SMART-SLAM: FEED-FORWARD SLAM WITH SELF-ADAPTIVE SUBMAP MEMORY

Anonymous authors
Paper under double-blind review

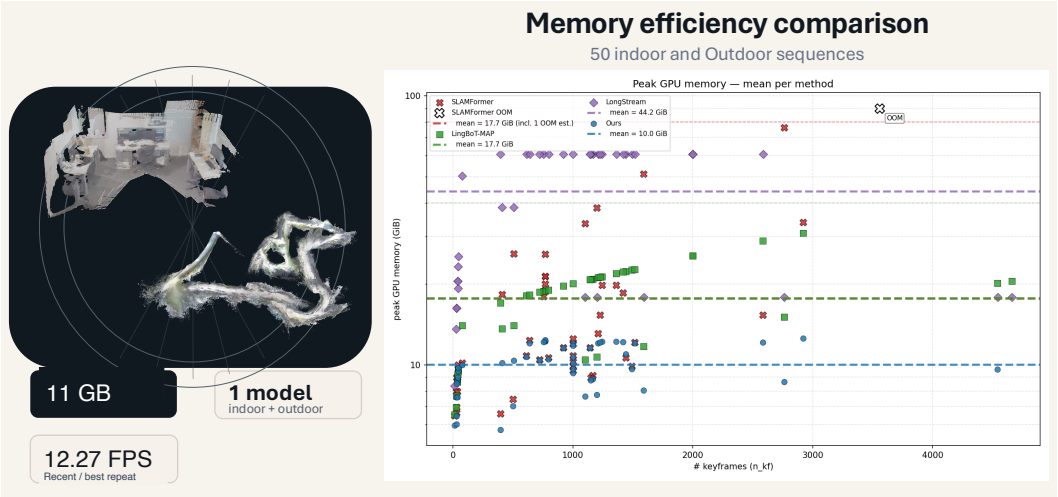


Figure 1: **System teaser.** Left: indoor/outdoor reconstructions; the 11 GB memory and 12.27 FPS speed badges refer to the H200 first-256-input screen (Tab. 9). The speed is its fastest complete-online Recent repeat, not a method mean; terminal BPBR is excluded. Right: an independent GPU-memory diagnostic versus accepted keyframes (Sec. H.3).

ABSTRACT

Long-stream visual geometry needs access to past observations without keeping all scene tokens in active attention. We introduce SMART-SLAM, a feed-forward SLAM system with self-adaptive submap memory. Finalized token banks remain in a host-side archive; each online forward activates only the current submap, its predecessor, and up to k retrieved historical submaps. A detached $Q^\top V$ feature-affinity signal adapts a lightweight retriever while the geometry backbone stays frozen. At end of sequence, Boundary-Pair Backend Re-query (BPBR) composes sparse adjacent relations into an open $\mathrm{SE}(3)$ chain for poses and point maps, without loop closure or iterative geometric optimization. On matched RTX A6000 prefixes, bounded-context online inference is $3.83\times$ faster and uses 60.8% less resident GPU memory than native full-history attention. BPBR reduces the online pose error by 53.2% on TUM and 39.7% on Oxford. Indoor reconstruction reaches F-scores of 74.70% on NRGBD and 83.80% on 7-Scenes. Host-side storage grows with the stream while active attention remains bounded.

1 INTRODUCTION

Visual simultaneous localization and mapping (SLAM) estimates camera motion and 3D structure from image streams. Learning-based systems have improved tracking, dense reconstruction, and map representations, while typically retaining geometric optimization (Teed & Deng, 2021; Zhu et al., 2022; Yugay et al., 2023; Sandström et al., 2025). Visual geometry foundation models now predict cameras and point maps directly from images (Wang et al., 2025a; 2026). Recent systems integrate

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

these models into local geometric modules or a unified SLAM processor (Yuan et al., 2025; Chen et al., 2026; Cheng et al., 2026; Ren et al., 2026; Maggio & Carlone, 2026; Deng et al., 2026).

Long streams expose a basic memory problem. Full-context attention retains historical evidence, but its cost grows rapidly with sequence length. Streaming systems manage this cost through trained implicit state or caches (Cheng et al., 2026; Chen et al., 2026; Zhuo et al., 2025); training-free accelerators prune, merge, or quantize cached representations (Yuan et al., 2026; Su et al., 2026; Fang et al., 2026). Discarded tokens cannot be directly queried on a later revisit. Submap systems preserve local geometry but often recover global consistency through overlapping chunks, registration, loop edges, or factor-graph optimization (Maggio & Carlone, 2026; Deng et al., 2026; Zhu et al., 2025; Yugay et al., 2025). This motivates a different interface: *keep historical evidence addressable, but activate it selectively.*

SMART-SLAM separates three decisions. **Store broadly, activate narrowly:** complete finalized token banks remain in host memory, while each event loads only the current submap, its predecessor, and up to k retrieved historical submaps. **Adapt retrieval, not the backbone:** detached $Q^\top V$ feature affinities supervise a lightweight retriever against fixed historical descriptors. **Separate streaming inference from terminal geometry:** the frozen processor decodes online poses and point maps; Boundary-Pair Backend Re-query (BPBR) then revisits sparse adjacent boundaries and composes an open $\mathrm{SE}(3)$ chain without loop closure, bundle adjustment, or pose-graph optimization.

