Reconsolidatable Streaming 3D Reconstruction

Yihang Chen* Lin-Zhuo Chen* Jian Gao* Jiacheng Dong* Yao Yao
Nan Xue Xing Zhu Yujun Shen Yinghao Xu†

*Equal Contribution †Project Lead

In this work, we enhance LingBot-Map with a ***reconsolidation mechanism***, yielding LingBot-Map 2.0. Underlying this design is the observation that popular streaming reconstruction systems process each frame in a single pass, which often causes errors to accumulate over time. Instead, we adopt a hierarchical strategy in which the lower layers inherit the streaming capability of LingBot-Map, while the upper layers can switch between a streaming mode and a reconsolidation mode. The two modes share identical weights and differ only in the use of causal versus bidirectional attention. As a result, reconsolidation can be triggered at any time step, upon which the upper layers jointly re-process all selected keyframes and update the KV cache with the refined results, ensuring that subsequent streaming steps always benefit from the most up-to-date context. This capability, termed Geometric Context Reconsolidation (GCR), makes drift correction a native part of the model design, obviating the need for offline refinement. Alongside GCR, we also strengthen the base model of LingBot-Map 2.0, which outperforms existing streaming reconstruction counterparts across nine settings on seven datasets. Notably, our proposed feed-forward reconsolidation alone, without iterative optimization or an external back-end, suffices to lift LingBot-Map 2.0 above loop-closure-equipped streaming methods and offline models on most benchmarks.

Website: https://technology.robbyant.com/lingbot-map-v2
Github: https://github.com/robbyant/lingbot-map-v2
Checkpoints: https://huggingface.co/collections/robbyant/lingbot-map-v2

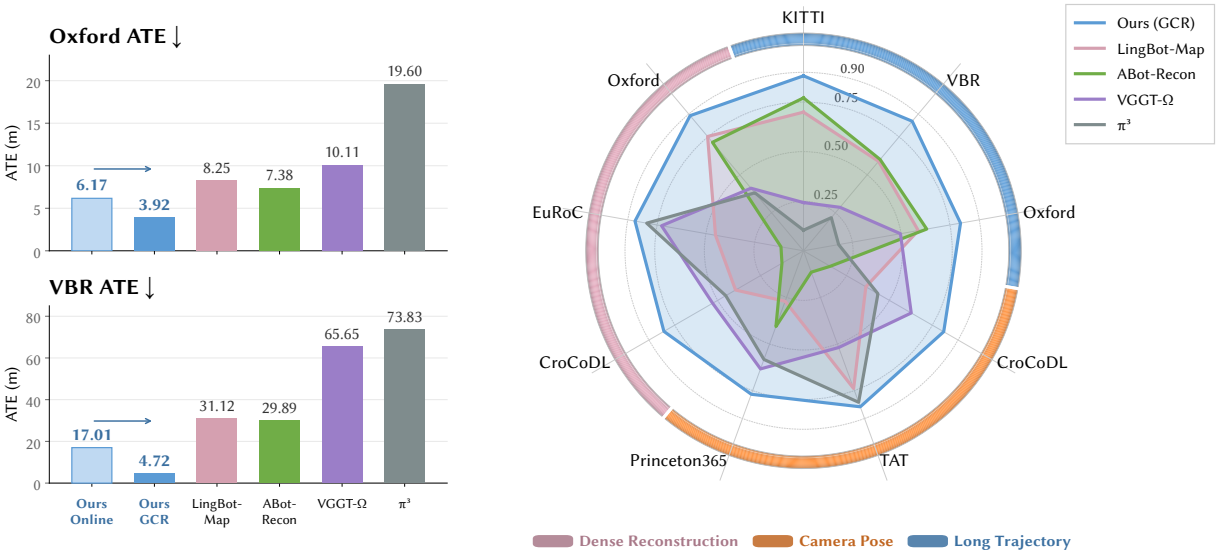


Figure 1. Comparison of LingBot-Map 2.0 with state-of-the-art streaming and offline 3D reconstruction methods.

1 Introduction

Human spatial memory is formed incrementally, as a continuous stream of visual input is consolidated into a persistent representation of the environment [50]. This representation, however, is not fixed once formed. In cognitive neuroscience, memory reconsolidation refers to the finding that a consolidated memory, upon retrieval, can be updated with new information [34, 52]. Later observations can therefore revise earlier impressions within the memory itself, so subsequent behavior relies on the updated representation rather than the outdated one. Can streaming 3D reconstruction models operate similarly, revising their consolidated context as later views arrive?

Streaming 3D reconstruction models [13, 14, 16, 17, 28, 33, 47, 74, 78, 89, 91, 95] consolidate past views into a persistent context, such as a recurrent state or a key-value cache, predicting camera poses and dense geometry for each incoming view. Since each view enters this context only once and is never revisited, errors from early partial observations accumulate as drift even when later views reveal them. Correcting them requires revising the consolidated context itself, exposing a tension: later views must reopen it, yet ongoing streaming relies on it as a stable reference.

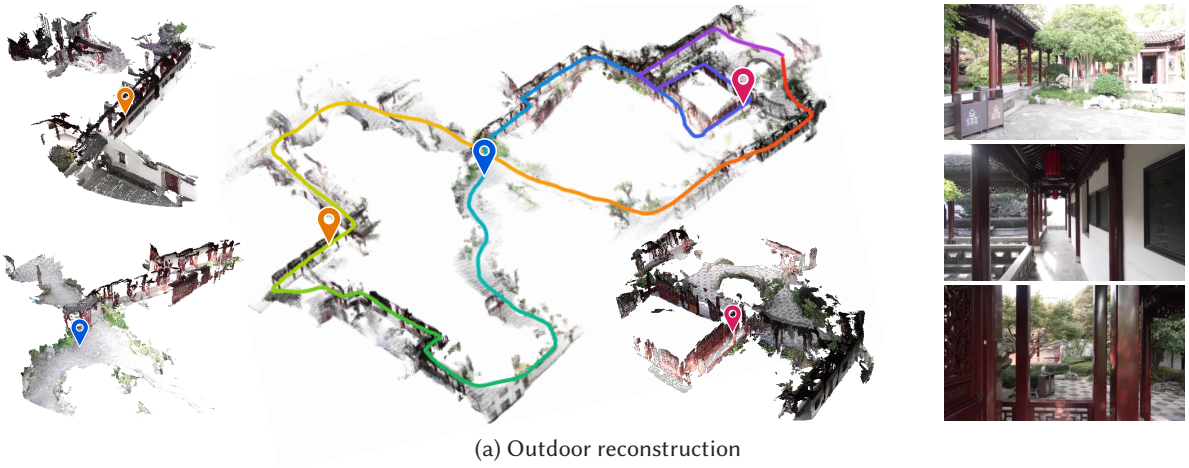
Existing methods correct streaming reconstructions in different ways, each involving distinct trade-offs. Hybrid systems such as MASt3R-SLAM [51] and VGGT-SLAM [48] couple learned offline 3D models with classical SLAM backends, but their corrections rely on hand-crafted loop detection and iterative pose-graph optimization outside the network. Streaming models can optionally employ external backends, including the loop-closure stages in HorizonStream [17] and ABot-Recon [28] as well as dedicated backends such as CLoSeR [36] and Map2SLAM [25], to refine the output trajectory, yet still leave the cached context unchanged, so subsequent views continue to read the drifted context. SLAM-Former [90] and SLAMFormer-∞ [22] revise the model state within a single network, but reopen every layer with full attention, treating per-view local geometry, which later views rarely change, the same as the cross-view pose relations that drift corrupts. These observations point to a key design principle: revision should act on the consolidated context rather than on its outputs, and it should reopen only the part that later views need to change, keeping the rest fixed as the stable reference that streaming requires, since consecutive views vary gradually unlike unordered image sets (Tab. 7).

To this end, we introduce `LingBot-Map 2.0`, a streaming 3D reconstruction model built upon Geometric Context Reconsolidation (GCR). Analogous to the consolidation of incoming views into a cached geometric context during online inference, GCR reconsolidates this context: once later views are available, it reopens the stored context, updates it jointly across the observed views, and writes the revised context back to the cache. Following this principle, GCR partitions the context by layer into a *committed context* and a *revisable context*. The committed context comes from the lower layers, which mainly capture per-view geometry and remain causal, so it is invariant to subsequent views and reused without modification. Only the revisable context, produced by the upper layers that align views with one another, is recomputed with bidirectional attention among the observed views. The recomputed context yields revised estimates for past views and updates the cache in place, so subsequent views are predicted from the corrected context.

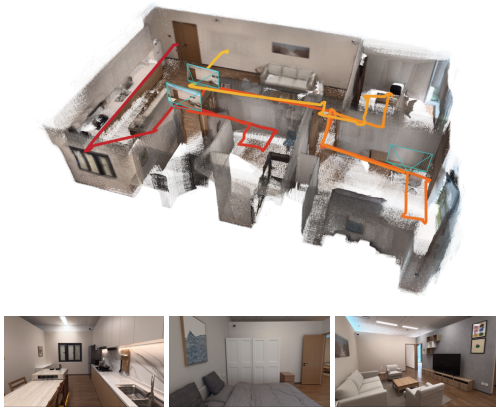
Because the revisable layers switch between causal and bidirectional attention, we train `LingBot-Map 2.0` on long sequences with a mixed-mode recipe spanning streaming, reconsolidation, and their transitions. Reconsolidation can be triggered at any point over ten-thousand-frame streams, suppressing drift and closing loops in one feed-forward pass without place retrieval or pose-graph optimization (Fig. 2).

In summary, our contributions are as follows:

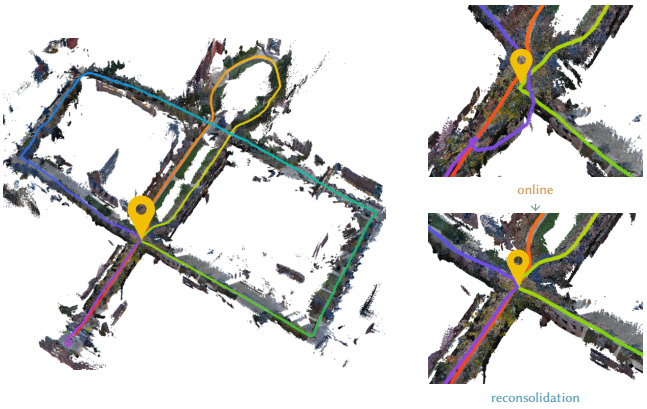
- We introduce `LingBot-Map 2.0`, built around Geometric Context Reconsolidation (GCR), which revises past estimates by selectively reconsolidating the geometric context that produced them: the committed context is reused as is, and only the revisable context is recomputed.
- We propose a mixed-mode training recipe that trains the revisable layers for causal streaming, bidirectional reconsolidation, and their transitions, with sequence scale normalization and adjacent relative pose supervision.
- We evaluate `LingBot-Map 2.0` on nine settings across seven datasets. In the streaming setting, it outperforms existing streaming methods on nearly all metrics; after reconsolidation, a purely feed-forward pass without an optimization backend, it surpasses streaming methods with loop-closure backends on long-trajectory and scene-level benchmarks, as well as offline models on most metrics (Fig. 1).



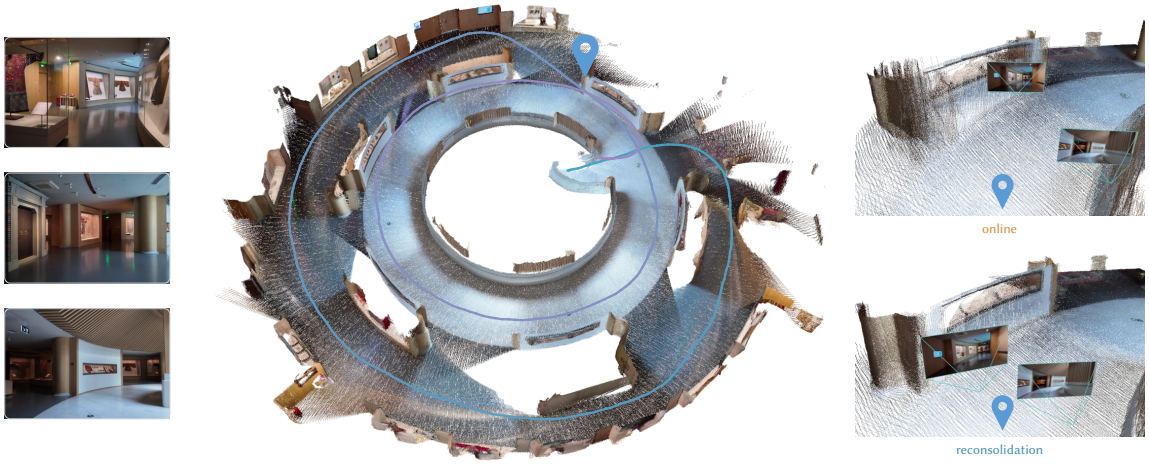
(a) Outdoor reconstruction



(b) Indoor reconstruction



(c) Trajectory refinement



(d) Point-cloud refinement

Figure 2. Reconstruction and reconsolidation with LingBot-Map 2.0. (a) Outdoor reconstruction with three local detail views and (b) indoor reconstruction with its camera trajectory and selected input-camera frusta, both after reconsolidation and shown alongside sampled input frames. (c) Trajectory refinement on a campus revisit and (d) point-cloud refinement in a museum compare online estimates with reconsolidated results.

2 Related Work

Traditional 3D Reconstruction. Classical pipelines recover camera motion and dense geometry from correspondences via structure from motion (SfM), visual SLAM, and multiview stereo (MVS) [11, 60, 61], where SfM reconstructs unordered photo collections offline, SLAM tracks video streams online, and MVS estimates dense surfaces from posed views. Learning-augmented systems such as MVSNet [86], DROID-SLAM [69], and DPV-SLAM [43, 70] replace hand-crafted stages with learned cost volumes or dense optical-flow updates, yet accurate reconstruction still hinges on per-scene iterative bundle adjustment and known camera intrinsics.

3D Foundation Models. Feed-forward 3D foundation models dispense with per-scene optimization by regressing scene geometry directly from unposed images. DUSt3R [79] predicts pairwise metric pointmaps in a shared coordinate frame but requires post-hoc global alignment beyond two views, whereas VGGT [75] jointly estimates cameras and dense geometry across arbitrary view sets through alternating frame and multiview attention. Subsequent works advance this paradigm through permutation-equivariant architectures in π^3 [81], stronger large-scale geometric accuracy in VGGT-Ω [76] and Depth Anything 3 [41], and token-efficient inference in VGG-T^3 and FastVGGT [21, 63]. Because these models attend globally over a fixed batch of views, however, memory scales poorly to multi-thousand-frame videos and ongoing streams cannot be updated incrementally without recomputing earlier views.

Streaming 3D Reconstruction. For online streams, sequential models instead consolidate incoming views into a persistent state as frames arrive, using fixed-capacity spatial or recurrent memory in Spann3R and CUT3R [74, 78] or causal KV caches in StreamVGGT and STream3R [33, 95]. To scale to long trajectories with bounded compute, subsequent works explore KV-cache pruning and compression [16, 47, 89], test-time state adaptation in TTT3R and Scal3R [14, 84], sliding-window or chunked relative alignment in LoGeR, HorizonStream, and ABot-Recon [17, 28, 91], and compact anchor-window-memory consolidation in LingBot-Map [13]. Across all of these causal designs, each view enters the context only once and is never revisited, so errors from early partial observations persist in the state and compound into long-range drift.

Global Consistency and Refinement. Overcoming such drift calls for revising earlier predictions as new evidence arrives. Classical SLAM enforces global consistency through visual loop detection and pose-graph or bundle adjustment [11, 69]. Building on offline 3D models, hybrid systems such as MASt3R-SLAM and VGGT-SLAM [20, 45, 48, 51] align local submaps via external optimization, while backend stages for streaming models such as CLoSeR and Map2SLAM [15, 17, 25, 28, 36, 84] refine output trajectories through place retrieval and graph optimization without updating the streaming model's cached context, leaving subsequent frames reading drifted states. Closest to our work, SLAM-Former [90] and SLAMFormer-∞ [22] revise states inside a shared transformer, respectively updating cached tokens via global attention and scaling through locally conditioned chunks and iterative graph optimization, yet both reopen every backbone layer and recompute per-view local geometry alongside cross-view alignment. In contrast, `LingBot-Map 2.0` partitions its context into an invariant *committed context* and an upper *revisable context*, reusing committed layers as is and reopening only revisable layers in one feed-forward pass to correct past estimates and reconsolidate the cache.

