

LongStream: 长序列流式自回归视觉几何

作者: Chong Cheng^{1,2} Xianda Chen⁴ Tao Xie^{2,3} Wei Yin²
Weiqiang Ren² Qian Zhang² Xiaoyang Guo^{2,†} Hao Wang^{1,‡}

机构: ¹ 香港科技大学 (广州) ² 地平线机器人 ³ 浙江大学 ⁴ 中南大学

[‡] 项目负责人 [†] 通讯作者

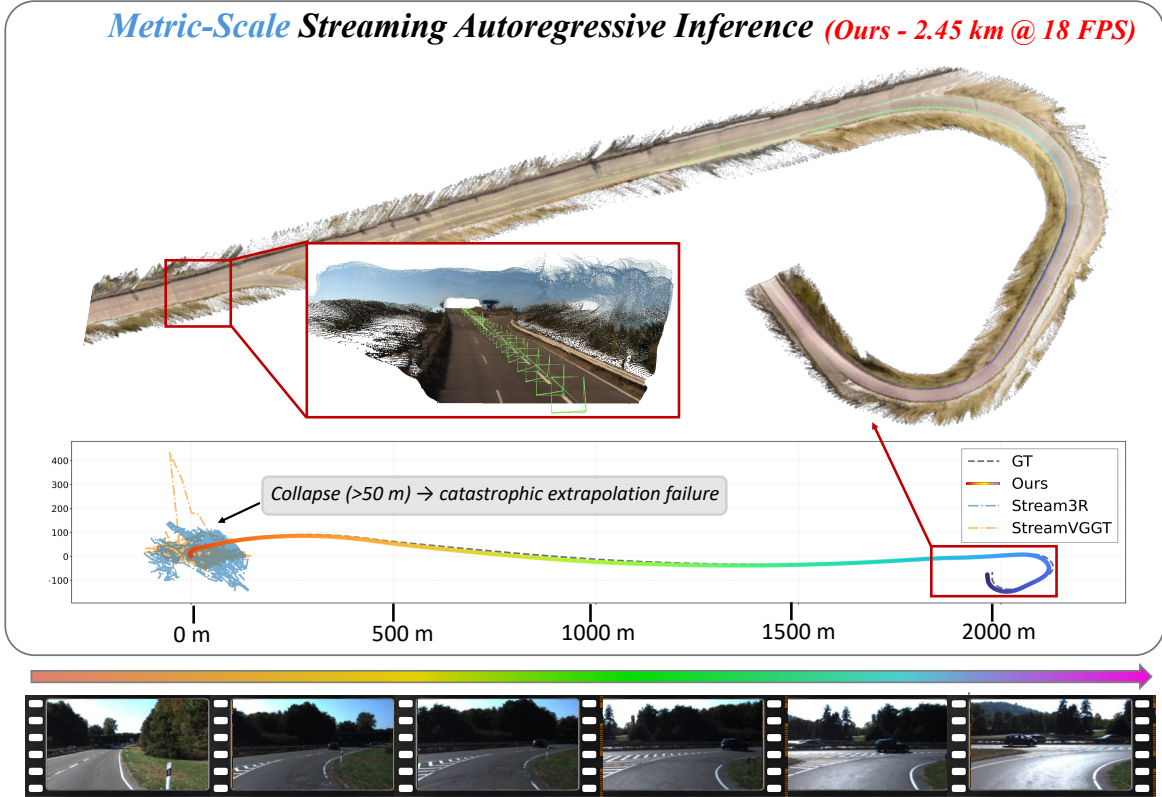


图 1. 面向度量尺度场景重建的流式自回归模型对比。现有流式模型 (e.g., Stream3R [1]、StreamVGGT [2]) 受外推误差影响, 在数十米范围内便会崩溃。相比之下, 本文提出的 LongStream 可稳定完成公里级重建。其规范解耦表述与缓存一致训练可保持度量精度和几何稳定性, 并在数公里长序列上维持 18 FPS 的处理速度。

摘要

长序列流式三维重建 (*Streaming 3D Reconstruction*) 仍是一项重要的开放性挑战。现有自回归模型在处理长序列时往往失效, 因为其将位姿锚定于第一帧, 进而导致注意力衰减、尺度漂移 (*Scale Drift*) 和外推误差。本文提出 *LongStream*, 这是一种规范解耦 (*Gauge-Decoupled*) 的流式视觉几何模型, 可在严格在线且未来帧不可见 (*Future-Invisible*) 的设定下, 对数千帧序列进行度量尺度 (*Metric Scale*) 场景重建。本文方法包含三个方面。首先, 舍弃第一帧锚点并预测

关键帧相对位姿 (*Keyframe-Relative Pose*), 将长距离外推重新表述为难度恒定的局部任务。其次, 提出正交尺度学习 (*Orthogonal Scale Learning*), 将几何与尺度估计完全解耦, 以抑制漂移。最后, 我们发现 *Transformer* 存在注意力偏置问题, 包括对注意力汇聚点 (*Attention Sink*) 的依赖和长期 *KV* 缓存 (*KV Cache*) 饱和。为此, 本文提出缓存一致训练 (*Cache-Consistent Training, CCT*) 并结合周期性缓存刷新 (*Periodic Cache Refresh*)。该方法抑制超长序列中的注意力偏置与污染, 缩小训练与推理之间的差距。实验表明, *LongStream* 达到当前最佳性能, 能够以 18

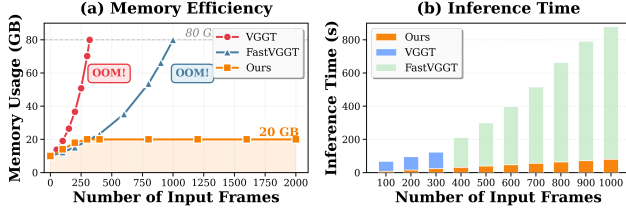


图 2. 内存与运行时间对比。本文方法的内存占用和延迟保持稳定，而 VGGT 与 FastVGGT 随序列增长迅速上升，并在长序列上发生内存溢出 (Out of Memory, OOM)。

FPS 对公里级序列进行稳定的度量尺度重建。项目主页: <https://3dagentworld.github.io/longstream/>

1. 引言

几何重建是从图像序列中联合估计相机位姿与稠密三维结构的技术 [3–8]，也是自动驾驶、增强现实/虚拟现实 (Augmented Reality/Virtual Reality, AR/VR) 和具身机器人等应用的基础技术 [9, 10]。这些领域要求系统能够实时、稳健地处理长序列视频流。

传统流程 [3, 4, 6, 11, 12] 与近期基于 Transformer 的方法 [13–18] 虽达到当前最佳精度，但本质上均为离线方法。每当加入新帧时，它们都需重新处理整个序列 [1, 2]，造成大量冗余计算，因而无法实时运行。为解决这一问题，研究者提出了流式模型。STream3R [1] 和 StreamVGGT [2] 等近期工作采用因果 Transformer 与 KV 缓存逐步构建重建结果。然而，这些模型处理长序列时会遭遇灾难性的外推失效。如图 1 和图 2 所示，现有流式方法虽能以线性时间逐步重建场景，但其轨迹在数十米内便会崩溃，最终导致跟踪完全失效。

我们认为，这种失效源于当前模型固有的“规范耦合 (Gauge-Coupled)”设计。这些模型锚定于第一帧坐标系，并通过训练回归绝对位姿。这种设计迫使模型学习位置固定的映射，使长序列预测愈加困难。受硬件限制，模型在训练批次中只能接触较小的帧索引，因此在推理时难以可靠外推至较大索引。由此，“短序列训练、长序列测试”偏差 [19, 20] 造成了域差距。

