Scal3R: Learning Efficient Multi-Relative Pose Query for Scalable Online 3D Reconstruction

Chin-Yang Lin[1,2] Yang-Che Sun[1] Cheng Sun[2] Fu-En Yang[2]
Min-Hung Chen[2] Yen-Yu Lin[1] Wei-Chen Chiu[1] Yu-Lun Liu[1]

[1]Department of Computer Science, National Yang Ming Chiao Tung University
[2]NVIDIA

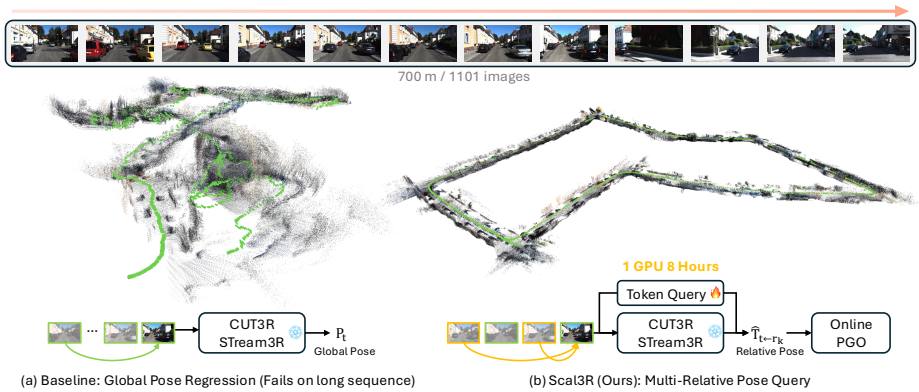


Fig. 1: Scal3R enables scalable online 3D reconstruction on long video streams. (a) Existing feed-forward models, such as CUT3R [76] and STream3R [38], regress absolute global poses (P_t) relative to a fixed first frame. This forces extrapolation far beyond its training distribution, resulting in catastrophic drift and geometric collapse on kilometer-scale sequences. **(b)** Scal3R reformulates the problem into a local multi-reference relative pose query ($\hat{\mathbf{T}}_{t\leftarrow r_k}$). By injecting lightweight learnable tokens into a frozen backbone and aggregating relative constraints via online Pose-Graph Optimization (PGO), Scal3R suppresses long-range drift and recovers globally consistent geometry. It requires only 8 hours of fine-tuning on a single GPU.

Abstract. Online 3D reconstruction models perform poorly on long videos. This happens because regressing poses relative to a fixed first-frame anchor forces extrapolation far beyond the training distribution. Small drifts accumulate and amplify into significant geometric collapse. However, we observe that per-frame depth remains stable throughout this failure. The backbone's local geometry remains intact; only the global pose head breaks down. Motivated by this decoupling, we introduce **Scal3R**. This approach reformulates online reconstruction as multi-reference relative pose querying. We use lightweight learnable tokens, which make up about ∼1% of the parameters, and inject them into a completely frozen

backbone via asymmetric attention. This setup queries poses relative to multiple past keyframes. An online pose-graph optimization system with loop closure suppresses long-range drift. Scal3R reaches convergence in 8 hours on a single GPU. It reduces the average ATE by over 60% on KITTI compared to the online baseline. It also achieves state-of-the-art performance across Virtual KITTI, Sintel, TUM-Dynamic, ScanNet, and 7-Scenes. Project page: `https://linjohnss.github.io/scal3r/`

Keywords: Online 3D reconstruction · Relative pose estimation · Prompt tuning · Pose graph optimization

1 Introduction

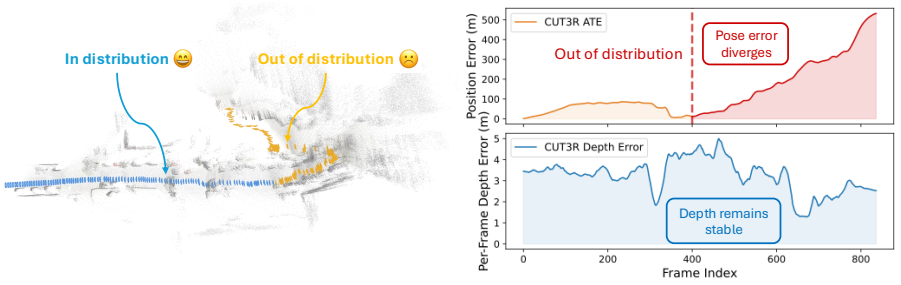
Feed-forward models such as CUT3R [76] and STream3R [38] enable real-time 3D reconstruction by decoding per-frame geometry into a unified global coordinate system via direct pose regression relative to the first frame. While this **global-anchor** paradigm works well for short sequences, it faces severe stability and scalability bottlenecks in long-range environments.

This failure stems from two fundamental issues. First, existing 3D datasets [79, 88] cover limited scene scales. This means that models trained on short sequences must extrapolate to global coordinates far outside their training distribution. When deployed on real-world trajectories spanning hundreds of meters, even minor feature drift is amplified into geometric collapse (Fig. 2a). Second, real-world camera motions are highly variable, and for unbounded video streams, the global coordinate system will inevitably encounter out-of-distribution trajectories, making feed-forward global pose regression theoretically unable to scale.

As shown in Fig. 2b, this failure is highly localized: while global pose errors diverge catastrophically, per-frame depth remains consistently stable. This indicates that the backbone's local geometric representations are intact and the failure is isolated to the global pose regression head. Motivated by this, we shift from *global pose regression* to *local relative querying*: rather than forcing the model to extrapolate a global mapping, we exploit its well-learned local geometry to estimate stable relative transformations via a visual query mechanism conditioned on reference viewpoints, with the backbone entirely frozen.

We present **Scal3R**, an efficient framework for scalable online 3D reconstruction (Fig. 1). A small set of learnable tokens is injected into the frozen backbone via *asymmetric attention*: pose tokens attend to image features as queries while image tokens compute self-attention exclusively among themselves, preserving the frozen representation space. Each pose token predicts the relative transformation between the current frame and a historical reference, constraining localization to the local viewpoint domain where the model is most reliable.

To maintain global consistency, Scal3R incorporates multi-reference relative querying and online pose-graph optimization: relative poses are queried against multiple dynamically selected reference frames simultaneously, aggregated via PGO into a drift-free trajectory, and supplemented by a loop closure mechanism that integrates naturally into the multi-reference pipeline. As shown in Fig. 1,



(a) Extrapolation failure (b) Error correlation analysis

Fig. 2: Global pose regression fails under out-of-distribution sequences, while local geometry remains reliable. (a) Feed-forward models like CUT3R [76] produce accurate reconstructions for in-distribution frames (**blue**). However, they suffer from severe geometric collapse when extrapolating to unseen long-range trajectories (**orange**). **(b)** Our error correlation analysis reveals a critical decoupling. The global position error diverges catastrophically in the out-of-distribution region (**red**), while per-frame depth remains stable (**blue**). This finding suggests that the backbone's local geometric representations are intact. This motivates Scal3R to freeze the base model and replace fragile global regression with stable multi-reference relative pose querying.