3 Method

`LingBot-Map 2.0` reconstructs a sequential RGB stream $\{I_t\}$ online, predicting a camera g_t and a depth map D_t for each view while revising earlier predictions as the stream progresses (Fig. 3). Our key idea is to revise past estimates by reconsolidating the geometric context that produced them, rather than correcting the estimates themselves. Updating this internal context corrects historical geometry and refreshes the state read by future views at once. Section 3.1 reviews how LingBot-Map consolidates observations into geometric context. Section 3.2 introduces Geometric Context Reconsolidation, which lets later views correct earlier entries of this context while the causal history that streaming builds on stays intact, Sec. 3.3 presents the full framework and its training, and Sec. 3.4 describes the inference pipeline.

3.1 Preliminaries: Geometric Context Attention

We build on the Geometric Context Transformer of LingBot-Map [13], which tokenizes each view into image tokens paired with context tokens (a camera token and register tokens), processes them through $L = 24$ layers alternating

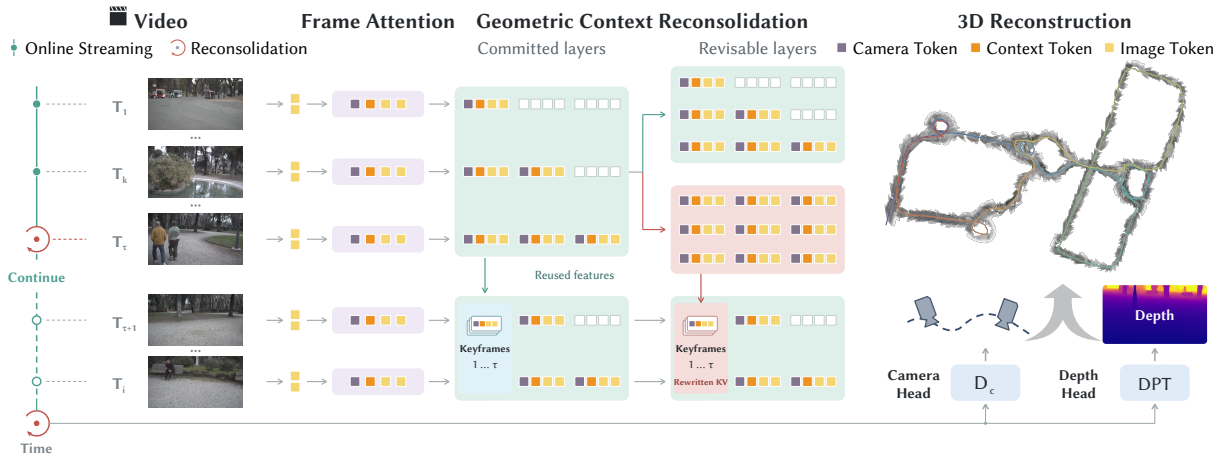


Figure 3. Method overview. Reconsolidation reuses the committed context and recomputes the revisable layers bidirectionally from stored tokens. Shared camera and depth heads update past estimates, and subsequent views read the rewritten context. Red arrows mark mid-stream and final reconsolidation; the point cloud shows VBR Pincio after reconsolidation (480 keyframes).

within-view frame attention and cross-view Geometric Context Attention (GCA), and decodes cameras and depth via a causal camera head and a multi-layer DPT [56] head.

Geometric Context Attention. GCA keeps cross-view attention causal and compact by drawing on three sources of context. Streaming retains a subset of frames as keyframes (Sec. 3.4). A fixed initialization prefix $\mathcal{A}$ establishes a shared coordinate frame and scale, a local window $\mathcal{W}_t$ of up to W views registers frame t against the most recent keyframes, and the earlier keyframes $\mathcal{K}_t$ contribute only their context tokens as a long-range trajectory memory. Following LingBot-Map [13], queries from frame t attend to

$$\mathcal{V}_{\mathrm{s}}(t) = \underbrace{\mathcal{T}(\mathcal{A})}_{\text{anchors}} \cup \underbrace{\mathcal{T}(\mathcal{W}_t)}_{\text{window}} \cup \underbrace{\mathcal{C}(\mathcal{K}_t)}_{\text{memory}}, \tag{1}$$

where $\mathcal{T}(\cdot)$ and $\mathcal{C}(\cdot)$ denote all tokens and context tokens of a frame set.

Consolidation. We refer to the streaming key-value (KV) cache as the *geometric context* $\mathcal{G}_t$, storing the keys and values visible under Eq. (1) at every GCA layer together with an analogous camera-head cache. During streaming inference, we maintain this cache through what we term *context consolidation*,

$$\mathcal{G}_t = \mathcal{G}_{t-1} \oplus \phi_\theta(I_t; \mathcal{G}_{t-1}), \tag{2}$$

where ϕ_θ encodes a new keyframe by attending to $\mathcal{G}_{t-1}$, and $\oplus$ appends its keys and values while the keyframe leaving the window exposes only its context tokens to later views. Intermediate frames read $\mathcal{G}_{t-1}$ without modifying it. Since consolidation only appends and compresses, each entry is computed once from the views available at the time and never revisited. Errors arising from these partial observations therefore persist in the context and propagate to every later prediction that attends to it, accumulating as drift over long sequences.

3.2 Geometric Context Reconsolidation

Consolidation thus poses a central challenge for streaming reconstruction: the geometric context must remain fixed to serve as a reference for subsequent views, yet the errors it contains are typically revealed only by those later views. Following the principle of memory reconsolidation (Sec. 1), Geometric Context Reconsolidation revises this context itself and reopens only the part that later views mainly need to correct. It partitions the geometric context by layer into two parts: a *committed context*, which is invariant to subsequent views and serves as a stable reference for continued streaming, and a *revisable context*, which is recomputed with bidirectional attention over the observed keyframes and written back to serve both past estimates and future predictions. We detail each part below and then describe the three modes that this design enables for streaming reconstruction.

Committed and revisable context. The first L_c layers always follow the causal visibility of Eq. (1), whether the model is streaming or reconsolidating, and we refer to them as *committed layers*. The remaining $L - L_c$ *revisable layers* follow the same rule during streaming but attend bidirectionally across the observed keyframes during reconsolidation. This split reflects that later views mainly correct the relative poses and alignment between views rather than the local geometry of each view: the lower layers mainly capture per-view geometry, whereas the upper layers align views with one another. We set $L_c = 12$ and study this choice in Tab. 7. Accordingly, the geometric context at time τ is split into the committed context $\mathcal{G}^{\mathrm{c}}_\tau$, which holds the keys and values of the committed layers, and the revisable context $\mathcal{G}^{\mathrm{r}}_\tau$, which holds those of the revisable layers together with the camera head cache, since the camera head reads the output of the revisable layers.

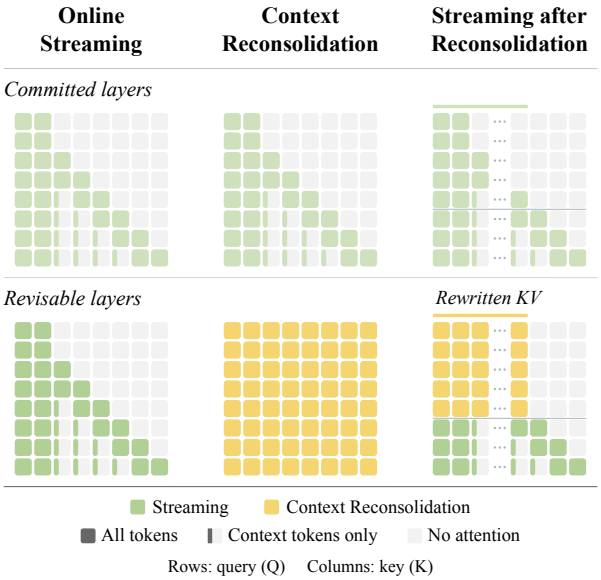


Figure 4. Attention across modes. Committed layers stay causal; revisable layers switch from causal streaming to bidirectional reconsolidation. Streaming then resumes with the rewritten context.

The committed context is invariant to subsequent views. Intuitively, a committed layer never attends to views that arrive after keyframe k, so nothing observed later can alter what it computes for k. Formally, let F^l_k denote the features of keyframe k at a committed layer $l \le L_c$; induction over l shows that F^l_k depends only on frame k, the initialization prefix $\mathcal{A}$, and the keyframes retained before k. Hence neither F^l_k nor $\mathcal{G}^{\mathrm{c}}_\tau$ changes when later views arrive or when the revisable layers switch visibility. In particular, the tokens output by the last committed layer, $H_k = F^{L_c}_k$, are fixed once keyframe k has been processed.

Let $\mathcal{K}_\tau$ be the keyframes retained up to time τ. Reconsolidation can therefore be selective, recomputing only $\mathcal{G}^{\mathrm{r}}_\tau$ from $H_{\mathcal{K}_\tau} = \{H_k : k \in \mathcal{K}_\tau\}$ while reusing $\mathcal{G}^{\mathrm{c}}_\tau$ as is:

$$\mathcal{G}^{\mathrm{r}}_\tau \leftarrow \rho_\theta(H_{\mathcal{K}_\tau}), \qquad \mathcal{G}^{\mathrm{c}}_\tau \text{ unchanged}, \tag{3}$$

where ρ_θ is described below.

The two parts of the context are maintained and used in three modes that share the same weights (Fig. 4): Online Streaming builds both parts, Context Reconsolidation rewrites the revisable part, and Streaming after Reconsolidation reads the two parts together.

Online Streaming. Here, each incoming keyframe passes through all L layers under the causal visibility of Eq. (1), and, following Eq. (2), the keys and values of its committed layers are appended to $\mathcal{G}^{\mathrm{c}}$, while those of its revisable layers and the camera head are appended to $\mathcal{G}^{\mathrm{r}}$. In addition, we store the tokens H_k, which are exactly the inputs that the revisable layers need to process this keyframe again, together with the committed-layer features used by the depth head. These tokens cannot be recovered from the cached keys and values, so they are stored separately; since H_k is fixed, storing them once suffices.

Context Reconsolidation. Context Reconsolidation realizes Eq. (3) at any time τ: it keeps $\mathcal{G}^{\mathrm{c}}_\tau$ and recomputes $\mathcal{G}^{\mathrm{r}}_\tau$ from the stored tokens $H_{\mathcal{K}_\tau}$, so no image is re-encoded and no committed layer is rerun. Specifically, the revisable layers reprocess the stored tokens of all observed keyframes, now attending bidirectionally across them. Writing $U_\theta(H; \mathcal{V})$ for the revisable layers, which map layer-L_c tokens H to final-layer features under visibility $\mathcal{V}$, this pass computes

$$Z^{\mathrm{r}}_{\mathcal{K}_\tau} = U_\theta\big(H_{\mathcal{K}_\tau}; \mathcal{T}(\mathcal{K}_\tau)\big), \tag{4}$$

where $\mathcal{T}(\mathcal{K}_\tau)$ contains all tokens of the observed keyframes. It uses the same weights as Online Streaming and accesses no frame after τ. Because every keyframe now attends to all others, earlier keyframes can draw on later views, including revisits.

The new features are used in two ways. For past estimates, the camera head is replayed over the keyframes in temporal order under its causal rule, decoding $Z^{\mathrm{r}}_{\mathcal{K}_\tau}$ into reconsolidated cameras, and the depth head is reapplied to each

keyframe using the stored committed-layer features and the new revisable-layer features. For future views, the keys and values produced in this pass by the revisable layers and the camera head replace those in $\mathcal{G}_\tau^{\mathrm{r}}$, yielding $\hat{\mathcal{G}}_\tau^{\mathrm{r}}$.

Streaming after Reconsolidation. Streaming after Reconsolidation reads the two parts together, resuming from $\mathcal{G}_\tau^{\mathrm{c}} \cup \hat{\mathcal{G}}_\tau^{\mathrm{r}}$. For each new view, the committed layers attend to the unchanged causal history in $\mathcal{G}_\tau^{\mathrm{c}}$, while the revisable layers and the camera head attend to the rewritten context in $\hat{\mathcal{G}}_\tau^{\mathrm{r}}$, and new keyframes are appended to both parts as in Online Streaming. The two parts stay aligned because both $\hat{\mathcal{G}}_\tau^{\mathrm{r}}$ and the new view's inputs to the revisable layers come from the unchanged committed layers, and mixed-mode training (Sec. 3.3) exposes the revisable layers to causal queries over bidirectionally computed context. Correcting past predictions alone would leave subsequent views reading stale context, whereas rewriting the context lets their predictions build on the corrected history.

3.3 LingBot-Map 2.0 Framework

Architecture. The overall architecture is illustrated in Fig. 3. Each input image is encoded by a ViT into image tokens, which are augmented with a camera token and register tokens as in Sec. 3.1. The tokens are processed by $L = 24$ alternating layers of frame attention and GCA, of which the first $L_c = 12$ are committed and the remaining 12 are revisable (Sec. 3.2). Frame attention operates within each view, whereas GCA operates across views under Eq. (1), or bidirectionally over the observed keyframes in the revisable layers during reconsolidation. Finally, a camera head refines the final-layer camera token over four iterations with causal attention across views to predict the absolute camera pose, whereas a frame-wise DPT depth head fuses image tokens across both committed and revisable layers (layers 5, 12, 18, and 24) without a cross-view cache to predict the depth map. All three modes share these weights.

Scale normalization. To supervise geometry consistently across modes, before computing the losses we normalize ground truth and predictions separately for each sequence. Let $\mathcal{P}$ be a randomly sampled temporal prefix and $M_{\mathcal{P}}$ its valid ground-truth depth mask. Using the same prefix and mask for both quantities, we compute

$$s_{\mathrm{gt}} = \mathrm{mean}_{M_{\mathcal{P}}}\, D_{\mathrm{gt}}, \qquad s_{\mathrm{pred}} = \mathrm{mean}_{M_{\mathcal{P}}}\, \hat{D}. \tag{5}$$

We divide ground-truth depths and camera translations by s_{gt}, and predicted depths and the translations from every camera prediction stage by s_{pred}. Each scale is shared across the sequence. Gradients propagate through s_{pred}; rotations and intrinsics are unaffected. This normalization removes the scale ambiguity at the training loss boundary. Inference requires no ground truth and reconstructs geometry up to an unknown scale.

Loss function. We supervise depth, absolute camera pose, and relative pose after the sequence scale normalization above,

$$\mathcal{L} = \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{abs}}\mathcal{L}_{\mathrm{abs}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{depth}}$ and $\mathcal{L}_{\mathrm{abs}}$ follow VGGT and LingBot-Map. LingBot-Map computes $\mathcal{L}_{\mathrm{rel}}$ over frame pairs at wide temporal offsets, leaving adjacent pairs as a small fraction of the loss. Yet errors between consecutive frames directly cause trajectory jitter and local drift. We instead compute $\mathcal{L}_{\mathrm{rel}}$ over every adjacent pair. Given camera poses $T_t \in \mathrm{SE}(3)$ with scale-normalized translations in the world frame, let $\Delta T_t = T_t^{-1} T_{t-1}$ be the ground truth motion increment with rotation ΔR_t and translation $\Delta \mathbf{t}_t$, and define $\Delta \hat{T}_t$ analogously from the predictions. For a sequence of N frames,

$$\begin{aligned}\mathcal{L}_{\mathrm{rel}} = \frac{1}{N-1} \sum_{t=2}^{N} \Big[& \lambda_{\mathrm{tr}} \big\| \Delta \hat{\mathbf{t}}_t - \Delta \mathbf{t}_t \big\|_1 \\ & + \lambda_{\mathrm{rot}}\, d_{\mathrm{rot}}(\Delta \hat{R}_t, \Delta R_t) \Big],\end{aligned} \tag{7}$$

where $d_{\mathrm{rot}}(\Delta \hat{R}_t, \Delta R_t) = \arccos\big((\mathrm{tr}(\Delta \hat{R}_t^{\top} \Delta R_t) - 1)/2\big)$ is the geodesic distance on $\mathrm{SO}(3)$, and $\lambda_{(\cdot)}$ denote scalar loss weights (Sec. 4.1). The translation term uses an L1 penalty. Supervising ground truth increments sharpens local step accuracy without penalizing genuine camera motion, while $\mathcal{L}_{\mathrm{abs}}$ and $\mathcal{L}_{\mathrm{depth}}$ supervise cameras and geometry in their normalized coordinate systems.