为此，本文提出 LongStream——一种“规范解耦”的流式自回归几何框架。该框架从理论上解耦全局 $SE(3)$ 坐标与度量尺度的规范自由度。

首先，为处理长序列中的坐标规范问题，我们移除固定的第一帧锚点，转而回归关键帧相对位姿。由此，欠定的长距离外推问题被重新表述为难度恒定的局部估计任务，使预测不受全局坐标选择影响。

其次，为解决 $Sim(3)$ 尺度漂移问题，本文引入正交尺度学习机制 [21]，在目标函数层面解耦几何学习与度量尺度估计。几何分支在尺度不变空间中优化形状，专用尺度头则独立预测全局尺度因子。这种设计有效减少尺度纠缠，确保稳定的度量输出。

除规范解耦表述外，流式 Transformer 架构本身也存在挑战。我们发现，注意力汇聚点 [22] (即对第一帧词元的偏置依赖) 与长期 KV 缓存污染是时序性能退化和漂移的主要原因。为解决这些问题，本文提出缓存

一致训练方案，在训练时显式传递并裁剪缓存，使训练与推理的上下文保持一致。此外，周期性缓存刷新会边缘化陈旧上下文、缓解长期记忆饱和并稳定几何估计。

在室外数据集 (KITTI、vKITTI、Waymo) 和室内数据集 (TUM-RGBD、ETH3D、7Scenes) 上的实验表明，LongStream 实现了当前最佳的流式重建性能，能够在公里级序列上进行实时 (18 FPS) 的度量尺度 AR 重建。本文的贡献总结如下：

- 本文提出 LongStream，这是一种以“规范解耦”设计为核心的流式几何基础模型。该模型预测关键帧相对位姿，并采用正交尺度解耦，从系统层面消除对第一帧锚点的依赖，有效缓解长序列外推失效。
- 本文发现，对注意力汇聚点的依赖与 KV 缓存污染是长时域 (Long-Horizon) 性能退化的主要原因。缓存一致训练方案与周期性缓存刷新可稳定时序注意力并减少几何漂移。
- 室内外基准上的实验表明，在严格在线且未来帧不可见的设定下，LongStream 达到当前最佳性能，并在长序列上保持实时吞吐率与稳定的度量尺度。

2. 相关工作

经典 SfM 与 MVS。运动恢复结构 (Structure-from-Motion, SfM) 与多视图立体 (Multi-View Stereo, MVS) 流程 [3–6, 11, 23, 24] 通过建模图像对应关系之间的几何联系重建三维场景。SfM 通过特征匹配与光束法平差 (Bundle Adjustment) 估计相机位姿和稀疏结构 [3, 4]，MVS 则利用平面扫描或代价体得到的逐像素深度对结构进行稠密化 [5, 6, 11, 23, 24]。这类依赖大量优化并使用手工特征的流程虽然精度较高且可解释性强，但难以扩展到大型或动态场景，也难以实时部署。

离线三维重建。早期端到端方法主要在图像对上训练 [25–28]。与体素、网格或隐式场相比，点图计算效率更高且适合实时应用 [29–33]，可用于同步定位与建图 (Simultaneous Localization and Mapping, SLAM) 和神经渲染 [34–38]。DUST3R [13] 无需相机内参，即可从两幅图像回归点图和相对位姿，但仍局限于成对处理并需要全局对齐。MASt3R [14] 引入稠密特征与双向匹配，以提升标定和匹配的稳健性，但仍以图像对为单位运行，处理多视图场景时需要代价高昂的融合。VGGT [15] 可通过一次前馈过程从多视图输入预测位姿、深度、点图和轨迹，但依赖固定参考帧与绝对位姿监督，因而产生参考帧偏置和尺度偏置。 π^3 [16] 通过置换等变设计消除参考视图偏置并预测局部点图，但其输出在全局相似变换下仍存在歧义，且不具备度量尺度。

流式三维重建。流式方法逐帧更新几何。经典单目 SLAM 与基于学习的变体 [39–44] 逐步恢复结构与运动。CUT3R [45] 维护循环状态以在线输出度量点图，但其循环神经网络 (Recurrent Neural Network, RNN) 骨干难以建模长期依赖，性能会在长序列上退化。Stream3R [1] 采用带 KV 缓存的因果 Transformer，虽可扩展至长视频流，但随着缓存词元占据主导，注意力会发生崩溃。StreamVGGT [2] 引入时序因果注意力

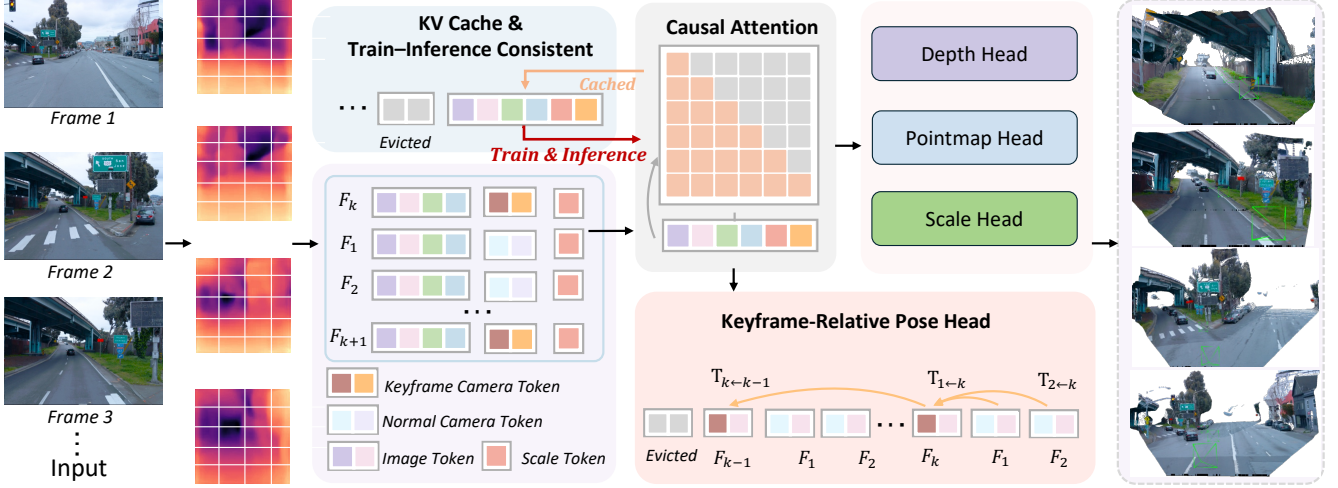


图 3. 本文提出的 LongStream 概览。对于流式输入，ViT 编码器提取图像块词元，并加入关键帧、普通帧和尺度词元。这些词元通过共享 KV 缓存的因果注意力进行融合；训练与推理阶段均以一致方式使用该缓存，从而实现缓存一致的流式建模。网络预测关键帧相对位姿 $\mathbf{T}_{i \leftarrow k}$ 、深度、点图（Pointmap）和全局尺度，从而在长序列上实现稳定的度量尺度重建。

和缓存更新，并通过蒸馏提升一致性；然而，缓存污染仍使长时域稳定性难以保证。总体而言，现有流式方法的性能会随序列增长而显著下降，且无法泛化至更长的视频流。

3. 方法

3.1. 整体框架

本文提出 LongStream，这是一种规范解耦（Gauge-Decoupled）的流式几何框架。如图 3 所示，该框架在统一的时空 Transformer 中联合预测位姿、深度和尺度。ViT 编码器生成图像块词元，并以相机（Camera）、寄存器（Register）和尺度（Scale）词元对其进行增强，以区分关键帧角色。这些词元由带共享 KV 缓存（KV Cache）的因果聚合器融合，从而支持长序列流式推理。对于每一帧，模型预测关键帧相对位姿（Keyframe-Relative Pose） $\mathbf{T}_{i \leftarrow k}$ 、深度、点图（Pointmap）以及全局尺度 s 。训练与推理采用相同的布局，从而实现规范解耦且稳定的度量尺度（Metric Scale）重建。下面详细介绍各个设计组件。

