RECALL-SLAM: LONG-HORIZON RGB-D SLAM THROUGH RECURRING LOCAL RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth provides RGB-D SLAM with explicit geometry and metric scale, yet brief blur, texturelessness, weak overlap, or corrupted depth can still break an otherwise long trajectory. This exposes a weakness in local multi-view geometry. More broadly, this local capability is required not only by the frontend but throughout SLAM. RGB-D SLAM can be viewed as repeatedly solving the same local problem at increasing map scales: nearby views recover camera motion and local structure, wider observations reconcile geometry over a region, and views around a place revisited much later relate distant parts of the trajectory. Each task jointly recovers cameras and scene structure from a bounded set of spatially related RGB-D observations, then applies the estimate at the appropriate map scale. This suggests that a long-horizon system should make a strong local solver available across its hierarchy. Based on this insight, we present RECALL-SLAM, a training-free system that retains observation history and repeatedly poses these bounded reconstruction problems to one pretrained Pi3X model. Every model call natively consumes RGB, calibrated depth, and camera intrinsics and predicts cameras, dense geometry, and confidence. Bounded regional graphs and a compact global graph integrate these predictions into a persistent, revisable map while all levels advance asynchronously. Experiments demonstrate stable, coherent reconstruction on sequences extending to tens of thousands of frames; RECALL-SLAM also improves the FastCaMo-Real F-score by 20.5 percentage points and reduces Oxford Spires ATE by 58.3%. Project page: https://anonymous-submission-x.github.io/ReCallSLAM/.

1 INTRODUCTION

Depth provides RGB-D SLAM with explicit geometry and metric scale, yet real capture remains vulnerable to brief local degradation. Textureless views, rapid motion blur, weak overlap, or corrupted depth can make a small set of observations difficult to relate. A few such episodes can break an otherwise long trajectory, limiting reliable deployment even when metric depth is available.

These failures expose a weakness in local multi-view geometry. More broadly, the same capability is required throughout a SLAM system, not only by its frontend. Nearby observations must recover camera motion and local structure; observations spanning a larger region must reconcile its geometry; views around a place revisited much later must estimate a relation between distant parts of the trajectory. Their temporal spans and effects on the map differ, but the underlying problem is the same: jointly recover cameras and scene structure from a bounded set of spatially related observations. We call each such problem a *geometry primitive*. A strong solver for this primitive should therefore remain available across the SLAM hierarchy.

Feed-forward geometry models provide a strong solver for this primitive. DUSt3R, MASt3R, VGGT, and π^3 infer cameras, pointmaps, depth, or correspondences directly from a view set using learned geometric priors (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026b). This suggests a system principle: preserve the model's strong bounded-view reasoning and invoke it wherever the geometry primitive recurs. The surrounding SLAM system should retain history, select coherent observations, and apply each answer at the appropriate map scale.

Existing feed-forward SLAM systems already demonstrate the value of learned geometric priors. MASt3R-SLAM builds matching, tracking, loop closure, and optimization around a two-view pri-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

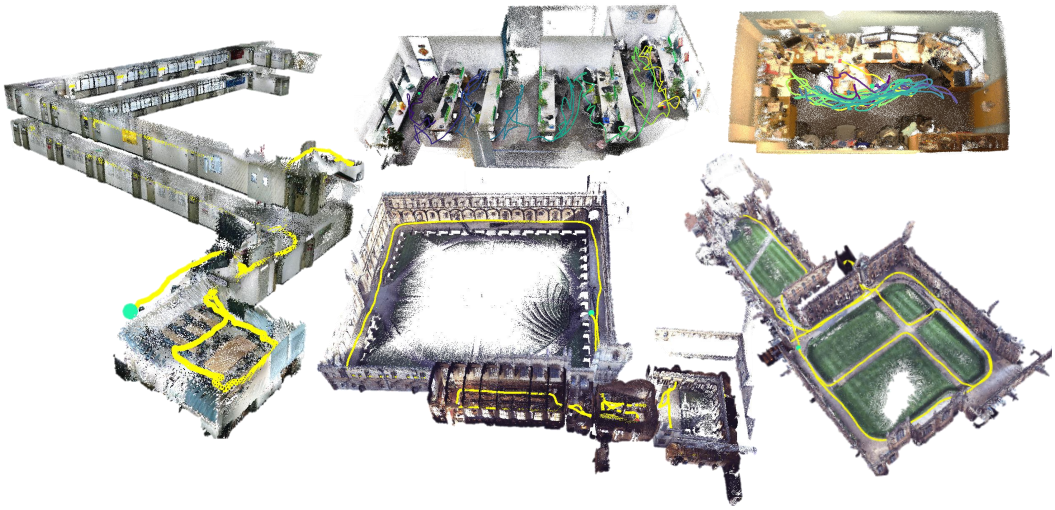

Figure 1: Long-horizon RGB-D maps produced by RECALL-SLAM on indoor and outdoor sequences with frequent revisits. The colored curves show the estimated camera trajectories. The examples motivate a system that repeatedly solves bounded local geometry while retaining a revisable global map over long streams.

or; PROFusion couples feed-forward geometry with RGB-D fusion; VGGT-SLAM and VGGT-SLAM 2.0 align feed-forward submaps in explicit graphs; and VGGT-Long reconstructs loop-centric chunks before registering them to the accumulated map (Murai et al., 2025; Dong et al., 2026; Maggio et al., 2025; Maggio & Carlone, 2026; Deng et al., 2026). Their interfaces commonly specialize the learned model to frontend estimation or individual chunks, while higher-level relations use additional matching, registration, or fusion procedures. This leaves an opportunity to expose the same feed-forward multi-view capability directly to every scale of SLAM.

We introduce RECALL-SLAM around three consistently defined levels. *Level 0: local* applies an RGB-D-conditioned Pi3X solver to overlapping local windows, producing the streaming trajectory, local geometry, and extended observations. *Level 1: context* organizes longer-span RGB-D view sets and uses their predictions to revise the internal arrangement of bounded submaps. *Level 2: submap* retrieves revisited places, jointly reconstructs views from both visits, and uses the fused observation to coordinate submap placement. A submap is a bounded group of local rigid pieces whose internal structure remains revisable. At every level, depth is part of the model input together with RGB and calibration; the resulting camera and geometry predictions become constraints in the corresponding graph. The levels advance asynchronously as observations arrive, so the system continually incorporates model predictions into a persistent map.

Experiments show accurate metric trajectories and stable, coherent reconstruction on sequences extending to tens of thousands of frames. RECALL-SLAM improves FastCaMo-Real F-score by 20.5 percentage points over PROFusion and reduces Oxford Spires ATE by 58.3% on the sequences completed by ORB-SLAM3 RGB-D. Our contributions are threefold. First, we formulate local, context, and revisit estimation through one bounded multi-view geometry primitive, providing a common account of the geometric capability required throughout RGB-D SLAM. Second, we realize this formulation as a training-free streaming system that uses one RGB-D-conditioned model at all three levels and integrates its predictions through bounded intra-submap and compact inter-submap graphs. Third, we evaluate the complete system on short and long indoor sequences, difficult real capture, and large outdoor routes using trajectory, reconstruction, and qualitative geometric evidence.

2 RELATED WORK

2.1 TRADITIONAL RGB-D SLAM

RGB-D SLAM combines metric depth with explicit estimation and mapping. KinectFusion established real-time dense tracking and volumetric fusion (Newcombe et al., 2011). Feature-based sys-

tems such as RGB-D SLAM and ORB-SLAM2 use image evidence, depth constraints, keyframes, and pose-graph or bundle adjustment (Endres et al., 2012; Mur-Artal & Tardós, 2017). Dense systems directly align intensities or surfaces: ElasticFusion maintains a deformable surfel map, BundleFusion jointly optimizes poses and reintegrates geometry, and BAD SLAM performs direct bundle adjustment with dense depth constraints (Whelan et al., 2015; Dai et al., 2017; Schöps et al., 2019). Neural-rendering systems instead optimize poses together with implicit, point-based, or Gaussian scene representations (Sucar et al., 2021; Zhu et al., 2022; Wang et al., 2023; Sandstr"om et al., 2023; Tang et al., 2023; Keetha et al., 2024; Yugay et al., 2023; Zhong et al., 2025).

These approaches obtain strong estimates when their matching, registration, or rendering objectives remain well constrained, but difficult local observations can affect tracking, regional correction, and loop estimation alike. RECALL-SLAM complements the explicit state and optimization structure of SLAM with a learned multi-view solver, and makes that local geometric capability available throughout the hierarchy rather than at one estimation stage.

2.2 FEED-FORWARD RECONSTRUCTION AND CONTEXT EXTENSION

Feed-forward models directly predict cameras and geometry from one or more views, including pointmap methods such as DUSt3R and MASt3R, joint camera-and-geometry predictors such as VGGT and π^3, and recent metric or modality-conditioned models including MapAnything, OmniVGGT, Depth Anything 3, and DAGE (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026b; Keetha et al., 2026; Peng et al., 2025; Lin et al., 2025; Ngo et al., 2026). Pi3X adds calibrated RGB-D conditioning to the released π^3 model (Wang et al., 2026a). These models make a bounded view set a powerful reconstruction unit; long-sequence methods then differ mainly in how they extend or organize that unit.

Spann3R and Point3R maintain explicit spatial memories, whereas CUT3R compresses history into a recurrent state (Wang & Agapito, 2024; Wu et al., 2025; Wang et al., 2025b). LingBot-Map uses a geometric context transformer, ABot-Recon deliberately keeps learned context local, TTT3R adapts recurrent updates at test time, and LoGeR combines short- and long-term memory (Chen et al., 2026a; Han et al., 2026; Chen et al., 2026b; Zhang et al., 2026a). These designs substantially improve streaming reconstruction, although preserving consistent old and new geometry remains difficult as context grows or is compressed. RECALL-SLAM keeps long-term state in an explicit SLAM hierarchy and repeatedly reconstructs selected observations with a fixed pretrained model. This requires no recurrent training or test-time parameter update and allows stored views to be reconsidered when new geometric evidence arrives.