Scal3R produces geometrically consistent reconstructions on sequences spanning hundreds of meters, finetuned in only 8 hours on a single GPU.

In summary, our contributions are as follows:

- We identify that global extrapolation instability and limited training data coverage cause long-sequence collapse in long-sequence reconstruction. We also show that local geometric representations remain reliable throughout.
- We propose Scal3R. It introduces multi-reference relative pose querying via visual prompt tuning on a frozen backbone. Asymmetric attention injection ensures that pose learning does not degrade point cloud quality.
- Integrating multi-reference querying with online PGO, Scal3R achieves low drift on kilometer-scale sequences. It converges in about 8 hours on a single GPU using only 4-view training samples.

2 Related Work

Multi-view 3D Reconstruction. Early reconstruction relied on offline Structure-from-Motion [54, 56] and Multi-View Stereo [57] pipelines. Feed-forward methods dramatically improved efficiency by predicting geometry in a single forward pass [6, 9, 12, 22, 23, 32, 39, 43, 65, 77, 87, 94], with recent transformer-based models scaling to large unordered collections via joint pose-and-geometry prediction [74, 75, 80] and efficient aggregation strategies [13, 16, 24, 25, 36, 44–46, 59, 63, 64, 68, 71, 73, 82, 86, 92]. A complementary trend adapts frozen 3D foundation models for downstream tasks without backbone retraining [33, 60]. Scal3R follows this paradigm but uniquely targets scalable pose estimation on unbounded video streams, where batch-processing systems fundamentally cannot operate.

Online 3D Reconstruction. Online methods shift from batch processing to incremental inference, processing video frame by frame, achieved through recurrent TSDF fusion [62] and differentiable bundle adjustment [69]. CUT3R [76] and STream3R [38] bring this to 3D foundation models via persistent state updates and causal Transformers, spawning a broad family of streaming systems [2,5,14,37, 42,51,58,66,72,83,89,95,98] and SLAM integrations [20,29,47,49,50,52,90,91,93]. However, all share a critical flaw in that poses are regressed relative to the first frame, anchoring the trajectory to a single global reference. As shown in Figs. 1 and 2, this strategy becomes increasingly fragile as sequences grow, where small feature drifts are amplified into catastrophic geometric collapse.

Long-sequence Streaming 3D Reconstruction. Suppressing drift over kilometer-scale sequences remains an open challenge, addressed through test-time gradient updates [10], training-free memory management [89, 95], long-range token pools [42], explicit spatial memory [83], stage-decoupled streaming [15], and offline global optimization [18, 19, 85], each trading off online capability against global consistency. A unifying insight from the visual odometry literature is that relative formulations generalize better than absolute ones [8, 21]. This insight guides Scal3R. Rather than improving global-anchor regression, we reformulate the problem as multi-reference relative pose querying on a frozen backbone, eliminating the root extrapolation failure with only $\sim$1% additional parameters.

Efficient Prompt Tuning. Parameter-efficient adaptation has shown that frozen pre-trained models need very little change to transfer well. Adapters [27], prefix tokens [41], soft prompts [40], low-rank perturbations [28], and parallel adapter modules [7] all match or exceed full fine-tuning at under 2% of parameters, a finding confirmed broadly across vision transformers [30, 34, 55, 70, 84]. This paradigm has since reached 3D vision, where geometry-aware prompts and low-rank adapters on frozen 3D transformers [1, 67, 78, 96] and reconstruction backbones [48, 81] consistently match full fine-tuning. In 3D reconstruction, Human3R [11] first demonstrates prompt tuning on a frozen CUT3R for joint human-scene reconstruction. Scal3R is the first to apply this paradigm to relative pose estimation, recasting it as a multi-reference prompt query task via asymmetric attention injection that preserves the backbone's pointmap quality while gaining globally consistent motion representations.

3 Method

3.1 Overview

Scal3R addresses online 3D reconstruction by reformulating global pose regression as a multi-reference relative pose query problem. Given a streaming sequence of images $\mathcal{I} = \{I_1, I_2, \ldots, I_T\}$, instead of directly regressing absolute camera poses in a unified world coordinate system, we query relative poses with respect to a set of maintained reference frames. This reformulation fundamentally eliminates

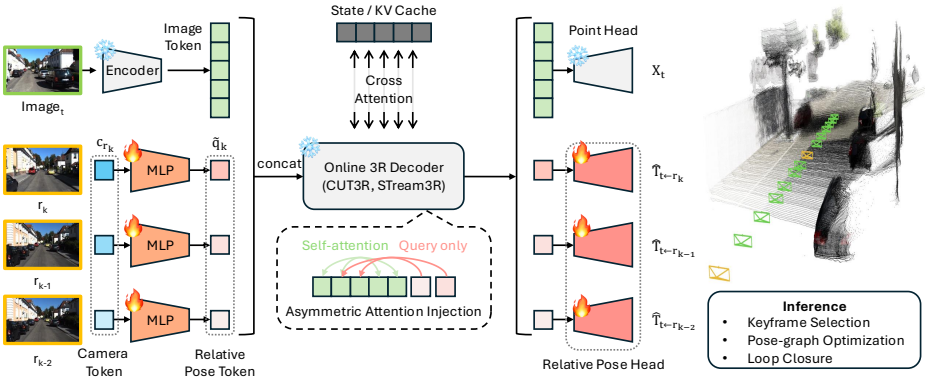


Fig. 3: Overview of the Scal3R framework. For an incoming frame Image$_t$, a frozen encoder extracts dense image tokens. At the same time, historical camera tokens from selected reference frames $(r_k, r_{k-1}, \dots)$ are projected using lightweight trainable MLPs to generate relative pose tokens. These tokens are concatenated and sent into a completely frozen 3D reconstruction decoder (*e.g.*, CUT3R [76] or STream3R [38]). Our **Asymmetric Attention Injection** mechanism is crucial as it ensures that relative pose tokens serve only as queries to extract geometric cues, while image tokens perform self-attention exclusively among themselves. This method preserves the original high-quality point cloud generation (X_t) through the frozen Point Head, while the trainable Relative Pose Head predicts robust multi-reference relative transformations $(\hat{\mathbf{T}}_{t\leftarrow r_k})$. Finally, an online inference backend (PGO and loop closure) aggregates these local constraints to produce a globally consistent trajectory.

the long-horizon extrapolation instability that plagues existing global-regression approaches.

Architecturally, Scal3R builds upon frozen pretrained online 3D reconstruction backbones (*e.g.*, CUT3R [76] or STream3R [38]), preserving their rich spatiotemporal geometric priors. A lightweight set of learnable tokens is injected into the frozen decoder via an asymmetric attention mechanism, enabling relative pose decoding without modifying the pretrained weights. At the backend, an online Pose-graph Optimization (PGO) module aggregates the predicted pairwise relative constraints into a globally consistent trajectory. An overview of the full pipeline is shown in Fig. 3.

3.2 Preliminaries: Online 3D Reconstruction Backbones

We briefly review the two representative backbone paradigms underlying Scal3R.

Persistent state model (CUT3R). At each timestep t, the frozen online 3D reconstruction backbone processes the current frame I_t together with a persistent hidden state S_{t-1} encoding the scene history, producing an updated state S_t and local geometry prediction G_t.