3.2. 规范解耦表述

本文旨在克服现有流式模型依赖规范耦合（Gauge-Coupled）设计所带来的局限。我们认为，稳健的几何学习系统在理论上必须对规范自由度保持不变，即不受任意全局坐标和度量尺度影响。为此，本文提出一个系统分离 $SE(3)$ 与 $Sim(3)$ 自由度的框架。

在 $SE(3)$ 规范中，我们舍弃绝对位姿回归，将位姿学习重新定义为规范不变（Gauge-Invariant）的相对位姿估计 [46, 47]。学习目标变为：

$$\mathbf{T}_{i \leftarrow k} = \mathbf{T}_i \circ \mathbf{T}_k^{-1}, \quad (1)$$

其中， $\mathbf{T}_i$ 和 $\mathbf{T}_k$ 表示世界坐标系到相机坐标系的绝对位姿，第 k 帧为前一个关键帧。对于任意世界坐标

系重参数化 $S \in SE(3)$ ，该表述在数学上均保持规范不变性。这一解耦设计同时实现两个目标：一是将大索引引发的分布外（Out-of-Distribution）长距离外推问题 [1, 2]，转化为索引间隔 $(i - k)$ 有界、难度恒定的分布内（In-Distribution）局部任务；二是缓解固定锚点偏置，后者会导致以首帧为锚点的因果模型不稳定。

在 $Sim(3)$ 规范中，我们采用尺度不变（Scale-Invariant, SI-Log）思想 [21] 解决混乱的尺度纠缠问题。本文采用正交尺度学习（Orthogonal Scale Learning）机制，在架构和目标函数两个层面分离几何学习与度量尺度估计。几何分支经过归一化，并依据尺度不变原则进行监督 [21]；专用尺度头（Scale Head）则预测全局尺度因子 s 。

3.3. 网络架构

给定输入图像 I_i ，模型为每一帧 i 预测其相对于参考关键帧 k 的相对位姿， $\mathbf{p}_{i \leftarrow k} = [\mathbf{t}_{i \leftarrow k}, \mathbf{q}_{i \leftarrow k}, f_{i \leftarrow k}]$ ，其中 $\mathbf{t}$ 为平移， $\mathbf{q}$ 为单位四元数， f 为焦距偏移。模型还输出深度图 D_i 、对应的世界坐标点云 X_i 、全局尺度因子 s 以及帧表示 h_i ：

$$\{h_i, \mathbf{p}_{i \leftarrow k}, D_i, X_i, s\} = F_\theta(I_i), \quad i = 1, \dots, S. \quad (2)$$

其中 h_i 表示帧级特征表示。

沿用 VGGT 类流式模型的架构 [1, 2, 15]，本文模型由基于 DINOv2 [48] 的词元化器、因果 Transformer 聚合器，以及面向相对位姿、深度、点图和尺度的任务专用预测头组成。几何采用关键帧相对方式建模：每一帧预测其相对于当前关键帧 k 的位姿和点图，使流式重建不再依赖固定的首帧坐标系。

对于每一帧 I_i ，词元化器提取图像块特征 $x_i^p \in \mathbb{R}^{P \times C}$ ，并使用相机词元和寄存器词元进行增强，其中关键帧与非关键帧使用不同的词元。此外，引入

共享尺度词元以实现 $Sim(3)$ 解耦。所有词元拼接为 $H^{(0)} \in \mathbb{R}^{B \times S \times (P+T) \times C}$ ，随后由一组 Transformer 块处理；这些块在严格因果掩码下交替执行帧内注意力和全局注意力：

$$H^{(l+1)} = \text{Block}^{(l)}(H^{(l)}, \text{AttnMask}), \quad (3)$$

其输出被送入位姿头、深度头和点图头进行迭代细化。**相对位姿头 (Relative Pose Head)**。该预测头从聚合器接收帧特征和关键帧特征，并显式预测当前帧 i 相对于参考关键帧 k 的相对变换 $\mathbf{T}_{i \leftarrow k}$ ：

$$\mathbf{p}_{i \leftarrow k} = [\mathbf{t}_{i \leftarrow k}, \mathbf{q}_{i \leftarrow k}, f_{i \leftarrow k}], \quad (4)$$

其中，平移 $\mathbf{t}_{i \leftarrow k} \in \mathbb{R}^3$ 、旋转 $\mathbf{q}_{i \leftarrow k} \in \mathbb{H}$ （单位四元数）和焦距偏移 $f_{i \leftarrow k} \in \mathbb{R}^2$ 共同构成相对相机位姿：

$$\mathbf{T}_{i \leftarrow k} = \mathbf{T}_i \mathbf{T}_k^{-1}. \quad (5)$$

对于世界坐标系的任意右乘变换，该定义均保持不变，从而消除对固定世界锚点的依赖。

为确保模型仅学习 (i, k) 之间的相对关系，本文采用参考帧感知注意力 (Reference-Aware Attention) 方案。对于分配给关键帧 k 的非关键帧 i ，在因果掩码或窗口掩码下，其词元仅关注来自 k 以及 k 到 i 之间各帧的词元。相应地，关键帧词元仅关注来自前一关键帧 $k-1$ 以及 $k-1$ 到 k 之间各帧的词元，而非完整历史。聚合完成后，我们通过拼接和线性投影融合当前帧及其参考关键帧的位姿词元：

$$\mathbf{h}_{\text{fused}} = \text{Proj}([\mathbf{h}_i, \mathbf{h}'_k]). \quad (6)$$

最后，沿用 RAFT [49] 的设计，该预测头采用 AdaLN 调制的 Transformer，通过增量更新 $\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} + \Delta \mathbf{p}^{(t)}$ 迭代预测相对位姿 $\mathbf{p}_{\text{rel}} = [\mathbf{t}, \mathbf{q}, f]$ 。

尺度头。为实现 $Sim(3)$ 规范不变性，本文设计了显式尺度头，它接收专用的尺度词元 (*Scale Token*)；这一设计与同期工作 MapAnything [50] 的尺度词元思想相似。尺度头预测不受约束的对数尺度变量 $x_s \in \mathbb{R}$ ，随后对其取指数，得到严格为正的尺度因子：

$$s = \exp(\mathbf{w}^\top \mathbf{h}_{\text{scale}}), \quad (7)$$

其中， $\mathbf{h}_{\text{scale}}$ 是聚合器最后一层中尺度词元的特征。尺度 s 仅作用于平移、深度和点图输出，旋转与视场角保持不变。尺度头仅在提供真实度量尺度的数据集上训练。

深度头与点图头。深度头 (Depth Head) 和点图头 (Pointmap Head) 接收聚合后的帧级与图像块级特征，为每一帧预测深度图 $D_i \in \mathbb{R}^{H \times W}$ 、对应的世界坐标点 $X_i \in \mathbb{R}^{H \times W \times 3}$ ，以及逐像素置信度分数。这些分支与尺度头协同工作：几何在尺度不变空间内优化，而 *ScaleToken* 独立学习全局缩放因子，从而实现完整的 $Sim(3)$ 规范解耦。

3.4. 概率框架与损失函数

为在统一概率框架内实现规范解耦设计，本文将总体目标表述为：给定输入序列，最大化几何、运动与尺度的联合似然。令 I 表示图像帧， X 表示三维点图， D 表示深度， p 表示相对位姿， s 表示全局尺度。本文最小化负对数后验 (Negative Log Posterior)：

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{geom}} + \mathcal{L}_{\text{depth}}}_{\text{Geometry \& Depth Likelihood}} + \underbrace{\mathcal{L}_{\text{pose}}}_{\text{Pose Likelihood}} + \underbrace{\mathcal{L}_{\text{scale}}}_{\text{Scale Prior}}, \quad (8)$$

其中，各项对应后验分解中的一个条件因子：

$$p(D, X, p, s | I) \propto p(D | X, I) \cdot p(X | p, s, I) \cdot p(p | I) \cdot p(s). \quad (9)$$