The resulting system trades growing host storage for bounded active context. Our matched efficiency study measures a $3.83\times$ speedup over native full-history attention, with only 0.47% lower throughput than fixed-budget Recent retrieval. BPBR supplies substantial drift reduction on TUM and Oxford. Indoor evaluations additionally report accurate poses and reconstruction F-scores of 74.70% on NRGBD and 83.80% on 7-Scenes.

Our contributions are:

- **Addressable submap memory.** A complete host-side token archive supports selective historical reuse by a frozen foundation-model SLAM processor, with a bounded online attention queue.
- **Label-free retrieval adaptation.** A detached QV-KL objective updates only the pooler, projection, and scorer, using the geometry processor's own feature affinities.
- **Sparse boundary correction.** BPBR re-queries a few frames per adjacent submap pair, extracts $\mathrm{SE}(3)$ relations, and consistently updates poses and point maps without iterative geometric optimization.

2 Related Work

Foundation-model SLAM. Visual geometry foundation models predict cameras, depth, and point maps from multi-view RGB inputs (Wang et al., 2025a; 2026). MASt3R-SLAM combines reconstruction priors with dense matching and geometric optimization (Murai et al., 2025). VGGT-SLAM 2.0 and VGGT-Long align overlapping chunks using image retrieval, loop edges, and geometric backends (Maggio & Carlone, 2026; Deng et al., 2026). M^3 adds a trained matching head to refine sliding-window outputs (Ren et al., 2026). SLAM-Former integrates frontend tracking and backend refinement in one Transformer (Yuan et al., 2025), but full-token attention becomes costly on long sequences. SMART-SLAM retains its token interface while storing history outside active GPU attention.

Streaming visual geometry. CUT3R updates a persistent recurrent state (Wang et al., 2025b); TTT3R adapts its state-update rate from alignment confidence without additional training (Chen et al., 2025). WinT3R combines sliding windows with a global camera-token pool (Li et al., 2025). StreamVGGT uses causal attention and cached memory (Zhuo et al., 2025), while LongStream uses keyframe-relative poses, orthogonal scale learning, and cache-consistent training (Cheng et al., 2026). LingBot-Map combines geometric context, a pose-reference window, and trajectory memory (Chen et al., 2026). Unlike recurrent-state updates or newly trained streaming models, our online adaptation changes only a submap retriever, not the geometry backbone.

Memory-efficient history and retrieval. InfiniteVGGT prunes a rolling history (Yuan et al., 2026); XStreamVGGT combines cache pruning and quantization (Su et al., 2026); IncVGGT fuses overlapping views and prunes context (Fang et al., 2026). SMART-SLAM instead retains complete

external submap token banks and adapts a lightweight scorer using detached $Q^\top V$ feature affinities. BPBR re-queries adjacent boundaries and composes an open $\mathrm{SE}(3)$ chain. SMART-SLAM** adds geometrically checked non-adjacent factors and $\mathrm{Sim}(3)$ pose-graph optimization as an optional control.

3 SMART-SLAM: Streaming Pipeline

SMART-SLAM stores history outside active attention and retrieves context for a frozen processor. Fig. 2 separates decoder-produced submap outputs from EOS boundary re-query and chaining. Notation and state definitions are collected in Sec. B.

3.1 Pipeline at a Glance

Submaps are indexed by $m = 0, \ldots, M-1$; M is their total number at end of sequence (EOS). Each ordered list $\mathcal{F}_m$ contains $n_m = |\mathcal{F}_m|$ distinct accepted frame IDs, with $1 \le n_m \le S$ and disjoint lists; S is the submap capacity. A submap closes at S keyframes or EOS and writes its raw record. The retriever selects history; current and predecessor submaps are added separately. The frozen processor decodes this queue, initializes the current refined snapshot, and supplies detached $Q^\top V$ feedback for at most one retriever update. Finalized records stay fixed. At EOS, BPBR (and optional PGO) re-expresses outputs, while admitted SMART-SLAM* independently decodes full history. State ownership is specified in Sec. B (Tab. S3).

3.2 Frozen Processor and Bounded Active Set

Frozen SLAM processor. The processor is a single feed-forward Transformer stack instantiated from the released 518-pixel SLAM-Former checkpoint Yuan et al. (2025) and then frozen during evaluation. Its modules are parameterized by the fixed collection $\phi^\star$ and form the processor

$$\mathcal{P}_{\phi^\star} := \left(F_{\phi^\star}, B_{\phi^\star}, H^{\mathrm{cam}}_{\phi^\star}, H^{\mathrm{pts}}_{\phi^\star}\right). \tag{1}$$

Here, $F_{\phi^\star}$ is the frontend image encoder, $B_{\phi^\star}$ is the backend refinement Transformer, and $H^{\mathrm{cam}}_{\phi^\star}$ and $H^{\mathrm{pts}}_{\phi^\star}$ are the camera and point-map heads. Following π^3 Wang et al. (2026), shared register tokens are reused across frames, without designating a canonical reference image. For accepted RGB