Causal Transformer model (STream3R). At each timestep t, the frozen online 3D reconstruction backbone processes the current frame I_t via causal attention over a sliding feature window $\{F_{t-k}, \ldots, F_t\}$ to perform cross-temporal geometric alignment in feature space, producing local geometry prediction G_t.

Both paradigms share a common decoding structure. At each frame, the decoder maintains image feature tokens $F_t \in \mathbb{R}^{H\times W\times C}$ and a dedicated camera token $\mathbf{c}_t \in \mathbb{R}^D$. Pointmaps are decoded from F_t for local geometry, while the global camera pose $P_t \in SE(3)$ relative to the first frame is regressed from $\mathbf{c}_t$.

Although effective for short sequences, global-reference regression degrades over long sequences: as the sequence grows, the model must align each new frame to an increasingly distant first-frame coordinate system, causing small feature drifts to be amplified into severe geometric collapse at the decoding stage. Scal3R retains the rich representations F_t and $\mathbf{c}_t$ learned by these backbones, while discarding their unstable global pose regression heads. Instead, we leverage historical camera tokens $\{\mathbf{c}_{r_k}\}$ stored in the pose token buffer as geometric conditioning signals to enable scalable relative pose queries (Sec. 3.3).

3.3 Multi-Reference Relative Pose Tuning

Our core contribution is a parameter-efficient prompt tuning mechanism that enables robust multi-reference relative pose prediction on a completely frozen backbone. The total number of newly introduced parameters accounts for $\sim$1% of the backbone's total parameter count. To endow the model with the ability to query multiple reference viewpoints simultaneously, we maintain a *pose token buffer* for storing the camera tokens of selected past keyframes. We learn a shared base query token $\mathbf{q} \in \mathbb{R}^D$ that serves as a query template directing the decoder to extract the geometric relationship between the current frame and a given reference frame. For each reference slot k, the corresponding reference frame features retrieved from the buffer are projected into feature space via a lightweight MLP and fused with the base token by additive injection:

$$\tilde{\mathbf{q}}_k = \mathbf{q} + \mathrm{MLP}(\mathbf{c}_{r_k}), \tag{1}$$

where $\mathbf{c}_{r_k}$ is the camera token of the k-th reference frame. This dynamic assembly allows the system to flexibly scale the number of active queries K based on available references, ensuring robustness during sequence initialization or buffer resets. Importantly, since each token queries independently, the number of reference frames can be freely extended at inference time without retraining.

Asymmetric Attention Injection. Naively inserting new tokens into the decoder's self-attention would perturb the attention distribution of image tokens, degrading pointmap reconstruction quality. We instead propose *asymmetric attention injection* (Fig. 4), where the pose query tokens $\{\tilde{\mathbf{q}}_k\}$ participate in decoder attention exclusively as *queries*, attending to all image tokens to extract geometric features, while image tokens compute their Keys and Values without

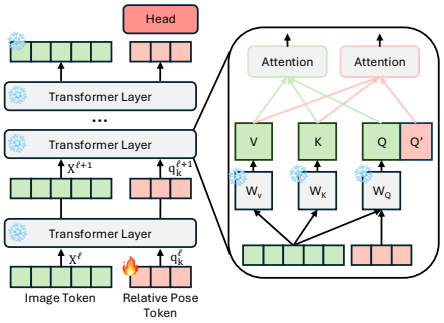


Fig. 4: Asymmetric Attention Injection. Relative pose tokens are injected only as additional queries (Q'), without modifying the keys and values of image tokens. This asymmetric design enables pose conditioning while preserving the pretrained image representations intact.

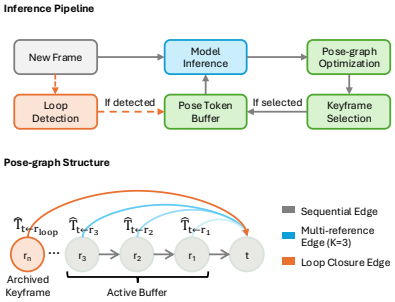


Fig. 5: Inference pipeline and pose graph structure. To suppress accumulative drift in long sequences, incoming frames are registered via keyframe selection and pose-graph optimization over sequential, multi-reference (K=3), and loop closure edges before being committed to the pose token buffer.

attending to the pose query tokens. For a decoder layer with image tokens $\mathbf{X}$:

$$\tilde{\mathbf{q}}_k^{\ell+1} = \text{Attention}\Big(\mathbf{Q} = \tilde{\mathbf{q}}_k^{\ell},\ \mathbf{K} = \mathbf{X}^{\ell},\ \mathbf{V} = \mathbf{X}^{\ell}\Big), \tag{2}$$

$$\mathbf{X}^{\ell+1} = \text{SelfAttention}\Big(\mathbf{X}^{\ell}\Big). \tag{3}$$

This one-directional information flow guarantees that the image feature representation space remains identical to that of the original frozen model, fully preserving pointmap reconstruction fidelity. No attention mask is needed, as pose tokens never enter the image K/V sequence.

Relative Pose Decoding and Loss. After multi-layer feature exchange, each pose query token $\tilde{\mathbf{q}}_k$ encapsulates the relative geometric constraint between the current frame t and its corresponding reference frame r_k. A lightweight MLP head decodes these tokens into relative poses. We adopt the 6D rotation representation [97] to ensure continuity in the rotation space, and output the relative transformation:

$$\hat{\mathbf{T}}_{t\leftarrow r_k} = \text{Head}(\tilde{\mathbf{q}}_k) \in SE(3). \tag{4}$$

The training loss supervises rotation $\mathbf{R}$ and translation $\mathbf{t}$ separately, where $\mathbf{R}_{pred}^k$ and $\mathbf{t}_{pred}^k$ denote the rotation matrix and translation vector decomposed from $\hat{\mathbf{T}}_{t\leftarrow r_k}$, and $\mathbf{R}_{gt}^k$, $\mathbf{t}_{gt}^k$ are the corresponding ground-truth components. To handle monocular scale ambiguity, translation vectors are scale-aligned before

loss computation. The total loss aggregates over all K reference links:

$$\mathcal{L} = \sum_{k=1}^{K} \left(\lambda_R \left\| \mathbf{R}_{pred}^{k} - \mathbf{R}_{gt}^{k} \right\|_F + \lambda_t \left\| \mathbf{t}_{pred}^{k} - \mathbf{t}_{gt}^{k} \right\|_2 \right), \tag{5}$$

where λ_R and λ_t are loss weighting hyperparameters.

3.4 Online Pose-graph Optimization

While multi-reference relative pose predictions provide accurate pairwise constraints, naively chaining them accumulates drift over long sequences. We therefore integrate an online pose-graph optimization (PGO) framework (Fig. 5) that uses the predicted relative poses as between-factors and performs incremental trajectory correction as new frames arrive.