Mixed-mode training. Supporting all three modes with shared weights θ requires the revisable layers to handle causal context, bidirectional context, and transitions between them. Continuing from the geometric streaming base model trained under the progressive view curriculum of LingBot-Map [13] (Sec. 4.1), mixed-mode continuous training jointly optimizes consolidation, reconsolidation, and their transitions (Sec. 4.2). At each batch, we sample one of three

attention patterns with equal probability and apply it only to the 12 revisable layers. The committed layers remain causal under Eq. (1) in every pattern, including the initialization prefix connections, so the invariance of Sec. 3.2 holds throughout training. Each pattern exposes these layers to a different kind of context: (i) a causal pattern, where views read consolidated context as in Online Streaming; (ii) a fully bidirectional pattern, as in Context Reconsolidation; and (iii) a bidirectional temporal prefix followed by a causal suffix, where the suffix reads rewritten context as in Streaming after Reconsolidation. All patterns are implemented by attention masks within a single forward pass, matching the three visibilities in Fig. 4. Causal regions use a local pose reference window sampled between 16 and 64 frames, while bidirectional regions attend globally over their permitted views.

3.4 Inference Pipeline

Like autoregressive LLMs, our model caches the keys and values of processed views rather than recomputing them, and we inherit the paged KV-cache runtime of LingBot-Map [13]. After jointly encoding the initialization prefix $\mathcal{A}$, we select one keyframe every s frames (Sec. 5.2): keyframes are consolidated into $\mathcal{G}$ with H_k stored, whereas intermediate frames are predicted without modifying $\mathcal{G}$ so every frame receives an online camera and depth. Context Reconsolidation requires no loop detection and can be triggered at any point, such as after the last frame (Sec. 6.1) or at 75% of the keyframes (Sec. 6.3) in our experiments. Since it revises keyframes only, intermediate frames retain their online depths while interpolating cameras from the reconsolidated keyframes, which also form the reconsolidated dense map (Sec. 5.2).

4 Training & Inference

Learning both causal online streaming and context reconsolidation within one set of parameters requires staging how sequential memory and cross-mode visibility are acquired. Directly optimizing cached streaming attention over thousand-frame trajectories from scratch suffers from compounding early-frame drift, while introducing context reconsolidation prematurely prevents the causal layers from establishing a stable online coordinate reference. We therefore train `LingBot-Map 2.0` in two main stages: (i) *Geometric Streaming Base Model Training* first builds multi-view 3D perception offline and progressively scales causal Geometric Context Attention over extended trajectories, producing the *geometric streaming base model* of `LingBot-Map 2.0` (Sec. 4.1); and (ii) *Mixed-Mode Continuous Training* continues training this streaming base model across causal, bidirectional, and continuation masks so that a single checkpoint supports Online Streaming, Context Reconsolidation, and Streaming after Reconsolidation (Sec. 4.2). Data curation and runtime implementation follow in Secs. 4.3 and 4.4.

4.1 Geometric Streaming Base Model Training

Multi-view offline training. The image encoder is initialized from LingBot-Vision-L [24], whose local geometric representations outperform DINOv3-L [66] under strict pose thresholds (Sec. 6.3). Input frames are resized so that the longer side equals 512 pixels under preserved aspect ratios and tokenized on a 16×16 patch grid before entering the 24-layer alternating frame- and cross-view backbone [75]. We first train the backbone offline with dense cross-view attention across all views in each batch without causal masking, enabling the network to learn view-order-agnostic correspondence, camera geometry, and depth structure jointly from unstructured photo collections and short video clips.

Long-horizon causal streaming. Initialized from this offline multi-view checkpoint, all 24 cross-view layers switch to the causal visibility of Eq. (1), registering each incoming view against the initialization anchors, a local patch window, and compact context tokens from earlier keyframes. Following LingBot-Map [13], sequence length expands progressively from short clips to long trajectories so that short-range registration converges before long-range memory dependencies take effect, while the local window size W is drawn from $[16, 64]$ and 16-way context parallelism shards sequence tokens across devices. Ground-truth and predicted geometry are normalized per sequence via Eq. (5), with the scale computed on the sampled prefix and applied to all views and all four camera-head refinement steps (weighted with decay $\gamma = 0.6$ [75]), and optimization minimizes Eq. (6) with $\lambda_{\text{depth}} = 1$, $\lambda_{\text{abs}} = 5$, and $\lambda_{\text{rel}} = 50$, pairing L1 translation ($\lambda_{\text{tr}} = 2$) and $\mathrm{SO}(3)$ geodesic ($\lambda_{\text{rot}} = 0.5$) penalties over every consecutive frame pair (Eq. (7)) to suppress local step jitter alongside global trajectory drift. Completing this curriculum yields the strengthened *geometric streaming base model* of `LingBot-Map 2.0` for purely causal online reconstruction.

4.2 Mixed-Mode Continuous Training

Initialization and mixed-visibility sampling. Reopening upper causal layers bidirectionally without adaptation severely degrades pose accuracy, as attention projections trained solely on past tokens suffer distribution shift when exposed to future keys and values (Tab. 5). Continuing from the geometric streaming base model (Sec. 4.1), mixed-mode continuous training starts a fresh optimization schedule using Muon [44] together with AdamW [46] at a peak learning rate of 5×10^{-5} under 16-way context parallelism. Across every batch, the lower $L_c = 12$ committed layers remain strictly causal under Eq. (1) so their cached per-view tokens H_k stay invariant, while the 12 upper revisable layers sample one of three attention masks with equal probability (Sec. 3.3 and Fig. 4): a causal mask (preserving Online Streaming with $W \in [16, 64]$), a global bidirectional mask over all views in the sequence (learning Context Reconsolidation on top of fixed H_k), or a bidirectional prefix followed by a causal suffix (simulating Streaming after Reconsolidation, where subsequent queries attend to the rewritten cache $\hat{\mathcal{G}}^{\mathrm{r}}_{\tau}$).

Unified cross-mode supervision. Because the camera and DPT depth heads are shared across all three modes, representations produced under causal, bidirectional, and prefix-to-causal visibilities must remain geometrically compatible. We therefore apply the same sequence-level scale normalization (Eq. (5)) and loss weights as in Sec. 4.1 across all three mask types, producing the single shared checkpoint evaluated throughout Sec. 6.

4.3 Training Data and Sampling

Dataset overview. Table 1 summarizes the 36 real and synthetic sources used for training, covering indoor rooms, outdoor streets, aerial scenes, and object-centric captures while excluding every evaluation benchmark in Sec. 5. Static multi-view collections and short clips serve offline training, whereas extended video trajectories support both causal streaming and mixed-mode continuous training.

Offline training mixture. Offline training uses unstructured multi-view datasets, including BlendedMVS [87], MegaDepth [39], MegaDepthX [38], and AerialMegaDepth [73], together with 3D object collections Co3D [57], Objaverse [19], and Texverse [92], as well as indoor and synthetic video sources including ScanNet [18], ScanNet++ [88], WildRGBD [83], Aria Synthetic [3], ADT [53], PointOdyssey [93], and Kubric [27]. This heterogeneous mix exposes the backbone to wide viewpoint baselines, diverse focal lengths, and fine object geometry.

Causal and mixed-mode mixture. For causal streaming and mixed-mode continuous training, unordered photo collections are dropped so every training sample reflects continuous camera motion (Tab. 1). This trajectory pool concentrates on multi-room indoor navigation rendered with Habitat [59] across HM3D [55], Matterport3D [12], and Gibson [82], alongside large-scale outdoor, driving, aerial, and game-engine streams from TartanAir/TartanAir2 [80], TartanGround [54], MidAir [23], MatrixCity [37], KITTI-360 [40], and the OmniWorld-Game split of OmniWorld [94].

Trajectory sampling. Ordered video clips for causal and mixed-mode training are drawn from random starting points with variable frame strides; whenever a sampled walk reaches the end of a recording before filling the target length, traversal reflects backward from the boundary [13] to form an unbroken pass with natural revisit overlap. The resulting sequences supervise causal tracking in the streaming base stage and, paired with the three visibility masks above, train context reconsolidation and post-reconsolidation continuation during mixed-mode continuous training.

4.4 Inference and Cache Implementation

Paged KV-cache streaming. Building on the paged KV-cache runtime of LingBot-Map [13], inference first encodes the initialization prefix $\mathcal{A}$ jointly to establish a shared coordinate frame and scale, and then processes incoming frames sequentially at the keyframe stride defined in Sec. 5. Image-patch keys and values occupy fixed-size pages for $\mathcal{A}$, the local pose-reference window $\mathcal{W}_t$, and a single scratch slot, whereas older keyframes $\mathcal{K}_t$ retain only their compact context tokens in packed pages. Consolidating a keyframe appends its keys and values to the persistent context, recycles the patch-token page of the keyframe leaving $\mathcal{W}_t$ in $O(1)$ time, and archives its layer-L_c tokens H_k together with committed DPT features from layers 5 and 12 once to host memory. Intermediate non-keyframes are decoded online against the active cache via the scratch page, which is immediately rolled back so that persistent state grows only at keyframes.

Context reconsolidation and stream continuation. Context Reconsolidation can be invoked at any point in the stream without a loop detector—such as at sequence end or mid-stream at 75% of the keyframes in our evaluations—by retrieving the archived H_k and committed DPT features from host memory and recomputing only the 12 revisable layers

Table 1. Training dataset composition. Modality indicates unstructured multi-view sets vs. ordered video sequences. Offline train denotes multi-view offline training; Causal/mix denotes long-sequence causal streaming and mixed-mode continuous training. On OmniWorld, causal/mix training uses the OmniWorld-Game subset; HM3D, Matterport3D, and Gibson are rendered with Habitat [59].

Dataset	Modality	Offline train	Causal/ mix	Dataset	Modality	Offline train	Causal/ mix	Dataset	Modality	Offline train	Causal/ mix
BlendedMVS [87]	Multi-view	✓	✓	SceneNet RGB-D [49]	Video	✓	✓	MatrixCity [37]	Video	✓	✓
HyperSim [58]	Multi-view	✓	✓	WildRGBD [83]	Video	✓	✓	MidAir [23]	Video	✓	✓
MegaDepth [39]	Multi-view	✓	–	UnrealStereo4K [71]	Video	✓	✓	MapFree [2]	Video	✓	✓
MegaDepthX [38]	Multi-view	✓	–	Aria Synthetic [3]	Video	✓	✓	SHIFT [67]	Video	✓	✓
AerialMegaDepth [73]	Multi-view	✓	–	ADT [53]	Video	✓	✓	OmniWorld [94]	Video	✓	✓
MVS-Synth [29]	Multi-view	✓	✓	DL3DV [42]	Video	✓	✓	PointOdyssey [93]	Video	✓	–
GTA-SfM [77]	Multi-view	✓	✓	DynamicReplica [30]	Video	✓	✓	Kubric [27]	Video	✓	–
Co3D [57]	Multi-view	✓	–	TartanAir [80]	Video	✓	✓	Robo3R [85]	Video	✓	–
Objaverse [19]	Multi-view	✓	–	TartanAir2 [80]	Video	✓	✓	HM3D [55]	Video	–	✓
Texverse [92]	Multi-view	✓	–	TartanGround [54]	Video	✓	✓	Matterport3D [12]	Video	–	✓
ScanNet [18]	Video	✓	✓	Virtual KITTI 2 [10]	Video	✓	✓	Gibson [82]	Video	–	✓
ScanNet++ [88]	Video	✓	✓	KITTI-360 [40]	Video	–	✓	Internal Data	Video	✓	✓

bidirectionally across the retained keyframes. The shared camera and depth heads decode updated keyframe trajectories and depth maps (Secs. 3.4 and 5.2), while the recomputed revisable-layer and camera-head KV pools replace $\mathcal{G}_\tau^{\mathrm{r}}$ in place and reuse the invariant committed cache $\mathcal{G}_\tau^{\mathrm{c}}$ unchanged. Subsequent incoming frames then attend directly to $\mathcal{G}_\tau^{\mathrm{c}} \cup \hat{\mathcal{G}}_\tau^{\mathrm{r}}$, so reconsolidation simultaneously corrects historical estimates and refreshes the state used for future predictions.

Accuracy–speed comparison. We benchmark inference speed on a single GPU using the 1,101-frame KITTI 07 sequence after warmup, excluding model loading and disk I/O while including loop-closure stages where applicable, and plot throughput against 11-sequence KITTI AUC@30 in Fig. 5.

Online streaming and loop-closure baselines process all 1,101 raw frames. Fixed-state recurrent models such as CUT3R and TTT3R achieve high raw frame rates because their recurrent state has constant size, but suffer severe trajectory drift over kilometer-scale driving. Among causal methods that preserve long-horizon geometry, Ours (Online) runs in real time at 27.51 views/s while achieving 77.28 AUC@30, outperforming LingBot-Map, ABot-Recon, and HorizonStream in both speed and accuracy by recycling evicted patch-token pages in constant time.

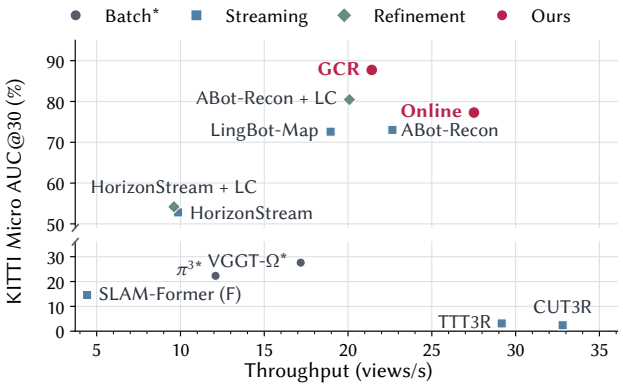


Figure 5. Accuracy–speed trade-off on KITTI. Eleven-sequence AUC@30 versus single-GPU throughput. *: offline keyframe inference; GCR: reconsolidation of cached keyframes. The vertical axis omits 36–49%.

Offline batch models and the SLAM-Former frontend are timed on the 373 retained keyframes. Starting from an existing causal cache over the same 373 retained keyframes without counting cache construction, Context Reconsolidation runs at 21.41 cached keyframes/s—reusing the committed layers without re-encoding images—and raises AUC@30 to 87.73, surpassing external loop-closure pipelines without place retrieval or pose-graph optimization. Combining full-stream online inference with sequence-end reconsolidation corresponds to an end-to-end throughput of 19.17 frames/s across the full sequence.

5 Evaluation Benchmark

Streaming 3D reconstruction must generalize from room-scale interiors and object-scale scans to multi-kilometer urban traversals under varying motion patterns, revisit intervals, and illumination conditions. To evaluate both online

consolidation and context reconsolidation across these regimes, we establish an evaluation benchmark spanning nine settings across seven primary datasets, described in Sec. 5.1 together with the evaluation protocols and metrics in Sec. 5.2.

5.1 Datasets

VBR [7] is a large-scale vision benchmark captured across historic urban streets and university campuses in Rome, featuring multi-kilometer loops, wide-baseline revisits, and centimeter-accurate RTK/LiDAR-inertial ground-truth trajectories. We evaluate on all seven sequences with public ground-truth trajectories: Colosseo-0, Campus-0, Campus-1, Pincio-0, Spagna-0, Diag-0, and Ciampino-1, spanning 8,815 to 18,846 frames and up to 5.20 km per trajectory, for both long-trajectory estimation and revisit consistency.

Oxford Spires [68] is a campus-scale dataset captured across historic Oxford courtyards, narrow passages, and indoor-outdoor transitions, providing millimeter-accurate LiDAR-inertial camera trajectories and laser-scanned reference geometry. Following LingBot-Map [13], we use all 10 sequences with ground-truth trajectories: Bodleian Library-02, Christ Church-02, 03, and 05, Keble College-02 through 05, and Observatory Quarter-01 and 02, ranging from 3,961 to 13,518 frames. We evaluate full-sequence tracking across all valid frames, dense 3D reconstruction at stride $s = 10$, and visual revisit alignment.

KITTI Odometry [26] provides city-scale autonomous driving trajectories across urban streets, rural roads, and highways. We evaluate long-trajectory tracking across all 11 sequences with ground-truth poses, namely `00` through `10`, spanning 271 to 4,661 frames and up to 5.10 km, and evaluate loop-closure consistency on the six sequences containing revisits: `00`, `02`, `05`, `06`, `07`, and `09`.