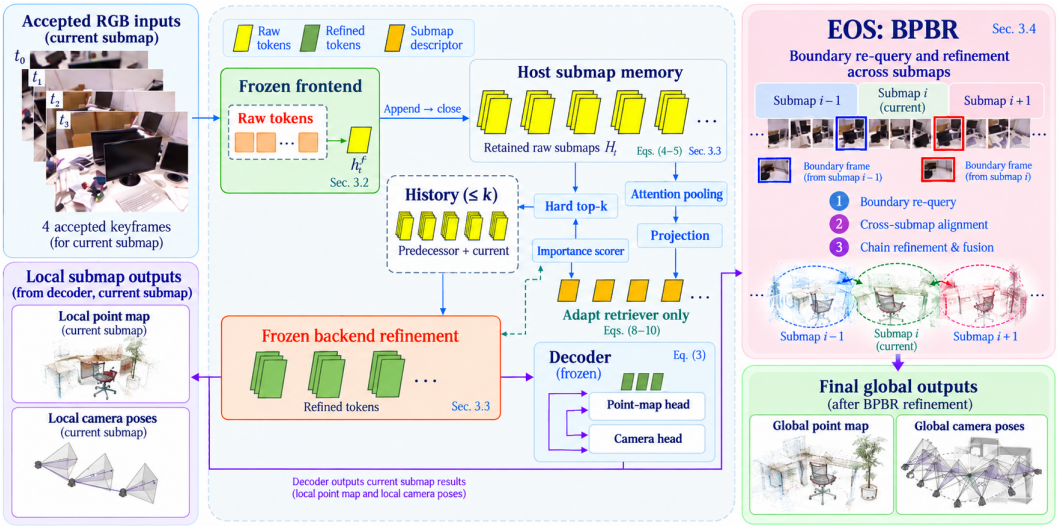


Figure 2: **Overview of SMART-SLAM.** A frozen frontend encodes accepted frames. Host-side submap memory supplies bounded retrieved context to the frozen backend; only the retriever adapts. Decoder heads produce local poses and point maps. At EOS, BPBR re-queries adjacent boundaries and re-expresses the decoded outputs in a shared frame (Sec. 3.4). Thumbnails and geometry are schematic.

frame I_t, the frontend produces $h_t^F = F_{\phi^\star}(I_t) \in \mathbb{R}^{P \times C_z}$, with P token positions and C_z channels. Stacking these in $\mathcal{F}_m$ order forms $H_m \in \mathbb{R}^{n_m \times P \times C_z}$.

Submap-conditioned active queue. For submap m, the ordered active list and its token sequence are

$$\mathcal{A}_m = \big([m-1]_{m>0},\, m,\, \mathrm{Sort}(\mathcal{K}_m)\big), \qquad \Xi_m = \mathtt{Concat}_{j \in \mathcal{A}_m} H_j, \tag{2}$$

Here, $[m-1]_{m>0}$ includes the predecessor when it exists. The eligible history is $\mathcal{U}_m = \{j : 0 \le j \le m-2\}$, empty for $m < 2$; $\mathcal{K}_m \subseteq \mathcal{U}_m$ contains at most k retrieved submaps. Sort orders their IDs increasingly. `Concat` joins banks along the frame dimension in queue order, giving $\Xi_m \in \mathbb{R}^{(\sum_{j \in \mathcal{A}_m} n_j) \times P \times C_z}$, or $L_m = P \sum_{j \in \mathcal{A}_m} n_j$ tokens after flattening the frame and patch dimensions. Only Ξ_m enters online backend attention; other banks stay in the host archive.

Prediction heads. The backend refines the active queue in one forward pass:

$$Y_{\mathcal{A}_m} = B_{\phi^\star}(\Xi_m), \qquad T_i^{\mathrm{head},m} = H_{\phi^\star}^{\mathrm{cam}}(Y_i^{(m)}), \qquad X_i^{\mathrm{head},m} = H_{\phi^\star}^{\mathrm{pts}}(Y_i^{(m)}), \tag{3}$$

$Y_{\mathcal{A}_m}$ is the refined active sequence; for $i \in \mathcal{F}_m$, $Y_i^{(m)}$ is its frame-i slice and $(T_i^{\mathrm{head},m}, X_i^{\mathrm{head},m})$ are camera-to-world (C2W) poses and point maps in the active pass's shared local world frame $w_{\mathcal{A}_m}$. The EOS equations instead use committed Online poses $T_{w \leftarrow i}^0 \in \mathrm{SE}(3)$ (rigid transforms) and world point maps $X_i^0 \in \mathbb{R}^{H_{\mathrm{img}} W_{\mathrm{img}} \times 3}$, where w is the committed world frame and $H_{\mathrm{img}}, W_{\mathrm{img}}$ are point-map height and width. The head-to-world handoff is separate from BPBR (Sec. H.4). For C2W poses, the relative-transform convention is $T_{a \leftarrow b} := (T_{w \leftarrow a})^{-1} T_{w \leftarrow b}$, mapping coordinates from camera b into camera a. The Online pre-EOS output is $\mathcal{O}^0 := \{(T_{w \leftarrow i}^0, X_i^0) : i \in \bigcup_m \mathcal{F}_m\}$.

Only the current-submap slice initializes the stored refined bank Y_m; other slices and query/key/value tensors Q, K, V are pass-local. Active input volume is $\mathcal{O}((k+2)SPC_z)$, with potentially quadratic attention workspace. Token and descriptor storage grow as $\mathcal{O}(MSPC_z)$ and $\mathcal{O}(MD)$, where $D = 128$ is the descriptor dimension. Scoring and dense output costs are detailed in Sec. B.1 and Sec. C.

3.3 Addressable Memory and Adaptive Retrieval

The submap module stores history and adapts context selection; θ_m denotes its retriever parameters before the update at submap event m.