2.3 FEED-FORWARD RECONSTRUCTION WITHIN SLAM SYSTEMS

A growing set of systems connects feed-forward geometry to persistent maps. SLAM3R predicts local pointmaps with feed-forward registration, while PROFusion combines feed-forward geometry with an RGB-D fusion and optimization pipeline (Liu et al., 2025; Dong et al., 2026). MASt3R-SLAM builds tracking, fusion, loop closure, and optimization around pointmap matching (Murai et al., 2025). VGGT-SLAM and VGGT-SLAM 2.0 optimize feed-forward submaps in explicit graphs, with the latter also using internal attention for revisit verification (Maggio et al., 2025; Maggio & Carlone, 2026). VGGT-Long, UniSim-SLAM, and AIM-SLAM study loop-centric chunks, unified Sim(3) optimization, and informative multi-view selection (Deng et al., 2026; Lee et al., 2026; Jeon et al., 2026). ScaRF-SLAM uses the reverse allocation: conventional visual SLAM estimates the trajectory and a feed-forward model reconstructs geometry (Zhang et al., 2026b).

These systems effectively combine learned geometry with long-term estimation, but their model interfaces are usually specialized to frontend motion, a local chunk, or final reconstruction. Higher-level relations may therefore return to separate matching, registration, or fusion procedures. RECALL-SLAM instead instantiates the same RGB-D geometry primitive at the *local* (Level 0), *context* (Level 1), and *submap* (Level 2) levels; the hierarchy changes the selected views and optimized state, while the feed-forward solver and its geometric output remain common.

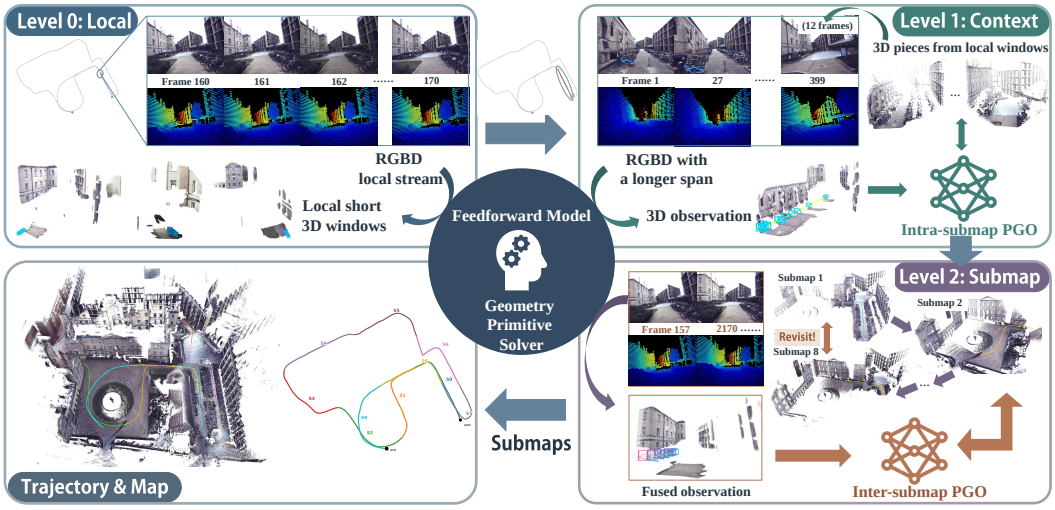

Figure 2: Overview of RECALL-SLAM. Every level forms a bounded set of calibrated RGB-D views, invokes the same pretrained Pi3X geometry solver, and routes its predicted cameras and geometry to the state at that level. The local level (Level 0) produces the streaming trajectory, local reconstruction, and extended observations. The context level (Level 1) uses longer-span predictions to correct provisional geometry and constrain intra-submap optimization. The submap level (Level 2) jointly reconstructs revisit views into fused observations that drive inter-submap optimization.

3 METHOD

3.1 SYSTEM OVERVIEW

RECALL-SLAM receives a calibrated RGB-D stream and uses one fixed feed-forward model as a *geometry primitive solver*. Each call consumes RGB, depth, and camera intrinsics, jointly predicts cameras and scene geometry, and returns confidence for the predicted evidence. The SLAM hierarchy decides which views form a call and which state its answer constrains. This separation lets the same learned multi-view capability act at three scales without asking one model invocation to retain the complete history.

Figure 2 summarizes the three levels. *Level 0: local* reconstructs overlapping local windows to extend the trajectory and collect supporting observations. *Level 1: context* uses wider temporal and regional contexts both to correct provisional geometry and to constrain bounded intra-submap optimization. *Level 2: submap* retrieves revisited places, jointly reconstructs views from both visits, and coordinates the submaps in a global graph. A model answer therefore becomes a local camera estimate, an intra-submap observation, or a fused inter-submap observation according to the level that posed the problem.

3.2 LOCAL GEOMETRY RECONSTRUCTION AS A PRIMITIVE

We define a *geometry primitive* as the joint reconstruction of a bounded, spatially related set of calibrated RGB-D observations. For an instance q, let

$$\mathcal{V}_q = \{(\mathbf{I}_i, \mathbf{D}_i, \mathbf{K}_i)\}_{i \in \mathcal{I}_q} \tag{1}$$

contain RGB image $\mathbf{I}_i$, measured depth $\mathbf{D}_i$, and camera intrinsics $\mathbf{K}_i$. The views need not be adjacent in time; they are local because they observe a common region that admits one bounded reconstruction. The pretrained solver Φ_θ predicts cameras, dense geometry, and confidence,

$$\Phi_\theta(\mathcal{V}_q) = \{\widehat{\mathbf{T}}_{q,i}, \widehat{\mathbf{X}}_{q,i}, \mathbf{c}_{q,i}\}_{i \in \mathcal{I}_q}. \tag{2}$$

We instantiate Φ_θ with Pi3X (Wang et al., 2026a;b). Pi3X natively conditions its joint prediction on RGB, measured depth, and calibration; depth is therefore part of the inference problem rather than a scale-only postprocessing signal. A small multiplicative bias can nevertheless remain in

the predicted metric gauge. As the final step of each primitive solve, we robustly estimate one scalar from valid measured and predicted depths and apply it jointly to the predicted translations and pointmaps. The resulting cameras and geometry remain one RGB-D-conditioned multi-view prediction in metric units.

The solver output is preserved together with the selected frames, calibration, confidence, and geometric consistency tests. Admission converts a valid prediction into camera observations with uncertainty derived from this evidence; rejection preserves the same record without allowing it to constrain the map. This distinction is important because overlapping primitive instances may support, refine, or contradict an earlier estimate.

3.3 Recursive SLAM hierarchy

Every accepted primitive answer is connected to a persistent geometric state. Repeatedly compatible Level 0 observations partition the stream into rigid *pieces*; each piece stores fixed camera offsets after context-level correction. A bounded group of neighboring pieces forms a *submap*, whose internal shape can still be revised. If frame i belongs to piece $p(i)$ and submap $s(i)$, its world pose is

$$\mathbf{T}_i^W = \mathbf{G}_{s(i)}\,\mathbf{P}_{p(i)}\,\mathbf{C}_i, \tag{3}$$

where $\mathbf{C}_i$ is the fixed pose inside its piece, $\mathbf{P}_p$ places a piece in its submap, and $\mathbf{G}_s$ places a submap in the world. A primitive instance has an auxiliary coordinate frame $\mathbf{Q}_q$; its prediction contributes robust constraints of the form $\mathbf{Q}_q\widehat{\mathbf{T}}_{q,i} \simeq \mathbf{P}_{p(i)}\mathbf{C}_i$. Level 1 updates the bounded piece poses and primitive frames, while Level 2 holds each internal shape fixed and updates only submap poses. This common interface lets answers acquired at different times remain usable as the map is revised.

3.3.1 Level 0: local

Level 0 reconstructs short overlapping windows as RGB-D geometry-primitive instances. Shared cameras place each new prediction in the gauge of the existing trajectory: the overlap determines one rigid transform, which is then applied consistently to the newly predicted cameras and pointmaps. Accepted cameras extend a provisional metric trajectory, while joint inference over the entire window provides multi-view support when an individual frame is blurred, textureless, or geometrically weak. Disagreement in the shared views prevents an unsupported window from silently redefining the local trajectory.

A dense-wide branch applies the same solver to longer continuous windows as their frames arrive. These predictions do not overwrite or gate the short-window trajectory. Instead, agreement among short and dense-wide observations provides evidence for rigid piece boundaries, and the accepted answers remain available to later optimization. Level 0 thus outputs both a streaming local estimate and a traceable pool of geometric observations.

3.3.2 Level 1: context

Wide-span correction. Level 1 first selects sparse RGB-D views across a longer interval than Level 0. A validated Pi3X prediction applies one continuous correction to the provisional trajectory, defines the fixed camera offsets $\mathbf{C}_i$ within the resulting pieces, and remains a graph observation.

Regional estimation. Parent questions summarize broader regions, while context questions connect separated parts or provisional boundaries. They are scheduled from available observations before piece association; after inference, a prediction spanning several pieces constrains their relative placement, while one contained within a piece cannot alter its fixed internal geometry.

Intra-submap optimization. The bounded private PGO combines Level 0 observations, accepted wide-span corrections, parent/context predictions, and relevant revisit evidence. It optimizes piece poses $\mathbf{P}_p$ and primitive frames $\mathbf{Q}_q$ while keeping $\mathbf{C}_i$ fixed. Residuals compare each predicted camera with the same camera composed through its piece, weighted by confidence and RGB-D consistency. One piece fixes the gauge, external pieces provide boundary conditions, and measured horizontal surfaces add soft gravity and height constraints.

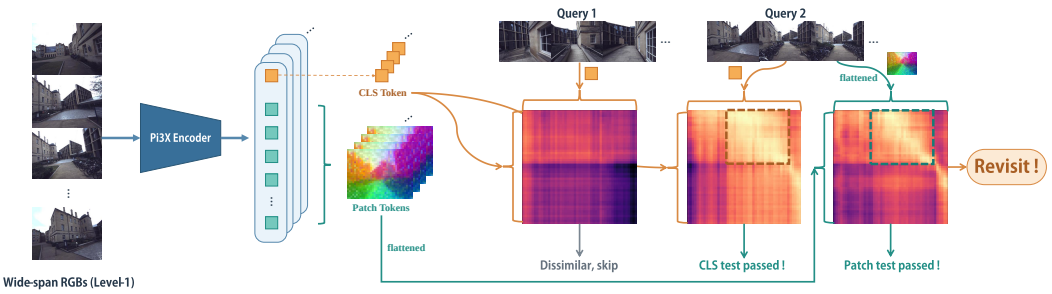

Figure 3: Level 2 revisit detection. Pi3X CLS descriptors retrieve historical candidates; spatially pooled patch descriptors verify multi-frame layout and localize the matching windows at both endpoints. Detection does not estimate their transformation.

