcal{I}_a \subseteq \mathcal{P}_a$ satisfying

$$|x_{a,p} - \ell_\star| \le \epsilon \quad (p \in \mathcal{I}_a), \qquad \sum_{p\in\mathcal{I}_a} w_{a,p} > \tfrac{1}{2}\sum_{p\in\mathcal{P}_a} w_{a,p}. \tag{26}$$

Then the estimate in Eq. (8) obeys

$$|\ell^\star - \ell_\star| \le \epsilon, \qquad e^{-\epsilon} \le \frac{\exp(\ell^\star)}{s_\star} \le e^{\epsilon}. \tag{27}$$

Proof. For each qualifying anchor, cumulative weight is below half to the left of $[\ell_\star - \epsilon, \ell_\star + \epsilon]$ and above half at its right endpoint. The weighted median therefore lies in this interval. A strict majority

of anchor estimates share the interval, placing their ordinary median inside it as well. Exponentiation gives the scale bound.

The inner median aggregates *retained weight*, while the outer median gives each *anchor* equal influence regardless of pixel count. Static weighting and MAD filtering select the support on which the majority condition is evaluated. With the default $A = 2$, the bound requires both anchors to be accurate. Tolerating an arbitrary corrupted anchor requires at least three anchors with a reliable majority. Anchor MAD measures disagreement, while a shared bias enters the error tolerance.

Information budget for correlated bridges. Anchor estimates share two inference windows. Inflating each deviation by $\sqrt{A}$ normalizes the weight of this correlated evidence. For A locally quadratic factors with identical unwhitened Jacobian $\mathbf{J}$ and baseline deviation $\widetilde{\sigma}$, setting $\sigma_a = \sqrt{A}\,\widetilde{\sigma}$ gives

$$\sum_{a=1}^{A} \frac{\mathbf{J}^\top \mathbf{J}}{\sigma_a^2} = \frac{\mathbf{J}^\top \mathbf{J}}{\widetilde{\sigma}^2}. \tag{28}$$

The aggregate information then equals that of one factor. The identity describes identical factors before robust downweighting or noise clipping. The general case uses the same inflation as a weighting heuristic. Anchor disagreement and low static support further increase the assigned deviations.

E.5 Local SL(4) optimization and bounded residual scores

Local coordinates and residual Jacobians. The Lie algebra of the graph's ambient group is

$$\mathfrak{sl}(4) = \{\mathbf{A} \in \mathbb{R}^{4\times 4} : \operatorname{tr}(\mathbf{A}) = 0\}. \tag{29}$$

Let $\mathbf{E}_{ab}$ denote the off-diagonal matrix units and define $\mathbf{B}_1 = \operatorname{diag}(1,-1,0,0)$, $\mathbf{B}_2 = \operatorname{diag}(0,1,-1,0)$, and $\mathbf{B}_3 = \operatorname{diag}(0,0,1,-1)$. The fifteen generators $\{\mathbf{G}_k\}_{k=1}^{15}$ consist of the twelve $\mathbf{E}_{ab}$, $a \neq b$, and these three diagonal matrices. Thus

$$\boldsymbol{\xi}^\wedge = \sum_{k=1}^{15} \xi_k \mathbf{G}_k, \qquad \operatorname{Exp}(\boldsymbol{\xi}) := \exp(\boldsymbol{\xi}^\wedge), \qquad \boldsymbol{\xi} \in \mathbb{R}^{15}. \tag{30}$$

The identity $\det\exp(\mathbf{A}) = \exp(\operatorname{tr}\mathbf{A})$ keeps these increments in SL(4). Within a local logarithm chart, we use right updates $\mathbf{H} \leftarrow \mathbf{H}\operatorname{Exp}(\delta\boldsymbol{\xi})$. For a fixed measurement $\mathbf{Z}_{ij} := \mathbf{H}_j^i$, the residual in Eq. (9) is

$$\mathbf{E}_{ij} = \mathbf{Z}_{ij}^{-1}\mathbf{H}_i^{-1}\mathbf{H}_j, \qquad \mathbf{e}_{ij} = \operatorname{Log}(\mathbf{E}_{ij}) \in \mathbb{R}^{15}. \tag{31}$$

Here Log includes conversion from a traceless matrix to the chosen 15 coordinates. Write $\mathbf{J}_l^{-1}(\mathbf{e})$ and $\mathbf{J}_r^{-1}(\mathbf{e})$ for the differentials defined by $\operatorname{Log}(\operatorname{Exp}(\boldsymbol{\eta})\operatorname{Exp}(\mathbf{e})) \simeq \mathbf{e} + \mathbf{J}_l^{-1}(\mathbf{e})\boldsymbol{\eta}$ and its right-multiplication counterpart (Solà et al., 2018). Right node updates give

$$\mathbf{E}'_{ij} = \operatorname{Exp}(-\operatorname{Ad}_{\mathbf{Z}_{ij}^{-1}}\delta\boldsymbol{\xi}_i)\mathbf{E}_{ij}\operatorname{Exp}(\delta\boldsymbol{\xi}_j), \tag{32}$$

$$\mathbf{e}'_{ij} \simeq \mathbf{e}_{ij} - \mathbf{J}_l^{-1}(\mathbf{e}_{ij})\operatorname{Ad}_{\mathbf{Z}_{ij}^{-1}}\delta\boldsymbol{\xi}_i + \mathbf{J}_r^{-1}(\mathbf{e}_{ij})\delta\boldsymbol{\xi}_j. \tag{33}$$