Submap construction and external memory. At closure, $d_m^{(m)} = g_{\theta_m}(H_m) \in \mathbb{R}^D$ summarizes the raw bank, where g_{θ_m} is attention pooling followed by projection. After decoding, the finalized record is $R_m = (m, \mathcal{F}_m, H_m, d_m^{(m)}, Y_m)$ (Sec. A.3). Its stored values remain fixed; the superscript (m) records the descriptor's closure event. Optional NVMe storage holds token banks only; descriptors, IDs, metadata, and map state remain in RAM.

Top-k retrieval and active queue assembly. The scorer s_{θ_m} assigns $q_{m,j} = s_{\theta_m}(d_m^{(m)}, d_j^{(j)})$ to each $j \in \mathcal{U}_m$. Hard top-k selection forms $\mathcal{K}_m$ (default $k = 8$), without a score threshold or a retrieval slot for the predecessor. The scorer and selection are specified in Sec. A.3 (Eqs. 13 and 14).

$Q^\top V$-KL test-time adaptation. For fetched $j \in \mathcal{K}_m$, $i \in \mathcal{F}_m$, and $f \in \mathcal{F}_j$, $Q_i, V_f \in \mathbb{R}^A$ are query/value vectors from a fixed backend attention layer, averaged over patch tokens and flattened over heads; A is their resulting common dimension. On valid fetched candidates $\mathcal{J}_m \subseteq \mathcal{K}_m$ with $|\mathcal{J}_m| \ge 2$, the teacher uses $a_{m,j} = |\mathcal{F}_j|^{-1} \sum_{f \in \mathcal{F}_j} \max_{i \in \mathcal{F}_m} Q_i^\top V_f$ and $\tilde{a}_{m,j} = (a_{m,j} - \mu_a)/\max(\sigma_a, 10^{-6})$; μ_a, σ_a are the mean and sample standard deviation over $\mathcal{J}_m$. At temperature $\tau = 0.5$, let $p_{m,j}^\star = \mathrm{softmax}_{j \in \mathcal{J}_m}(\tilde{a}_{m,j}/\tau)$ and $p_{m,j}^{\theta_m} = \mathrm{softmax}_{j \in \mathcal{J}_m}(q_{m,j}/\tau)$. Both are normalized over $\mathcal{J}_m$; only teacher affinities are standardized. Only the retriever minimizes

$$\mathcal{L}_{\mathrm{QV}}^{(m)} = -\tau^2 \sum_{j \in \mathcal{J}_m} \mathrm{sg}[p_{m,j}^\star] \log p_{m,j}^{\theta_m}, \tag{4}$$

where sg stops teacher gradients. This is τ^2 times the Kullback–Leibler (KL) divergence plus teacher-only entropy (Sec. B.1). The current query $g_{\theta_m}(H_m)$ is differentiable; historical descriptors, the selected support, and the frozen processor receive no gradients. One Adam step updates θ_m to θ_{m+1}; with fewer than two valid candidates, the update is skipped and the state is unchanged. Details are in Sec. A.2.

3.4 Boundary-Pair Backend Re-query

The relative geometry recovered between adjacent submaps can still drift on long sequences Maggio & Carlone (2026); Deng et al. (2026). Default BPBR reuses the same backend on sparse adjacent-boundary windows. Write $\mathcal{F}_m = [f_{m,1}, \ldots, f_{m,n_m}]$, where $f_{m,r}$ is the physical ID of the r-th accepted frame in submap m. With $\ell_m = \min(K^\star, n_m)$, for $0 \le m < M-1$ the boundary window is

$$\mathcal{B}_m = (f_{m,n_m-\ell_m+1}, \ldots, f_{m,n_m}, f_{m+1,1}, \ldots, f_{m+1,\ell_{m+1}}), \tag{5}$$

The default is $K^\star = 2$ frames per side; a submap shorter than $K^\star$ contributes all available frames. The offline EOS stage makes one extra backend call on each $\mathcal{B}_m$ to obtain its boundary relation:

$$\begin{aligned} H_{\mathcal{B}_m} &= \texttt{Concat}_{i\in\mathcal{B}_m} h_i^F, & Y_{\mathcal{B}_m} &= B_{\phi^\star}(H_{\mathcal{B}_m}), \\ T^{\mathcal{B}}_{w_{\mathcal{B}_m}\leftarrow i} &= H^{\mathrm{cam}}_{\phi^\star}(Y_i^{\mathcal{B}_m}), & \Delta_m^{\mathrm{BPBR}} &= (T^{\mathcal{B}}_{w_{\mathcal{B}_m}\leftarrow f_{m,n_m}})^{-1}\, T^{\mathcal{B}}_{w_{\mathcal{B}_m}\leftarrow f_{m+1,1}}. \end{aligned} \tag{6}$$

$Y_{\mathcal{B}_m}$ is the refined boundary sequence and $Y_i^{\mathcal{B}_m}$ its frame-i slice; $H_{\mathcal{B}_m}$ is the ordered raw boundary bank. Each $T^{\mathcal{B}}_{w_{\mathcal{B}_m}\leftarrow i}$ is a C2W pose in the shared boundary-pass world frame $w_{\mathcal{B}_m}$. The relation $\Delta_m^{\mathrm{BPBR}} = T^{\mathcal{B}}_{f_{m,n_m}\leftarrow f_{m+1,1}} \in \mathrm{SE}(3)$ maps the first camera of submap $m{+}1$ into the last camera of submap m. BPBR assumes compatible translation scales between the boundary pass and Online outputs and performs no scale fitting (Sec. H.4).