该因子分解使学习过程与第 3.2 节介绍的 $SE(3)$ 和 $Sim(3)$ 规范不变表述保持一致。

相对位姿损失。相对位姿损失 $\mathcal{L}_{\text{pose}}$ 对应于 $p(p | I)$ ，并在多次迭代更新中监督 RelPoseHead 的输出 $\mathbf{p}_{\text{rel}} = [\mathbf{t}, \mathbf{q}, f]$ ：

$$\mathcal{L}_{\text{pose}} = \sum_{t=1}^T \gamma^{t-1} \left(\ell(\hat{q}^{(t)}, q_{i \leftarrow k}) + \ell_t(\hat{t}^{(t)}, t_{i \leftarrow k}) + \ell(\hat{f}^{(t)}, f_{i \leftarrow k}) \right), \quad (10)$$

其中， ℓ 表示 L1 损失， $f_{i \leftarrow k}$ 为焦距偏移。为保持规范解耦，平移项 ℓ_t 在归一化坐标空间内计算，以确保平移监督不会隐式编码全局尺度。

几何损失。受 SI [21] 启发，几何损失 $\mathcal{L}_{\text{geom}}$ (形状优化) 对应于 $p(X | p, s, I)$ ，并在归一化空间内计算：

$$\tilde{X}_{\text{pred}} = \frac{\hat{X}_{\text{raw}}}{\text{Norm}(\hat{X}_{\text{raw}})}, \quad \tilde{X}_{\text{gt}} = \frac{X}{\text{Norm}(X)}, \quad (11)$$

$$\mathcal{L}_{\text{geom}} = \|\tilde{X}_{\text{pred}} - \tilde{X}_{\text{gt}}\|_1.$$

归一化消除了显式尺度依赖，确保 $\partial \mathcal{L} / \partial s = 0$ 。因此， $\mathcal{L}_{\text{geom}}$ 仅监督主干网络学习正确的三维结构。

尺度损失。尺度损失 $\mathcal{L}_{\text{scale}}$ 对全局尺度 s 进行正则化。由于尺度是乘性的，我们在对数空间比较预测尺度与真实尺度，以衡量相对误差并稳定梯度：

$$\mathcal{L}_{\text{scale}} = \|\log \hat{s} - \log s_{\text{gt}}\|_1. \quad (12)$$

其中， s_{gt} 是根据标定后的真实深度计算得到的度量尺度。该损失仅用于经过度量标定的样本；对于非度量数据，仅使用几何损失和深度损失进行训练。

3.5. KV 缓存与训练--推理一致性

流式 Transformer 的既有研究表明，模型往往依赖注意力汇聚点 [22, 52–54] 来稳定注意力；缺少该汇聚点时，滑动窗口越过首帧便可能触发模型崩溃 (Model Collapse)。

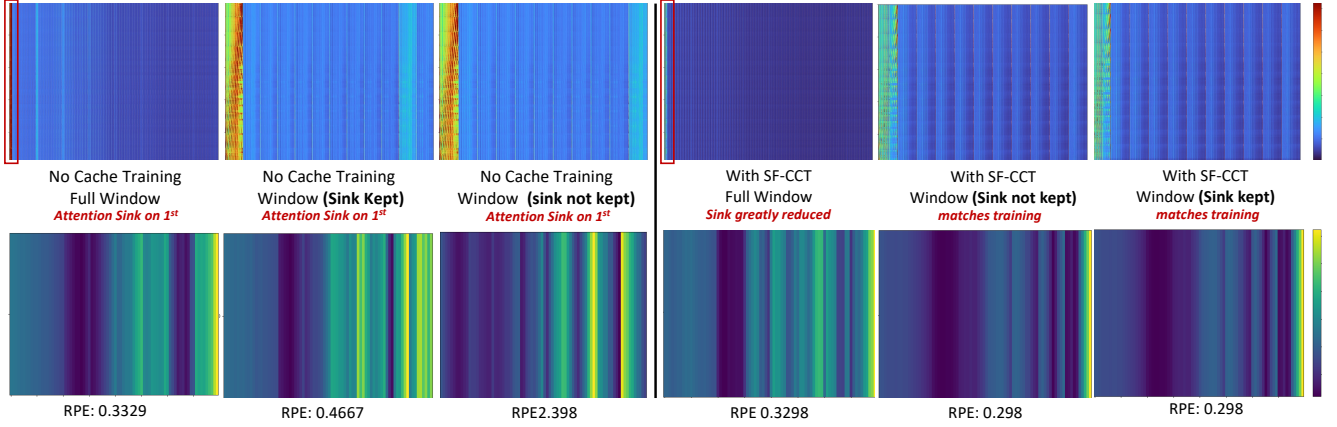


图 4. 缓存一致训练 (Cache-Consistent Training, CCT)。图中展示不同训练-推理设置下的注意力图 (上) 和相对位姿误差 (Relative Pose Error, RPE) 热力图 (下)。不采用 CCT 时 (左), 因果推理会形成明显的注意力汇聚点 (Attention Sink); 窗口式推理在保留该汇聚点时会将其放大, 而移除时则会崩溃。采用 CCT 后 (右), 因果模式中的汇聚点受到显著抑制, 在两种窗口模式中同样得到抑制, 由此获得稳定且最优的精度。浅蓝色表示对关键帧的注意。

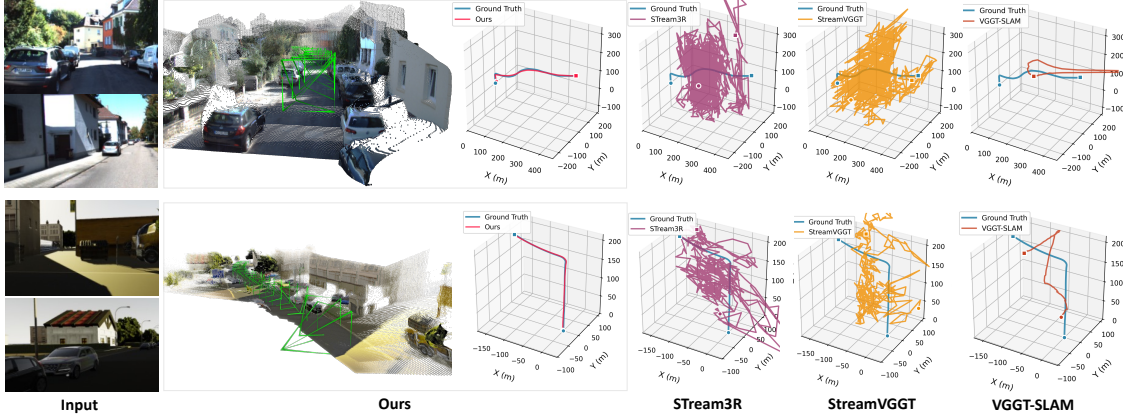


图 5. 长序列位姿估计的定性对比。我们在跨度达数百米的 KITTI 和 vKITTI 序列上, 将 LongStream 与流式方法和 SLAM 基线进行比较。Stream3R 与 StreamVGGT 在长轨迹上不断累积漂移, 而 VGGT-SLAM 在第二个 vKITTI 序列上耗尽内存。LongStream 在所有场景中均保持稳定且一致的位姿, 即使存在大幅度回环运动也能维持轨迹连续性。

Algorithm 1 缓存一致训练

```

1: 输入: 数据块  $\{c_1, \dots, c_N\}$ , 初始缓存  $KV^{(0)} = \emptyset$ 
2: for  $i = 1$  to  $N$  do
3:    $(out_i, KV^{new}) = \text{model}(c_i, KV^{(i-1)})$ 
4:    $KV^{(i)} = \text{trim}(KV^{new}, \text{window\_size})$ 
5: end for