Keyframe Selection. Including every frame in the pose graph introduces numerical redundancy and unnecessary computation. We adopt an online 3D overlap-based keyframe selection strategy. A KD-tree spatial index maintains the reconstructed 3D point cloud; the visible point set for each frame is determined by projecting predicted 3D points onto a unit sphere and computing the angular overlap with past keyframes, following [5], to avoid interference from geometrically non-adjacent regions. For each incoming frame t, the predicted 3D points are transformed to world coordinates via the current pose estimate. If the depth-normalized overlap score falls below a threshold τ_{overlap} and the median depth confidence exceeds τ_{conf}, the frame is designated as a keyframe, indicating novel geometry with reliable prediction quality. Only keyframes update the frozen decoder's KV cache and enter the pose token buffer. Non-keyframe KV states are discarded by restoring the pre-forward snapshot, keeping the streaming decoder state clean. The first N_{init} frames are unconditionally treated as keyframes to initialize the system.

Pose-graph Optimization. We model the trajectory as a factor graph where each camera pose $T_t \in SE(3)$ is a variable node. The multi-reference relative poses form between-factors connecting the current frame to its K_t references:

$$E_{\text{between}} = \sum_{t} \sum_{k=1}^{K_t} \rho\left(\left\| \log\left(T_{r_k}^{-1} T_t \cdot \hat{T}_{t \leftarrow r_k} \right) \right\|_{\Sigma_{t,k}}^{2} \right), \tag{6}$$

where $\hat{T}_{t \leftarrow r_k}$ is the predicted relative pose, $\Sigma_{t,k}$ is a diagonal noise covariance, and $\rho(\cdot)$ is the Huber robust kernel to downweight outlier constraints. To account for higher uncertainty in predictions between temporally distant frame pairs, we adopt a gap-dependent noise model where the standard deviation scales as $\sigma = \sigma_{\text{base}} \cdot \Delta^{0.5}$ with frame gap Δ. We employ iSAM2 [35] for incremental optimization. Upon each new frame arrival, the factor graph is updated and efficiently re-optimized via the Bayes tree structure. Optimized poses are written back to the buffer so that subsequent frames use corrected references.

For long sequences, the frozen decoder's streaming state is reset every N_{reset} frames to prevent memory overflow and feature degradation. To maintain pose-graph connectivity across resets, the last frame of each segment is re-fed as the first frame of the next segment, with a tight identity constraint imposed between the two corresponding nodes in the factor graph.

Loop Closure. Despite PGO continuously correcting local drift, long-term trajectory consistency requires explicitly detecting and closing loops when the camera revisits previously observed regions. A key advantage of our multi-reference design is that loop closure integrates naturally into the existing inference pipeline. When a loop candidate is detected between the current frame t and a past keyframe r_{loop}, the archived camera token of r_{loop} is simply re-injected into the pose token buffer as an additional reference slot. The frozen model then predicts a long-range relative pose constraint $\hat{\mathbf{T}}_{t \leftarrow r_{\text{loop}}}$ without any architectural modification, which is added to the pose graph as a high-confidence edge with a tight Gaussian noise model.

For loop detection, we employ a pretrained DINOv2 [53] backbone with a SALAD aggregation layer [31] to produce discriminative scene-level descriptors, indexed online via FAISS over keyframes only. Candidates are filtered by cosine similarity threshold τ_{sim}, minimum temporal gap τ_{gap}, and non-maximum suppression within a window w_{nms} to suppress redundant detections.

4 Experiments

4.1 Implementation Details

Model Configurations. We build upon two representative online 3D reconstruction backbones: CUT3R, which centers on persistent state updates, and STream3R, which is based on causal Transformers. In our experiments, we employ their 24-layer Transformer backbones (comprising a DINOv2 encoder and a Transformer decoder) and keep them entirely frozen to leverage their strong spatiotemporal geometric priors. For each incoming frame, we introduce a set of lightweight, learnable relative pose query tokens, which account for only approximately 1% of the total model parameters. These tokens are injected into the decoder layers via asymmetric attention injection to extract geometric constraints of the current frame relative to the reference frames in the pose token buffer. A pose decoding head then maps these features into $SE(3)$ space, predicting the 6D rotation and translation vectors.

Training. We fine-tune our model on the TartanAir [79] dataset, which provides diverse scenarios and precise trajectory ground truth. To ensure robustness to varying motion velocities and baseline lengths, we adopt a Random Interval Sampling strategy during training. For each training sample, we select 4 views from a sequence, with one serving as the current frame I_t and the remaining three as reference frames ($K = 3$), and randomly perturb the temporal intervals

Table 1: Camera pose estimation on KITTI (ATE ↓). The upper block lists offline baselines, and the lower block reports online methods. Scal3R reduces average ATE by over 60% compared to the strongest online competitor TTT3R, with particularly pronounced gains on long-range sequences. **-** denotes OOM or tracking failure.

Methods	KITTI (ATE ↓)											Avg.
	00	01	02	03	04	05	06	07	08	09	10	
	4542× 3.7km	1101× 2.5km	4661× 5.1km	801× 0.6km	271× 0.4km	2761× 2.2km	1101× 1.2km	1101× 0.7km	4071× 3.2km	1591× 1.7km	1201× 0.9km	
VGGT [74]	-	-	-	-	-	-	-	-	-	-	-	-
Fast3R [86]	-	723.1	-	166.9	112.2	-	137.4	90.9	-	225.6	211.7	238.3
DA3 [45]	-	-	-	-	12.6	-	-	-	-	-	-	12.6
π^3 [80]	-	-	-	-	2.3	-	-	-	-	-	-	2.3
MASt3R-SLAM [52]	188.5	562.9	282.4	121.7	92.6	-	57.2	77.0	263.6	184.1	179.1	200.9
MUSt3R [5]	-	490.1	-	121.5	58.1	-	100.6	66.8	-	-	-	167.4
CUT3R [76]	191.0	644.1	295.4	150.1	20.1	155.6	132.7	72.8	231.8	206.2	179.6	207.2
Point3R [83]	-	-	-	-	-	-	-	-	-	-	-	-
STream3R [38]	189.0	694.9	302.6	162.7	100.5	159.1	122.7	87.1	264.5	220.9	195.9	227.3
WinT3R [42]	-	698.1	-	144.5	89.4	150.8	135.8	76.0	242.8	200.9	210.3	216.5
StreamVGGT [98]	-	-	-	-	98.8	-	-	-	-	-	-	98.8
TTT3R [10]	178.4	529.1	280.8	98.0	11.4	147.9	132.1	70.2	240.2	191.0	125.2	182.2
Ours (CUT3R)	45.3	165.0	139.9	33.9	9.9	29.4	47.4	6.4	227.0	29.6	33.2	69.7
Ours (STream3R)	57.9	176.4	170.8	10.9	9.3	39.1	18.1	15.2	173.8	73.9	33.9	70.8

between frames. This mechanism forces the model to extract stable relative pose representations under varying levels of geometric constraint. Since each pose query token attends independently, the number of reference frames can be freely scaled at inference time without retraining; we use $K = 12$ during inference. For the long outdoor benchmarks (KITTI and vKITTI), we reset the frozen decoder's streaming state every $N_{\text{reset}} = 10$ frames; all other datasets use no reset. We train with a batch size of 8 for 40 epochs using the AdamW optimizer with a learning rate of 1×10^{-4}. Thanks to the frozen backbone and lightweight query tokens, the entire fine-tuning converges in approximately 8 hours on a single NVIDIA A100 GPU, avoiding the collapse of geometric priors commonly observed in full-parameter fine-tuning on small-scale datasets.