The adjoint is defined by $(\operatorname{Ad}_{\mathbf{Z}}\boldsymbol{\eta})^\wedge = \mathbf{Z}\boldsymbol{\eta}^\wedge\mathbf{Z}^{-1}$. Thus the residual Jacobians are

$$\mathbf{J}_{ij,i} = -\mathbf{J}_l^{-1}(\mathbf{e}_{ij})\operatorname{Ad}_{\mathbf{Z}_{ij}^{-1}}, \qquad \mathbf{J}_{ij,j} = \mathbf{J}_r^{-1}(\mathbf{e}_{ij}). \tag{34}$$

The logarithm differentials account for the current residual and reduce to identities at zero residual. The factor Jacobians define the local robust least-squares system solved by Levenberg-Marquardt with a fixed graph gauge.

Bounded factor gradients. Let $\mathbf{u} = \boldsymbol{\Sigma}_{ij}^{-1/2}\mathbf{e}_{ij}$ and $r = \|\mathbf{u}\|_2$. For adjacent and loop factors, the Huber loss acts on the *vector norm* and has gradient

$$\rho_\delta(r) = \begin{cases} \frac{1}{2}r^2, & 0 \le r \le \delta, \\ \delta r - \frac{1}{2}\delta^2, & r > \delta, \end{cases} \qquad \nabla_{\mathbf{u}}\rho_\delta(\|\mathbf{u}\|_2) = \begin{cases} \mathbf{u}, & r \le \delta, \\ \delta\mathbf{u}/r, & r > \delta. \end{cases} \tag{35}$$

The score norm is bounded by δ. **Proposition 2.** For unwhitened residual Jacobian $\mathbf{J}_{ij}$ with respect to the free graph increments and positive-definite $\boldsymbol{\Sigma}_{ij}$, the factor gradient satisfies

$$\|\nabla_{\boldsymbol{\xi}}\rho_\delta(\|\mathbf{u}\|_2)\|_2 \le \delta\|\mathbf{J}_{ij}^\top\boldsymbol{\Sigma}_{ij}^{-1/2}\|_2 = \frac{\delta}{\sigma_{ij}}\|\mathbf{J}_{ij}\|_2 \quad \text{when } \boldsymbol{\Sigma}_{ij} = \sigma_{ij}^2\mathbf{I}_{15}. \tag{36}$$

Proof. Apply the chain rule to the whitened residual and multiply the score bound by the operator norm of $\mathbf{J}_{ij}^{\top}\boldsymbol{\Sigma}_{ij}^{-1/2}$.

Eq. (10) ensures $\sigma_{\text{loop}} \geq \sigma_{\min} > 0$. For fixed geometry, lower quality decreases both the quadratic information weight $1/\sigma_{\text{loop}}^2$ and the gradient bound $\delta\|\mathbf{J}_{ij}\|_2/\sigma_{\text{loop}}$. Quality weighting controls the strength of accepted factors, while Huber loss limits their residual scores during optimization. During local Levenberg-Marquardt optimization, trajectory updates depend on these factor gradients, graph conditioning, and consistency among factors. The assigned deviations act as weights in this nonconvex objective.

E.6 Memory and retrieval complexity

The compact interface stores dense geometry and retrieval history separately. For fp16 topology tokens and frame keys, the retrieval payload per submap is

$$B_k = 2(Md_{\text{r}} + N_k d_{\text{f}}) \text{ bytes.} \tag{37}$$

For $M = N_k = 16$, $d_{\text{r}} = 2048$, and $d_{\text{f}} = 1024$, this is 98,304 bytes, or 96 KiB. Graph state, intrinsics, depth samples, paths, and source images occupy additional storage. For K active submaps of P dense points, S archived submaps of at most N frames, and graph sets $\mathcal{V}, \mathcal{E}$, the principal memory terms are

$$\mathcal{M}_{\text{dense}} = O(KP), \qquad \mathcal{M}_{\text{retrieval}} = O(S(Md_{\text{r}} + Nd_{\text{f}})), \qquad \mathcal{M}_{\text{graph}} = O(|\mathcal{V}| + |\mathcal{E}|). \tag{38}$$

Fixed windows and eviction bound resident dense geometry. The compact archive and graph grow with the trajectory. Solver workspace and inference activations contribute additional peak memory.