```

然而, 这种锚定方式会使模型脆弱地依赖首帧。在长序列中, 它会导致几何饱和 (Geometric Saturation) [53], 具体表现为注意力不稳定、关键帧跳变以及位姿误差持续增长。即使采用相对位姿监督, 这些问题仍然存在, 说明汇聚点锚定本身会引入非对称的位置偏差。

我们认为, “短时域崩溃 (Short-Horizon Collapse)” 并非由移除汇聚点直接导致, 而是训练-推理不匹配的症状。因此, 本文提出算法 1 所示的缓存一致训练 (CCT), 通过在数据块之间显式传递并裁剪 KV 缓存, 使训练与推理的可见范围保持一致。训练期间, 我们移

除恒定的汇聚点词元, 并在滑动窗口中采用纯因果掩码; 同时, 在数据块之间显式传递并裁剪 KV 缓存, 使缓存可见性与推理过程一致。

如图 4 所示, CCT 使分块训练与逐帧推理之间的注意力模式在数学上等价, 迫使模型在没有持久锚点的纯滑动窗口中运行, 从而消除对汇聚点的依赖。

对于超长序列, 累积的 KV 仍会导致长期记忆饱和和几何漂移。因此, 本文采用周期性缓存刷新 (Periodic Cache Refresh): 每隔 N 个关键帧重置汇聚点帧与 KV 缓存, 对陈旧上下文执行硬边缘化, 这类似于 SLAM 中的状态边缘化。

该策略在清除退化特征的同时保持几何连续性, 无需额外计算即可周期性重置记忆; 由于整个模型在关键帧相对坐标系中运行, 因此可以在任意关键帧刷新缓存, 而不会破坏一致性或降低精度。

结合 CCT 与周期性缓存刷新后, 模型能够在数千帧上实现稳定且具备泛化能力的流式处理, 同时保持一致的几何精度和行为良好的注意力分布。

方法	KITTI [51] (ATE ↓)										
	00	01	02	03	04	05	06	07	08	09	10
	4542x, 3.7km	1101x, 2.5km	4661x, 5.1km	801x, 0.6km	271x, 0.4km	2761x, 2.2km	1101x, 1.2km	1101x, 0.7km	4071x, 3.2km	1591x, 1.7km	1201x, 0.9km
FastVGGT	*	705.39	*	62.38	10.27	157.74	124.43	69.27	*	190.10	194.75
MASt3R-SLAM	*	530.37	*	18.87	88.98	159.430	92.00	*	*	*	*
VGGT-SLAM	*	607.16	*	169.83	13.12	*	*	*	*	*	*
CUT3R	185.89	651.52	296.98	148.06	22.17	155.61	132.54	77.03	238.39	205.94	193.39
TTT3R	190.93	546.84	218.77	105.28	11.62	153.12	132.94	70.95	180.57	211.01	133.00
STream3R	190.98	681.95	301.40	158.25	102.73	159.85	135.03	90.37	261.15	216.31	207.49
StreamVGGT	191.93	653.06	303.35	157.50	108.24	160.46	133.71	89.00	263.95	216.69	209.80
本文方法	92.55	46.01	134.70	3.81	1.95	84.69	23.12	14.93	62.07	85.61	21.48

表 1. KITTI 数据集上的 ATE 定量对比。上半部分列出基于优化的基线，下半部分给出流式方法。本文方法在所有序列上均取得最高精度。

方法	TUM [55]	Oxford Spires [56]	Waymo [57]
	ATE ↓	ATE ↓	ATE ↓
FastVGGT	0.418	36.577	1.281
MASt3R-SLAM	0.082	37.728	7.625
VGGT-SLAM	0.123 (0.053 [†])	31.003	7.431
CUT3R	0.542	32.440	9.396
TTT3R	0.308	36.214	3.486
STream3R	0.633	37.569	42.203
StreamVGGT	0.627	37.255	45.101
本文方法	0.076	19.815	0.737

表 2. 在 TUM [55]、Oxford Spires [56] 和 Waymo [57] 上的定量对比。上部：基于优化的方法；下部：流式方法。本文方法在这些小尺度轨迹上表现出良好的鲁棒性，并在所有在线基准上取得最佳性能。[†] 表示结果引自 [58]。

方法	vKITTI [59] (ATE ↓)					平均
	场景 01 447×, 332m	场景 02 223×, 113m	场景 06 270×, 51m	场景 18 339×, 254m	场景 20 837×, 711m	
FastVGGT	3.435	0.311	0.120	2.050	101.667	31.427
MASt3R-SLAM	83.771	20.206	3.840	68.875	231.064	98.714
VGGT-SLAM	25.128	0.237	0.281	1.641	68.840	23.667
CUT3R	50.968	29.913	0.820	29.012	127.583	55.276
TTT3R	29.877	11.785	0.598	7.445	71.208	28.099
STream3R	68.280	26.450	8.185	43.597	198.279	82.815
StreamVGGT	71.616	15.349	10.274	23.900	221.407	83.916
本文方法	1.422	0.185	0.303	0.683	4.030	1.610

表 3. 在 vKITTI 上的定量对比。上部：基于优化的方法；下部：流式方法。本文方法在所有序列上均取得最高精度。

4. 实验

4.1. 实现细节

模型配置。LongStream 以 VGGT 初始化，并保留其由全局注意力与帧级注意力交替组成的 24 层骨干网络，参数量约为 1.3B。训练时采用固定的可见性布局、一致的滑动窗口以及大小为十的关键帧间隔，使每个批次包含多次关键帧切换。优化器采用 AdamW 和余弦衰减，峰值学习率为 4×10^{-6} ，并设置 1k 步预热。所有图像、深度和点图均缩放至最长边不超过 518 像素，同时采用宽高比抖动、间隔采样和跨块打乱。

训练方法。LongStream 采用两阶段训练方案：第一阶段在 32 块 A100 GPU 上以批大小 22 进行 50k 次批次独立训练，耗时三天；第二阶段采用与流式推理匹配

的缓存一致训练 (Cache-Consistent Training, CCT)，序列长度从 10 到 80 之间采样，缓存窗口大小设为 10。仅在具有已标定真值时使用度量尺度监督。推理时，LongStream 在单块 GPU 上可达到 18 FPS。更多实现细节见附录。

训练数据。LongStream 使用多领域数据集集合进行训练，包括 Kubric [60]、WildRGB [61]、ScanNet [62]、HyperSim [63]、Mapillary [64]、Replica [65]、MVS-Synth [66]、PointOdyssey [67]、Virtual KITTI [59]、Aria Synthetic Environments [68]、Aria Digital Twin [68]、Objaverse [69]、Spring [70] 和 Waymo Open [57]。BlendedMVS [71]、Co3Dv2 [72]、MegaDepth [73] 和 DL3DV [74] 不提供度量尺度真值，因而不参与尺度训练；其余训练数据集均提供度量监督。

基线方法。我们将 LongStream 与离线 Transformer、流式模型和 SLAM 类系统进行比较。由于 VGGT [15] 和 π^3 [16] 在长视频片段上均超出内存限制，因此使用 FastVGGT [75] 作为可实际运行的替代方案。流式基线包括 CUT3R [45]、TTT3R、STream3R [1] 和 StreamVGGT [2]。MASt3R-SLAM [34] 与 VGGT-SLAM [58] 是基于优化的 SLAM 基线。VGGT-SLAM 以 16/32 帧为块进行推理，因此可以访问每个块内的未来帧。所有方法均采用官方默认设置，并在统一协议下评估。

4.2. 定量结果

相机位姿估计。我们在 vKITTI [59] (训练集)、Waymo [57] (留出集) 以及训练中未见的 KITTI [35]、TUM-RGBD [55] 和 Oxford Spires [56] 数据集上评估绝对轨迹误差 (Absolute Trajectory Error, ATE)。如表 1–3 所示，随着序列长度增加，流式基线的误差呈非线性增长，这可能源于长时域历史信息的饱和与污染，而 LongStream 始终保持稳定。离线模型在长视频片段上经常出现内存不足 (Out of Memory, OOM) 和跟踪丢失。尽管完全以流式方式运行，LongStream 仍以 18 FPS 达到当前最佳精度，并在不同环境间展现出稳健的泛化能力。