Baselines. We compare Scal3R with offline transformers, streaming models, and SLAM-style systems. Offline transformers include VGGT [74], π^3 [80], Fast3R [86], and DA3 [45]. Streaming baselines include CUT3R [76], MUSt3R [5], TTT3R [10], STream3R [38], WinT3R [42], StreamVGGT [98], and Point3R [83]. MASt3R-SLAM [52] is included as an incremental SLAM counterpart. All methods operate in an intrinsic-free setting, taking only RGB input without known camera intrinsics, and are evaluated with official default settings under a unified protocol.

4.2 Quantitative Results

Camera Pose Estimation. We evaluate ATE across multiple benchmarks, including KITTI [26] and Virtual KITTI (vKITTI) [4] for outdoor driving sce-

Table 2: Camera pose estimation on Virtual KITTI (ATE ↓). The upper block lists offline baselines, the middle block reports online methods, and the lower block presents our Scal3R variants. Scal3R surpasses all streaming methods by a large margin and approaches the accuracy of offline approaches on long-range sequences. **-** denotes OOM or tracking failure.

Methods	vKITTI (ATE ↓)					Avg.
	Scene01 447x, 1.3km	Scene02 233x, 0.6km	Scene06 270x, 0.6km	Scene18 339x, 1.1km	Scene20 837x, 2.5km	
VGGT [74]	-	0.22	-	-	-	0.22
Fast3R [86]	72.71	23.72	6.05	58.93	159.40	64.16
DA3 [45]	25.22	0.22	0.17	12.53	182.89	44.20
π^3 [80]	4.63	0.33	0.18	3.04	-	2.04
MASt3R-SLAM [52]	75.16	-	7.056	-	-	41.11
MUSt3R [5]	77.23	7.39	0.28	48.53	170.91	60.87
CUT3R [76]	59.76	29.73	0.65	44.89	146.92	56.39
Point3R [83]	75.70	30.44	4.81	62.85	138.24	62.41
STream3R [38]	66.83	25.47	1.52	68.18	223.07	77.01
WinT3R [42]	59.02	28.61	1.10	42.89	203.16	66.96
StreamVGGT [98]	55.09	32.91	0.57	57.18	-	36.44
TTT3R [10]	29.05	12.15	0.54	6.75	77.90	25.28
Ours (CUT3R)	4.50	0.72	0.66	5.56	16.72	5.63
Ours (STream3R)	5.39	3.26	3.21	6.40	21.32	7.92

Table 3: Camera pose estimation on Sintel, TUM-Dynamic, and ScanNet (ATE ↓). Scal3R achieves state-of-the-art performance across all three benchmarks, demonstrating strong generalization to diverse unseen indoor and synthetic environments.

Methods	Sintel	TUM	ScanNet
	ATE ↓	ATE ↓	ATE ↓
MASt3R-SLAM [52]	0.233	0.103	0.081
MUSt3R [5]	0.242	0.048	0.046
CUT3R [76]	0.210	0.049	0.095
Point3R [83]	0.375	0.067	0.120
STream3R [38]	0.214	0.026	0.052
WinT3R [42]	0.225	0.074	0.062
StreamVGGT [98]	0.394	0.057	0.120
TTT3R [10]	0.210	0.028	0.064
Ours (CUT3R)	0.168	0.033	0.092
Ours (STream3R)	0.171	0.018	0.049

narios, as well as Sintel [3], TUM-Dynamic [61], and ScanNet [17] for diverse indoor and synthetic environments. Following CUT3R [76] and STream3R [38], we apply Sim(3) alignment to the ground truth before computing ATE. As shown in Tabs. 1 to 3, Scal3R consistently outperforms both offline and online baselines. On KITTI (Tab. 1), our method achieves an average ATE of 69.7, reducing error by over 60% compared to the strongest online competitor TTT3R (182.2), with particularly pronounced gains on long-range sequences such as Seq. 00 and Seq. 02. On vKITTI (Tab. 2), Scal3R (CUT3R) attains an average ATE of 5.63, surpassing all streaming methods by a large margin and approaching the accuracy of the offline method π^3, while remaining fully online. On Sintel, TUM-Dynamic, and ScanNet (Tab. 3), Scal3R generalizes robustly to unseen domains: Scal3R (CUT3R) achieves the best ATE of 0.168 on Sintel, and Scal3R (STream3R) attains state-of-the-art ATE of 0.018 on TUM-Dynamic and 0.049 on ScanNet, demonstrating strong performance across both large-scale outdoor and dense indoor environments without sacrificing online processing.

3D Reconstruction. We evaluate 3D reconstruction quality on the 7-Scenes dataset using 300-frame sequences, reporting Accuracy (Acc.), Completeness (Comp.), and Normal Consistency (NC). As shown in Tab. 4, Scal3R variants consistently enhance the geometric consistency of their frozen backbones. Ours (STream3R) achieves the best performance across all metrics, attaining an NC mean of 0.579 and median of 0.622, surpassing the STream3R backbone by a clear margin. Notably, while competing streaming methods struggle to improve geometric consistency beyond their base backbone, our scale-decoupled formulation yields consistent gains in NC without sacrificing accuracy or completeness.

Table 4: 3D reconstruction on 7-Scenes (300 frames). We report Accuracy (Acc.), Completeness (Comp.), and Normal Consistency (NC), each with mean and median values. Scal3R consistently improves upon both CUT3R and STream3R backbones, achieving the best NC across all sequences.

Methods	7-Scenes (300 frames)					
	Acc. ↓		Comp. ↓		NC ↑	
	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R [76]	0.130	0.090	0.062	0.023	0.544	0.565
STream3R [38]	0.095	0.040	0.030	0.006	0.560	0.590
Ours (CUT3R)	0.069	0.034	0.035	0.010	0.566	0.601
Ours (STream3R)	0.052	0.012	0.020	0.004	0.579	0.622

4.3 Qualitative Results

Figs. 6 and 7 visualize 3D reconstruction and trajectory estimation on outdoor long-sequence benchmarks, confirming stable reconstruction and pose prediction across varying spatial extents. In terms of 3D reconstruction (Fig. 6), we compare our method against CUT3R and STream3R on Virtual KITTI Scene 01 (332 m) and Scene 02 (113 m). While CUT3R produces severely distorted point clouds with substantial geometric drift, and STream3R collapses into degenerate reconstructions that deviate considerably from the ground-truth layout, both Ours (CUT3R) and Ours (STream3R) recover scene geometry that closely matches the ground truth, with clean structural boundaries and well-preserved spatial extent. For long-sequence pose estimation (Fig. 7), we visualize estimated trajectories on KITTI Seq. 00 and Seq. 05 against CUT3R, WinT3R, and STream3R. All three baselines suffer from catastrophic trajectory collapse, producing chaotic, self-intersecting paths that bear no resemblance to the ground-truth loop structure. In contrast, Ours (CUT3R) faithfully traces the full loop trajectory, maintaining metric accuracy and geometric coherence across hundreds of meters.