For a query submap with N_q keys, exact evaluation of Eq. (1) compares every query key with every archived key. Eq. (2) then forms an $M \times M$ token-similarity matrix for each of the L shortlisted submaps. The respective dot-product costs are

$$C_{\text{coarse}} = O(SN_q N d_{\text{f}}), \qquad C_{\text{fine}} = O(LM^2 d_{\text{r}}). \tag{39}$$

Coarse scoring grows linearly with the archive for fixed window size. Fine scoring depends on the shortlist size L. Selection, sorting, and bounded-window verification add overhead. Retrieval payload, peak memory, and throughput measure these storage and computation costs.

F Limitations

Ω-SLAM relies on the camera and depth predictions of VGGT-Ω for submap alignment and loop verification. Under pure rotation, rigid reprojection loses depth sensitivity. Weak parallax also limits its ability to distinguish depth errors. Shared depth or calibration bias can affect multiple anchors, leaving their scale estimates mutually consistent but inaccurate. Loop verification faces a related difficulty in repeated structures and scenes dominated by coherent foreground motion, where incorrect predictions can remain geometrically consistent. Multi-view support improves dynamic-scene ATE, but raises error on 7-Scenes. Adding context stability further increases precision while slightly worsening ATE on all three datasets (Table 3). These results show that additional verification has different trajectory effects across scenes. At the system level, retaining a fixed number of dense submaps bounds resident dense-map storage, while the compact archive and factor graph continue to grow with the trajectory. Exact retrieval also becomes more expensive as the archive expands. Verification that accounts for correlated prediction errors, together with indexed retrieval and incremental graph optimization, offers directions for improving reliability and extending operation to longer sequences.

G Implementation details

Retrieval features. Frame keys average the 1024-dimensional image-encoder patch tokens before cross-frame attention and are L2-normalized. Cached topology features concatenate the 1024-dimensional outputs of the frame and inter-frame blocks, giving $d_{\text{r}} = 2048$. Partial cosine scoring uses $\rho = 0.25$. Production farthest-point sampling starts nearest the normalized register centroid.

Static evidence and scale support. For projected optical-axis depth z and nearest sampled target depth d, we use $r = |z - d| / \max(\min(z, d), 10^{-6})$ and $\tau_{\mathrm{s}} = 0.08$ in the Cauchy weight. Evidence is averaged over immediate temporal neighbours on a stride-two grid, then bilinearly upsampled. A projection with $z > 1.05d$ is treated as occluded. A relative depth jump above 0.10 marks a discontinuity. Samples without evidence retain unit static weight and undergo the same confidence and validity filtering. Dense export retains weights of at least 0.50. Scale fitting uses the geometric mean of the two shared-frame weights. With fewer than 100 samples above the static threshold, fitting retains the highest-weight available candidates, up to 100 points. Log-ratio outliers beyond 3.5 scaled MADs are removed. A zero MAD uses a 0.10 log-ratio tolerance.

Loop gates and factor weights. Retrieval retains three candidates. Mutual patch matches require cosine similarity at least 0.35 and at least 12 matches. Epipolar and homography support each require an inlier ratio of 0.60, at 16- and 12-pixel thresholds, respectively. The selected 2-D model is checked jointly with rigid reprojection: at least 25% of mutual matches must meet the 24-pixel threshold, and at least eight must also be 2-D inliers. Strict mode requires two supported pairs covering two frames per side, context overlap of 0.50, inlier-ratio change at most 0.25, and residual change at most 16 pixels. All pixel thresholds use inference-image coordinates.

Using the normalized margin m from Sec. 3.4, the pairwise quality is

$$\begin{aligned} q = {} & 0.15m(s_{\mathrm{r}}; 0.35) + 0.15m(s_{\mathrm{p}}; 0.35) + 0.30m(v_{\mathrm{g}}; 0.60) \\ & + 0.25m(v_{\mathrm{d}}; 0.25) + 0.15 \min(n/36, 1). \end{aligned} \tag{40}$$

Here $s_{\mathrm{r}}, s_{\mathrm{p}}$ are retrieval and mean patch similarities, $v_{\mathrm{g}}, v_{\mathrm{d}}$ are 2-D and rigid inlier ratios, and n counts mutual matches. We set $(\sigma_{\min}, \sigma_{\max}) = (0.03, 0.20)$. For accepted strict-mode loops, the coverage gates are satisfied and q is multiplied by $[0.75 + 0.25 \min(n_{\mathrm{pair}}/3, 1)](0.5 + 0.5\kappa)$, where n_{pair} counts supported pairs.

Adjacent bridges use

$$\sigma_{\mathrm{adj}} = \min\Big(0.20, (0.05 + 2\sigma_{\ell})\sqrt{A}(2 - \bar{w})\Big). \tag{41}$$

Here σ_{ℓ} combines within-anchor uncertainty and between-anchor dispersion in quadrature, floored at 10^{-6}. Here $\bar{w}$ is the mean static support. Each loop-submap endpoint shares one image with its original submap, giving a single-anchor scale estimate. The query attachment is quadratic with deviation 0.05, as are within-window factors. Adjacent and revisit factors use Huber $\delta = 1.345$, with the latter weighted by $\sigma_{\mathrm{loop}}(q)$. The first-node prior has deviation 10^{-6}.