Using $\mathbf{\Omega}_{q,i}$ for the information of an admitted camera observation, the central private objective is

$$E_1(\mathbf{P}, \mathbf{Q}) = \sum_q \sum_{i \in \mathcal{I}_q} \rho\left(\left\| \mathrm{Log}\left[(\mathbf{Q}_q \widehat{\mathbf{T}}_{q,i})^{-1} \mathbf{P}_{p(i)} \mathbf{C}_i \right] \right\|^2_{\mathbf{\Omega}_{q,i}} \right). \tag{4}$$

Additional depth constraints enter the same objective. Optimizing $\mathbf{Q}_q$ absorbs each prediction's local gauge, while optimizing $\mathbf{P}_p$ transfers consistent multi-view relations into the submap shape.

3.3.3 Level 2: submap

Revisit detection. For each stable submap, frames are sampled across its non-overlapping core to form a retrieval bank (Figure 3). Each sampled RGB image passes through the fixed Pi3X encoder: its normalized CLS token is the image descriptor for coarse retrieval, while spatial-pyramid pooling of patch tokens provides a second descriptor that retains coarse image layout. No separately trained place recognizer or matcher is used. After excluding nearby history, bidirectional cosine coverage ranks submap pairs. High-scoring one-to-one frame matches must have reciprocal or temporally ordered multi-frame support and agree under the patch descriptor. Supported matches are grouped into bounded intervals in both submaps; these paired intervals are the matching windows passed to geometric estimation. All rejected pairs remain indexed.

Joint revisit estimation. From each matching-window pair, Level 2 selects bounded primary RGB-D views at both endpoints and jointly reconstructs their union in one Pi3X call. This directly places both visits in a shared frame rather than registering separate chunks. Disjoint validation views and reversed input order test the result. An accepted answer becomes a fused observation: its endpoint cameras can refine the private graphs, and its induced submap relation enters the global graph. Rejections remain linked to the original detection.

Alternating optimization. Accepted fused observations enter the inter-submap PGO, an SE(3) graph over submap poses $\mathbf{G}_s$; sequential edges retain connectivity and revisit edges distribute long-range corrections. Inter- and intra-submap PGO alternate over shared evidence. The global step fixes internal shapes, updates submap placement, and passes these placements to Level 1 as boundary conditions. The private step fixes those boundaries, refines the piece arrangement, and returns the updated shapes to the next global step. Individual cameras never become global variables, and Pi3X need not be rerun at every cycle.

For a sequential or fused relation $\mathbf{Z}_{ab}$ between submaps a and b, the global objective has the compact form

$$E_2(\mathbf{G}) = \sum_{(a,b) \in \mathcal{E}} \rho\left(\left\| \mathrm{Log}\left[\mathbf{Z}_{ab}^{-1} \mathbf{G}_a^{-1} \mathbf{G}_b \right] \right\|^2_{\mathbf{\Lambda}_{ab}} \right), \tag{5}$$

where $\mathbf{\Lambda}_{ab}$ reflects the support and consistency of the underlying primitive answer. Robust estimation limits the influence of a geometrically plausible but conflicting revisit while retaining it in the observation record.

For ambiguous cross-floor revisits, ReCall-SLAM may retain verified rotation and horizontal translation while leaving vertical displacement unobserved, using depth-derived gravity as the common axis rather than imposing an unsupported full SE(3) loop.

3.4 Asynchronous streaming execution

The hierarchy advances through three concurrent activities. Frame arrival immediately schedules Level 0 reconstruction and publishes its observations. Once sufficient temporal support exists, Level 1 contexts and Level 2 revisit proposals become eligible without waiting for the sequence to end. Each accepted answer is committed to the relevant observation graph, whose CPU optimizer can reuse all available evidence while later model calls are pending. A single shared Pi3X instance serializes neural inference, while input preparation, validation, private optimization, and global optimization overlap on the CPU. The result is a continuously revisable stream rather than an offline assembly of independently reconstructed chunks.

4 Experiments

We evaluate metric trajectory accuracy, indoor reconstruction quality, robustness to degraded capture, large-scale outdoor tracking, and the contribution of the three-level hierarchy.

4.1 Benchmarks and evaluation protocol

Benchmarks. TUM RGB-D (Sturm et al., 2012) provides 12 standard indoor tracking sequences. Seven BundleFusion sequences (Dai et al., 2017) evaluate trajectory and reconstruction during long indoor capture with frequent revisits. The 12 FastCaMo-Real sequences (Tang et al., 2023) evaluate reconstruction under severe fast-motion blur. Oxford Spires (Tao et al., 2025) evaluates large-scale outdoor trajectories. Dataset selection and processing are detailed in Appendix D.

Baselines. RGB-only comparisons include LingBot-Map and ABot-Recon (Chen et al., 2026a; Han et al., 2026). RGB-D comparisons include 2DGS-SLAM, PROFusion, VGGT-SLAM 2.0 under matched RGB-D conditions, and ScaRF-SLAM (Zhong et al., 2025; Dong et al., 2026; Maggio & Carlone, 2026; Zhang et al., 2026b). PROFusion connects feed-forward geometry to an RGB-D fusion pipeline. ScaRF-SLAM makes the opposite allocation to ours: ORB-SLAM3 RGB-D supplies its trajectory, while a feed-forward backend reconstructs the map. We therefore report ORB-SLAM3 for its trajectory and the complete ScaRF-SLAM pipeline only for reconstruction. Baseline variants are specified in Appendix D.

Metrics. We report ATE for TUM, BundleFusion, and Oxford trajectories, and precision, recall, and F-score at 5 cm for BundleFusion and FastCaMo-Real surfaces. RGB-only outputs use scale alignment; RGB-D outputs preserve their native scale and use only rigid alignment.

4.2 Indoor trajectory and reconstruction benchmarks

Table 1 combines the three indoor benchmarks. RGB-only and RGB-D trajectories use Sim(3) and SE(3) alignment, respectively. ORB-SLAM3 supplies ScaRF-SLAM's trajectory; its marked means include incomplete prefixes and therefore favor ORB-SLAM3 (10/12 TUM and 2/7 BundleFusion runs reach the final frame). The ScaRF-SLAM rows contain only reconstructions. Pale red and yellow mark the best and second-best results.

ReCall-SLAM achieves the lowest TUM ATE and trails VGGT-SLAM 2.0-Pi3X by only 0.30 cm on BundleFusion. Although BundleFusion contains long trajectories and frequent revisits, its motion remains inside compact rooms with strong spatial overlap, favoring chunk-based alignment. On wider-area motion, ReCall-SLAM is clearly stronger: over the 11 complete Oxford routes, it reduces ATE from VGGT-SLAM 2.0-Pi3X's 510.43 to 76.11 cm and retains more coherent global structure in the long-sequence comparisons. It also raises FastCaMo-Real F-score from PROFusion's 46.29% to 66.82%.

4.3 Outdoor large-scale evaluation

Our Oxford Spires selection starts from 13 captures. Eleven have continuous usable depth and are evaluated as complete sequences. The remaining two contain depth gaps longer than one second

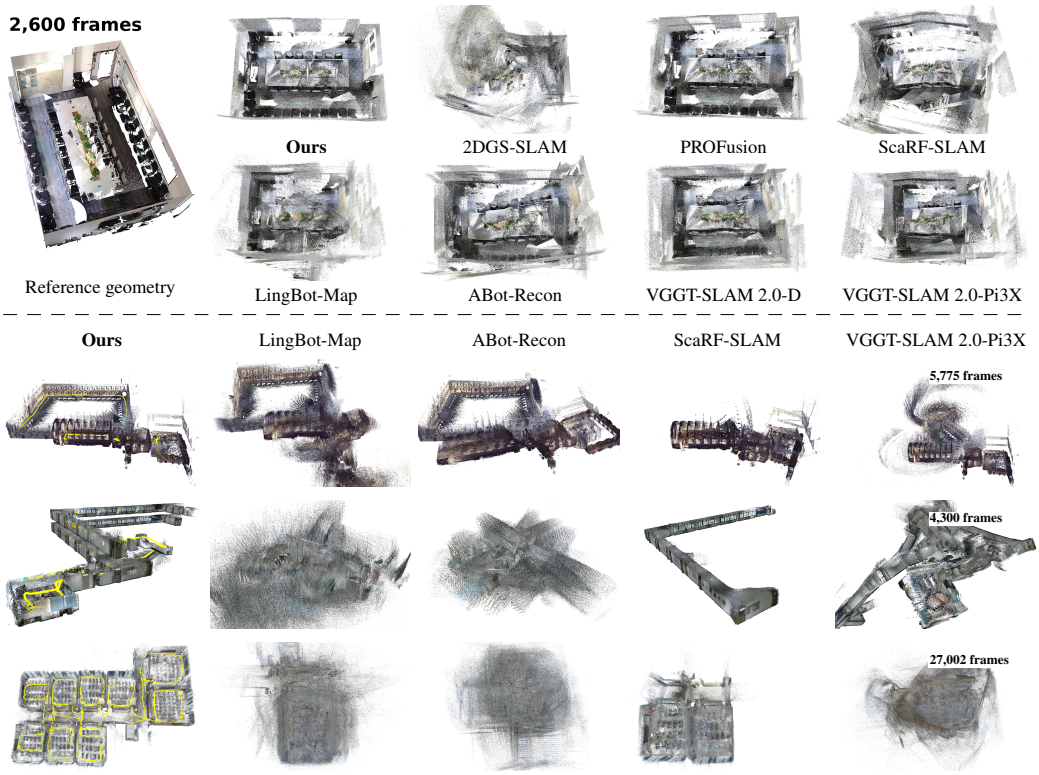

Figure 4: Qualitative reconstruction comparison. The upper panel places the FastCaMo-Real meeting-room reference beside eight reconstructions over 2,600 frames. The lower rows show Oxford Christ Church 2 (5,775 frames), a multi-floor corridor (4,300 frames), and classrooms (27,002 frames). Each row uses a shared fixed view and identical display processing across methods.