4.4 Ablation Study

We conduct ablation studies on vKITTI and KITTI to validate the key components of our method, covering model training strategy, inference-time design choices, and loop closure. Results are summarized in Tabs. 5 to 7.

Model Training Strategy. As shown in Tab. 5, removing reference-frame supervision entirely (*w/o reference*) sharply degrades $\text{RPE}_{\text{trans}}$ to 3.336, while a single reference reduces ATE to 15.764. Our full multi-reference training achieves the best ATE of 5.632, confirming that denser reference supervision is critical for robust long-sequence pose estimation.

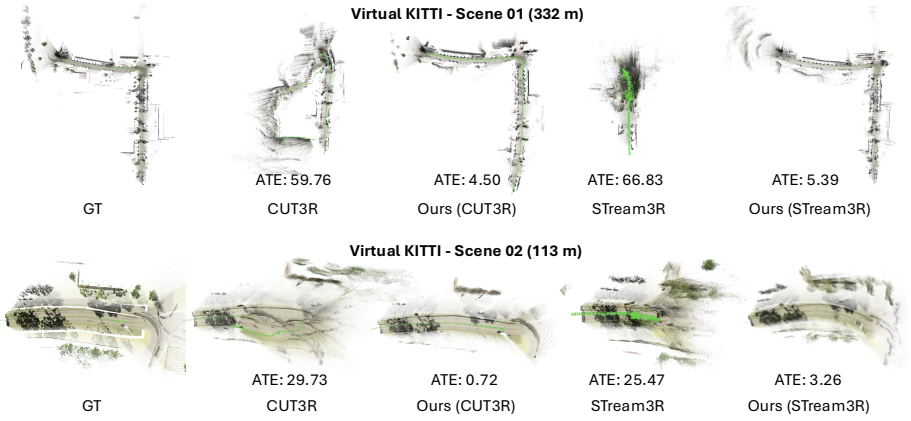


Fig. 6: Qualitative 3D reconstruction on Virtual KITTI. Comparison against CUT3R and STream3R on Scene 01 (332 m) and Scene 02 (113 m). While both baselines produce distorted or collapsed point clouds, Scal3R recovers scene geometry closely matching the ground truth across both sequences.

Table 5: Ablation on training strategy (vKITTI). We vary the number of reference frames used for supervision during training. Denser reference supervision consistently lowers ATE, while removing references entirely causes severe RPE degradation.

Method	ATE ↓	RPE_{trans} ↓	RPE_{rot} ↓
Baseline (CUT3R)	56.390	1.678	0.342
w/o reference	33.318	3.336	0.842
1 reference	15.764	0.217	0.448
Ours	**5.632**	**0.177**	**0.485**

Table 6: Ablation on inference-time components (vKITTI). Both keyframe selection and PGO are essential, and scaling the reference count from 4 to 12 at inference time consistently reduces ATE without retraining.

Method	ATE ↓	RPE_{trans} ↓	RPE_{rot} ↓
w/o keyframe	38.258	0.787	1.388
w/o PGO	20.089	0.203	0.544
$K = 4$	15.748	0.188	**0.470**
$K = 8$	7.362	0.180	**0.470**
Full ($K = 12$)	**5.632**	**0.177**	0.485

Inference-Time Components. Tab. 6 ablates the inference-time pipeline. Removing keyframe selection or PGO each leaves a large gap to the full system (ATE: 38.258 without keyframe selection). Progressively increasing reference count from 4 to 12 then consistently reduces ATE from 15.748 to 5.632.

Runtime Analysis. Fig. 8 breaks down the latency on KITTI (K=12). The frozen forward pass dominates on both backbones (86.3% on CUT3R, 91.2% on STream3R), so keyframe selection, PGO, and loop detection add little: the full pipeline runs at 14.4 and 7.95 FPS, versus 15.9 and 9.1 FPS for the backbones.

Loop Closure. As shown in Tab. 7 and Fig. 10, loop closure reduces average ATE on KITTI from 143.45 to 75.01 (48% improvement), with particularly large

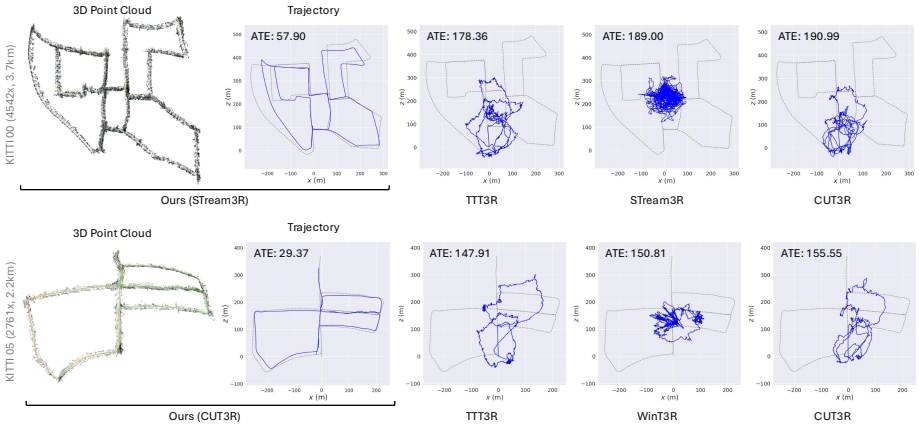


Fig. 7: Qualitative trajectory comparison on KITTI long sequences. We visualize estimated camera trajectories on Seq. 00 and Seq. 05 against CUT3R, WinT3R, and STream3R. All baselines suffer from catastrophic drift, while Scal3R faithfully recovers the full loop structure with metric accuracy.

Table 7: Ablation on loop closure (KITTI, ATE ↓). Loop closure reduces average ATE by 48%, with the largest gains on loop-heavy sequences (*e.g.*, Seq. 00 and 05).

Method	LC	KITTI (ATE ↓)							Avg.
		00	02	05	06	07	08	09	
Ours (CUT3R)		185.70	235.28	108.73	84.86	49.83	241.86	97.86	143.45
	✓	45.34	139.88	29.37	47.40	6.44	227.02	29.63	75.01
Ours (STream3R)		166.37	239.36	134.08	35.95	60.68	195.20	74.18	129.40
	✓	57.90	170.75	39.05	18.07	15.19	173.83	73.91	78.39

gains on sequences containing large loops (*e.g.*, Seq. 00 and Seq. 05), where globally consistent trajectory correction is most needed.

Robustness Analysis. Fig. 9 presents per-frame ATE curves sorted in ascending order across all KITTI sequences. Methods such as MUSt3R and Point3R suffer catastrophic failures at moderate sequence lengths, while our method maintains the lowest per-frame ATE throughout the entire evaluation range, confirming superior robustness under challenging long-sequence conditions.