CroCoDL [6] is a cross-device collaborative localization and mapping benchmark registered against NavVis LiDAR-scanned 3D reference models, spanning unstructured indoor and outdoor ARCHE environments—including dimly lit basements, semi-collapsed structures, open-sky ruins, and outdoor plazas—captured under varying illumination and scene layouts. After excluding two sequences with inconsistent image-size metadata or mixed orientations, we evaluate both camera pose estimation and dense 3D reconstruction on 62 handheld sequences sampled from frame 0 with stride $s = 5$.

Tanks and Temples [32] is a multi-view benchmark of indoor and outdoor large-scale scenes and structures with laser-scanned ground truth. Following LingBot-Map [13], we evaluate camera pose estimation on six scenes—Barn, Caterpillar, Church, Ignatius, Meetingroom, and Truck—using all 2,185 images at stride $s = 1$.

Princeton365 [31] is a diverse RGB video benchmark spanning object-scanning, indoor, and outdoor scenes, with ground-truth six-degree-of-freedom camera poses calibrated via a synchronized 360° camera rig and multi-board markers. We evaluate camera pose estimation across all 57 validation scanning sequences, sampling from frame 0 with stride $s = 10$.

EuRoC Vicon [8] comprises the six indoor Vicon Room micro-aerial-vehicle sequences from the EuRoC MAV benchmark, featuring rapid six-degree-of-freedom motion paired with motion-capture poses and millimeter-accurate Leica 3D laser scans. We evaluate dense 3D reconstruction across all six sequences sampled with stride $s = 5$.

5.2 Evaluation Protocols and Metrics

Long Trajectory Estimation. Across VBR, Oxford Spires, and KITTI, uniform keyframe sampling [13] retains the first eight frames for $\mathcal{A}$ and selects subsequent frames with stride $s = \lceil N/B \rceil$, using $B = 480$ in the main tables and $B = 320$ in lighter ablations. After Context Reconsolidation, non-keyframe camera centers are linearly interpolated and rotations in Tab. 2 are recovered from the nearest retained keyframe k via $R_i^{\mathrm{r}} = R_k^{\mathrm{r}} (R_k^{\mathrm{s}})^{-1} R_i^{\mathrm{s}}$. Streaming baselines other than SLAM-Former process the full RGB stream directly, as does Scal3R [84] on the shorter KITTI sequences via its official disk offloading. By contrast, offline models including VGGT-Ω [76], π^3 [81], and DA3 [41], together with SLAM-Former [90] and Scal3R on the longer VBR and Oxford Spires streams, exceed GPU or host memory on multi-thousand-frame inputs; we therefore evaluate them on the same uniformly selected keyframes and recover intermediate frames via linear camera-center interpolation and SLERP [64]. All methods are scored over the full timeline with valid

ground truth using sequence-averaged Sim(3)-aligned Absolute Trajectory Error (ATE, in meters) and AUC@5 and AUC@30 over all valid within-sequence camera pairs. For KITTI ablations, we additionally report sequence-averaged translational Relative Pose Error (RPE-t, in meters), computed as RMSE over consecutive full-timeline frame pairs after Sim(3) alignment. For layer-split ablations, depth Absolute Relative Error (AbsRel) is averaged over valid pixels, evaluated keyframes, and sequences after per-sequence median scale alignment.

Camera Pose Estimation. On CroCoDL, Tanks and Temples, and Princeton365, poses are scored directly on the sampled views without interpolating skipped frames. We report AUC@15 and AUC@30 by integrating the cumulative accuracy of the per-pair maximum of relative rotation and translation angular errors up to thresholds of 15° and 30°, weighted by valid within-sequence camera-pair counts, alongside sequence-averaged Sim(3)-aligned ATE.

Dense Reconstruction. On CroCoDL, EuRoC Vicon, and Oxford Spires, reconstructions are evaluated directly on the sampled views without depth interpolation or confidence filtering. Predicted and reference point clouds are aligned via Umeyama Sim(3) [72] followed by rigid ICP [5] within the visible region shared across methods, and uniformly voxel-downsampled before computing Accuracy (Acc), Completeness (Comp), and F1 score. We use an F1 threshold of 0.25 m on CroCoDL and EuRoC Vicon, and 4 m on Oxford Spires.

Revisit Consistency. Following NetVLAD [1], detector-independent ground-truth revisits are defined by matching each view to its spatially nearest earlier camera within 5 m and 45° after at least 500 frames and 100 m of travel; matches separated by at most 20 frames are merged into a single revisit event. Under full-trajectory Sim(3) alignment, we report whole-trajectory ATE alongside the local translational error e_t in meters and rotational error e_R in degrees of the relative pose $T_i^{-1}T_j$ across matched revisit pairs.

6 Experiments

We compare `LingBot-Map 2.0` with state-of-the-art streaming, global-refinement, and offline methods across pose and trajectory estimation (Sec. 6.1), dense 3D reconstruction (Sec. 6.2), ablation studies (Sec. 6.3), and revisit consistency (Sec. 6.4), following the training setup in Sec. 4 and benchmark protocols in Sec. 5. Unless otherwise specified, all results—including both Ours (Online) and Ours (Reconsolidation)—are evaluated using the single final checkpoint from mixed-mode continuous training (Sec. 4.2).

6.1 Pose and Trajectory Estimation

Across long-trajectory benchmarks (Sec. 6.1.1) and scene-level camera pose benchmarks (Sec. 6.1.2), we report pose AUC over all valid within-sequence camera pairs alongside sequence-averaged Sim(3)-aligned ATE for Online Streaming and Context Reconsolidation.

6.1.1 Long Trajectory Estimation

We evaluate long-horizon tracking on VBR [7], Oxford Spires [68], and KITTI [26], whose trajectories span urban driving loops, indoor-outdoor campus transitions, and multi-kilometer traversals of up to nearly 19,000 frames. As shown in Tab. 2, even when evaluated on keyframes, offline foundation models degrade substantially over multi-kilometer horizons: trained primarily on bounded view sets, global attention without sequential grounding struggles to preserve consistent scale and orientation across wide-baseline transitions and sparse loops. Among causal streaming approaches, local-window pipelines such as HorizonStream and ABot-Recon maintain smooth short-range motion, yet progressively accumulate drift as older views leave the active window without a global reference. By consolidating incoming keyframes into a unified world frame anchored by $\mathcal{A}$ and compact trajectory-memory tokens, Ours (Online) curbs long-range drift and outperforms all competing streaming methods on every dataset and metric, nearly halving ATE on VBR from 29.21 m to 17.01 m.

Building on this causal state, sequence-end Context Reconsolidation reopens the revisable layers bidirectionally across the retained keyframes in a single feed-forward pass, reducing ATE to 4.72 m on VBR, 3.92 m on Oxford Spires, and 9.42 m on KITTI without place retrieval or pose-graph optimization. It consistently outperforms both full-backbone refinement in SLAM-Former and streaming pipelines paired with external loop-closure backends, including CLoSeR, Scal3R w/ LC, and ABot-Recon w/ LC. The advantage is even more pronounced at the strict AUC@5 threshold: external

Table 2. Long-trajectory estimation. AUC@5 and AUC@30 (%) over valid within-sequence camera pairs and sequence-averaged Sim(3)-aligned ATE (m). Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color). LC denotes loop closure. Non-keyframe poses linearly interpolate camera centers, with SLERP [64] (†) or nearest-keyframe relative online rotation (‡); Scal3R [84] runs on all KITTI frames.

		VBR			Oxford Spires			KITTI		
LC	Method	AUC@5↑	AUC@30↑	ATE↓	AUC@5↑	AUC@30↑	ATE↓	AUC@5↑	AUC@30↑	ATE↓
Offline										
✗	VGGT-Ω†	**4.91**	**38.41**	**65.65**	**14.59**	**63.96**	**10.11**	**4.23**	**27.61**	**122.19**
✗	π^3†	2.36	24.95	73.83	5.78	37.44	19.60	2.94	22.30	166.69
✗	π^3X†	4.49	32.90	66.16	10.04	51.09	13.23	3.82	26.03	155.29
✗	DA3†	2.20	23.48	81.69	11.17	45.99	19.33	1.60	15.31	183.50
Streaming										
✗	SLAM-Former (frontend)†	0.68	13.66	87.28	1.16	14.28	28.81	0.78	14.57	199.42
✗	LingBot-Map	8.66	50.74	31.12	29.85	68.23	8.25	17.24	72.59	24.07
✗	HorizonStream	5.51	42.52	29.21	1.09	20.37	18.39	6.52	52.81	22.50
✗	ABot-Recon	5.10	42.46	29.89	5.13	46.83	7.38	21.77	73.04	18.32
✗	**Ours (Online)**	**14.74**	**63.57**	**17.01**	**43.41**	**78.59**	**6.17**	**22.42**	**77.28**	**17.56**
Global refinement										
✗	SLAM-Former (w/ backend)†	1.04	15.07	87.41	4.14	26.54	26.56	0.91	13.56	211.62
✓	CLoSeR	17.76	71.30	11.16	29.94	81.16	4.45	21.07	79.20	12.79
✓	Scal3R w/ LC†	2.14	15.73	60.22	11.33	49.92	11.08	39.85	84.95	13.72
✓	HorizonStream w/ LC	4.56	43.98	17.78	1.21	18.97	15.32	6.06	54.18	17.98
✓	ABot-Recon w/ LC	10.92	64.80	9.98	5.82	58.72	5.37	24.77	80.45	13.56
✗	**Ours (Reconsolidation)**‡	**37.94**	**85.47**	**4.72**	**74.71**	**93.16**	**3.92**	**41.00**	**87.73**	**9.42**

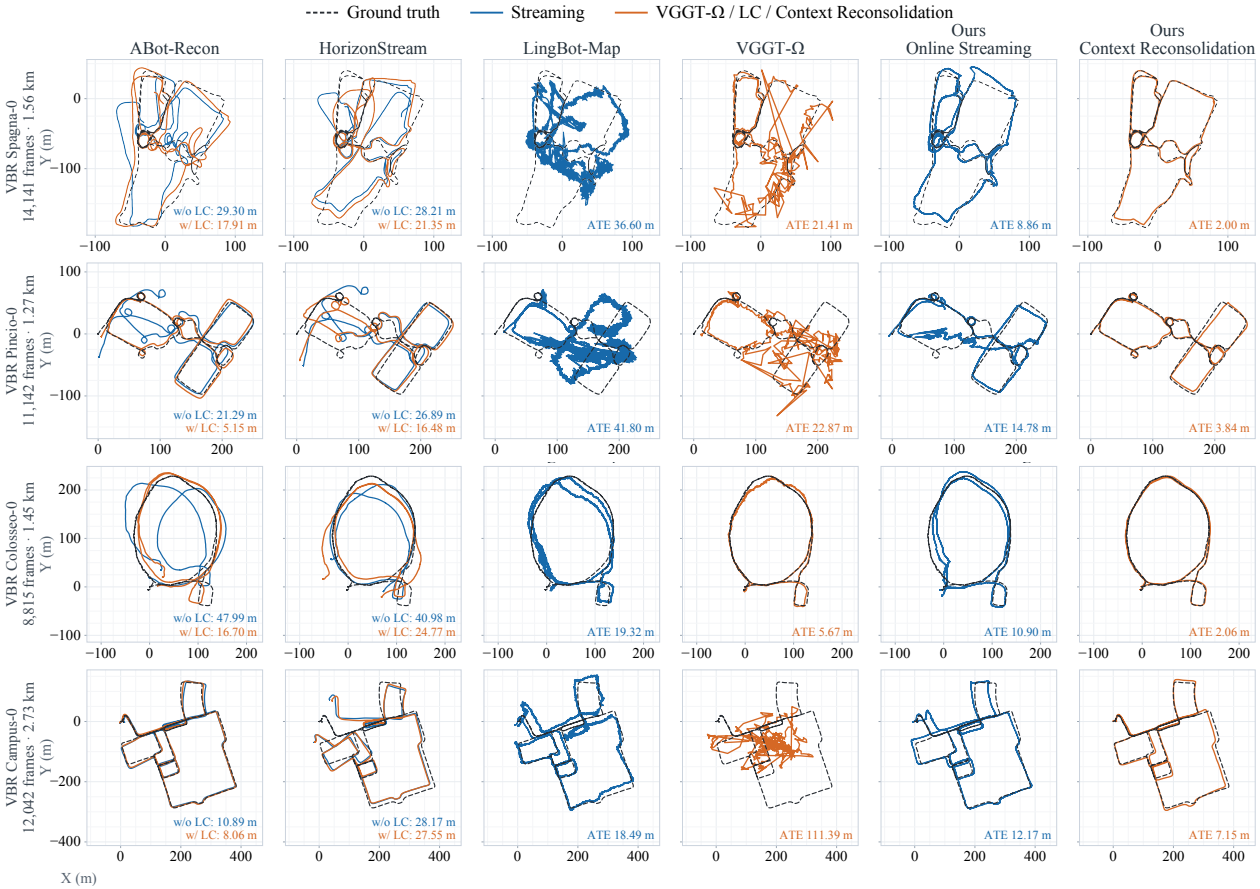


Figure 6. Long-trajectory comparisons on four VBR sequences. Sim(3)-aligned to GT; units in meters.

pose-graph backends redistribute coarse loop residuals across trajectory nodes post hoc without updating internal visual features, leaving residual angular discrepancies across wide-baseline camera pairs. By re-aligning keyframe representations directly in the revisable layers, Ours (Online) already surpasses every loop-closure baseline in AUC@5 on Oxford Spires at 43.41, and Context Reconsolidation more than doubles AUC@5 over the strongest loop-closure competitors on VBR (37.94) and Oxford Spires (74.71), while still leading on KITTI at 41.00. Qualitative comparisons in Fig. 6 and additional trajectories in Fig. S1 illustrate this behavior across VBR, Oxford Spires, and KITTI: whereas offline and sliding-window baselines exhibit severe scale drift or unclosed loops over multi-building paths, Online Streaming preserves global trajectory structure and Context Reconsolidation brings revisited loops into tight alignment with ground truth.

Per-sequence results and extended baselines. Appendix A (Tabs. S1 to S3) reports per-sequence and mean ATE across all 7 VBR, 10 Oxford Spires, and 11 KITTI trajectories against an extended set of baselines spanning efficient offline foundation models, geometric optimization SLAM systems—several using ground-truth camera intrinsics—and cache-pruning, recurrent, or periodically reset streaming models. Across these extended comparisons, recurrent-state and token-pruning caches progressively lose early geometric context over multi-thousand-frame horizons; periodic 100-frame resets curb state saturation within short segments, yet re-initializing the state breaks global frame continuity across sub-trajectories and leaves a wide gap behind continuous context consolidation in Ours (Online). Compared with multi-stage geometric optimization systems that rely on iterative bundle or pose-graph optimization, Ours (Reconsolidation) achieves substantially lower mean ATE on VBR and KITTI and outperforms all uncalibrated baselines on Oxford Spires in a single feed-forward pass.

6.1.2 Camera Pose Estimation

We evaluate scene-level camera pose estimation on CroCoDL [6], Tanks and Temples [32], and Princeton365 [31] (Sec. 5.2 and Tab. 3). On Tanks and Temples and Princeton365, where the full view set fits in GPU memory, offline foundation models provide strong baselines by attending jointly across all views, whereas sliding-window streaming methods lose global reference as the camera orbits around foreground structures. Maintaining long-range context tokens in a unified world frame allows Ours (Online) to preserve cross-orbit registration, nearly doubling AUC@15 over competing streaming baselines on Princeton365 to 41.40 while already achieving a lower 0.15 m ATE than full-attention offline models. Context Reconsolidation further tightens multi-view alignment across both benchmarks, edging past π^3 and DA3 on Tanks and Temples at 93.33 AUC@15 and establishing a clear lead across all 57 Princeton365 scanning sequences with 53.17 AUC@15 and 0.09 m ATE.

Table 3. Camera pose estimation. AUC@15 and AUC@30 (%) are weighted by valid camera-pair counts; ATE (m) is averaged equally across sequences. Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color). LC denotes loop closure. "–" indicates that LingBot-Map did not report AUC@15 on TAT [13].