Table 1: Indoor comparison on TUM RGB-D, BundleFusion, and FastCaMo-Real. Trajectory ATE is in cm; reconstruction precision, recall, and F-score are percentages at a 5 cm threshold. Lower ATE and higher F-scores are better. † denotes a mean that includes incomplete returned trajectories.

	Trajectory ATE ↓		BundleFusion reconstruction			FastCaMo-Real reconstruction		
Method	TUM	BundleFusion	P@5 ↑	R@5 ↑	F@5 ↑	P@5 ↑	R@5 ↑	F@5 ↑
LingBot-Map	5.949	59.432	26.36	71.95	38.40	12.63	28.91	17.14
ABot-Recon	17.727	41.179	20.88	69.94	32.12	14.00	25.42	17.88
2DGS-SLAM	25.890	198.755	8.05	34.88	12.71	12.15	19.52	14.41
PROFusion	14.017	47.556	21.75	48.69	29.88	45.65	47.68	46.29
ORB-SLAM3 RGB-D	7.097†	8.680†	–	–	–	–	–	–
ScaRF-SLAM (sensor D)	–	–	68.55	89.92	77.68	33.30	24.44	26.77
ScaRF-SLAM (DA3)	–	–	58.15	80.91	67.54	22.05	25.67	21.56
VGGT-SLAM 2.0-D	4.316	8.484	67.35	90.78	77.29	17.22	28.92	21.22
VGGT-SLAM 2.0-Pi3X	3.165	**6.504**	**78.79**	**91.92**	**84.82**	33.42	34.73	33.67
ReCall-SLAM	**2.969**	6.806	73.71	89.02	80.63	**71.07**	**64.15**	**66.82**

and are split at those gaps into 24 independently evaluated segments; full protocols appear in Appendix D. Among methods producing complete outputs, RECALL-SLAM has the lowest ATE on both groups, with 76.11 cm over the 11 complete sequences and 16.01 cm over the 24 segments.

ORB-SLAM3 terminates early on every complete sequence and on part of the segmented set, so averaging its returned prefixes against complete outputs would be misleading. We instead compare it only on the common cases that ORB-SLAM3 finishes, all of which are segments. On this matched subset, RECALL-SLAM reduces ATE from 36.96 to 15.43 cm, a 58.3% reduction. Table 2 keeps this comparison separate from the complete-output means.

Table 2: Oxford Spires trajectory ATE in cm over 11 complete sequences, 24 segments, and the 14-segment subset completed by ORB-SLAM3; lower is better.

Method	Complete ↓	Segments ↓	ORB subset ↓
LingBot-Map	494.21	29.56	–
ABot-Recon	383.14	20.14	–
ORB-SLAM3 RGB-D	–	–	36.96
VGGT-SLAM 2.0-D	297.03	22.23	–
VGGT-SLAM 2.0-Pi3X	510.43	17.26	–
ReCall-SLAM	**76.11**	**16.01**	**15.43**

Table 3: Hierarchy ablation on 12 TUM RGB-D sequences. Entries are mean trajectory ATE RMSE in cm after rigid SE(3) alignment; lower values are better.

Variant	ATE ↓
Level 0	5.583
Level 0 + Level 1	3.459
Level 0 + Level 2	3.004
Full	**2.969**
Full without intra-submap PGO	3.280

4.4 Long-sequence qualitative results

Figure 4 shows reconstruction behavior that short-sequence trajectory averages do not capture directly. The FastCaMo-Real case tests recovery under blur and weak overlap; the Oxford route, multi-floor corridor, and 27,002-frame classroom sequence test repeated traversal and preservation of large connected structure. Across shared views, ReCall-SLAM retains recognizable regional geometry and global arrangement, whereas several baselines produce locally plausible surfaces but collapse or separate distant regions. Figure 1 provides additional final maps and trajectories.

4.5 Ablation

Table 3 isolates the three levels on TUM RGB-D. Level 1 and Level 2 independently reduce the Level 0 ATE from 5.583 cm to 3.459 and 3.004 cm, and the full hierarchy reaches 2.969 cm. In the final variant, Level 1 long-span predictions directly reassemble the pieces but the intra-submap pose graph is removed; its 3.280 cm error shows the additional value of jointly reconciling those observations inside each submap.

5 Limitations and conclusion

ReCall-SLAM inherits failures from its RGB-D input and pretrained geometry solver. Missing or miscalibrated depth, blur, dynamics, weak overlap, and distribution shift can still produce incorrect geometry, and reuse of one solver can correlate errors across levels. The hierarchy also stores observations and spends computation on repeated inference, retrieval, and optimization; storage grows with mapped area, while compact submap variables limit how freely corrections propagate across boundaries. Confidence checks, retained negative proposals, robust graph objectives, and qualitative inspection reduce these risks without eliminating them. The evaluation further mixes trajectory references, reconstructed surfaces, and surface-only supervision, so no single metric captures every long-horizon failure.

Within these limits, ReCall-SLAM shows that a feed-forward model can serve as a common RGB-D geometry solver throughout SLAM. Local windows, wider contexts, and revisited regions are posed as the same bounded reconstruction primitive, while intra-submap and inter-submap graphs determine where each prediction acts. This design produces accurate indoor trajectories, a 20.5-point FastCaMo-Real F-score gain over PROFusion, complete trajectories on all 11 selected Oxford sequences, and a 58.3% ATE reduction on the Oxford subset completed by ORB-SLAM3 RGB-D. More broadly, it offers a direct way to turn strong bounded-view reconstruction into a persistent, revisable long-horizon map without additional training.

AI USE STATEMENT

Generative AI tools assisted literature discovery, language editing, and LaTeX organization. The authors verified citations and numerical results against primary papers and experiment records, reviewed all AI-assisted content, and take responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The supplementary material provides implementation details and key parameters for the two fixed indoor and Oxford configurations, together with model and checkpoint identities, data processing, evaluation protocols, per-sequence results, failure cases, and resource measurements. Each configuration is shared across its datasets without per-sequence tuning. Code and evaluation artifacts will be released upon publication.

REFERENCES

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3d reconstruction, 2026a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3d reconstruction as test-time training, 2026b.

Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, 36(4):1–18, 2017.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, loop it, align it—pushing VGGT's limits on kilometer-scale long RGB sequences, 2026.

Siyan Dong, Zijun Wang, Lulu Cai, Yi Ma, and Yanchao Yang. PROFusion: Robust and accurate dense reconstruction via camera pose regression and optimization, 2026.

Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the RGB-D SLAM system. In *IEEE International Conference on Robotics and Automation*, pp. 1691–1696, 2012. doi: 10.1109/ICRA.2012.6225199.

Jiarong Han, Jincheng Xiong, Yuzhou Liu, Linzhe Shi, Changjie Wu, Ning Guo, Mu Xu, Hang Zhang, and Ming Qian. Revisiting local context for long-horizon streaming 3d reconstruction, 2026.

Jinwoo Jeon, Dong-Uk Seo, Eungchang Mason Lee, and Hyun Myung. AIM-SLAM: Dense monocular SLAM via adaptive and informative multi-view keyframe prioritization with foundation model, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.02018.

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

Inha Lee, Dongjae Jeong, Junhee Lee, and Kyungdon Joo. UniSim-SLAM: Feed-forward SLAM with unified Sim(3) optimization, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R, 2024.

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

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos, 2025.

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

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

Raul Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. doi: 10.1109/TRO.2017.2705103.

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

Richard A. Newcombe, Andrew Fitzgibbon, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, and Steve Hodges. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136, 2011. doi: 10.1109/ISMAR.2011.6092378.

Tuan Duc Ngo, Jiahui Huang, Seoung Wug Oh, Kevin Blackburn-Matzen, Evangelos Kalogerakis, Chuang Gan, and Joon-Young Lee. DAGE: Dual-stream architecture for efficient and fine-grained geometry estimation, 2026.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, and Ziwei Liu. OmniVGGT: Omni-modality driven visual geometry grounded transformer, 2025.

Erik Sandstr"om, Yue Li, Luc Van Gool, and Martin R. Oswald. Point-SLAM: Dense neural point cloud-based SLAM, 2023.

Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 134–144, 2019. doi: 10.1109/CVPR.2019.00022.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773.

Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. iMAP: Implicit mapping and positioning in real-time, 2021.

Yijie Tang, Jiazhao Zhang, Zhinan Yu, He Wang, and Kai Xu. MIPS-Fusion: Multi-implicit-submaps for scalable and robust online neural RGB-D reconstruction, 2023.

Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The Oxford Spires dataset: Benchmarking large-scale LiDAR-visual localisation, reconstruction and radiance field methods, 2025.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory, 2024.

Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer, 2025a.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3X engineering update and official implementation. `https://github.com/yyfz/Pi3`, 2026a. Accessed 2026-09-23.

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, 2026b.

Thomas Whelan, Stefan Leutenegger, Renato Salas-Moreno, Ben Glocker, and Andrew Davison. ElasticFusion: Dense SLAM without a pose graph. In *Robotics: Science and Systems*, 2015. doi: 10.15607/RSS.2015.XI.001.

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

Vladimir Yugay, Yue Li, Theo Gevers, and Martin R. Oswald. Gaussian-SLAM: Photo-realistic dense SLAM with gaussian splatting, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. LoGeR: Long-context geometric reconstruction with hybrid memory, 2026a.

Yuhao Zhang, Yifu Tao, Frank Dellaert, and Maurice Fallon. ScaRF-SLAM: Scale-consistent reconstruction with feed-forward models and classical visual SLAM, 2026b.

Xingguang Zhong, Yue Pan, Liren Jin, Marija Popović, Jens Behley, and Cyrill Stachniss. Globally consistent RGB-D SLAM with 2d gaussian splatting, 2025.

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. NICE-SLAM: Neural implicit scalable encoding for SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12776–12786, 2022. doi: 10.1109/CVPR52688.2022.01245.