三维重建。我们在 7Scenes 和 TUM 的测试序列上进行不经下采样的完整序列重建，并报告 CD 和 F1@0.25。

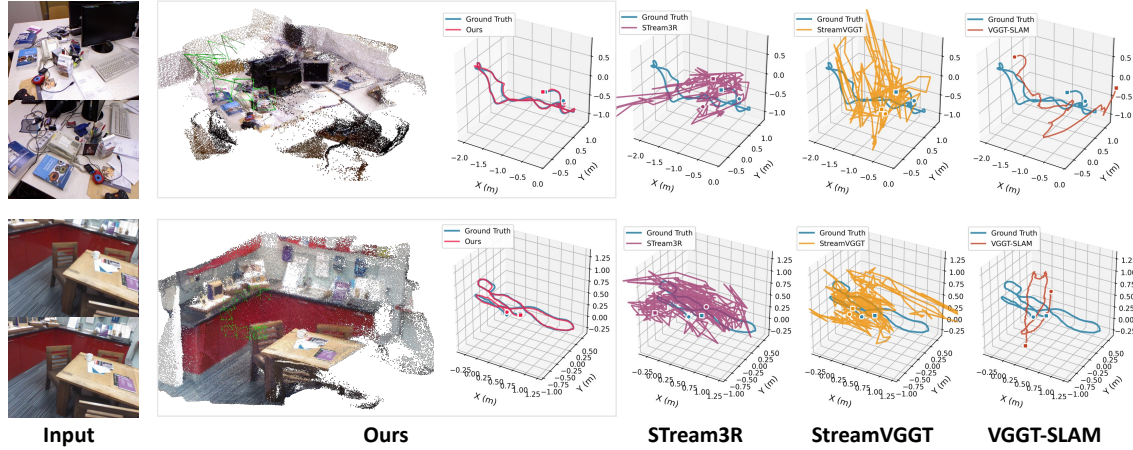


图 6. 室内序列的定性对比。我们在具有剧烈视角变化、遮挡和反复回退的挑战性场景上进行评估。Stream3R、StreamVGGT 和 VGGT-SLAM 在这些高度曲折的轨迹上产生漂移，而 LongStream 在整个序列中均能保持稳定位姿和一致的三维结构。符号 * 表示运行过程中出现内存不足错误或超过三次跟踪失败。

方法	7Scenes		TUM	
	CD ↓	F1@0.25 ↑	CD ↓	F1@0.25 ↑
FastVGGT	6.373	0.710	0.104	0.926
MASt3R-SLAM	5.987	0.691	0.057	0.954
VGGT-SLAM	6.306	<u>0.696</u>	1.993	0.633
CUT3R	6.281	0.274	0.474	0.533
TTT3R	6.231	0.260	0.249	0.792
STream3R	<u>6.353</u>	0.479	1.126	0.444
StreamVGGT	6.630	<u>0.483</u>	0.680	0.402
本文方法	2.260	0.641	0.225	0.673

表 4. 在 7Scenes 和 TUM 上的定量对比。评估指标为 Chamfer 距离 (Chamfer Distance, CD, 越低越好) 和 F1@0.25 (越高越好)。最佳结果以**粗体**标出，次佳结果以下划线标出。

相对位姿	尺度头	CCT	缓存刷新	ATE ↓	RPE ↓	尺度误差 ↓
×	×	×	×	8.043	2.207	-
✓	×	×	×	2.819	0.750	-
✓	✓	×	×	2.645	0.484	0.010
✓	✓	✓	×	0.984	0.454	0.032
✓	✓	✓	✓	0.115	0.126	0.035

表 5. 相对位姿、尺度头、CCT 和缓存刷新的消融研究。绿色表示启用，红色表示禁用。第 2 行与第 3 行之间的 ATE 差距由少数较大的轨迹离群值造成。尺度误差表示绝对尺度偏差，越低越好。

如表 4 和图 6 所示，LongStream 在两个基准上均具备与离线方法相当的性能。这些数据集的空间范围较小且帧覆盖密集，导致指标大体处于饱和状态。在 7Scenes 上，部分方法采用全部帧评估时会在某些场景中失败，所得 CD 值高出数个数量级，从而抬高平均 CD。
尺度估计。我们在 vKITTI 上评估恢复出的度量尺度精度。LongStream 在整个序列中均能生成稳定的尺度估计，相对于真值的尺度比达到 0.9905。

4.3. 定性结果

图 5 和图 6 分别可视化了公里级与房间尺度序列上的轨迹，验证了模型在大小两类空间范围内均能稳定预测位姿。在室外场景中（图 5），STream3R 和 StreamVGGT 等现有流式方法在长轨迹上出现累积漂移，而基于优化的 VGGT-SLAM 在更长序列上受到内存限制并发生跟踪丢失。相比之下，LongStream 在数百米范围内保持轨迹连续性与度量精度，并在没有显式回环闭合模块的情况下成功闭合大型回环。

在室内环境中（图 6），面对具有剧烈视角变化、遮挡和反复回退的高度曲折相机轨迹，模型同样表现出良好的鲁棒性。基线方法在这些不规则运动模式下往往产生不稳定位姿或明显漂移，而 LongStream 能保持连贯的全局轨迹。

4.4. 消融研究

我们在单个 vKITTI 序列上进行消融实验，以验证 LongStream 的四个核心组件：关键帧相对位姿头、尺度分支、缓存一致训练 (CCT) 和周期性缓存刷新。如表 5 所示，组合所有模块可将 ATE 从 8.043 降至 0.115，降幅接近两个数量级。

规范解耦的位姿与尺度。从绝对位姿回归切换至本文的规范解耦形式可带来最大性能增益（第 1 行 → 第 2 行），表明将局部几何与全局坐标分离对于泛化到训练窗口之外至关重要。尺度分支是保持度量一致性的必要条件；移除该分支后，模型无法在整个序列中生成稳定的全局尺度。

时序缓存一致性。周期性缓存刷新可防止无限流中的长期记忆饱和，而专用尺度分支对于保持度量一致性至关重要。

与基线相比，组合所有组件可使最终 ATE 降低近两个数量级。更详细的分析见附录。

5. 结论

LongStream 可在超长序列上实现稳定的度量尺度重建，克服现有方法的漂移与外推失效问题。其规范解耦位姿设计与缓存一致训练可在数千帧范围内保持几何和尺度一致，并在多种室内外数据集上取得较高精度与实时性能。

局限性。该模型仍假设场景大体静态，依赖启发式关键帧调度，且在极长窗口中点图一致性会轻微下降。这些局限为未来提升模型的稳健性与通用性指明了方向。

6. 致谢

本研究得到国家自然科学基金 (编号 62406267)、广东省项目 (编号 2024QN11X072)、广州市—香港科技大学 (广州) 联合资助计划 (编号 2025A03J3956) 以及广州市教育项目 (编号 2024312122) 的支持。