		CroCoDL			Tanks and Temples			Princeton365		
LC	Method	AUC@15↑	AUC@30↑	ATE↓	AUC@15↑	AUC@30↑	ATE↓	AUC@15↑	AUC@30↑	ATE↓
Offline										
✗	VGGT-Ω	**58.08**	**70.18**	**1.26**	82.13	84.20	0.65	48.32	56.51	**0.17**
✗	π^3	41.83	54.67	2.11	91.86	**95.08**	0.15	44.20	52.58	0.22
✗	DA3 (Nested 1.1)	40.62	52.90	2.83	**92.72**	95.01	**0.14**	**48.85**	**57.35**	**0.17**
Streaming										
✗	SLAM-Former (frontend)	6.95	15.39	5.95	50.03	60.77	0.83	15.99	28.04	0.33
✗	LingBot-Map	34.42	49.05	2.12	–	92.34	0.21	22.03	34.77	0.20
✗	HorizonStream	11.44	20.92	4.48	49.66	68.16	0.67	15.88	29.84	0.28
✗	ABot-Recon	17.91	32.49	3.41	49.50	69.24	0.55	22.38	42.16	0.21
✗	**Ours (Online)**	**50.70**	**66.82**	**1.11**	**87.67**	**92.99**	**0.19**	**41.40**	**53.09**	**0.15**
Global refinement										
✗	SLAM-Former (w/ backend)	15.13	26.39	4.45	84.61	91.22	0.17	36.62	47.17	0.25
✓	CLoSeR	15.72	27.68	4.42	63.09	77.25	0.45	41.71	61.75	0.16
✓	HorizonStream w/ LC	9.47	18.57	4.37	48.12	67.84	0.75	10.18	20.29	0.31
✓	ABot-Recon w/ LC	17.92	32.51	3.41	49.72	69.47	0.55	22.57	42.67	0.21
✗	**Ours (Reconsolidation)**	**74.09**	**85.30**	**0.44**	**93.33**	**95.97**	**0.14**	**53.17**	**67.10**	**0.09**

The gain from context reconsolidation is especially pronounced on CroCoDL, whose handheld sequences traverse dimly lit basements, semi-collapsed ruins, and open plazas under harsh lighting transitions and rapid viewpoint changes. In these cluttered trajectories, external loop-closure backends relying on visual place retrieval rarely establish reliable loop constraints, leaving ABot-Recon w/ LC and HorizonStream w/ LC nearly unchanged from their streaming frontends. Even in causal mode, Ours (Online) reaches 1.11 m ATE, outperforming both retrieval-based refinement pipelines and offline VGGT-Ω. Bypassing retrieval gates, Context Reconsolidation attends directly across all retained keyframes in the revisable layers, boosting AUC@15 from 50.70 to 74.09 and AUC@30 from 66.82 to 85.30, cutting ATE to 0.44 m, and surpassing VGGT-Ω by over 15 AUC@30 points.

6.2 Dense Reconstruction

We evaluate dense 3D reconstruction on CroCoDL, EuRoC Vicon [8], and Oxford Spires (Sec. 5.2 and Tab. 4), spanning unstructured indoor–outdoor ruins, agile room-scale micro-aerial flights with millimeter-accurate laser ground truth, and multi-building campus traversals. Because unprojecting per-view depths into a shared 3D space couples local depth precision with cross-view pose alignment, accumulated drift in sliding-window baselines projects repeated structures to offset locations, producing duplicated walls and coverage gaps. Consolidating keyframes in a unified world frame online allows Ours (Online) to reach 67.03 F1 on CroCoDL and 94.63 F1 on Oxford Spires—already surpassing both streaming and offline baselines on these extended trajectories—while matching the best 0.03 m completeness on EuRoC.

Table 4. Dense reconstruction. Sequence-averaged accuracy (Acc, m), completeness (Comp, m), and F1 (%). F1 thresholds are 0.25 m for CroCoDL and EuRoC and 4 m for Oxford. Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color).

	CroCoDL			EuRoC			Oxford		
Method	Acc↓	Comp↓	F1↑	Acc↓	Comp↓	F1↑	Acc↓	Comp↓	F1↑
Streaming									
SLAM-Former (frontend)	0.61	1.13	39.53	0.20	0.05	81.15	3.72	4.56	68.50
LingBot-Map	0.44	0.66	52.95	**0.16**	**0.03**	**88.02**	1.78	0.95	92.30
HorizonStream	0.55	1.65	32.20	0.25	0.07	72.02	1.58	1.32	91.70
ABot-Recon	0.63	1.18	37.28	0.24	0.05	74.61	1.67	0.98	91.43
Ours (Online)	**0.38**	**0.34**	**67.03**	0.18	**0.03**	85.72	**1.35**	**0.65**	**94.63**
Offline and reconsolidation									
VGGT-Ω	0.43	0.45	60.69	0.10	0.05	94.96	2.24	3.03	84.47
π^3	0.45	0.60	56.25	**0.09**	**0.04**	96.63	2.32	1.87	83.77
DA3	0.45	0.60	59.33	0.13	**0.04**	91.69	2.29	1.36	87.70
Ours (Reconsolidation)	**0.31**	**0.27**	**77.14**	**0.09**	**0.04**	**97.71**	**1.03**	**0.66**	**95.46**

Context Reconsolidation re-applies the shared multi-scale DPT head to fuse the invariant committed features from lower layers, which preserve sharp per-view local geometry without re-encoding images, with the bidirectionally updated revisable features from upper layers, which resolve global pose and scale drift. Examining Accuracy and Completeness separately clarifies how reconsolidation improves surface geometry across scales. On room-scale EuRoC, short indoor trajectories already let online streaming cover the room densely, and reconsolidation halves Acc from 0.18 m to 0.09 m to reach 97.71 F1, edging past π^3 and VGGT-Ω. On campus-scale Oxford Spires, offline models suffer degraded completeness over multi-thousand-frame horizons, whereas Ours (Reconsolidation) reduces Acc to 1.03 m while preserving 0.66 m completeness at 95.46 F1. On CroCoDL, resolving multi-view drift across cluttered, poorly lit scenes lowers Acc to 0.31 m and Comp to 0.27 m, reaching 77.14 F1 and surpassing VGGT-Ω by over 16 points.

Qualitative comparisons. Figure 7 and Fig. S2 (App. B) compare dense point clouds and camera trajectories from `LingBot-Map 2.0`, ABot-Recon, and VGGT-Ω across 12 outdoor, mixed indoor–outdoor, and indoor sequences spanning urban driving loops, campus courtyards and arcades, narrow street corridors, gardens, and multi-room interiors. Across multi-block loops and indoor–outdoor transitions, ABot-Recon accumulates drift once early keyframes leave its sliding window, leaving revisited streets unclosed and splitting repeated facades into double surfaces, while VGGT-Ω exhibits global scale warping and fragmented sub-trajectories over long horizons. In contrast, `LingBot-Map 2.0` closes large-scale loops cleanly without pose-graph optimization, preserving planar ground surfaces, crisp architectural facades, and coherent 3D structure across scales.

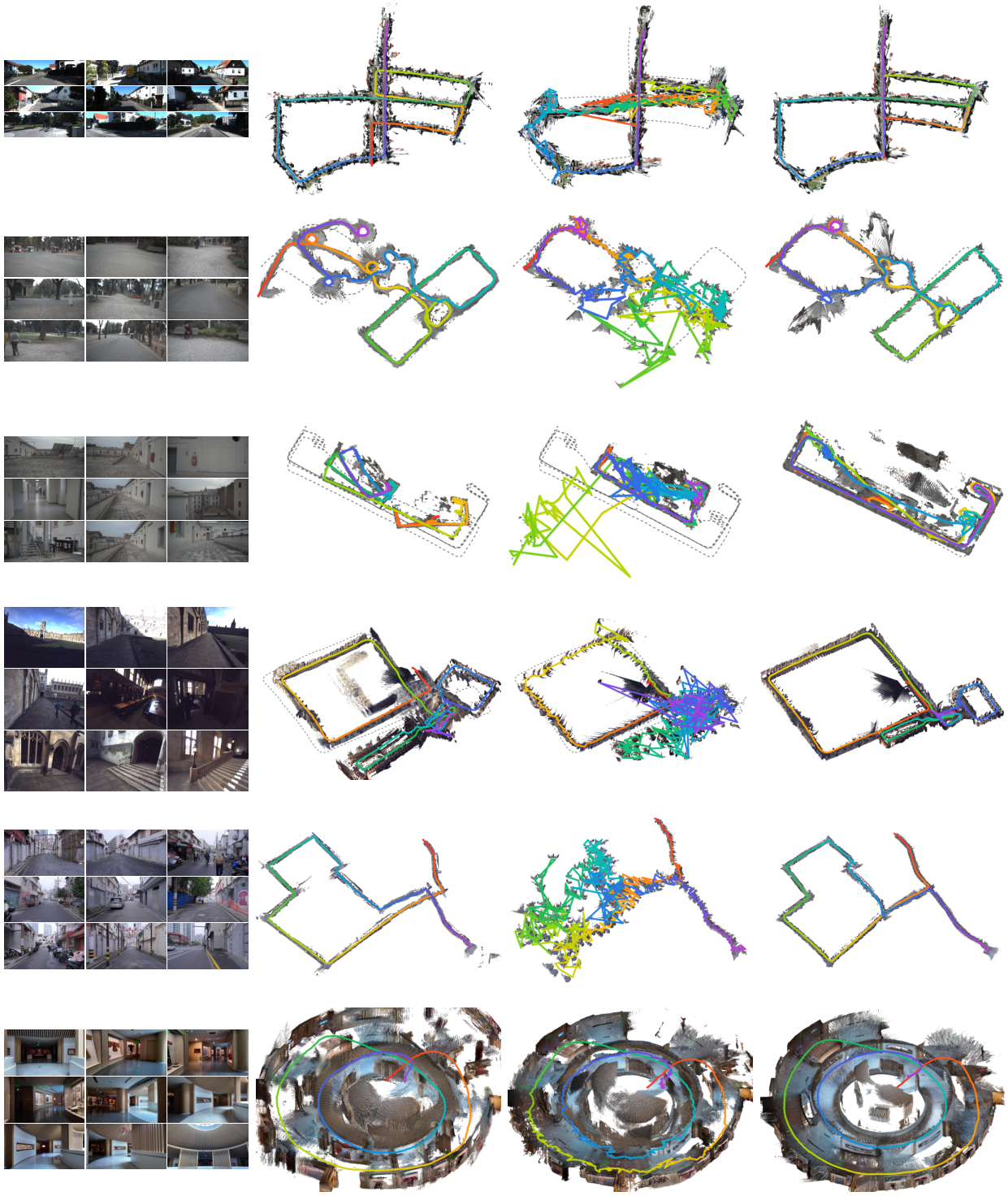


Figure 7. Qualitative 3D reconstruction comparisons. Sampled input frames (left) and reconstructed 3D point clouds overlaid with camera trajectories for ABot-Recon, VGGT-Ω, and `LingBot-Map 2.0` across large-scale outdoor driving/campus sequences and a multi-loop indoor scene.

6.3 Ablation Studies

We conduct ablation studies to examine our training design, mid-stream context write-back, the layer split between committed and revisable stages, and visual encoder initialization (Tabs. 5 to 8).

Training components. Using dedicated checkpoints trained under a lighter schedule—distinct from our full-schedule main model—we start from LingBot-Map [13] as the causal baseline and evaluate progressive training choices on KITTI in Tab. 5. Applying bidirectional reconsolidation directly to the purely causal baseline weights causes trajectory accuracy to collapse, with ATE rising from 24.04 m to 72.28 m: upper layers trained solely under causal masks suffer severe distribution mismatch when exposed to future keys and values at inference time. Extending causal training to longer sequences substantially curbs online drift, and adding dense adjacent relative-pose supervision in Eq. (7) cuts frame-to-frame translational jitter RPE-t nearly in half from 1.38 m to 0.81 m by constraining consecutive motion increments alongside global scale-normalized poses. Even so, causal streaming training alone leaves zero-shot reconsolidation degraded. Incorporating mixed-mode continuous training (Sec. 3.3) exposes the revisable layers to causal, bidirectional, and reconsolidated-prefix attention patterns within a single set of weights, reducing GCR ATE to 14.96 m and RPE-t to 0.20 m while also achieving the lowest online trajectory error.

Table 5. Cumulative training components on KITTI. Long: long-sequence training; Adj.: adjacent-pair supervision; Mix: mixed-mode training. Components are added progressively from the baseline (first row) under the lighter training schedule. ATE and RPE-t are in meters; AUC@30 is in percent. Bold/underline mark best/second best.

Training components			Online Streaming			Context Reconsolidation		
Long	Adj.	Mix	ATE↓	AUC@30↑	RPE-t↓	ATE↓	AUC@30↑	RPE-t↓
–	–	–	24.04	73.56	2.84	72.28	44.58	0.98
✓	–	–	16.50	**78.57**	1.38	<u>62.02</u>	<u>48.52</u>	0.57
✓	✓	–	<u>16.42</u>	76.52	<u>0.81</u>	63.62	47.28	<u>0.49</u>
✓	✓	✓	**15.49**	<u>78.00</u>	**0.76**	**14.96**	**78.53**	**0.20**

Streaming after Reconsolidation. Table 6 evaluates whether writing the revised context $\hat{\mathcal{G}}_{\tau}^{\mathrm{r}}$ back into the active cache benefits subsequent views across all 7 VBR, 10 Oxford Spires, and 11 KITTI sequences. Each trajectory is split at 75% of its retained keyframes under one full-trajectory Sim(3) alignment per mode. Streaming after GCR triggers one reconsolidation at the split and then resumes causal inference without further revision. Continuation ATE decreases from 23.58 m to 14.87 m on VBR, from 6.96 m to 4.33 m on Oxford Spires, and from 21.21 m to 14.22 m on KITTI. On VBR, history ATE also falls from 13.57 m to 6.63 m. These continuation frames are inferred causally after the trigger, so the comparison evaluates the effect of conditioning future predictions on the rewritten context. Full GCR provides a sequence-end reference that also uses observations from the final quarter.

Table 6. Streaming after Context Reconsolidation across long-trajectory benchmarks. History and continuation ATE (m, ↓), averaged equally over 7 VBR, 10 Oxford Spires, and 11 KITTI sequences. The split is at 75% of retained keyframes, with one full-trajectory Sim(3) alignment per sequence and mode. Streaming after GCR revises once at the split and resumes online; Full GCR revises at sequence end. Bold marks the lowest ATE in each column.

	VBR		Oxford Spires		KITTI	
Mode	History	Continuation	History	Continuation	History	Continuation
Online	13.57	23.58	5.52	6.96	15.77	21.21
Streaming after GCR	6.63	14.87	4.56	4.33	9.89	14.22
Full GCR	**4.98**	**3.66**	**4.21**	**2.85**	**9.46**	**8.40**

Number of revisable layers. Table 7 studies where to partition the 24-layer backbone into committed and revisable stages on KITTI under the same lighter schedule. Leaving only 8 revisable layers restricts cross-view alignment capacity and increases trajectory error to 17.40 m ATE. Conversely, reopening all 24 layers slightly improves per-keyframe depth AbsRel from 0.077 to 0.074, yet degrades trajectory ATE from 14.96 m to 16.06 m while doubling reconsolidation compute and forfeiting committed-cache invariance. This trade-off reflects the specialization of layers across depth: lower causal layers already extract reliable local geometry across smoothly varying video streams and provide a stable coordinate anchor, whereas upper layers govern global cross-view registration where drift accumulates. Splitting the

backbone evenly into 12 committed and 12 revisable layers achieves the lowest trajectory error while reusing half of the backbone cache without recomputation.

Visual encoder initialization. We compare DINOv3-L [66] and LingBot-Vision-L [24] as visual encoder initializations for multi-view offline training under an identical 16k-iteration schedule with a longer side of 512 and patch size 16. Across 7-Scenes, NRGBD, ETH3D, HiRoom, and Sintel, we report single-scale-aligned per-sequence AbsRel and pose AUC over all valid within-sequence camera pairs in Tab. 8. LingBot-Vision-L improves depth accuracy and yields a clear 5.02-point gain at the strict AUC@3 threshold, rising from 50.77 to 55.79. Because small angular errors in local multi-view perception compound over thousand-frame streams, this stronger fine-grained geometric prior provides a better starting point for subsequent streaming and mixed-mode continuous training.