A IMPLEMENTATION DETAILS

A.1 SHARED SETTINGS AND GEOMETRY SOLVER

We use exactly two algorithm configurations. One is shared by all 12 TUM RGB-D, seven BundleFusion, and 12 FastCaMo-Real sequences; a second is shared by every Oxford Spires sequence and continuous fragment. Input data, calibration, and sequence length vary, but parameters are not tuned per sequence.

All levels use the same fixed Pi3X model (Wang et al., 2026a;b). We use official source revision `9fa3ddb3` and public checkpoint. Its weights are not fine-tuned for SLAM, for a particular level, or for any evaluation dataset. A single model instance on one GPU serves the local, context, and revisit questions.

Every input view consists of RGB, registered metric camera-z depth, its validity mask, and calibrated intrinsics. RGB is resized to width 518 while preserving aspect ratio and a height divisible by the 14-pixel patch size; the $4{:}3$ benchmark images therefore become 518×392. Dense depth is resized by nearest-neighbor sampling, and the intrinsics undergo the identical image transformation. Oxford sparse depth uses calibrated forward rasterization instead, as detailed in Appendix D. Pi3X jointly predicts camera poses, depth, pointmaps, and confidence from this RGB-D input. Its depth normalization leaves a scale gauge, which we recover with a robust median metric factor from valid measured depth and predicted camera-frame geometry. We apply this factor jointly to predicted translations and pointmaps. This restores the gauge of an RGB-D-conditioned prediction rather than introducing depth only after RGB-only inference.

Table 4 lists the parameters that determine the hierarchy.

Table 4: Key parameters of the shared indoor configuration. The same values are used for TUM RGB-D, BundleFusion, and FastCaMo-Real.

Component	Setting
Level 0 short windows	11 consecutive views, advanced by 10 frames; adjacent windows share one view.
Level 0 extended observations	31 consecutive views, evaluated on the same 10-frame schedule.
Level 1 wide-span correction	12 representative views; candidate spans are considered from 40 to 320 frames at 40-frame intervals.
Piece and submap formation	One-window overlap between consecutive corrected regions. Logical submap changes use a 0.10 m occupancy grid, 0.80 new-space criterion, 160-frame minimum extent, 80-frame look-ahead, and 20-frame persistence.
Regional observations	Bounded temporal domains of at most 321 frames, with distinct primary, validation, and held-out view roles.
Revisit retrieval	At most 48 core-frame descriptors per submap, a 12-view pose-diverse summary, the top three candidates from each active score, and a minimum 96-frame separation.
Joint revisit reconstruction	Six primary views from each endpoint, with disjoint validation and held-out views and a reversed-order consistency check.
Graph optimization	At most 80 neural questions per private graph and 30 iterations per global update; measured horizontal-plane evidence is sampled every 20 frames.

A.2 LEVEL 0: LOCAL

Level 0 begins a short-window reconstruction whenever ten new frames have arrived. Its one-frame overlap with the previous window fixes the rigid transform that places the new cameras and pointmap in the current trajectory gauge. The complete multi-view answer is retained together with its confidence and consistency measurements. In parallel, the 31-view branch produces extended local observations. These answers do not replace or veto the short-window stitch; agreement among repeated short and extended observations supports the later piece partition, and their original predictions remain available to the private graph. No separate feature tracker, local bundle adjustment, ICP, or RGB-D registration module is used.

A.3 LEVEL 1: CONTEXT

Level 1 first revisits longer temporal spans using 12 representative RGB-D views. Candidate spans are evaluated in increasing physical extent, and the selected views balance camera displacement, appearance diversity, and newly observed geometry. An accepted answer applies one continuous correction to the provisional Level 0 trajectory over its support. The corrected camera arrangement then defines fixed offsets inside rigid pieces, while the original Pi3X answer remains an observation rather than being discarded after initialization.

Regional parent and boundary-spanning context questions provide the second Level 1 source of evidence. Their domains are scheduled from the available stream rather than from a completed piece partition, and selected frames are associated with pieces only after inference. Consequently, one prediction can relate several pieces or cross a provisional submap boundary. Primary, validation, and held-out roles use distinct frames so that the answer used as a graph observation is not also its only consistency check.

Piece boundaries require compatible support from overlapping Level 0 observations. Logical submaps are then separated when an occupancy transition introduces sufficient persistent new space. Each submap owns a bounded private graph whose variables are piece poses and auxiliary primitive frames; camera offsets within each piece remain fixed. The graph combines admitted short-window, extended-local, wide-span, regional, and relevant revisit observations. Measured horizontal surfaces contribute soft gravity-direction and height evidence. Updates use robust losses and preserve the original observation records, so later optimization changes their joint interpretation rather than rewriting the model outputs.

A.4 LEVEL 2: SUBMAP

For retrieval, each submap stores Pi3X image descriptors from two encoder depths. Image-level tokens provide the primary similarity ranking, while a spatial pyramid of patch tokens checks whether the matched frames retain a compatible local layout. Descriptor aggregation uses at most 48 frames from the submap core, together with a 12-view pose-diverse summary. We retain the top three candidates from each active retrieval score after requiring temporally separated evidence. Frame-level support passes with either at least two mutual nearest-neighbor matches or a cluster of at least four matches whose absolute temporal-order correlation is at least 0.25. Up to three disjoint support regions may be queried for the same submap pair. Retrieval only proposes geometry; it never supplies the transformation used by the pose graph.

For each retained proposal, six views are selected around each endpoint and jointly reconstructed in one Pi3X call. Separate validation and held-out selections use the same per-endpoint budget, and the primary input order is reversed for an additional consistency test. A relation is rejected when the primary and validation estimates disagree by more than 0.15 m or 7.5°. Corrections that depart substantially from the current graph are checked with path-length-aware innovation limits, and the reconstructed geometry may not increase its 90th-percentile inconsistency by more than 1%. We retain all proposals and their outcomes, including rejections, for analysis.

Accepted relations enter a global $\mathrm{SE}(3)$ graph whose variables are submap poses only. Sequential relations maintain connectivity, while revisit relations distribute long-range corrections. For the indoor configuration, sequential submap relations use nominal uncertainties of 0.002 m and 0.5°; revisit uncertainty inferred from model disagreement is clipped to 0.05–0.15 m and 1–5°. A global update keeps each submap's current internal shape fixed. Its new placements become boundary conditions for subsequent private updates, so private and global optimization alternate over the accumulated observations without invoking Pi3X again for every optimization cycle.

A.5 ASYNCHRONOUS EXECUTION AND OXFORD CONFIGURATION

The hierarchy runs as an asynchronous stream. Frame decoding and preparation, Level 0 inference, context construction, retrieval, and the two graph optimizers make progress as soon as their dependencies are available. One shared GPU worker executes Pi3X calls sequentially, while bounded CPU workers prepare later questions and update private and global graphs concurrently with GPU inference. Completed observations and graph states are atomically published for resumption, while

rejected proposals remain in the evidence record. Runtime and resource measurements are reported in Appendix B.

Oxford uses the same model, image resolution, window schedules, view selection, submap construction, retrieval gates, closure validation, optimization budget, and asynchronous execution as the indoor setting. We rectify its fisheye RGB and project LiDAR returns into calibrated camera-z depth without densification. The algorithmic changes are an increase from 80 to 128 neural questions allowed in each private graph, a revisit translation uncertainty of 2 m instead of the indoor 0.05–0.15 m range, and a rotation-uncertainty lower bound of 2° instead of 1°; the 5° upper bound is unchanged. One shared Oxford setting is used for all 11 complete sequences and all 24 continuous fragments. These two settings produce all results reported for RECALL-SLAM, without per-sequence parameter selection.

B RUNTIME AND RESOURCE USE

We profile a complete streaming run of RECALL-SLAM on FastCaMo-Real *gym*, the longest sequence in that benchmark (4,922 RGB-D frames). The run uses one NVIDIA RTX PRO 6000 Blackwell GPU and 48 affinity-bound logical CPUs, with input and output on NVMe. The process takes 1,417.2 s end to end, or 3.47 frames/s, including an 11.5 s model load and file I/O. The frontend reaches its final frame at 1,036.3 s (4.77 frames/s during its active interval), leaving a 380.8 s processing tail. Peak sampled GPU memory is 50,993 MiB.

Module	Total	Overlap	Alone
L0: streaming local geometry			
Pi3X (GPU)	671.9	582.4	89.5
Window CPU	253.7	213.0	40.7
L1: context-level submap refinement			
Pi3X (GPU)	137.6	122.8	14.9
Intra-submap graph CPU	813.8	599.9	213.9
L2: revisits and global coordination			
Joint Pi3X (GPU)	179.1	173.4	5.6
Global SE(3) graph CPU	425.4	302.5	122.9

Table 5: Selected task times on *gym* (s). Overlap is concurrent CPU/GPU task time; Alone is the rest. This is scheduling overlap, not time saved.

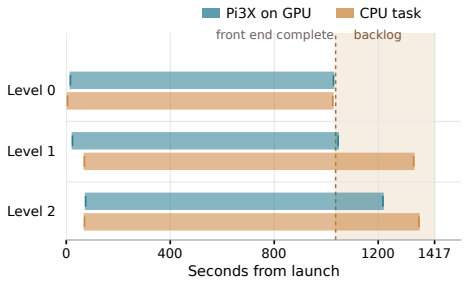

Figure 5: Activity spans by level. The shaded 381 s tail is unfinished streaming work after the front end, not a separate postprocessing stage.

Table 5 groups all 1,618 Pi3X window calls by functional level: 671.9 s at L0, 137.6 s at L1, and 179.1 s for joint inter-submap estimation. Pi3X time is measured with CUDA events; the selected CPU rows use task-wall intervals, including preparation and waits. The last two columns split each logged interval by whether a task interval on the other device is also present. For example, 599.9 s of L1 intra-submap graph task time overlaps GPU inference and 213.9 s does not. Overlap is observable scheduling, not proof that removing the task would leave total wall time unchanged; the shared GPU worker and queue dependencies may still extend it. Solver execution alone is 274.1 s for the intra-submap graph and 197.1 s for the global graph. Figure 5 shows each lane's first-to-last activity span, including gaps. Of 968 graph updates, 880 start before the front end finishes. The remaining 380.8 s is backlog from the same stream: some questions and graph updates finish after the front end, rather than in a separate postprocessing phase.