Submap 0 anchors the world frame. For each subsequent submap, BPBR-chain positions its first frame using the predicted boundary relation while preserving the Online intra-submap motion encoded by $T^0_{w\leftarrow i}$:

$$\begin{aligned} \hat{T}_{w\leftarrow f_{0,r}} &= T^0_{w\leftarrow f_{0,r}}, && r = 1, \ldots, n_0, \\ \hat{T}_{w\leftarrow f_{m+1,1}} &= \hat{T}_{w\leftarrow f_{m,n_m}}\, \Delta_m^{\mathrm{BPBR}}, && m = 0, \ldots, M-2, \\ \hat{T}_{w\leftarrow f_{m+1,r}} &= \hat{T}_{w\leftarrow f_{m+1,1}}\, (T^0_{w\leftarrow f_{m+1,1}})^{-1}\, T^0_{w\leftarrow f_{m+1,r}}, && r = 1, \ldots, n_{m+1}. \end{aligned} \tag{7}$$

For $i \in \mathcal{F}_m$, define the shared submap coordinate change $C_m^{\mathrm{BPBR}} := \hat{T}_{w\leftarrow f_{m,1}}(T^0_{w\leftarrow f_{m,1}})^{-1} = [R_m^{\mathrm{B}}\ t_m^{\mathrm{B}}; 0\ 1]$. Here $R_m^{\mathrm{B}} \in \mathrm{SO}(3)$ and $t_m^{\mathrm{B}} \in \mathbb{R}^3$ are rotation and translation. Then

$$\hat{T}_{w\leftarrow i} = C_m^{\mathrm{BPBR}} T^0_{w\leftarrow i}, \qquad \hat{X}_i = X_i^0 (R_m^{\mathrm{B}})^\top + \mathbf{1}(t_m^{\mathrm{B}})^\top, \quad i \in \mathcal{F}_m, \tag{8}$$

where point maps store 3D coordinates as rows and $\mathbf{1} \in \mathbb{R}^{H_{\mathrm{img}}W_{\mathrm{img}}}$ is the all-ones column. For submap 0, $C_0^{\mathrm{BPBR}} = I_4$, the 4×4 identity. Camera-local points and intra-submap relative poses are preserved (Sec. B.1, Eq. 22). The default output is $\mathcal{O}^{\mathrm{BPBR}} := \{(\hat{T}_{w\leftarrow i}, \hat{X}_i) : i \in \bigcup_m \mathcal{F}_m\}$.

3.5 Endpoint Semantics and Optional Geometry

The online QV-TTA memory loop supports three terminal policies.

Default SMART-SLAM. The default applies the BPBR open chain in Eqs. 5 and 7. It uses adjacent factors only and performs no iterative geometric optimization, producing $\mathcal{O}^{\mathrm{BPBR}}$ as defined above.

SMART-SLAM*. A fixed resource rule admits one terminal full-history frozen-backbone decode when $N_{\mathrm{KF}} = \sum_{m=0}^{M-1} n_m < 52$; otherwise it returns BPBR. This resource rule was fixed before evaluation and uses the same checkpoint.

SMART-SLAM (optional Verified-PGO).** This control adds native $\mathrm{Sim}(3)$ (similarity-transform) pose-graph optimization to BPBR, using retrieved non-adjacent pairs screened by GT-free geometric checks. The pose-control backend is described in Sec. A.7; ETH3D's ** (verified-PGO) row uses a distinct reconstruction backend (Sec. G).

4 Experiments

We evaluate pose accuracy, reconstruction, and streaming efficiency. Comparisons separate shared-input controls, method-native outputs, and published results.

4.1 EXPERIMENTAL SETUP

We use TUM RGB-D (Sturm et al., 2012), NRGBD (Azinović et al., 2022), 7-Scenes (Shotton et al., 2013), KITTI (Geiger et al., 2012), and Oxford Spires (Tao et al., 2025). The default $(S, k, K^\star) = (4, 8, 2)$ uses the released 518-pixel SLAM-Former checkpoint (Yuan et al., 2025), input stride 1, and at most one retriever Adam step per finalized submap (learning rate 10^{-4}; gradient clip 1.0). The backbone stays frozen. Matched method-family and runtime controls use an RTX A6000; the indoor pose/reconstruction and Nearest-7 ablations use an RTX 4090. Frame support is in Sec. A.6 and Tab. S1; RTX 4090 evaluation conventions are in Sec. H.1.

4.2 RETRIEVAL AND TERMINAL CORRECTION

Adaptive retrieval and boundary correction. QV-TTA improves over Recent on four datasets, including 25.5% on NRGBD and 24.2% on Oxford (Tab. 1). BPBR further reduces online error by 53.2% on TUM, 47.4% on KITTI, and 39.7% on Oxford, but worsens NRGBD. Native full history is strongest on TUM/7-Scenes; * is lower on NRGBD, and BPBR is lower on Oxford. All KITTI/Oxford streams use *'s BPBR fallback. PGO improves the BPBR parent on four datasets but can worsen the Online parent. Adjacent-chain ICP controls are reported in Tab. S7.

QV-TTA lowers Oxford/KITTI ATE relative to Nearest-7 by $4.8\%/5.5\%$; BPBR supplies a further $52.2\%/44.6\%$ reduction (Tab. 2). Full-history decoding exceeds memory on this RTX 4090 allocation.