Table 7. Committed and revisable layer splits on KITTI. The 24-layer backbone is evaluated under the lighter training schedule. Reconsolidated trajectory ATE is in meters. Lower is better.

Layers		GCR	
Committed	Revisable	ATE	AbsRel
16	8	17.40	0.077
12	12	**14.96**	0.077
0	24	16.06	**0.074**

Table 8. Visual initialization for multi-view offline training. AbsRel and pose AUC (%) are averaged over five equally weighted datasets: 7-Scenes [65], NRGBD [4], ETH3D [62], HiRoom [41], and Sintel [9].

Initialization	AbsRel↓	AUC@3↑	AUC@30↑
DINOv3	0.0732	50.77	85.72
LingBot-Vision	**0.0726**	**55.79**	**86.55**

6.4 Revisit Consistency Evaluation

Whole-trajectory Sim(3) ATE distributes error across an entire sequence and can therefore mask local discontinuities where a camera returns to a previously visited location. We therefore evaluate local pose consistency directly across the ground-truth revisit pairs defined in Sec. 5.2.

Table 9. Pose accuracy at shared GT-defined revisits. ATE and e_t are in meters, e_R in degrees. VBR uses all seven sequences; KITTI uses the six sequences containing revisits (`00`, `02`, `05`, `06`, `07`, and `09`).

	VBR			KITTI		
Method	ATE↓	e_t↓	e_R↓	ATE↓	e_t↓	e_R↓
HorizonStream	29.21	32.81	28.87	30.36	72.63	17.33
ABot-Recon	29.89	36.45	24.07	22.34	34.50	8.85
Ours (Online)	**17.01**	**14.56**	**11.76**	**16.98**	**24.29**	**1.77**
HorizonStream w/ LC	17.78	16.31	22.46	22.08	15.72	13.68
ABot-Recon w/ LC	9.98	7.02	12.14	13.61	7.42	5.76
Ours (Reconsolidation)	**4.72**	**1.56**	**5.47**	**7.26**	**5.66**	**0.93**

Loop-closure accuracy. Under full-trajectory Sim(3) alignment, Tab. 9 reports whole-trajectory ATE alongside the local translational error e_t and rotational error e_R of the relative pose $T_i^{-1}T_j$ across matched revisit pairs on VBR and KITTI. Even in causal Online mode, consolidating observations in a shared world frame yields far smaller revisit discrepancies than sliding-window baselines, whose window-to-window stitching accumulates large offsets before the camera returns; notably, Ours (Online) already keeps revisit rotation error to $1.77°$ on KITTI, beating both competing pipelines even with loop closure enabled. Optional retrieval and pose-graph stages in HorizonStream w/ LC and ABot-Recon w/ LC reduce drift, yet local misalignment persists at revisit junctions. Without place recognition or pose-graph optimization, Context Reconsolidation links distant views directly in the revisable layers, cutting revisit error over fourfold vs. ABot-Recon w/ LC on VBR to 1.56 m e_t and reaching $0.93°$ e_R on KITTI.

Alignment across revisits. Figure 8 reprojects source views into revisited target views on VBR Spagna-0 (273 to 6,755), Oxford Christ Church-02 (489 to 4,929), and KITTI 00 (44 to 4,492), using predicted depths and relative poses directly. Because 3D reprojection couples per-view depth scale with long-range relative pose accuracy, any drift in translation, rotation, or scale appears immediately as shifted contours. Context Reconsolidation brings archways, rooflines, and

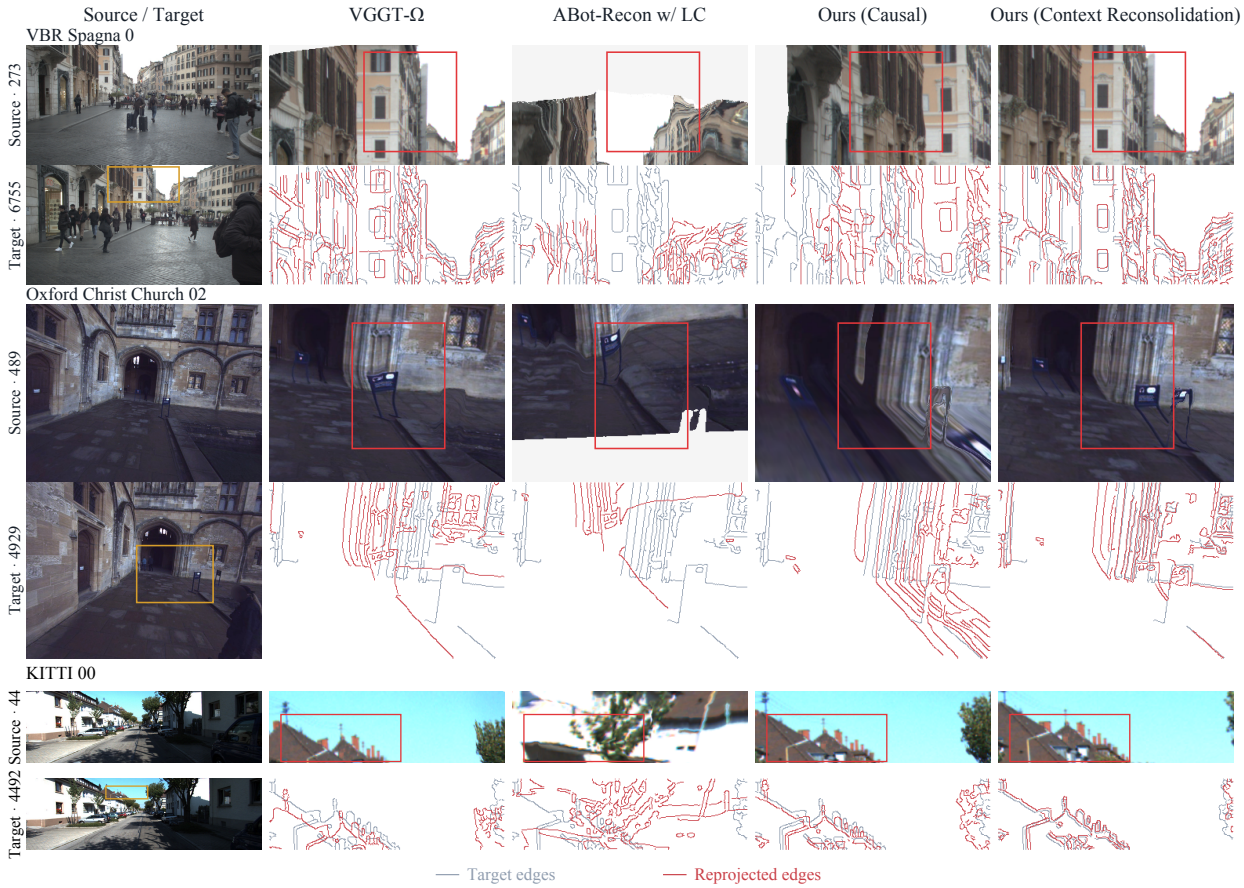


Figure 8. Alignment across revisits on VBR, Oxford, and KITTI. Reprojections of source images into revisited target views. Yellow boxes locate the crops; red boxes highlight corresponding structures. Gray and red edges denote target and reprojected contours.

window frames into clean register compared with causal streaming, whereas VGGT-Ω and ABot-Recon w/ LC retain noticeable double edges and structural offsets.

Emergent revisit attention. Figure 9 compares ground-truth camera proximity with head-averaged layer-16 bidirectional GCR attention weights on VBR Spagna-0, Campus-0, Ciampino-1, and Oxford Keble College-04. All four scenes use the main evaluation checkpoint, input preprocessing, and retained keyframes, and the extracted run reproduces the archived reconsolidated poses. Even without explicit loop-closure or place-retrieval supervision, sharp off-diagonal bands emerge wherever the camera returns to spatially proximate locations thousands of frames later. Mirroring ground-truth proximity across multi-pass loops, these bands show that the revisable layers associate co-visible geometry directly from token representations to resolve accumulated drift.

7 Conclusion and Discussion

We have presented `LingBot-Map 2.0`, a feed-forward foundation model for 3D reconstruction from continuous visual streams. At its core is Geometric Context Reconsolidation (GCR), which resolves accumulated streaming drift by selectively reconsolidating the cached geometric context that produced past estimates rather than post-processing output trajectories externally. Partitioning the backbone into an invariant lower *committed context* and a selectively reopened upper *revisable context* allows `LingBot-Map 2.0` to reuse causal per-view geometry without re-encoding images or rerunning lower layers, while recomputing cross-view alignment bidirectionally in a single feed-forward pass and writing the updated state back into the active cache. Trained with our geometric streaming base and mixed-mode continuous training recipe, alongside sequence scale normalization and dense adjacent relative-pose supervision, a single

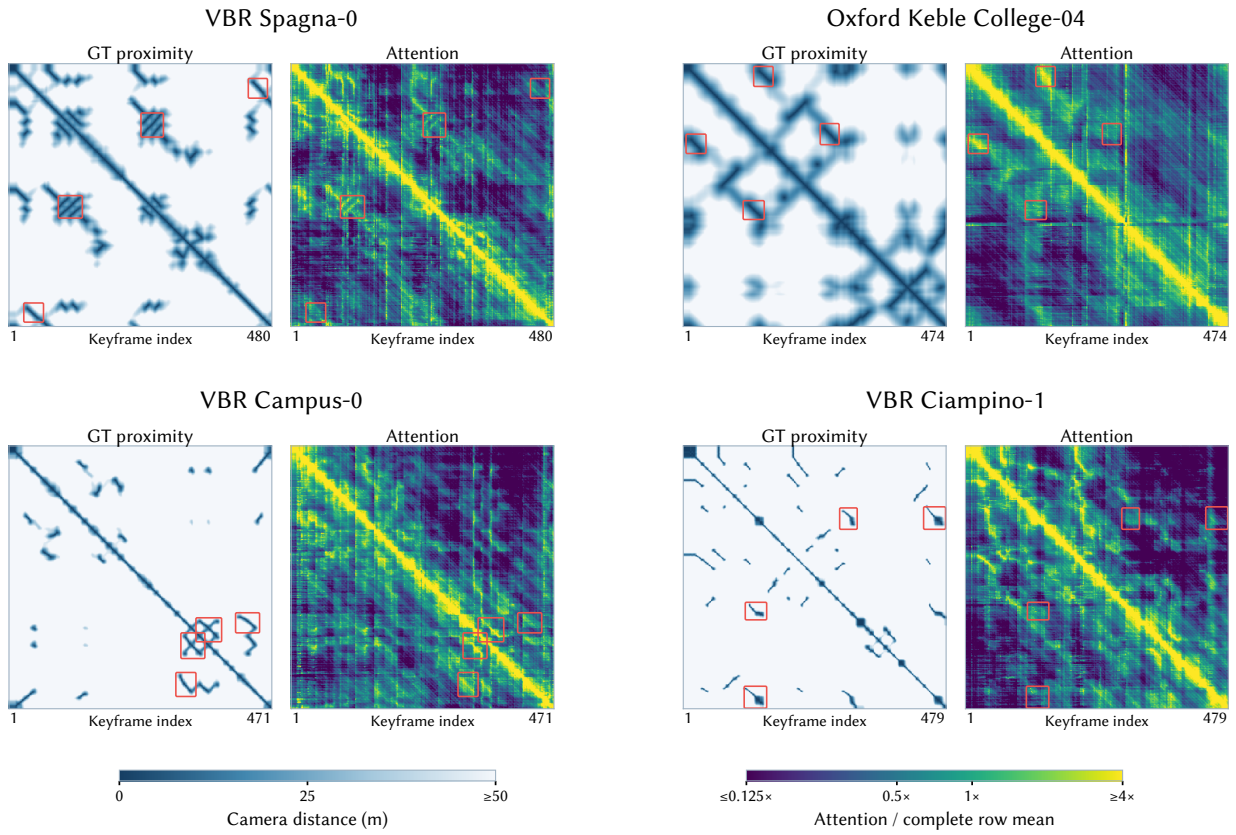


Figure 9. Attention across revisits on VBR and Oxford. Full-sequence GT proximity and layer-16 bidirectional attention from the same checkpoint and keyframe sets as the main evaluation; corresponding GT-defined revisit regions are boxed in red. Attention is divided by the complete row mean and displayed on a shared log scale clipped to $[0.125, 4]$.

set of weights supports Online Streaming, Context Reconsolidation, and Streaming after Reconsolidation, delivering real-time online inference alongside state-of-the-art long-trajectory tracking, scene-level pose estimation, detector-free loop closure, and dense 3D reconstruction.

Limitations and Future Work. While `LingBot-Map 2.0` performs strongly across diverse indoor and outdoor benchmarks, several limitations remain. First, although image-patch pages outside the local window are recycled in constant space, the retained keyframe context tokens and archived committed features H_k still grow linearly with the number of consolidated keyframes, and full bidirectional attention during reconsolidation scales quadratically with $|\mathcal{K}_\tau|$. Combining GCR with bounded or hierarchical memory management, such as merging redundant revisits or restricting reconsolidation to topologically relevant sub-contexts, is a promising path toward lifelong operation on unbounded streams. Second, `LingBot-Map 2.0` currently operates solely on uncalibrated monocular RGB streams, recovering geometry up to an unknown global scale and relying purely on visual cues under rapid motion or low texture; integrating multi-modal inputs such as IMU or range measurements to anchor metric scale and motion priors, and coupling the reconsolidated spatial context with downstream embodied navigation, are natural directions for future work.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307. IEEE, 2016.

[2] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Prisacariu, Daniyar

Turmukhambetov, and Eric Brachmann. Map-Free Visual Relocalization: Metric Pose Relative to a Single Image. In *Computer Vision – ECCV 2022*, pages 690–708. Springer Nature Switzerland, 2022.

[3] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. SceneScript: Reconstructing Scenes with an Autoregressive Structured Language Model. In *Computer Vision – ECCV 2024*, pages 247–263. Springer Nature Switzerland, 2024.

[4] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6280–6291, 2022.

[5] P.J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[6] Hermann Blum, Alessandro Mercurio, Joshua O'Reilly, Tim Engelbracht, Mihai Dusmanu, Marc Pollefeys, and Zuria Bauer. CroCoDL: Cross-device Collaborative Dataset for Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27424–27434, 2025.

[7] Leonardo Brizi, Emanuele Giacomini, Luca Di Giammarino, Simone Ferrari, Omar Salem, Lorenzo De Rebotti, and Giorgio Grisetti. VBR: A Vision Benchmark in Rome. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15868–15874. IEEE, 2024.

[8] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[9] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *European Conference on Computer Vision (ECCV)*, pages 611–625, 2012.

[10] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. arXiv:2001.10773, 2020.

[11] Carlos Campos, Richard Elvira, Juan J. Gomez Rodriguez, Jose M. M. Montiel, and Juan D. Tardos. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[12] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017.

[13] 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. arXiv:2604.14141, 2026.

[14] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D Reconstruction as Test-Time Training. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 50694–50718, 2026.

[15] Yihang Chen, Bin Tan, Yujun Shen, Yinghao Xu, and Nan Xue. PosePlaner: Denoising Any Feedforward Pose Predictor from Pairwise Planar Geometry. In *Advances in Neural Information Processing Systems*, 2026.

[16] Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyang Guo, and Hao Wang. LongStream: Long-Sequence Streaming Autoregressive Visual Geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 272–283, 2026.

[17] Chong Cheng, Peilin Tao, Nanjie Yao, Guanzhi Ding, Xianda Chen, Yuansen Du, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, Zhengqing Chen, and Hao Wang. HorizonStream: Long-Horizon Attention for Streaming 3D Reconstruction. arXiv:2605.23889, 2026.

[18] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023.

[20] 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 *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5446–5453. IEEE, 2026.

[21] Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline Feed-Forward 3D Reconstruction at Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 36464–36474, 2026.

[22] Zhijian Fang, Weicheng Zheng, Yijun Yuan, Weibang Wang, Zhuoguang Chen, Chang Sun, Junhao Huang, Kenan Li, Minghui Qin, and Hang Zhao. SLAMFormer-∞: Infinite SLAM Transformer for Unbounded Frontend and Backend Processing. arXiv:2608.03429, 2026.

[23] Michael Fonder and Marc Van Droogenbroeck. Mid-Air: A Multi-Modal Dataset for Extremely Low Altitude Drone Flights. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 553–562. IEEE, 2019.