The measured 3.47 frames/s is slow for online use. Reducing Pi3X calls, simplifying the workflow, and accelerating the remaining backend are future work.

C CHOICE OF FEED-FORWARD MODEL

We compared eight released feed-forward models on the same 18 eleven-view windows from six TUM RGB-D, BundleFusion, and Oxford Spires sequences, with windows near 30%, 50%, and

70% of each sequence. Pi3X, OmniVGGT, and MapAnything used measured depth during RGB-D inference; Pi3, DA3 Giant, DA3 Nested, VGGT, and DAGE inferred from RGB alone. For a common metric evaluation, we estimated one scale factor from measured and predicted depth after inference for every model, then applied rigid SE(3) trajectory alignment. For the five RGB models, this post-inference scale recovery was the only use of measured depth. No reference-pose scale fit, stitching, or SLAM optimization was used.

Table 6: Eight-model local trajectory comparison on 18 matched windows. The input column distinguishes native RGB-D inference from RGB-only inference; the RGB models use measured depth only afterward for metric scale recovery. ATE values are medians in cm; ATE/path is the median per-window ratio in percent. Each dataset contributes six windows; lower is better.

Model	Inference input	Overall	ATE/path (%)	TUM	BundleFusion	Oxford
Pi3X	RGB-D	1.40	**1.09**	1.19	1.80	1.34
OmniVGGT	RGB-D	2.23	2.34	1.20	2.23	7.52
MapAnything	RGB-D	2.61	1.96	1.40	3.29	3.12
Pi3	RGB	1.25	1.51	1.00	1.74	1.21
DA3 Giant	RGB	1.52	1.23	0.96	2.68	1.87
DA3 Nested	RGB	1.43	1.22	0.96	2.32	1.26
VGGT	RGB	1.98	1.71	1.33	2.31	3.01
DAGE	RGB	1.81	1.44	1.07	1.91	2.02

Pi3X has the lowest path-normalized median ATE across all eight models and the lowest overall median ATE among those conditioned on RGB-D. RGB-only Pi3 has a lower overall median ATE after post-inference depth scaling (1.25 versus 1.40 cm), so local ATE alone does not uniquely select Pi3X. Pi3X directly uses the available measured depth and provides the poses, depth, point maps, and neural descriptors required by the local, regional, and revisit questions in RECALL-SLAM. We therefore use Pi3X as the shared feed-forward model.

D BENCHMARK PROTOCOLS AND BASELINE ADAPTATIONS

D.1 DATASET SELECTION AND CONSTRUCTION

We evaluate on four quantitative RGB-D benchmarks. TUM RGB-D is represented by a fixed 12-sequence subset. Running every method on every TUM recording would add substantial repeated computation, especially for the full-frame feed-forward baselines, while many recordings repeat closely related rooms and motion patterns. The selected subset therefore keeps the comparison tractable while retaining deliberate variation in scene content, sequence length, camera calibration, and motion, including translation-dominated motion, rotation-dominated motion, desk-scale capture, room traversal, low-texture floor observations, and object-centered sequences. The same subset is used for every method. BundleFusion contributes seven sequences. We exclude the *apt2* recording because its released reference trajectory contains abrupt single-frame position discontinuities; we remove the complete sequence rather than repair or selectively mask the reference trajectory. FastCaMo-Real contributes all 12 released real sequences.

For Oxford Spires (Tao et al., 2025), we make the RGB-D SLAM benchmark from its RGB–LiDAR package. Eleven recordings have no measured-depth gap longer than 0.3 s and are evaluated as complete sequences. The remaining two recordings contain long interruptions in the LiDAR depth stream even though RGB acquisition continues. We split a recording whenever consecutive measured-depth timestamps differ by more than 1 s, producing seven segments from the Christ Church 05 recording and 17 from the Bodleian Library 02 recording. This threshold preserves ordinary missed depth frames: lowering it to 0.6 s produces the same boundaries, whereas 0.3 s fragments normal acquisition gaps. All 1,940 and 2,387 measured-depth frames from the two recordings are retained. Segment boundaries use timestamps alone, without trajectory error, ground truth, or model output, and no pose or loop edge is allowed to cross a boundary. We consequently report the 11 complete sequences and 24 fragments as separate populations.

Table 7 lists the original recordings used for evaluation. The two segmented Oxford recordings appear as individual fragments in Table 13.

D.2 RGB-D PREPROCESSING FOR PI3X

For each dataset, we pair RGB with registered camera-z depth and calibrated intrinsics, preserving capture timestamps or original frame numbers. TUM RGB and depth are associated by a globally greedy one-to-one match with a strict 20 ms limit; its released depth values are divided by 5,000 to obtain metres. BundleFusion and FastCaMo-Real provide registered 640×480 RGB-D images with millimetre depth. Oxford begins with 1440×1080 fisheye RGB and Euclidean LiDAR range projected into the camera. We rectify the RGB image to the calibrated pinhole camera, unproject each LiDAR return through the fisheye model, convert range to camera-z, and rasterize it in the same pinhole image with a nearest-pixel foreground z-buffer. Oxford depth values are converted from units of $1/256$ m. Missing returns remain zero; we do not complete, interpolate, or densify LiDAR depth.

The same model-input transform is then applied to RGB, depth, and intrinsics. RGB is resized bicubically to width 518, with the aspect-preserving height rounded to a multiple of the 14-pixel patch size and a center crop only when the height would exceed 518. Thus the $4{:}3$ inputs used by TUM, BundleFusion, FastCaMo-Real, and rectified Oxford become 518×392. Dense depth uses nearest-neighbor resampling. Sparse Oxford depth is instead forward-rasterized under the exact half-pixel RGB resize transform, retaining only measured returns and the nearest z value on collisions. Intrinsics undergo the same resize and crop transformation. Pi3X receives the resized RGB, the valid-depth mask and metric depth values, and calibrated intrinsics. Valid measured depths condition the model and provide the per-window metric gauge; one robust scale is applied jointly to predicted pointmaps and camera translations. Sparse holes therefore remain missing observations rather than fabricated surfaces.

D.3 BASELINE ADAPTATIONS

ScaRF-SLAM deliberately separates tracking and mapping and use ORB-SLAM3 for camera estimation. With RGB-D data, we use ORB-SLAM3 in its native RGB-D mode for the camera trajectory, including its normal local mapping, bundle adjustment, relocalization, and loop-closing threads. The resulting poses are then fixed while the released ScaRF feed-forward mapping backend predicts and fuses dense geometry. Accordingly, trajectory tables name this result *ORB-SLAM3 RGB-D*; ScaRF-SLAM appears only in reconstruction tables. We report the released DA3 reconstruction as the full ScaRF-SLAM output. Where useful, we additionally show a sensor-depth back-projection using the same ORB-SLAM3 poses and label it explicitly, so that it cannot be mistaken for ScaRF-SLAM's model reconstruction.

VGGT-SLAM 2.0 is released as an RGB-only, keyframe-based system, so we construct two controlled RGB-D adaptations while retaining its submap organization, SALAD retrieval, and graph optimization. In VGGT-SLAM 2.0-D, VGGT inference remains RGB-only. Registered sensor depth is transformed to the VGGT image grid and a robust, frame-balanced median log-depth ratio estimates one positive scale for each submap or loop prediction. That scale is applied to predicted depth, pointmaps, and camera translations, calibrated intrinsics are used for back-projection, and the persistent graph is optimized on $\mathrm{SE}(3)$ rather than the original scale-ambiguous manifold. No reference trajectory or surface enters this calibration. In VGGT-SLAM 2.0-Pi3X, every VGGT window inference is replaced by RGB-D-conditioned Pi3X inference. Pi3X receives RGB, measured depth, and calibration. Its metric finalization estimates one sensor-depth scale for the predicted depth head and a second sensor-depth scale shared by pointmaps and camera translations before conversion to the unchanged VGGT-SLAM submap interface. The surrounding retrieval and pose graph remain those of VGGT-SLAM 2.0, and no component of the RECALL-SLAM hierarchy is used.

The released VGGT-SLAM 2.0 evaluates only optical-flow-selected keyframes. Comparing those sparse poses with a method that outputs every frame would change the evaluated population. We therefore use an explicit full-frame protocol for both adaptations: every input frame is processed in windows containing 16 new frames plus one overlap frame. No missing pose is interpolated, and every input frame must have an estimated pose. This changes only frame selection; the VGGT-SLAM map, retrieval, and graph remain active.

2DGS-SLAM and PROFusion require registered dense depth and allocate dense indoor map representations. Their intended operating regime is bounded indoor capture, and their memory cost grows poorly for long outdoor routes. We therefore evaluate them on TUM RGB-D, BundleFusion,

and FastCaMo-Real, but do not apply them to sparse-LiDAR Oxford Spires or the self-captured long sequences. This scope is fixed by sensor compatibility and operating regime rather than by per-sequence accuracy.

D.4 METRICS, ALIGNMENT, AND FAILURES

We report absolute trajectory error (ATE) and relative pose error (RPE) for TUM RGB-D, BundleFusion, and Oxford Spires. BundleFusion and FastCaMo-Real reconstruction are evaluated by precision, recall, and F-score at a 5 cm distance threshold. Dataset means are arithmetic means over sequences, so a long sequence does not receive more weight than a short one. BundleFusion poses and surfaces are outputs of the BundleFusion reconstruction system and are treated as common references rather than independent motion-capture ground truth; FastCaMo-Real provides a reference surface but no reference trajectory.

Scale alignment is used only for methods whose algorithmic input is RGB alone. For LingBot-Map and ABot-Recon, we estimate one evaluation-only Sim(3) transform from trajectory correspondences and apply the same transform to the trajectory and exported reconstruction. Every method that consumes measured depth, including RECALL-SLAM, 2DGS-SLAM, PROFusion, ORB-SLAM3 RGB-D, VGGT-SLAM 2.0-D, and VGGT-SLAM 2.0-Pi3X, keeps its estimated metric scale and receives at most one rigid SE(3) alignment. No method receives a separately fitted reconstruction scale.