4.3 POSE COMPARISONS

SMART-SLAM gives ATE below 0.053 m on all three indoor datasets and lower NRGBD error than the local LongStream reproduction (Tab. 3).

Tab. 4 reports study-specific streaming results;

Controlled streaming comparison. On the shared KITTI inputs, SMART-SLAM has 45.1% lower error than the lowest of the four streaming controls under the selected-input protocol in Tab. 5.

Comparisons with geometric backends. On native outputs, SMART-SLAM has lower indoor errors than VGGT-Long and both MASt3R-SLAM configurations, but VGGT-SLAM2 is lower on TUM/NRGBD (Tab. 6). On common poses, * improves over VGGT-SLAM2 without LC on TUM, whereas LC gives the lowest error. Oxford nearly ties VGGT-Long; KITTI has higher error (Tab. 7).

4.4 COMPLETE-ONLINE EFFICIENCY

QV-TTA is $3.83\times$ faster and uses 60.8% less resident GPU memory than full history (Tab. 8); relative to Recent, throughput is 0.47% lower and memory 1.89% higher. Host storage grows by 12.8685 MiB/submap in the scaling diagnostic (Sec. C, Fig. S1).

Table 1: **Method-family ablation.** ATE-RMSE [m] on common accepted frames. Native uses full history; its KITTI aggregate is unavailable. Default includes BPBR; * uses full-history EOS with BPBR fallback; ** adds $\mathrm{Sim}(3)$ PGO. Bold marks the best SMART-SLAM variant.

Endpoint	TUM	NRGBD	7-Scenes	KITTI	Oxford
SLAM-Former Native	.03987	.0398	.0545	–	20.3392
Recent (S=4, k=8)	.1208	.0612	.0624	167.4706	33.4626
Online QV-TTA	.1166	.0456	.0642	149.0634	25.3563
QV-TTA → PGO	.1205	.0432	.0589	149.0750	25.3563
SMART-SLAM (BPBR)	.0546	.0724	.0630	78.4595	**15.2905**
SMART-SLAM*	**.0492**	**.0321**	.0581	78.4595	**15.2905**
SMART-SLAM**	.0515	.0509	**.0568**	**78.2696**	**15.2905**

Table 2: **Outdoor retrieval ablation on RTX 4090.** ATE [m]; fixed checkpoint and keyframe policy. Nearest-7 uses a different recency budget from Recent in Tab. 1.

Dataset	Sequences	Nearest-7	QV-TTA	+ BPBR	Full history
Oxford Spires	4	35.19	33.49	**16.01**	OOM
KITTI	6	167.81	158.65	**87.87**	OOM

Table 3: **Indoor pose evaluation on RTX 4090.** Errors [m]; AUC [%]. LongStream is a local reproduction. Evaluation details are in Sec. H.1.

Dataset	Method	ATE↓	ATE-RMSE↓	AUC@15↑	AUC@30↑
TUM fr1	SMART-SLAM	.0479	.0552	71.86	84.71
NRGBD	SMART-SLAM	.0525	.0697	88.30	94.00
	LongStream (repro.)	.2993	.3543	45.09	68.70
7-Scenes	SMART-SLAM	.0453	.0534	72.81	85.63

Table 4: **Published streaming references.** Pose errors [m]; AUC [%]. Each row retains its study-specific metric and protocol. External values follow Chen et al. (2026); Cheng et al. (2026) (Sec. H.2).

Dataset	Method	Metric	Error↓	AUC@30↑	AUC@3↑
7-Scenes	SMART-SLAM	ATE	.0453	85.63	24.32
	LingBot-Map (Chen et al., 2026)	ATE	.0800	79.06	13.20
	InfiniteVGGT (Yuan et al., 2026)	ATE	.1200	74.58	8.45
	STream3R (Lan et al., 2026)	ATE	.1100	74.81	9.84
TUM	SMART-SLAM (A6000)	ATE-RMSE	.0546	–	–
	LongStream (Cheng et al., 2026)	ATE	.0760	–	–
Oxford	SMART-SLAM (A6000)	ATE-RMSE	15.2905	–	–
	LongStream (Cheng et al., 2026)	ATE	19.8150	–	–

Table 5: **KITTI comparison on identical selected inputs.** All methods use the same ordered 4,236 SMART-SLAM-selected images across six sequences and the same evaluated frame IDs. CUT3R/TTT3R use no refresh. ATE-RMSE [m]; settings are specified in Sec. A.6.

Method	ATE-RMSE↓	Method	ATE-RMSE↓
CUT3R (Wang et al., 2025b)	142.822	WinT3R (Li et al., 2025)	143.493
TTT3R (Chen et al., 2025)	142.810	SMART-SLAM	**78.459**
InfiniteVGGT (Yuan et al., 2026)	146.751		

Table 6: **External SLAM methods on native pose outputs.** ATE-RMSE [m]. VGGT variants enable loop closure; MASt3R-SLAM uses default subsample 2 or subsample 1. Output frame sets differ across methods; common-pose comparisons appear in Tab. 7.