[24] Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision Pretraining for Dense Spatial Perception. arXiv:2607.05247, 2026.

[25] Jian Gao, Lin-Zhuo Chen, Yihang Chen, Jiagang Chen, Jiacheng Dong, Xun Cao, Nan Xue, Yinghao Xu, and Yao Yao. Map2SLAM: From Streaming Map Prediction to Real-Time SLAM. arXiv:2611.11169, 2026.

[26] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[27] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3739–3751. IEEE, 2022.

[28] 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. arXiv:2608.27529, 2026.

[29] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning Multi-view Stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2821–2830. IEEE, 2018.

[30] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent Dynamic Depth from Stereo Videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13229–13239. IEEE, 2023.

[31] Karhan Kayan, Stamatis Alexandropoulos, Rishabh Jain, Yiming Zuo, Erich Liang, and Jia Deng. Princeton365: A Diverse Dataset with Accurate Camera Pose. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10. IEEE, 2025.

[32] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):1–13, 2017.

[33] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 42746–42768, 2026.

[34] Jonathan L.C. Lee, Karim Nader, and Daniela Schiller. An Update on Memory Reconsolidation Updating. *Trends in Cognitive Sciences*, 21(7):531–545, 2017.

[35] Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. DROID-SLAM in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 36498–36508, 2026.

[36] Moyang Li, Zihan Zhu, Wei Zhang, Marc Pollefeys, and Daniel Barath. CLoSeR: Closing the Loop for Long-Context Streaming Reconstruction. In *Advances in Neural Information Processing Systems*, 2026.

[37] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3182–3192. IEEE, 2023.

[38] Yuan Li, Yuanbo Xiangli, Hadar Averbuch-Elor, Noah Snavely, and Ruojin Cai. Long-Tail Internet Photo Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 417–426, 2026.

[39] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050. IEEE, 2018.

[40] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023.

[41] 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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 141261–141285, 2026.

[42] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169, 2024.

[43] Lahav Lipson, Zachary Teed, and Jia Deng. Deep Patch Visual SLAM. In *Computer Vision – ECCV 2024*, pages 424–440. Springer Nature Switzerland, 2024.

[44] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is Scalable for LLM Training. arXiv:2502.16982, 2025.

[45] Mingkai Liu, Hao Zhao, and Xingxing Zuo. SURE-Map: Self-Correcting Streaming Geometric Foundation Models. arXiv:2609.15795, 2026.

[46] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[47] Si-Yu Lu, Po-Ting Chen, Hui-Che Hsu, Sin-Ye Jhong, Wen-Huang Cheng, and Yung-Yao Chen. OVGGT: O(1) Constant-Cost Streaming Visual Geometry Transformer. In *Computer Vision – ECCV 2026*, pages 112–129. Springer Nature Switzerland, 2026.

[48] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 129839–129867. Curran Associates, Inc., 2025.

[49] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2697–2706. IEEE, 2017.

[50] James L. McGaugh. Memory–a Century of Consolidation. *Science*, 287(5451):248–251, 2000.

[51] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16695–16705. IEEE, 2025.

[52] Karim Nader, Glenn E. Schafe, and Joseph E. Le Doux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726, 2000.

[53] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria Digital Twin: A New Benchmark Dataset for Egocentric 3D Machine Perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20076–20086. IEEE, 2023.

[54] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A Large-Scale Dataset for Ground Robot Perception and Navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20524–20531. IEEE, 2025.

[55] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[56] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168. IEEE, 2021.

[57] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10881–10891. IEEE, 2021.

[58] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10892–10902. IEEE, 2021.

[59] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9338–9346. IEEE, 2019.

[60] Johannes L. Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[61] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Computer Vision – ECCV 2016*, pages 501–518. Springer International Publishing, 2016.

[62] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269, 2017.

[63] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Fast Visual Geometry Transformer. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 98040–98062, 2026.

[64] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques - SIGGRAPH '85*, pages 245–254. ACM Press, 1985.

[65] 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 *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937. IEEE, 2013.

[66] 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:2508.10104, 2025.

[67] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21371–21382, 2022.

[68] Yifu Tao, Miguel Ángel Muñoz Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The Oxford Spires Dataset: Benchmarking large-scale LiDAR-visual localisation, reconstruction and radiance field methods. *The International Journal of Robotics Research*, 45(6):839–857, 2026.

[69] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16558–16569. Curran Associates, Inc., 2021.

[70] Zachary Teed, Lahav Lipson, and Jia Deng. Deep Patch Visual Odometry. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39033–39051. Curran Associates, Inc., 2023.

[71] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo Mixture Density Networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8938–8948. IEEE, 2021.

[72] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[73] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. AerialMegaDepth: Learning Aerial-Ground Reconstruction and View Synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21674–21684. IEEE, 2025.

[74] Hengyi Wang and Lourdes Agapito. 3D Reconstruction with Spatial Memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[75] 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 (CVPR)*, pages 5294–5306, 2025.

[76] 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 (CVPR)*, pages 21486–21499, 2026.

[77] Kaixuan Wang and Shaojie Shen. Flow-Motion and Depth Network for Monocular Stereo and Beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020.

[78] 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 (CVPR)*, pages 10510–10522, 2025.

[79] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709. IEEE, 2024.

[80] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[81] 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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 10481–10497, 2026.

[82] Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-World Perception for Embodied Agents. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9079. IEEE, 2018.

[83] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22378–22389, 2024.

[84] Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable Test-Time Training for Large-Scale 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21760–21771, 2026.

[85] Sizhe Yang, Linning Xu, Hao Li, Juncheng Mu, Jia Zeng, Dahua Lin, and Jiangmiao Pang. Robo3R: Enhancing Robotic Manipulation with Accurate Feed-Forward 3D Reconstruction. arXiv:2602.10101, 2026.

[86] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Computer Vision – ECCV 2018*, pages 785–801. Springer International Publishing, 2018.

[87] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1787–1796. IEEE, 2020.

[88] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–22. IEEE, 2023.

[89] Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual Geometry Grounded Transformer for Endless Streams. arXiv:2601.02281, 2026.

[90] Yijun Yuan, Zhuoguang Chen, Kenan Li, Weibang Wang, Minghui Qin, Zhijian Fang, Weicheng Zheng, and Hang Zhao. SLAM-Former: Putting SLAM Into One Transformer. In *Computer Vision – ECCV 2026*, pages 56–74. Springer Nature Switzerland, 2026.

[91] Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. LoGeR: Long-Context Geometric Reconstruction with Hybrid Memory. arXiv:2603.03269, 2026.

[92] Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. TexVerse: A Universe of 3D Objects with High-Resolution Textures. arXiv:2508.10868, 2025.

[93] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19855–19865, 2023.

[94] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. OmniWorld: A Multi-Domain and Multi-Modal Dataset for 4D World Modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 66716–66743, 2026.

[95] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming Visual Geometry Transformer. In *International Conference on Learning Representations*, 2026.

A Additional Trajectory Results

Tables S1 to S3 report per-sequence and mean Sim(3)-aligned ATE across all 7 VBR [7], 10 Oxford Spires [68], and 11 KITTI [26] trajectories, expanding the summary comparison in Sec. 6.1.1 with additional baselines across three method families. For offline foundation models, we include VGGT [75], accelerated variants VGG-T^3 [21] and FastVGGT [63], and the updated π^3 release π^3X [81], all evaluated on the same sampled keyframes with released weights, native preprocessing, and linear camera-center plus SLERP [64] interpolation. For geometric optimization systems, we evaluate submap and SLAM pipelines built on 3D foundation models—VGGT-SLAM [48], MASt3R-SLAM [51], VGGT-Long [20], and SURE-Map [45]—alongside classical deep visual odometry and SLAM systems that utilize known ground-truth camera intrinsics (*), namely DROID-SLAM [69], DPV-SLAM [43], DPVO [70], and DROID-W [35]. For causal streaming models, we additionally compare token-pruning and windowed-attention pipelines InfiniteVGGT [89], OVGGT [47], and Stream3R-w [33], recurrent-state models CUT3R [78] and TTT3R [14] together with their periodic 100-frame reset variants, and scale-stabilized LongStream [16].

Examining per-sequence accuracy clarifies how memory design governs long-horizon stability. While offline models and recurrent or token-pruning streaming baselines achieve reasonable accuracy on short sub-kilometer trajectories such as KITTI `03` and `04`, their error grows sharply on multi-kilometer sequences spanning thousands of frames, such as VBR Ciampino-1, Oxford Keble College-04 and Christ Church-02, and KITTI `00`, `02`, and `08`. Periodic 100-frame resets in CUT3R and TTT3R alleviate recurrent state saturation locally, reducing mean ATE on VBR from nearly 100 m to 70.40 m and 65.50 m, yet re-initializing the hidden state severs global coordinate continuity across segments. By maintaining an anchored world coordinate frame with paged local patch windows and compact keyframe context tokens, Ours (Online) sustains coherent tracking over multi-kilometer horizons, reaching 17.01 m mean ATE on VBR, 6.17 m on Oxford Spires, and 17.56 m on KITTI—already outperforming calibrated DROID-W and uncalibrated SURE-Map on VBR as well as all uncalibrated geometric optimization baselines on Oxford Spires.

Finally, comparing Ours (Reconsolidation) against multi-stage geometric optimization and external loop-closure pipelines highlights the effectiveness of revising internal geometric context directly. Although calibrated optimization systems such as DROID-W and submap-alignment pipelines such as SURE-Map and CLoSeR leverage known intrinsics, explicit loop retrieval, or iterative bundle and pose-graph optimization, a single uncalibrated feed-forward reconsolidation pass achieves the lowest mean ATE across all evaluated methods on both VBR at 4.72 m and KITTI at 9.42 m. On Oxford Spires, Ours (Reconsolidation) reaches 3.92 m mean ATE—the best among all uncalibrated methods—and achieves sub-meter trajectory error on eight of the ten campus sequences.

Table S1. Trajectory estimation on VBR. Sequence and mean ATE (m). Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color). LC denotes loop closure, and * indicates known intrinsics.

Method	Per-sequence ATE (m) ↓							
	Colosseo-0 8815f 1.45 km	Campus-0 12042f 2.73 km	Campus-1 11671f 2.95 km	Pincio-0 11142f 1.27 km	Spagna-0 14141f 1.56 km	Diag-0 10021f 1.02 km	Ciampino-1 18846f 5.20 km	Avg.
Offline								
VGGT	39.01	113.09	85.20	76.94	61.31	34.44	198.49	86.93
VGGT-Ω	**5.67**	111.39	**77.30**	**22.87**	**21.41**	26.97	193.90	**65.65**
DA3	24.41	125.95	93.47	73.31	30.28	24.78	199.67	81.69
π^3	14.62	111.64	90.78	47.43	58.03	19.96	174.38	73.83
π^3X	5.69	**100.63**	78.62	26.31	58.43	**19.24**	**174.18**	66.16
VGG-T^3	94.97	138.36	119.83	75.60	64.02	35.15	200.26	104.03
FastVGGT	25.48	117.33	82.23	74.38	57.52	35.03	195.38	83.91
Geometric optimization								
VGGT-SLAM	98.06	138.49	120.12	78.17	64.52	35.56	202.60	105.36
MASt3R-SLAM	99.16	133.77	119.92	46.81	58.27	32.41	202.22	98.94
VGGT-Long	44.13	132.68	115.62	64.30	59.05	33.66	179.42	89.84
DROID-SLAM*	49.50	125.06	71.68	49.15	60.72	35.81	144.59	76.64
DPV-SLAM*	74.14	106.73	57.16	48.04	51.67	35.75	87.32	65.83
DPVO*	75.08	111.22	63.01	55.05	54.22	34.69	85.84	68.44
DROID-W*	**13.54**	30.94	25.34	**13.19**	**14.74**	**23.52**	60.58	**25.98**
SURE-Map	19.69	**16.06**	**14.55**	32.19	35.23	24.41	**57.86**	28.57
Streaming								
SLAM-Former (frontend)	50.31	116.33	102.59	74.77	60.84	22.90	183.25	87.28
InfiniteVGGT	89.94	136.82	116.67	77.72	63.35	33.94	202.12	102.94
OVGGT	95.39	136.92	116.46	77.65	61.63	34.67	200.35	103.30
Stream3R-w	92.51	138.38	116.38	77.43	64.57	35.64	202.14	103.87
CUT3R	78.18	135.81	117.45	75.90	62.49	33.34	193.70	99.55
CUT3R w/ reset	75.42	58.36	61.87	48.91	48.62	25.86	173.80	70.40
TTT3R	99.12	130.01	108.55	75.77	59.51	35.51	187.04	99.36
TTT3R w/ reset	73.48	56.42	52.21	37.03	34.91	31.11	173.36	65.50
LongStream	83.00	120.29	106.92	71.39	60.81	30.58	177.51	92.93
LingBot-Map	19.32	18.49	11.45	41.80	36.60	22.54	67.64	31.12
HorizonStream	40.98	28.17	24.54	26.89	28.21	25.41	**30.29**	29.21
ABot-Recon	47.99	**10.89**	17.33	21.29	29.30	28.02	54.38	29.89
Ours (Online)	**10.90**	12.17	**9.25**	**14.78**	**8.86**	**11.62**	51.49	**17.01**
Global refinement								
SLAM-Former (w/ backend)	31.77	132.20	101.17	67.71	55.48	32.09	191.44	87.41
CLoSeR	14.22	7.43	3.71	9.81	13.04	12.94	16.94	11.16
Scal3R w/ LC	6.81	67.08	85.18	27.20	30.40	20.73	184.17	60.22
HorizonStream w/ LC	24.77	27.55	8.81	16.48	21.35	7.36	18.12	17.78
ABot-Recon w/ LC	16.70	8.06	4.30	5.15	17.91	9.68	**8.07**	9.98
Ours (Reconsolidation)	**2.06**	**7.15**	**3.46**	**3.84**	**2.00**	**6.07**	8.48	**4.72**

Table S2. Trajectory estimation on Oxford Spires. Sequence and mean ATE (m), with frame counts shown in gray. Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color). LC denotes loop closure, and * indicates known intrinsics. BL: Bodleian Library, CC: Christ Church, KC: Keble College, OQ: Observatory Quarter.