An incomplete ORB-SLAM3 trajectory remains a failed complete-sequence result. For the indoor tables, we nevertheless expose its returned-prefix ATE with a † and report a prefix-inclusive mean so that the available trajectory estimate is not hidden; this easier evaluation is explicitly favorable to ORB-SLAM3. For Oxford, incomplete trajectories remain excluded, and the paired comparison requires at least 99% measured-frame coverage and a pose for the final measured-depth frame; 14 of the 24 fragments satisfy this rule, whereas none of the 11 complete sequences does. Missing results are never replaced by zeros or interpolated poses.

E PER-SEQUENCE QUANTITATIVE RESULTS

The following tables give every quantitative sequence used in the main paper. "VGGT-D" and "VGGT-P" denote VGGT-SLAM 2.0-D and VGGT-SLAM 2.0-Pi3X, respectively. LingBot-Map and ABot-Recon are RGB-only and use the Sim(3) protocol described in Appendix D; all remaining trajectory columns use SE(3). For the indoor ORB-SLAM3 results, † marks ATE evaluated on an incomplete returned prefix; the marked means include these prefixes and are therefore favorable to ORB-SLAM3. Oxford retains its stricter paired-completion rule. A dash denotes failure or an unavailable result, not a zero-valued error. Because the two alignment protocols make different scale assumptions, the tables expose the values without ranking RGB and RGB-D methods in one shared best/second-best order. For reconstruction, "ScaRF sensor" is the explicit sensor-depth back-projection using the ORB-SLAM3 trajectory, while "ScaRF DA3" is the released feed-forward ScaRF-SLAM reconstruction; these outputs are reported separately. Oxford sequence names are shortened only in the result tables, with the exact names listed in Table 7. All 11 ORB-SLAM3 complete-sequence runs terminate with incomplete trajectories. Among the fragments, only the 14 runs satisfying the paired coverage rule receive an ORB-SLAM3 score.

Table 7: Original recordings used in the quantitative benchmarks. The two Oxford recordings in the final row contribute seven and 17 independently evaluated, continuous RGB-D segments, respectively. The table uses three sequence columns to keep the inventory compact.

Dataset	Sequence 1	Sequence 2	Sequence 3
TUM RGB-D	rgbd_dataset_freiburg1_360 rgbd_dataset_freiburg1_floor rgbd_dataset_freiburg1_rpy rgbd_dataset_freiburg2_desk	rgbd_dataset_freiburg1_desk rgbd_dataset_freiburg1_plant rgbd_dataset_freiburg1_teddy rgbd_dataset_freiburg2_xyz	rgbd_dataset_freiburg1_desk2 rgbd_dataset_freiburg1_room rgbd_dataset_freiburg1_xyz rgbd_dataset_freiburg3_long_office_household
BundleFusion	apt0 office0 office3	apt1 office1	copyroom office2
FastCaMo-Real	apartment_1 lab meeting_room studio	apartment_2 lounge_1 office workshop_1	gym lounge_2 stairwell workshop_2
Oxford complete	2024-03-12-keble-college-02 2024-03-12-keble-college-05 2024-03-14-blenheim-palace-01 2024-03-18-christ-church-02	2024-03-12-keble-college-03 2024-03-13-observatory-quarter-01 2024-03-14-blenheim-palace-02 2024-03-18-christ-church-03	2024-03-12-keble-college-04 2024-03-13-observatory-quarter-02 2024-03-14-blenheim-palace-05
Oxford segmented	2024-03-20-christ-church-05	2024-05-20-bodleian-library-02	

Table 8: Per-sequence TUM RGB-D trajectory ATE RMSE in cm; lower is better. 2DGS-SLAM retains its native every-second-frame sampling. † marks an incomplete ORB-SLAM3 prefix; its mean includes both marked prefixes.

Sequence	2DGS	PROFusion	LingBot	ABot	ORB-SLAM3	VGGT-D	VGGT-P	RECALL-SLAM
fr1/360	20.318	8.640	9.055	8.367	42.914	7.004	4.921	4.426
fr1/desk	4.009	3.304	4.029	6.192	2.040	2.475	2.226	2.786
fr1/desk2	8.913	4.902	4.911	5.487	3.169	4.468	3.376	3.234
fr1/floor	23.445	77.146	8.873	24.075	7.302†	4.122	2.822	2.734
fr1/plant	6.951	4.613	4.066	17.485	2.710	4.706	1.892	2.016
fr1/room	68.153	21.458	12.038	14.583	6.381	4.291	4.796	4.922
fr1/rpy	4.397	3.623	2.523	4.799	2.781	3.018	2.081	1.992
fr1/teddy	48.883	14.145	5.095	21.935	4.499†	4.386	3.268	3.099
fr1/xyz	1.953	1.869	1.847	13.779	1.364	1.290	0.996	0.953
fr2/desk	12.669	14.584	5.440	29.478	8.550	9.270	7.401	4.690
fr2/xyz	2.721	3.501	2.392	34.523	1.842	2.512	1.234	1.393
fr3/office	108.264	10.420	11.121	32.017	1.607	4.255	2.970	3.381
Mean	25.890	14.017	5.949	17.727	7.097†	4.316	3.165	2.969

Table 9: Per-sequence BundleFusion trajectory ATE RMSE in cm; lower is better. † marks an incomplete ORB-SLAM3 prefix; its mean includes all five marked prefixes.

Sequence	2DGS	PROFusion	LingBot	ABot	ORB-SLAM3	VGGT-D	VGGT-P	RECALL-SLAM
apt0	245.631	57.276	129.588	62.006	10.791	9.447	6.700	6.361
apt1	337.644	18.030	65.202	33.919	9.861†	9.167	6.461	6.447
copyroom	22.080	25.144	24.588	31.800	6.095	6.951	5.294	5.187
office0	235.493	55.808	52.660	51.837	9.507†	8.141	6.426	7.077
office1	320.966	32.797	33.507	33.120	8.406†	8.832	6.479	7.433
office2	55.081	92.592	71.087	37.092	9.154†	9.242	8.596	9.579
office3	174.389	51.248	39.393	38.479	6.948†	7.608	5.570	5.558
Mean	198.755	47.556	59.432	41.179	8.680†	8.484	6.504	6.806

Table 10: Per-sequence BundleFusion reconstruction, reported as P/R/F at 5 cm in percent; higher is better.

Sequence	2DGS	PROFusion	LingBot	ABot	ScaRF sensor	ScaRF DA3	VGGT-D	VGGT-P	RECALL-SLAM
apt0	4.3/34.6/7.7	26.3/58.0/36.2	12.4/39.1/18.9	17.5/60.7/27.2	61.2/88.1/72.3	51.1/80.4/62.5	67.2/93.5/78.2	76.1/91.6/83.1	77.1/91.5/83.7
apt1	3.4/19.5/5.8	40.0/74.3/52.0	27.8/78.8/41.2	26.9/82.9/40.6	69.1/91.0/78.5	52.9/85.3/65.3	67.3/90.0/77.0	78.9/94.4/86.0	77.2/92.0/83.9
copyroom	26.6/77.3/39.6	24.4/54.3/33.7	45.8/95.4/61.9	22.7/75.3/34.9	81.1/98.0/88.7	68.5/91.3/78.3	75.8/97.9/85.4	78.5/95.8/86.3	78.5/95.7/86.3
office0	2.2/19.9/4.0	11.0/40.2/17.3	26.5/80.0/39.8	20.3/72.6/31.8	67.3/95.0/78.8	59.1/80.8/68.2	65.3/90.8/76.0	81.0/91.6/86.0	72.7/89.3/80.1
office1	3.3/23.5/5.7	24.2/53.4/33.3	21.9/70.4/33.4	18.3/65.1/28.6	81.2/96.3/88.1	68.5/88.0/77.0	74.3/95.7/83.7	90.0/97.8/93.7	69.8/89.1/78.2
office2	10.6/32.1/15.9	8.8/21.0/12.4	18.6/51.1/27.3	21.3/62.5/31.7	44.2/65.0/52.6	38.6/54.6/45.2	46.0/69.3/55.3	64.8/75.9/69.9	58.2/69.1/63.2
office3	5.9/37.3/10.2	17.5/39.6/24.2	31.5/88.8/46.5	19.1/70.4/30.1	75.8/96.0/84.7	68.4/86.2/76.3	75.5/98.4/85.5	82.2/96.4/88.7	82.5/96.5/89.0
Mean	8.0/34.9/12.7	21.8/48.7/29.9	26.4/72.0/38.4	20.9/69.9/32.1	68.5/89.9/77.7	58.2/80.9/67.5	67.4/90.8/77.3	78.8/91.9/84.8	73.7/89.0/80.6

Table 11: Per-sequence FastCaMo-Real reconstruction, reported as P/R/F at 5 cm in percent; higher is better.