Method	TUM	NRGBD	7-Scenes	KITTI	Oxford
VGGT-SLAM2	.0369	.0448	.0742	36.2281	13.5430
VGGT-Long	.0823	.1319	.1131	13.9930	15.4782
MASt3R-SLAM default	.0608	.1061	.0792	36.2820	6.7135
MASt3R-SLAM subsample-1	.0594	.1144	.0812	22.4134	16.2636
SMART-SLAM	.0546	.0724	.0630	78.4595	15.2905
SMART-SLAM*	.0492	.0321	.0581	78.4595	15.2905

4.5 Reconstruction and qualitative results

SMART-SLAM gives F-scores of 74.70% on NRGBD and 83.80% on 7-Scenes (Tab. 10). Accuracy error is lower and F-score higher than the published LingBot-Map values, while completeness is worse. Reconstruction metrics use evaluator-side ICP alignment.

Table 7: **External comparisons on common predicted poses.** ATE-RMSE [m]; exact frame intersection without interpolation. LC denotes loop closure. Checkpoints and execution protocols remain method-specific.

Dataset	External method	LC off	LC on	SMART-SLAM	SMART-SLAM*
TUM	VGGT-SLAM2	.04790	.03266	.04811	.04421
Oxford	VGGT-Long	15.3216	15.3216	15.2905	15.2905
KITTI	VGGT-Long	45.4841	13.6779	78.4595	78.4595

Table 8: **Matched complete-online efficiency.** First 1,024 raw images of KITTI 00/06/09/10; same checkpoint, input order, poses, and RTX A6000. Memory is mean 1-Hz NVML resident memory. Timing excludes terminal BPBR.

Online mode	FPS↑	GPU memory [GiB]↓
Native full history	1.2371	16.2943
Recent (S=4, k=8)	**4.7588**	**6.2740**
QV-TTA	4.7364	6.3928

Table 9: **H200 short-prefix efficiency screen.** First 256 raw images of KITTI 00/06/09/10, with two complete-online repeats per mode and sequence (24 runs). FPS and mean 1-Hz NVML resident GPU memory exclude terminal BPBR. This screen is separate from the RTX A6000 1,024-input controls in Tab. 8.

Online mode	FPS↑	GPU memory [GiB]↓
Native full history	8.40	**7.03**
Recent (S=4, k=8)	**10.88**	10.01
QV-TTA	10.68	10.14

Table 10: **NRGBD and 7-Scenes reconstruction.** Acc/Comp/Overall [m]; F-score [%] at 0.05 m. LingBot-Map results are from Chen et al. (2026); protocols are not matched across papers (Sec. H.1).

Dataset	Method	Acc↓	Comp↓	Overall↓	F-score↑
NRGBD	SMART-SLAM	.0331	.1594	.0962	74.70
	LingBot-Map (paper)	.0700	.0300	–	65.10
7-Scenes	SMART-SLAM	.0124	.0566	.0345	83.80
	LingBot-Map (paper)	.0300	.0400	–	82.38

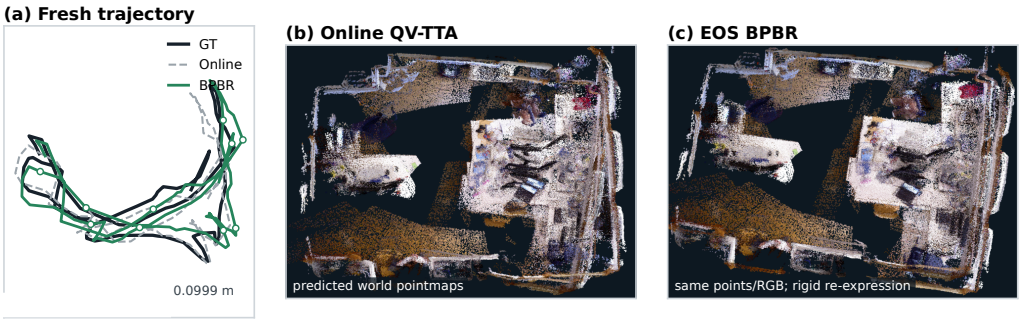


Figure 3: **Trajectory and point-map outputs on TUM `fr1_room`.** A separate $S = 8, k = 4$ run with 82 accepted keyframes gives Online/BPBR ATE-RMSE of $0.1412/0.0999$ m. Both point-map panels show the same 600,076 paired point/RGB samples.

Fig. 3 illustrates the joint pose/point-map update in Eq. 8; Fig. 4 includes both gains and regressions. Context and boundary-width sweeps are non-monotonic (Sec. D.1): at $(S, k) = (8, 4)$, every tested BPBR width worsens NRGBD/7-Scenes.