	Per-sequence ATE (m) ↓										
Method	KC-02 5969f	KC-03 5682f	KC-04 13518f	KC-05 11544f	OQ-01 5746f	OQ-02 5469f	CC-02 11775f	CC-03 6204f	CC-05 3961f	BL-02 4966f	Avg.
Offline											
VGGT	18.26	32.08	18.49	28.62	11.07	9.34	32.20	8.86	42.27	71.72	27.29
VGGT-Ω	**2.19**	**2.63**	**1.78**	**4.66**	**1.62**	**1.03**	**14.07**	**0.69**	35.37	37.04	**10.11**
DA3	3.12	10.27	6.07	21.28	8.25	8.68	40.55	1.49	36.02	57.57	19.33
π^3	5.81	13.59	4.76	20.30	4.61	3.81	39.14	2.70	**32.32**	68.98	19.60
π^3X	2.98	6.49	2.18	24.61	2.70	2.27	24.90	3.15	41.73	**21.30**	13.23
VGG-T^3	28.87	32.34	42.49	41.46	27.52	26.70	43.42	16.96	40.88	73.54	37.42
FastVGGT	18.42	32.22	14.52	24.47	10.53	7.18	34.97	7.98	41.64	75.42	26.73
Geometric optimization											
VGGT-SLAM	25.86	2.27	15.05	41.07	5.61	3.82	43.12	17.62	42.59	82.34	27.94
MASt3R-SLAM	5.06	31.40	**2.39**	37.37	2.25	1.67	16.84	0.70	37.73	34.11	16.95
VGGT-Long	1.76	14.74	32.51	4.60	3.69	5.66	11.59	3.99	13.51	16.16	10.82
DROID-SLAM*	6.71	31.03	2.73	16.42	**0.17**	**0.21**	14.39	0.47	17.18	**0.61**	8.99
DPV-SLAM*	8.77	4.27	15.88	27.94	0.32	11.75	**1.02**	0.93	39.73	3.17	11.38
DPVO*	7.30	25.84	19.12	26.11	0.80	4.84	1.13	**0.43**	42.61	3.12	13.13
DROID-W*	**0.82**	**1.30**	4.49	**2.67**	2.21	2.87	2.71	0.72	**1.92**	6.73	**2.64**
SURE-Map	2.05	3.93	3.47	9.14	3.54	3.71	13.60	3.88	31.92	8.92	8.42
Streaming											
SLAM-Former (frontend)	28.20	25.20	26.12	28.40	16.39	15.39	34.29	12.21	32.82	69.06	28.81
InfiniteVGGT	26.51	28.87	39.27	36.14	28.67	27.39	39.20	16.99	38.61	77.72	35.94
OVGGT	31.09	30.85	39.76	31.85	28.38	27.79	38.97	17.31	40.43	77.36	36.38
Stream3R-w	31.25	31.44	41.47	36.86	28.58	26.46	40.50	17.18	41.85	79.01	37.46
CUT3R	26.76	27.01	40.04	36.41	27.40	25.37	39.46	16.27	41.97	76.45	35.71
CUT3R w/ reset	20.11	13.88	19.55	32.17	18.19	19.23	28.39	6.09	31.64	29.62	21.89
TTT3R	23.18	21.74	33.79	39.50	29.01	28.77	35.89	12.47	33.74	49.98	30.81
TTT3R w/ reset	14.75	8.65	12.79	26.01	11.91	12.67	16.36	2.38	22.71	22.62	15.08
LongStream	20.02	17.69	30.00	33.65	19.73	18.95	35.60	11.01	34.43	42.69	26.38
LingBot-Map	3.40	3.25	2.99	9.33	2.80	2.49	14.15	2.95	31.81	9.35	8.25
HorizonStream	13.59	8.70	20.29	30.82	9.66	9.27	31.05	6.11	26.12	28.25	18.39
ABot-Recon	5.28	4.26	10.30	8.47	6.01	5.28	**11.60**	1.98	**9.30**	11.34	7.38
Ours (Online)	**1.43**	**1.60**	**1.91**	**1.97**	**1.66**	**1.62**	13.41	**0.51**	28.47	**9.11**	**6.17**
Global refinement											
SLAM-Former (w/ backend)	6.25	30.70	17.00	39.93	11.86	12.00	39.38	3.28	38.07	67.17	26.56
CLoSeR	2.60	2.32	5.58	4.58	4.32	4.04	5.57	1.18	**6.82**	7.46	4.45
Scal3R w/ LC	9.72	4.50	2.09	12.98	0.95	0.51	22.81	9.07	36.75	11.43	11.08
HorizonStream w/ LC	6.44	8.69	19.16	21.47	5.75	1.52	30.01	5.76	26.14	28.25	15.32
ABot-Recon w/ LC	2.46	4.26	5.83	4.68	3.34	2.73	7.87	1.92	9.30	11.34	5.37
Ours (Reconsolidation)	**0.44**	**0.42**	**0.55**	**0.69**	**0.32**	**0.40**	**1.00**	**0.24**	32.29	**2.88**	**3.92**

Table S3. Trajectory estimation on KITTI. Sequence and mean ATE (m). Within each group, bold/underline mark best/second best; across all methods, blue/pale blue mark best/second best (ties share color). LC denotes loop closure, and * indicates known intrinsics.

Method	Per-sequence ATE (m) ↓											
	00 4541f 3.70 km	01 1101f 2.50 km	02 4661f 5.10 km	03 801f 0.60 km	04 271f 0.40 km	05 2761f 2.20 km	06 1101f 1.20 km	07 1101f 0.70 km	08 4071f 3.20 km	09 1591f 1.70 km	10 1201f 0.90 km	Avg.
Offline												
VGGT	187.55	604.34	297.93	22.74	6.90	140.79	32.18	51.82	248.30	205.46	186.73	180.43
VGGT-Ω	**170.72**	**310.95**	**286.99**	10.06	**1.35**	**55.97**	10.00	**10.72**	210.02	**104.21**	**173.12**	**122.19**
DA3	181.61	610.23	300.62	48.49	12.49	154.91	53.26	33.30	256.72	191.21	175.70	183.50
π^3	178.24	511.47	301.96	7.45	2.25	148.55	29.36	47.74	220.68	208.11	177.80	166.69
π^3X	174.13	503.62	295.08	**3.17**	1.70	118.38	**5.12**	22.09	**200.64**	207.49	176.78	155.29
VGG-T^3	191.88	701.05	303.77	110.63	112.99	158.82	134.42	85.41	251.33	201.59	208.35	223.66
FastVGGT	188.42	543.88	302.07	24.92	9.92	133.82	45.93	50.46	252.72	206.02	186.31	176.77
Geometric optimization												
VGGT-SLAM	193.30	507.63	269.06	168.98	3.38	159.92	40.80	23.25	264.27	219.41	209.65	187.24
MASt3R-SLAM	145.43	556.15	279.09	6.00	97.80	144.41	130.59	77.43	263.65	162.09	151.77	183.13
VGGT-Long	10.83	66.50	55.69	5.58	3.95	10.13	6.95	**2.83**	57.55	31.22	22.15	24.85
DROID-SLAM*	92.66	87.09	140.40	2.39	0.97	54.13	64.87	21.11	63.39	73.22	15.88	56.01
DPV-SLAM*	113.28	**11.21**	117.29	2.28	**0.94**	57.32	55.96	18.70	109.26	74.80	13.26	52.21
DPVO*	113.23	12.49	117.38	**2.26**	**0.94**	57.10	56.06	18.99	109.41	74.66	13.28	52.35
DROID-W*	**8.24**	212.74	**14.88**	2.82	4.64	**8.40**	4.69	3.97	17.08	**15.00**	**5.50**	27.09
SURE-Map	20.80	39.53	53.47	2.28	1.34	18.44	**4.00**	3.99	**14.38**	21.14	10.19	**17.23**
Streaming												
SLAM-Former (frontend)	190.12	567.72	283.74	130.02	100.80	154.24	94.15	55.17	220.71	202.64	194.29	199.42
InfiniteVGGT	186.41	623.20	289.31	166.86	68.12	143.86	117.34	85.29	221.41	215.35	156.84	206.73
OVGGT	184.76	717.61	304.78	157.05	65.34	156.04	107.28	86.29	247.45	219.27	178.49	220.40
Stream3R-w	188.66	695.02	301.28	161.68	98.99	159.85	122.82	87.19	264.61	219.90	194.32	226.76
CUT3R	189.88	648.13	276.51	148.03	22.61	151.43	132.07	79.85	235.75	176.04	188.03	204.39
CUT3R w/ reset	191.08	101.71	266.74	20.74	7.66	106.04	69.25	23.51	160.85	71.78	37.46	96.07
TTT3R	177.72	539.47	272.82	100.90	10.95	147.40	131.13	65.99	230.32	179.40	119.81	179.63
TTT3R w/ reset	113.17	100.15	224.31	15.90	4.50	31.01	41.15	11.27	93.45	87.63	33.88	68.76
LongStream	90.02	63.45	236.44	5.03	2.57	80.62	12.77	13.84	89.59	95.06	25.65	65.00
LingBot-Map	23.41	88.76	71.51	**2.80**	0.88	17.39	4.82	5.11	**14.04**	21.64	14.45	24.07
HorizonStream	29.52	24.09	97.40	5.55	**0.72**	13.90	5.49	6.10	22.59	29.77	12.39	22.50
ABot-Recon	22.43	**21.59**	70.06	4.36	2.91	**13.38**	12.05	**1.98**	22.40	14.12	16.27	18.32
Ours (Online)	**20.13**	64.35	**46.49**	4.54	1.09	15.46	**2.87**	3.92	14.57	**12.99**	**6.70**	**17.56**
Global refinement												
SLAM-Former (w/ backend)	189.14	687.55	297.39	133.26	58.92	153.69	100.98	69.68	247.02	180.58	209.66	211.62
CLoSeR	6.55	41.78	28.53	4.86	1.82	6.55	4.76	4.87	26.51	**5.66**	8.79	12.79
Scal3R w/ LC	**4.17**	50.47	37.76	5.21	1.05	**3.84**	3.36	1.37	25.29	12.43	**5.94**	13.72
HorizonStream w/ LC	15.37	24.09	71.07	5.55	0.72	7.83	7.63	3.14	22.59	27.45	12.38	17.98
ABot-Recon w/ LC	11.68	**21.59**	39.83	4.36	2.91	5.89	8.65	1.65	22.40	13.98	16.27	13.56
Ours (Reconsolidation)	5.64	39.77	**22.97**	**1.16**	**0.58**	5.49	**1.19**	**1.30**	**9.21**	6.97	9.36	**9.42**

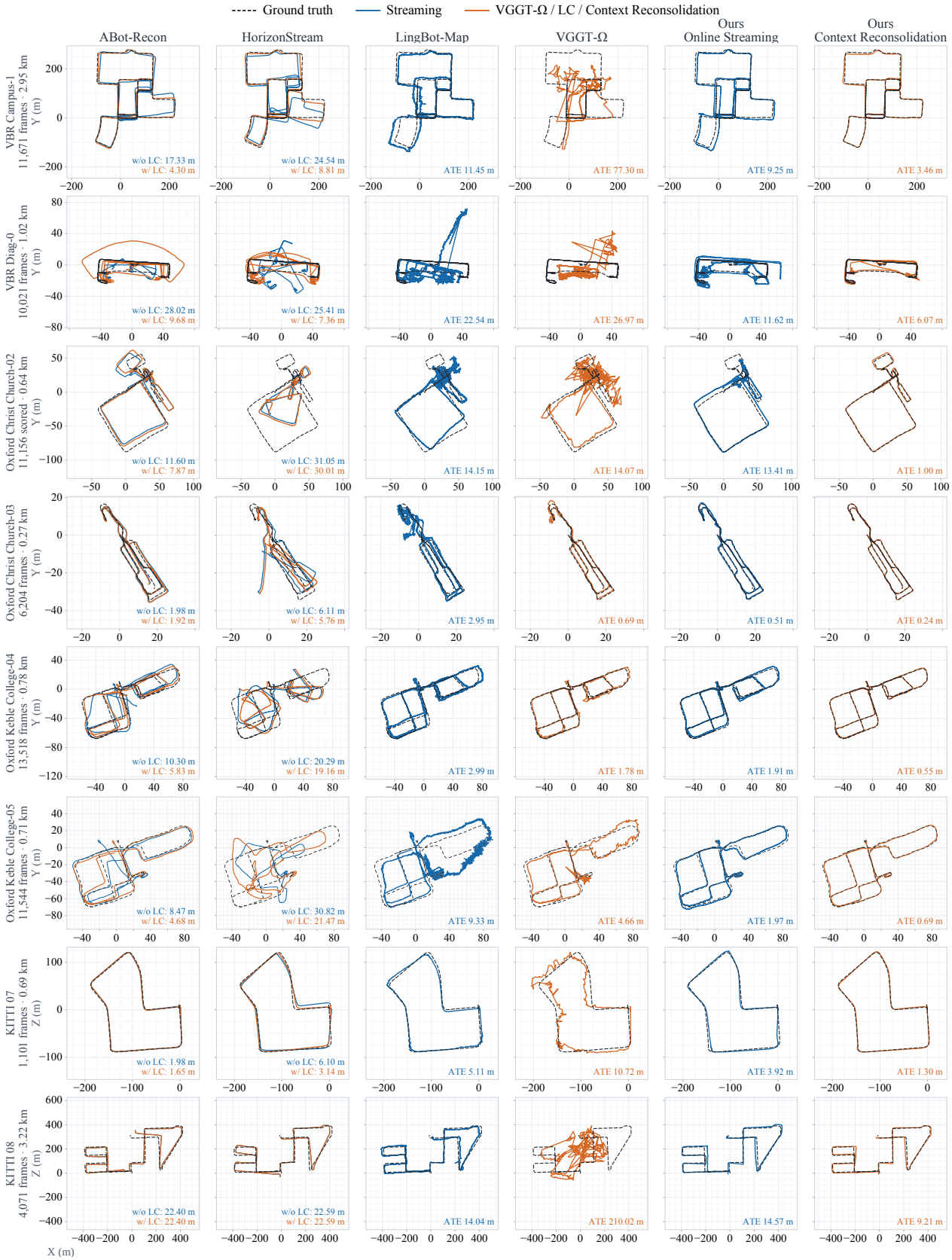


Figure S1. Additional long-trajectory comparisons. Independently Sim(3)-aligned to GT; shared XY (VBR/Oxford) or XZ (KITTI) axes per row. Axes and ATE are in meters; LC denotes loop closure.

B Additional Qualitative Results

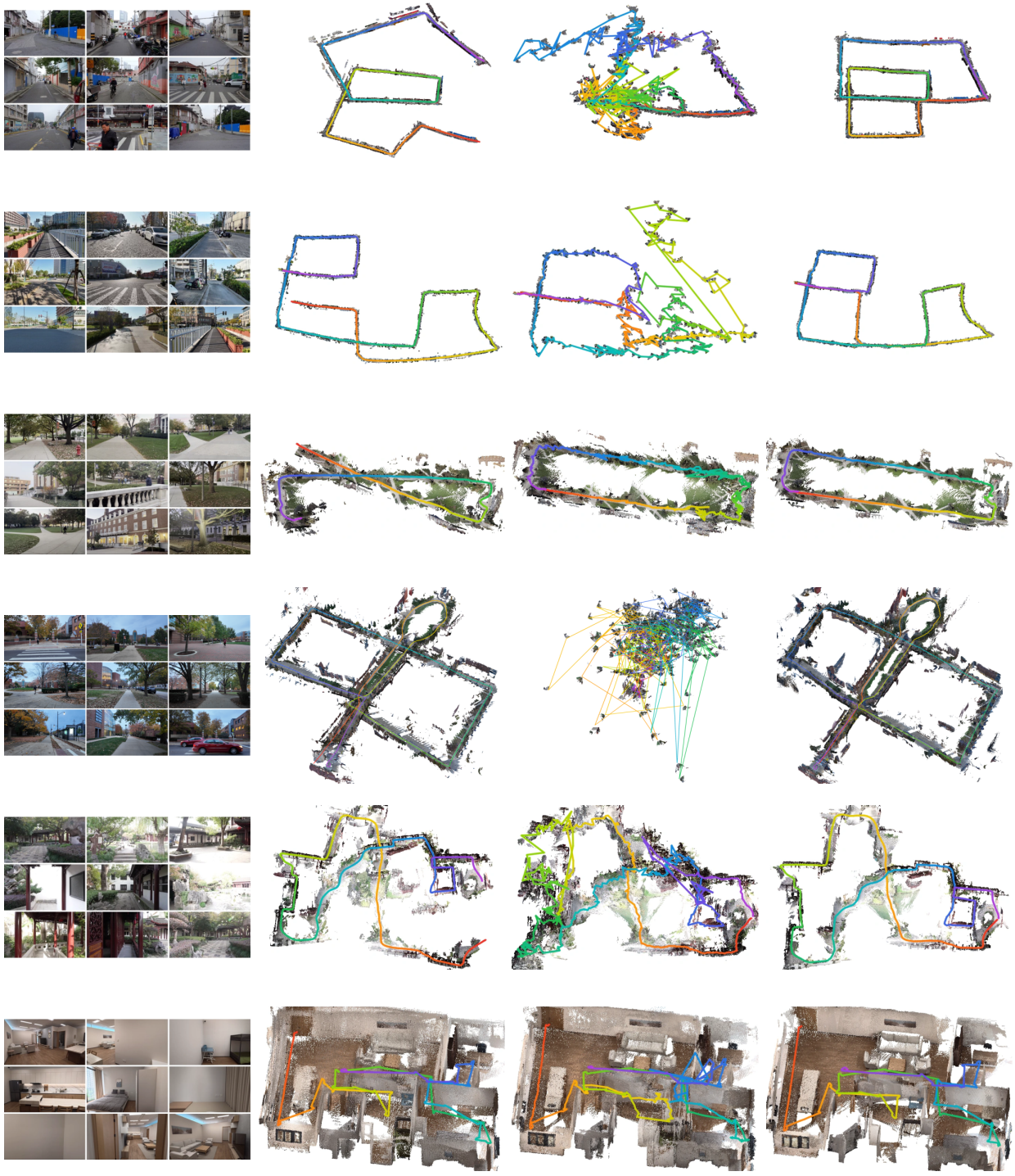


Figure S2. Qualitative 3D reconstruction comparisons (continued). Six additional sequences, with sampled input frames and reconstructions from ABot-Recon, VGGT-Ω, and `LingBot-Map 2.0` in the same column order as Fig. 7.