Sequence	2DGS	PROFusion	LingBot	ABot	ScaRF sensor	ScaRF DA3	VGGT-D	VGGT-P	RECALL-SLAM
apartment_1	9.9/18.4/12.9	70.6/72.4/71.5	18.0/36.1/24.1	11.1/22.5/14.9	73.5/61.0/66.7	57.3/57.4/57.4	14.4/32.1/19.9	53.6/49.1/51.2	91.1/86.5/88.7
apartment_2	13.3/16.9/14.9	56.5/55.7/56.1	14.2/31.0/19.5	11.3/18.5/14.0	21.7/12.8/16.1	12.5/11.4/12.0	8.4/15.1/10.8	56.6/58.4/57.5	77.3/73.3/75.3
gym	7.4/14.4/9.8	29.9/25.8/27.7	15.6/36.2/21.8	6.2/13.2/8.4	10.5/8.6/9.4	6.4/15.7/9.1	9.3/15.4/11.6	61.8/60.3/61.0	34.9/32.7/33.8
lab	7.4/21.1/11.0	34.9/33.7/34.3	10.0/28.0/14.7	13.6/23.4/17.2	57.2/55.2/56.2	29.2/60.4/39.4	12.7/15.2/13.9	14.7/25.0/18.6	86.4/70.9/77.9
lounge_1	12.1/18.7/14.7	66.9/69.0/67.9	7.6/10.0/8.6	10.0/19.7/13.2	17.7/16.2/16.9	10.5/19.4/13.7	12.2/29.3/17.3	54.2/46.0/49.7	53.6/55.0/54.3
lounge_2	6.0/22.9/9.5	27.4/46.8/34.6	5.4/25.4/8.9	8.6/27.6/13.1	13.4/14.2/13.8	9.6/25.8/14.0	14.2/31.0/19.5	8.1/12.1/9.7	27.2/46.0/34.2
meeting_room	8.0/18.4/11.1	50.9/51.5/51.2	9.2/30.0/14.1	11.9/25.6/16.2	14.2/25.5/18.3	8.3/31.4/13.1	15.5/30.4/20.5	15.0/16.9/15.9	78.0/76.3/77.1
office	15.9/13.1/14.4	35.8/26.6/30.5	12.6/18.5/15.0	13.3/18.4/15.4	68.8/14.4/23.8	33.6/10.3/15.8	9.5/21.0/13.1	11.2/18.2/13.9	77.7/45.7/57.5
stairwell	28.2/27.1/27.6	62.5/57.6/60.0	12.9/31.7/18.3	20.1/35.7/25.7	27.5/21.0/23.8	19.7/10.0/13.3	31.8/49.6/38.8	17.2/19.4/18.3	62.8/62.9/62.8
studio	17.0/25.6/20.4	50.5/60.8/55.2	24.2/32.4/27.7	37.0/52.9/43.5	47.5/37.7/42.0	37.6/36.2/36.9	49.1/53.4/51.1	74.0/62.7/67.9	78.3/65.1/71.1
workshop_1	11.2/19.1/14.1	32.0/36.3/34.0	13.9/43.6/21.1	13.3/26.3/17.6	22.1/13.1/16.4	15.6/11.8/13.4	18.1/31.8/23.1	22.5/33.0/26.8	92.9/78.4/85.0
workshop_2	9.3/18.5/12.3	29.8/36.0/32.6	7.8/23.8/11.8	11.8/21.5/15.2	25.6/13.6/17.7	24.3/18.1/20.7	11.3/22.5/15.0	12.0/15.7/13.6	92.6/76.9/84.0
Mean	12.1/19.5/14.4	45.6/47.7/46.3	12.6/28.9/17.1	14.0/25.4/17.9	33.3/24.4/26.8	22.1/25.7/21.6	17.2/28.9/21.2	33.4/34.7/33.7	71.1/64.2/66.8

Table 12: Per-sequence Oxford Spires ATE RMSE for the 11 complete sequences, in cm; lower is better. Every ORB-SLAM3 trajectory is incomplete and is therefore shown as "–".

Sequence	LingBot	ABot	ORB-SLAM3	VGGT-D	VGGT-P	RECALL-SLAM
Keble-02	323.87	247.31	–	143.30	61.52	55.20
Keble-03	324.40	316.80	–	98.44	1622.62	48.53
Keble-04	293.53	386.18	–	124.81	116.70	102.83
Keble-05	328.98	388.55	–	314.93	123.13	89.31
Observatory-01	302.27	339.12	–	160.40	53.75	44.37
Observatory-02	291.38	286.12	–	162.39	66.49	48.75
Blenheim-01	769.50	584.03	–	378.63	138.71	112.32
Blenheim-02	523.55	394.88	–	860.73	145.50	105.57
Blenheim-05	371.38	396.12	–	534.15	144.70	112.32
Christ Church-02	1668.06	691.07	–	444.79	1987.45	77.02
Christ Church-03	239.34	184.37	–	44.79	1154.21	41.04
Mean	494.21	383.14	–	297.03	510.43	76.11

Table 13: Per-fragment Oxford Spires ATE RMSE in cm; lower is better. CC05 and BL02 denote the Christ Church 05 and Bodleian Library 02 recordings, with segments numbered in temporal order from 1. The last row compares ORB-SLAM3 and RECALL-SLAM only on the same 14 ORB-complete fragments.

Sequence	LingBot	ABot	ORB-SLAM3	VGGT-D	VGGT-P	RECALL-SLAM
CC05 1	120.27	27.97	–	25.60	23.05	18.13
CC05 2	45.09	60.22	96.05	143.28	64.93	45.66
CC05 3	56.51	24.34	–	47.80	38.36	37.93
CC05 4	38.65	23.93	–	28.03	23.43	25.14
CC05 5	23.02	68.43	–	22.77	17.51	15.96
CC05 6	183.08	57.01	–	26.84	30.36	26.86
CC05 7	16.45	17.82	47.47	4.81	8.86	13.77
BL02 1	5.89	11.57	20.75	11.59	11.57	6.84
BL02 2	6.77	7.51	30.89	7.32	9.79	9.66
BL02 3	8.60	19.57	10.65	21.79	15.55	12.31
BL02 4	8.75	5.38	16.52	6.73	13.37	15.39
BL02 5	9.01	5.90	–	11.38	12.06	12.46
BL02 6	12.53	8.52	–	19.74	12.08	7.26
BL02 7	13.67	11.17	–	28.02	13.77	13.42
BL02 8	21.64	18.00	25.14	6.46	16.71	13.51
BL02 9	5.39	4.35	–	7.79	10.18	6.66
BL02 10	7.72	6.12	26.99	12.19	11.05	10.45
BL02 11	5.30	6.38	38.44	5.85	7.25	8.03
BL02 12	14.26	13.21	21.01	30.26	14.85	12.73
BL02 13	37.34	14.89	19.76	10.69	19.06	25.76
BL02 14	16.05	4.80	124.07	12.36	13.84	19.84
BL02 15	6.94	15.29	–	9.55	8.12	4.28
BL02 16	21.04	30.29	6.76	12.52	8.38	11.98
BL02 17	25.38	20.59	32.96	20.22	10.02	10.13
Mean, all 24	29.56	20.14	–	22.23	17.26	16.01
Mean, paired 14	–	–	36.96	–	–	15.43

F ADDITIONAL QUALITATIVE RESULTS

We provide the remaining per-sequence trajectory and reconstruction results for RECALL-SLAM. Each panel is rendered from the complete delivered trajectory or map; no sequence prefix is used as a successful result.

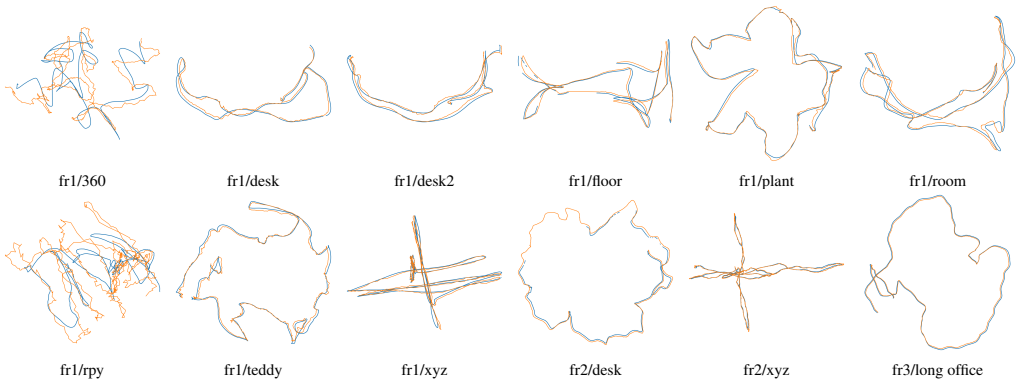

Figure 6: Per-sequence TUM RGB-D trajectory comparisons for the 12-sequence evaluation subset. Blue denotes ground truth and orange denotes RECALL-SLAM after the fixed-scale SE(3) alignment used for evaluation. Every delivered pose is shown.

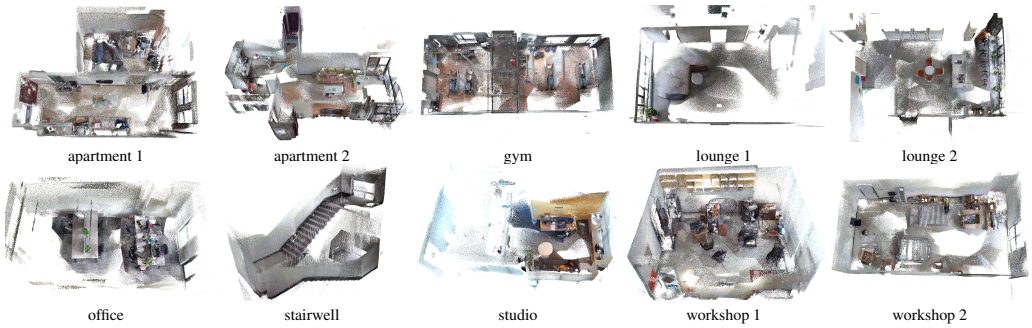

Figure 7: Additional FastCaMo-Real reconstructions. Together with the *lab* and *meeting-room* sequences presented in main-paper Fig. 4, these panels cover all 12 sequences.



Figure 8: BundleFusion reconstructions for six retained sequences. The seventh, *office3*, appears in main-paper Fig. 1.

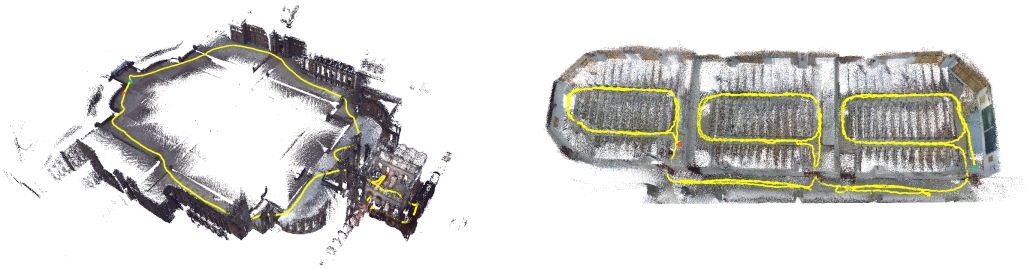

Figure 9: Additional large-scale reconstructions. Left: Blenheim Palace, the only one of the four Oxford Spires complete-sequence locations not visualized in the main paper. Right: our self-captured tiered lecture hall. Its **30,079-frame** stream follows a 344 m trajectory through three seating blocks and connecting aisles, making it substantially longer than the indoor benchmark sequences. Yellow curves show the estimated trajectories.