5 Conclusion

We presented Scal3R, which tackles the instability of global pose regression on long sequences by reformulating camera localization as multi-reference relative pose querying on a frozen backbone. Lightweight tokens (∼1% of parameters)

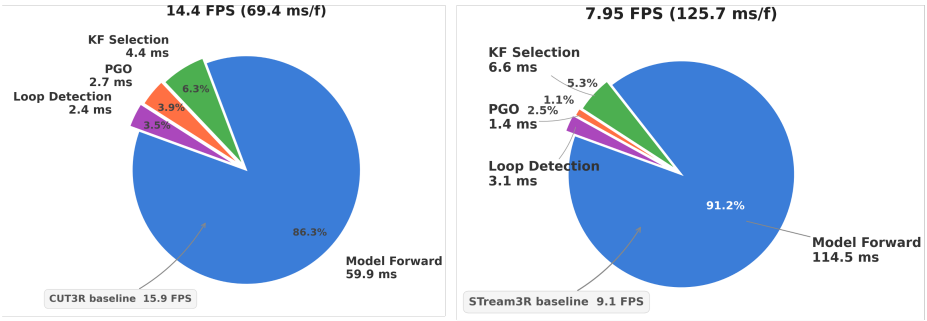


Fig. 8: Per-frame runtime breakdown on KITTI. On both the CUT3R (left) and STream3R (right) backbones, the frozen forward pass dominates latency; keyframe selection, PGO, and loop detection together add only a small fraction.

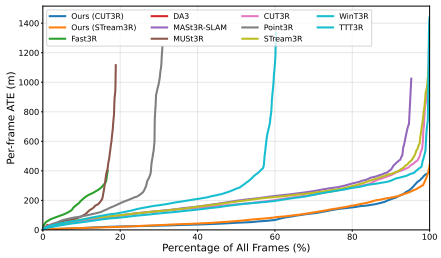


Fig. 9: Robustness Analysis on KITTI. Scal3R maintains the lowest error across the full evaluation range, while competing methods suffer catastrophic divergence at moderate sequence lengths.

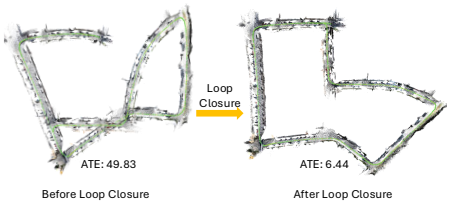


Fig. 10: Qualitative effect of loop closure on KITTI 07. Without loop closure, accumulated drift causes the trajectory to deviate from the ground-truth loop structure. With loop closure, Scal3R recovers a globally consistent trajectory.

query relative poses that an online pose-graph backend aggregates into a globally consistent trajectory. Trained in 8 hours on a single GPU, Scal3R enables accurate online reconstruction of long video streams.

Limitations. First, performance is bounded by the frozen backbone, degrading when it fails under occlusion or textureless regions. Second, the online backend has its own weaknesses: appearance-based loop closure can miss revisits under extreme viewpoint or illumination change, and keyframe selection and loop detection rely on hand-set thresholds. Improving both remains future work.

Acknowledgements. This work was supported by NVIDIA Taiwan AI Research & Development Center (TRDC). This research was funded by the National Science and Technology Council, Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2, 113-2628-E-A49-023-, 115-2628-E-A49-024-, and 111-2628-E-A49-018-MY4. Yu-Lun Liu acknowledges the Yushan Young Fellow Program by the MOE in Taiwan.

References

1. Ai, Z., Liu, Z., Lei, Y., Cui, Z., Zou, X., Zhou, J.: Gaprompt: Geometry-aware point cloud prompt for 3d vision model. arXiv preprint arXiv:2505.04119 (2025)
2. Antsfeld, L., Chidlovskii, B., Cabon, Y., Leroy, V., Revaud, J.: S-must3r: Sliding multi-view 3d reconstruction. arXiv preprint arXiv:2602.04517 (2026)
3. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) European Conf. on Computer Vision (ECCV). pp. 611–625. Part IV, LNCS 7577, Springer-Verlag (Oct 2012)
4. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
5. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1050–1060 (2025)
6. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19457–19467 (2024)
7. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
8. Chen, W., Chen, L., Wang, R., Pollefeys, M.: Leap-vo: Long-term effective any point tracking for visual odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19844–19853 (2024)
9. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Easi3r: Estimating disentangled motion from dust3r without training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9158–9168 (2025)
10. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645 (2025)
11. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Pons-Moll, G.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025)
12. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European conference on computer vision. pp. 370–386. Springer (2024)
13. Chen, Y., Qiu, Y., Li, R., Agha, A., Omidshafiei, S., Patrikar, J., Scherer, S.: Co-me: Confidence-guided token merging for visual geometric transformers. arXiv preprint arXiv:2511.14751 (2025)
14. Chen, Z., Qin, M., Yuan, T., Liu, Z., Zhao, H.: Long3r: Long sequence streaming 3d reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5273–5284 (2025)
15. Cheng, C., Chen, X., Xie, T., Yin, W., Ren, W., Zhang, Q., Guo, X., Wang, H.: Longstream: Long-sequence streaming autoregressive visual geometry (2026)
16. Cong, Z., Zhao, Q., Jeon, M., Tulsiani, S.: Flow3r: Factored flow prediction for scalable visual geometry learning. arXiv preprint arXiv:2602.20157 (2026)
17. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)

18. Dai, W., Su, W., Kong, D., Ming, Y., Kong, W.: Keyframe-based feed-forward visual odometry. arXiv preprint arXiv:2601.16020 (2026)
19. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences (2025)
20. Ding, T., Xie, Y., Liang, Y., Chatterjee, M., Miraldo, P., Jiang, H.: Laser: Layer-wise scale alignment for training-free streaming 4d reconstruction. arXiv preprint arXiv:2512.13680 (2025)
21. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16739–16752 (2025)
22. Du, Z., Danier, D., Lenssen, J.E., Bilen, H.: Moonseg3r: Monocular online zero-shot segment anything in 3d with reconstructive foundation priors. arXiv preprint arXiv:2512.15577 (2025)
23. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: International Conference on 3D Vision 2025 (2025)
24. Elflein, S., Li, R., Agostinho, S., Gojcic, Z., Leal-Taixé, L., Zhou, Q., Osep, A.: VGG-T^3: Offline feed-forward 3d reconstruction at scale. arXiv preprint arXiv:2602.23361 (2026)
25. Elflein, S., Zhou, Q., Leal-Taixé, L.: Light3r-sfm: Towards feed-forward structure-from-motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16774–16784 (2025)
26. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
27. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
28. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
29. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
30. Huang, L., Mao, J., Yi, J., Tao, Z., Wang, Y.: Cvpt: Cross visual prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 848–858 (2025)
31. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 17658–17668 (2024)
32. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1071–1081 (2025)
33. Jena, S., Ouasfi, A., Younes, M., Boukhayma, A.: Sparfels: Fast reconstruction from sparse unposed imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27476–27487 (2025)
34. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)