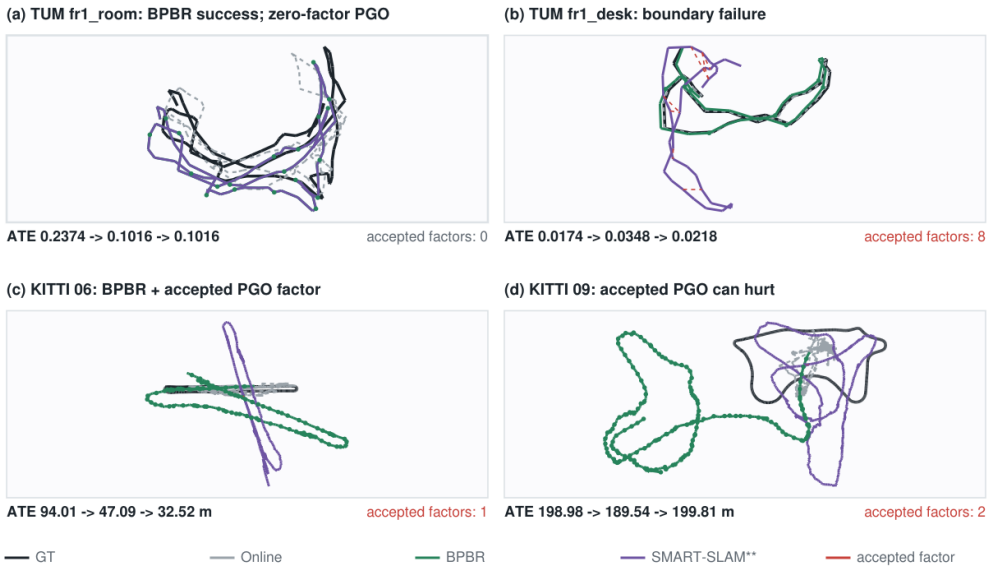


Figure 4: **Trajectory improvements and counterexamples.** Four cases are projected onto the GT PCA plane. BPBR and geometrically checked PGO can help or harm a trajectory. Green dots mark submap boundaries; red dashed chords mark accepted non-adjacent factors.

5 Conclusion

We presented SMART-SLAM, a feed-forward SLAM system separating persistent scene memory from online inference context. Complete finalized submap tokens remain addressable in host memory, while a bounded active set supplies historical evidence to a frozen geometry processor. A detached $Q^\top V$ feature-affinity signal adapts only the lightweight retriever, enabling label-free retrieval adaptation without updating the geometry backbone. At sequence completion, BPBR sparsely re-queries adjacent submap boundaries and jointly re-expresses camera poses and point maps through an open $\mathrm{SE}(3)$ chain, without loop closure or iterative geometric optimization.

On matched RTX A6000 prefixes, bounded-context inference provides a $3.83\times$ online speedup and uses 60.8% less resident GPU memory than native full-history attention. BPBR further reduces online pose error by 53.2% on TUM and 39.7% on Oxford, while indoor evaluations demonstrate accurate trajectory estimation and dense reconstruction. These results show that addressable history and sparse geometric re-query can complement bounded-context inference without full-history attention.

Reproducibility Statement

The supplement describes memory and adaptation, geometric endpoints, evaluation protocols, and per-sequence frame support. Comparisons distinguish shared inputs, method-native outputs, common predicted frames, and published results. Sec. H.5 summarizes the accompanying materials and implementation availability. We plan to release the code after acceptance.

AI Use Statement

AI assists with prose editing, notation checks, literature review, and figure preparation. The authors are responsible for the final content and its verification.

REFERENCES

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*, pp. 6290–6301, 2022.

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 preprint arXiv:2604.14141*, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyang Guo, and Hao Wang. Longstream: Long-sequence streaming autoregressive visual geometry. *arXiv preprint arXiv:2602.13172*, 2026.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Keyu Fang, Changchun Zhou, Yuzhe Fu, Hai Helen Li, and Yiran Chen. Incvggt: Incremental vggt for memory-bounded long-range 3d reconstruction. In *ICLR*, 2026.

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

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 *ICLR*, 2026.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool, 2025. URL `https://arxiv.org/abs/2509.05296`.

Dominic Maggio and Luca Carlone. Vggt-slam 2.0: Real-time dense feed-forward scene reconstruction. *arXiv preprint arXiv:2601.19887*, 2026.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kerui Ren, Guanghao Li, Changjian Jiang, Yingxiang Xu, Tao Lu, Linning Xu, Junting Dong, Jiangmiao Pang, Mulin Yu, and Bo Dai. M3: Dense matching meets multi-view foundation models for monocular gaussian splatting slam. *arXiv preprint arXiv:2603.16844*, 2026.

Erik Sandström, Ganlin Zhang, Keisuke Tateno, Michael Oechsle, Michael Niemeyer, Youmin Zhang, Manthan Patel, Luc Van Gool, Martin Oswald, and Federico Tombari. Splat-slam: Globally optimized rgb-only slam with 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1686–1697, 2025.

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 *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2013. URL `https://www.microsoft.com/en-us/research/publication/scene-coordinate-regression-forests-for-camera-relocalization-in-rgb-d-images/`.

Juergen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012.

Zunhai Su, Weihao Ye, Hansen Feng, Keyu Fan, Jing Zhang, Dahai Yu, Zhengwu Liu, and Ngai Wong. Xstreamvggt: Extremely memory-efficient streaming vision geometry grounded transformer with kv cache compression. *arXiv preprint arXiv:2602.21780*, 2026.

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. *International Journal of Robotics Research*, 2025.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

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)*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b.

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

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

Yijun Yuan, Zhuoguang Chen, Kenan Li, Weibang Wang, Minghui Qin, Zhijian Fang, Weicheng Zheng, and Hang Zhao. Slam-former: Putting slam into one transformer. *arXiv preprint arXiv:2509.16909*, 2025.

Vladimir Yugay, Yue Li, Theo Gevers, and Martin R Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. *arXiv preprint arXiv:2312.10070*, 2023.

Vladimir Yugay, Theo Gevers, and Martin R Oswald. Magic-slam: Multi-agent gaussian globally consistent slam. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6741–6750, 2025.

Liyuan Zhu, Yue Li, Erik Sandström, Shengyu Huang, Konrad Schindler, and Iro Armeni. Loopsplat: Loop closure by registering 3d gaussian splats. In *2025 International Conference on 3D Vision (3DV)*, pp. 156–167. IEEE, 2025.

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12786–12796, 2022.

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