参考文献

- [1] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025. 1, 2, 3, 6
- [2] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 1, 2, 3, 6
- [3] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2
- [4] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2
- [5] Chong Cheng, Gaochao Song, Yiyang Yao, Qinzhen Zhou, Gangjian Zhang, and Hao Wang. Graph-guided scene reconstruction from images with 3d gaussian splatting, 2025. 2
- [6] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 785–801, Cham, 2018. Springer International Publishing. 2
- [7] Gaochao Song, Chong Cheng, and Hao Wang. Gvxf: Gaussian voxel kernel functions for highly efficient surface reconstruction in open scenes, 2024.
- [8] Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4129–4137, New York, NY, USA, 2024. Association for Computing Machinery. 2
- [9] Wenhao Shen, Wanqi Yin, Hao Wang, Chen Wei, Zhongang Cai, Lei Yang, and Guosheng Lin. Hmr-adapter: A lightweight adapter with dual-path cross augmentation for expressive human mesh recovery. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6093–6102, New York, NY, USA, 2024. Association for Computing Machinery. 2
- [10] Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization, 2025. 2
- [11] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnets for high-resolution multi-view stereo depth inference. *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [12] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, page 4976–4982. IEEE, May 2024. 2
- [13] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2
- [14] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2
- [15] 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*, 2025. 2, 3, 6
- [16] 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. *arXiv preprint arXiv:2507.13347*, 2025. 2, 6
- [17] Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction, 2025.
- [18] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025. 2
- [19] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. 2
- [20] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019. 2
- [21] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. 2, 3, 4
- [22] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. 2, 4

- [23] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2
- [24] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction, 2022. 2
- [25] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 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*, pages 19457–19467, 2024. 2
- [26] Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25442–25452, 2025.
- [27] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion, 2025.
- [28] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2
- [29] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [30] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019.
- [31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tatcenik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [33] Chong Cheng, Yu Hu, Sicheng Yu, Beizhen Zhao, Zijian Wang, and Hao Wang. Reggs: Unposed sparse views gaussian splatting with 3dgs registration, 2025. 2
- [34] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 2, 6
- [35] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 6
- [36] Chong Cheng, Zijian Wang, Sicheng Yu, Yu Hu, Nanjie Yao, and Hao Wang. Unposed 3dgs reconstruction with probabilistic procrustes mapping, 2025.
- [37] Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps, 2025.
- [38] Sicheng Yu, Chong Cheng, Yifan Zhou, Xiaojun Yang, and Hao Wang. Rgb-only gaussian splatting slam for unbounded outdoor scenes, 2025. 2
- [39] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 2
- [40] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16651–16662, 2025.
- [41] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *2024 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2024.
- [42] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.
- [43] Alex Yu, Vickie Ye, Matthew Tatcenik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.
- [44] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 2
- [45] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 2, 6
- [46] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. 3

- [47] Zichao Zhang, Guillermo Gallego, and Davide Scaramuzza. On the comparison of gauge freedom handling in optimization-based visual-inertial state estimation. *IEEE Robotics and Automation Letters*, 3(3):2710–2717, 2018. 3
- [48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. 3
- [49] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision (ECCV)*, 2020. 4
- [50] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zaus, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Buló, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kotschieder. Mapanything: Universal feed-forward metric 3d reconstruction, 2026. 4
- [51] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 6
- [52] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvln: Real-time understanding for infinite video streams, 2025. 4
- [53] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation, 2025. 5
- [54] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls, 2026. 4
- [55] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 6
- [56] Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods, 2025. 6
- [57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset, 2020. 6
- [58] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl(4) manifold, 2025. 6
- [59] Johann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 6
- [60] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *CVPR*, 2022. 6
- [61] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgb-d objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024. 6
- [62] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6
- [63] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 6
- [64] Manuel Lopez-Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Buló², Yubin Kuang, and Peter Kotschieder. Mapillary planet-scale depth dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6
- [65] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The replica dataset: A digital replica of indoor spaces, 2019. 6
- [66] Po-Han Huang, Kevin Matzen, Johannes Kopf, Nendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

- [67] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 6
- [68] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, October 2023. 6
- [69] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 6
- [70] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo, 2023. 6
- [71] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blended-mvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 6
- [72] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 6
- [73] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 6
- [74] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 6
- [75] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer, 2025. 6

LongStream: 长序列流式自回归视觉几何

补充材料

7. 相对位姿与尺度的规范不变性

本附录给出简要证明：(1) *LongStream* 使用的关键帧相对位姿严格不受全局坐标系选择影响；(2) 几何目标与尺度目标正交解耦。

7.1. 关键帧相对位姿的 $SE(3)$ 规范不变性

本文证明学习目标

$$\mathbf{T}_{i \leftarrow k} = \mathbf{T}_i \mathbf{T}_k^{-1}, \quad (13)$$

在任意全局 $SE(3)$ 规范变换下均保持不变，从而保证训练不受世界坐标系任意选择的影响。

规范变换。 令 $\mathbf{G} \in SE(3)$ 将世界坐标系 $\mathcal{W}$ 重新参数化为 $\mathcal{W}'$ 。对于任意三维点 $\mathbf{x}$ ：

$$\mathbf{x}_{\mathcal{W}'} = \mathbf{G} \mathbf{x}_{\mathcal{W}}. \quad (14)$$

绝对位姿的变换。 对于世界坐标系到相机坐标系的位姿 $\mathbf{T}$ ，其在 $\mathcal{W}'$ 中的对应位姿为

$$\mathbf{T}' = \mathbf{T} \mathbf{G}^{-1}. \quad (15)$$

该关系源于相机坐标系下的坐标必须保持不变。

关键帧相对位姿的不变性。 将式 (15) 应用于帧 i 和 k ：

$$\mathbf{T}'_i = \mathbf{T}_i \mathbf{G}^{-1}, \quad \mathbf{T}'_k = \mathbf{T}_k \mathbf{G}^{-1}. \quad (16)$$

则 $\mathcal{W}'$ 中的相对位姿为

$$\begin{aligned} \mathbf{T}'_{i \leftarrow k} &= \mathbf{T}'_i (\mathbf{T}'_k)^{-1} \\ &= (\mathbf{T}_i \mathbf{G}^{-1}) (\mathbf{T}_k \mathbf{G}^{-1})^{-1} \\ &= \mathbf{T}_i (\mathbf{G}^{-1} \mathbf{G}) \mathbf{T}_k^{-1} = \mathbf{T}_i \mathbf{T}_k^{-1}. \end{aligned} \quad (17)$$

因此，

$$\mathbf{T}'_{i \leftarrow k} = \mathbf{T}_{i \leftarrow k}, \quad (18)$$

这表明该目标严格满足 $SE(3)$ 规范不变性。

7.2. 尺度的 $Sim(3)$ 正交解耦

下面证明归一化几何目标与全局尺度因子无关，从而保证形状与尺度通过不同的梯度路径优化。

令预测的度量点云为

$$\hat{\mathbf{X}} = s \hat{\mathbf{X}}_{\text{raw}}, \quad (19)$$

其中， $s > 0$ 为尺度头预测的全局尺度。

令 $\text{Norm}(\cdot)$ 为一阶齐次函数：

$$\text{Norm}(\alpha \mathbf{X}) = \alpha \text{Norm}(\mathbf{X}). \quad (20)$$

几何损失中使用的归一化预测为

$$\tilde{\mathbf{X}}_{\text{pred}} = \frac{\hat{\mathbf{X}}}{\text{Norm}(\hat{\mathbf{X}})} = \frac{s \hat{\mathbf{X}}_{\text{raw}}}{s \text{Norm}(\hat{\mathbf{X}}_{\text{raw}})} = \frac{\hat{\mathbf{X}}_{\text{raw}}}{\text{Norm}(\hat{\mathbf{X}}_{\text{raw}})}. \quad (21)$$

因此，几何损失

$$\ell_{\text{geom}} = \|\tilde{\mathbf{X}}_{\text{pred}} - \tilde{\mathbf{X}}_{\text{gt}}\|_1 \quad (22)$$

与 s 无关，因而有

$$\frac{\partial \ell_{\text{geom}}}{\partial s} = 0. \quad (23)$$

这证实全局尺度与形状优化完全解耦，并且仅通过专用尺度目标学习。

综上，关键帧相对位姿具有严格的 $SE(3)$ 规范不变性，归一化几何则保证 $Sim(3)$ 尺度正交性。二者共同构成面向长序列流式重建、具有明确理论依据的规范一致训练目标。