35. Kaess, M., Johannsson, H., Roberts, R., Ila, V., Leonard, J.J., Dellaert, F.: isam2: Incremental smoothing and mapping using the bayes tree. The International Journal of Robotics Research **31**(2), 216–235 (2012)
36. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
37. Khafizov, R., Komarichev, A., Rakhimov, R., Wonka, P., Burnaev, E.: G-cut3r: Guided 3d reconstruction with camera and depth prior integration. arXiv preprint arXiv:2508.11379 (2025)
38. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: Stream3r: Scalable sequential 3d reconstruction with causal transformer. arXiv preprint arXiv:2508.10893 (2025)
39. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024)
40. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 3045–3059 (2021)
41. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021)
42. Li, Z., Zhou, J., Wang, Y., Guo, H., Chang, W., Zhou, Y., Zhu, H., Chen, J., Shen, C., He, T.: Wint3r: Window-based streaming reconstruction with camera token pool. arXiv preprint arXiv:2509.05296 (2025)
43. Lin, C.Y., Sun, C., Yang, F.E., Chen, M.H., Lin, Y.Y., Liu, Y.L.: Longsplat: Robust unposed 3d gaussian splatting for casual long videos. In: ICCV (2025)
44. Lin, C.Y., Wu, C.H., Yeh, C.H., Yen, S.H., Sun, C., Liu, Y.L.: Frugalnerf: Fast convergence for few-shot novel view synthesis without learned priors. In: CVPR (2025)
45. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
46. Liu, S., Li, W., Qiao, P., Dou, Y.: Regist3r: Incremental registration with stereo foundation model. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4484–4493 (2025)
47. Liu, Y., Dong, S., Wang, S., Yin, Y., Yang, Y., Fan, Q., Chen, B.: Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16651–16662 (2025)
48. Lu, Z., Yang, H., Xu, D., Li, B., Ivanovic, B., Pavone, M., Wang, Y.: Lora3d: Low-rank self-calibration of 3d geometric foundation models. arXiv preprint arXiv:2412.07746 (2024)
49. Maggio, D., Carlone, L.: Vggt-slam 2.0: Real time dense feed-forward scene reconstruction. arXiv preprint arXiv:2601.19887 (2026)
50. Maggio, D., Lim, H., Carlone, L.: VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. arXiv preprint arXiv:2505.12549 (2025)
51. Mahdi, S., Ayar, F., Javanmardi, E., Tsukada, M., Javanmardi, M.: Evict3r: Training-free token eviction for memory-bounded streaming visual geometry transformers. arXiv preprint arXiv:2509.17650 (2025)
52. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16695–16705 (2025)

53. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
54. Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: European Conference on Computer Vision. pp. 58–77. Springer (2024)
55. Ren, L., Chen, C., Wang, L., Hua, K.: Da-vpt: Semantic-guided visual prompt tuning for vision transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4353–4363 (2025)
56. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
57. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
58. Shen, G., Deng, T., Wang, Y., Chen, Y., Shen, Y., Liu, J., Wang, J.: Grs-slam3r: Real-time dense slam with gated recurrent state. arXiv preprint arXiv:2509.23737 (2025)
59. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
60. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024)
61. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
62. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15598–15607 (2021)
63. Sun, X., Zhu, Z., Lou, Z., Yang, B., Tang, J., Zhang, L., Wang, H., Zhang, J.: Avggt: Rethinking global attention for accelerating vggt. arXiv preprint arXiv:2512.02541 (2025)
64. Sun, X., Jiang, H., Liu, L., Nam, S., Kang, G., Wang, X., Sui, W., Su, Z., Liu, W., Wang, X., et al.: Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. arXiv preprint arXiv:2508.03643 (2025)
65. Sun, Y.C., Sun, C., Lin, C.Y., Yang, F.E., Chen, M.H., Lin, Y.Y., Liu, Y.L.: 3am: Segment anything with geometric consistency in videos. arXiv preprint arXiv:2601.08831 (2026)
66. Taher, M., Alzugaray, I., Mazur, K., Kong, X., Davison, A.J.: Kv-tracker: Real-time pose tracking with transformers. arXiv preprint arXiv:2512.22581 (2025)
67. Tang, Y., Zhang, R., Guo, Z., Ma, X., Zhao, B., Wang, Z., Wang, D., Li, X.: Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 5171–5179 (2024)
68. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5283–5293 (2025)
69. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems **34**, 16558–16569 (2021)
70. Tu, C.H., Mai, Z., Chao, W.L.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7725–7735 (2023)

71. Wang, C.S.B., Schmidt, C., Piekenbrinck, J., Leibe, B.: Faster vggt with block-sparse global attention. arXiv preprint arXiv:2509.07120 (2025)
72. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024)
73. Wang, H., Agapito, L.: Amb3r: Accurate feed-forward metric-scale 3d reconstruction with backend. arXiv preprint arXiv:2511.20343 (2025)
74. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
75. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21686–21697 (2024)
76. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
77. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20697–20709 (June 2024)
78. Wang, S., Liu, X., Kong, L., Xu, J., Hu, C., Fang, G., Li, W., Zhu, J., Wang, X.: Pointlora: Low-rank adaptation with token selection for point cloud learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6605–6615 (2025)
79. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
80. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
81. Wang, Z., Cao, A., Wang, L.J., Park, J.J.: Moe3d: A mixture-of-experts module for 3d reconstruction. arXiv preprint arXiv:2601.05208 (2026)
82. Wang, Z., Xu, D.: Flashvggt: Efficient and scalable visual geometry transformers with compressed descriptor attention. arXiv preprint arXiv:2512.01540 (2025)
83. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. arXiv preprint arXiv:2507.02863 (2025)
84. Xin, Y., Yang, J., Luo, S., Du, Y., Qin, Q., Cen, K., He, Y., Zhang, Z., Fu, B., Yang, X., et al.: Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. arXiv preprint arXiv:2402.02242 (2024)
85. Xiong, Z., Zhang, C., Xu, Q., Tao, W.: Vggt-motion: Motion-aware calibration-free monocular slam for long-range consistency. arXiv preprint arXiv:2602.05508 (2026)
86. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025)
87. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024)
88. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)

89. Yuan, S., Yang, Y., Yang, X., Zhang, X., Zhao, Z., Zhang, L., Zhang, Z.: Infinitevggt: Visual geometry grounded transformer for endless streams. arXiv preprint arXiv:2601.02281 (2026)
90. Yuan, Y., Chen, Z., Li, K., Wang, W., Zhao, H.: Slam-former: Putting slam into one transformer. arXiv preprint arXiv:2509.16909 (2025)
91. Yugay, V., Nguyen, D.K., Gevers, T., Snoek, C.G.M., Oswald, M.R.: Visual odometry with transformers (2025)
92. Zhang, C., Le Moing, G., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., Ghahramani, Z., Zisserman, A., Zhang, J., Sajjadi, M.S.M.: Efficiently reconstructing dynamic scenes one d4rt at a time. arXiv preprint arXiv:2512.08924 (2025)
93. Zhang, G., Qian, S., Wang, X., Cremers, D.: Vista-slam: Visual slam with symmetric two-view association. arXiv preprint arXiv:2509.01584 (2025)
94. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
95. Zheng, Z., Xiang, X., Zhang, J.: Ttsa3r: Training-free temporal-spatial adaptive persistent state for streaming 3d reconstruction. arXiv preprint arXiv:2601.22615 (2026)
96. Zhou, X., Liang, D., Xu, W., Zhu, X., Xu, Y., Zou, Z., Bai, X.: Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14707–14717 (2024)
97. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)
98. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)