8. 补充注意力可视化分析

如图 7 所示，我们通过可视化帧级注意力，分析模型在流式推理期间如何将注意力分配给历史帧。通过对目标帧词元求和并对源帧词元取平均，将词元间注意力聚合为 $S \times S$ 的帧间矩阵。因果全窗口视图最多包含 80 个可见帧，而滑动窗口视图仅包含 10 个，这一差异也反映在可视化结果中。

批训练基线表现出明显的时序偏差：模型对第一帧（即“汇聚点”）和更远处的帧分配了不成比例的高注意力，却对与局部几何一致性最相关的近期帧关注不足。直观而言，几何模型应主要依赖时间上相邻的帧；然而，这种不平衡会使 RPE 快速增长，并导致远距离预测不稳定。在窗口化推理中，保留汇聚点会加速性能退化，移除汇聚点则会导致崩溃，表明该基线严重依赖初始帧。

采用本文的 KV 缓存一致训练 (CCT) 后，注意力分布更加均衡。模型降低了对第一帧的依赖，并将更多的注意力分配给邻近帧，从而在全窗口与滑动窗口推理中均表现得更加稳定。不过，当序列长度接近 ~ 80 帧时，仍可观察到注意力逐渐转向更早的历史帧，这与缓存饱和效应一致。

总体而言，这些可视化揭示了长序列性能退化的内在机制：基线模型会形成强烈的首帧吸引与远距离偏差，而 CCT 则促使注意力模式更符合时序几何一致性。

9. 长序列稳定性分析

表 6 报告了 Waymo #520018670 (135 m) 和 KITTI #03 (561 m) 上的长序列结果。随着序列长度增加，

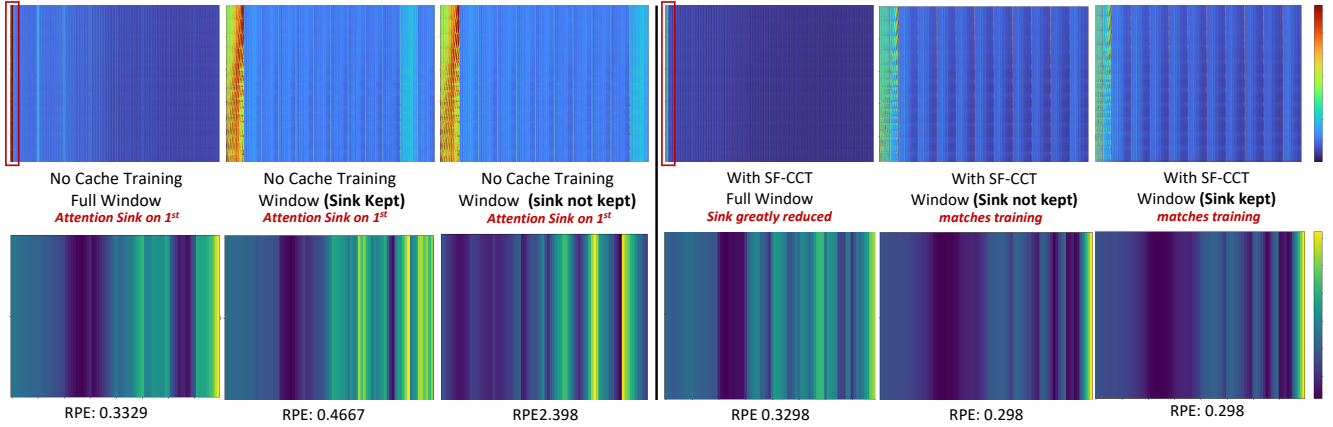


图 7. 缓存一致训练 (Cache-Consistent Training, CCT)。图中展示了不同训练—推理设置下的注意力图 (上) 和相对位姿误差 (Relative Pose Error, RPE) 热力图 (下)。不使用 CCT 时 (左), 因果推理会形成强烈的注意力汇聚点; 窗口化推理若保留该汇聚点, 会进一步将其放大, 若移除则会崩溃。使用 CCT 时 (右), 该汇聚点在因果模式下受到显著抑制, 在两种窗口模式下也同样受到抑制, 从而获得稳定且最佳的精度。浅蓝色表示对关键帧的注意力。

方法	15x	30x	60x	199x	30x	801x
CUT3R	0.159	1.421	2.505	8.591	3.867	148.06
TTT3R	0.153	0.873	1.127	5.505	1.957	105.28
Stream3R	0.181	1.329	2.998	21.440	3.412	158.25
VGGT-SLAM	0.172	0.518	0.992	3.740	0.929	169.83
本文方法 (w/o SW)	0.151	0.192	0.477	1.699	0.347	20.83
本文方法	0.151	0.183	0.343	0.723	0.164	3.81

表 6. 随着序列长度增加, Waymo #520018670 (左, 135 m) 和 KITTI #03 (右, 561 m) 上的 ATE (m)。w/o SW 表示不使用缓存刷新与滑动窗口的变体。

包括基线方法和未使用缓存刷新的 LongStream 在内的流式方法均表现出非线性误差增长。这表明在严格在线约束下, 更长的历史信息或重访不一定能提升性能, 反而可能随序列增长放大长时域效应。相比之下, LongStream 通过消除首帧锚定, 并采用带周期性缓存刷新的缓存一致训练来缓解长历史效应, 从而在长序列上保持稳定。

10. 补充超参数分析

本节给出详细的超参数消融研究。这些实验在 vKITTI 数据集上进行, 用于验证本文的设计选择。

10.1. 关键帧间隔的影响

我们首先考察模型对关键帧间隔 N 的敏感性。如表 7 所示, 采用 $N = 1$ 这类极短间隔会使系统退化为逐帧跟踪, 导致误差快速累积。相反, 将间隔增大至 15 同样会降低性能。这是因为训练块固定为 22 帧。关键帧切换过于稀疏时, 模型获得的监督信号太少, 无法可靠学习切换行为。

10.2. 缓存窗口大小的影响

我们进一步研究缓存窗口大小 W 的影响。如表 8 所示, 窗口大小为 10 已足以维持上下文, 而增至 30 会

间隔 N	ATE ↓	RPE ↓
1	4.047	0.565
3	3.384	0.514
8	0.122	0.131
10	0.115	0.126
15	1.398	0.412

表 7. 关键帧间隔的影响。 $N = 10$ 在漂移累积与训练动态之间取得最佳权衡。

窗口 W	ATE ↓	RPE ↓
10	0.115	0.126
20	0.119	0.129
30	0.516	0.293

表 8. 缓存窗口大小的影响。 $W = 10$ 可在维持充足上下文的同时防止几何饱和。

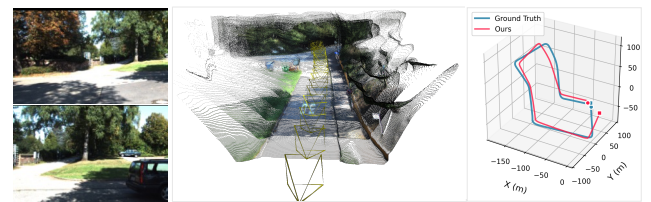


图 8. 在没有回环闭合校正时, LongStream 重访同一位置会出现轻微漂移。加入在线回环闭合线索是提升全局一致性的潜在方向。

显著降低精度, 使 ATE 升至 0.516。这一实验证据支持“几何饱和”理论: 过长的历史缓存会积累过时特征, 进而污染注意力机制。因此, 本文采用大小为 10 的窗口, 在防止长期漂移的同时降低计算成本。

11. 补充局限性

如图 8 所示，LongStream 未进行显式回环闭合优化，因此无法利用离线全局光束法平差所能实现的强轨迹校正。尽管本文提出的相对位姿形式和缓存一致训练已能在长时域内保持稳定的漂移特性，但引入轻量级在线回环闭合线索有望进一步提升全局一致性，尤其是在大型回环中。本文将其留作未来工作。