

LongStream: Long-Sequence Streaming Autoregressive Visual Geometry

Supplementary Material

7. Gauge Invariance of Relative Pose and Scale

This appendix provides concise proofs that (1) the keyframe–relative pose used in *LongStream* is strictly invariant to the choice of global coordinate frame, and (2) the geometry and scale objectives are orthogonally decoupled.

7.1. SE(3) Gauge Invariance of Keyframe Relative Pose

We show that our learning target

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

is invariant under any global $SE(3)$ gauge transformation. This guarantees that training is not affected by arbitrary choices of world coordinates.

Gauge transformation. Let $\mathbf{G} \in SE(3)$ re-parameterize the world frame \mathcal{W} into \mathcal{W}' . For any 3D point \mathbf{x} :

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

Transformation of absolute pose. For a world-to-camera pose \mathbf{T} , the corresponding pose in \mathcal{W}' is

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

This follows from enforcing that camera-frame coordinates remain unchanged.

Invariance of keyframe–relative pose. We apply Equation (15) to frames i and k :

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

Then the relative pose in \mathcal{W}' becomes

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

Thus,

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

showing that the target is strictly $SE(3)$ gauge-invariant.

7.2. Sim(3) Orthogonal Decoupling of Scale

We now show that our normalized geometry objective is independent of the global scale factor, ensuring that shape and scale are optimized through separate gradient paths.

Let the predicted metric point cloud be

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

where $s > 0$ is the global scale predicted by the scale head.

Let $\text{Norm}(\cdot)$ be homogeneous of degree one:

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

The normalized prediction used in the geometry loss is

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

Hence the geometry loss

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

is independent of s , and thus

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

This confirms that global scale is fully decoupled from shape optimization, and is learned solely through the dedicated scale objective.

In summary, keyframe–relative poses provide strict $SE(3)$ gauge invariance, while normalized geometry ensures $Sim(3)$ scale orthogonality. Together they yield a principled gauge-consistent training objective for long-sequence streaming reconstruction.

8. Additional Attention Visualization Analysis

As shown in Figure 7, we visualize *frame-level* attention to analyze how the model distributes focus over historical frames during streaming inference. Token–token attention is aggregated into an $S \times S$ frame–frame matrix by summing over target-frame tokens and averaging over source-frame tokens. The causal full-window view contains up to 80 visible frames, while the sliding-window view contains only 10, which is reflected in the visualization.

The batch-trained baseline exhibits a clear temporal bias: the model assigns disproportionately high attention to the first frame (the “sink”) and to more distant frames, while under-attending the recent frames that are most relevant for local geometric consistency. Intuitively, a geometry model

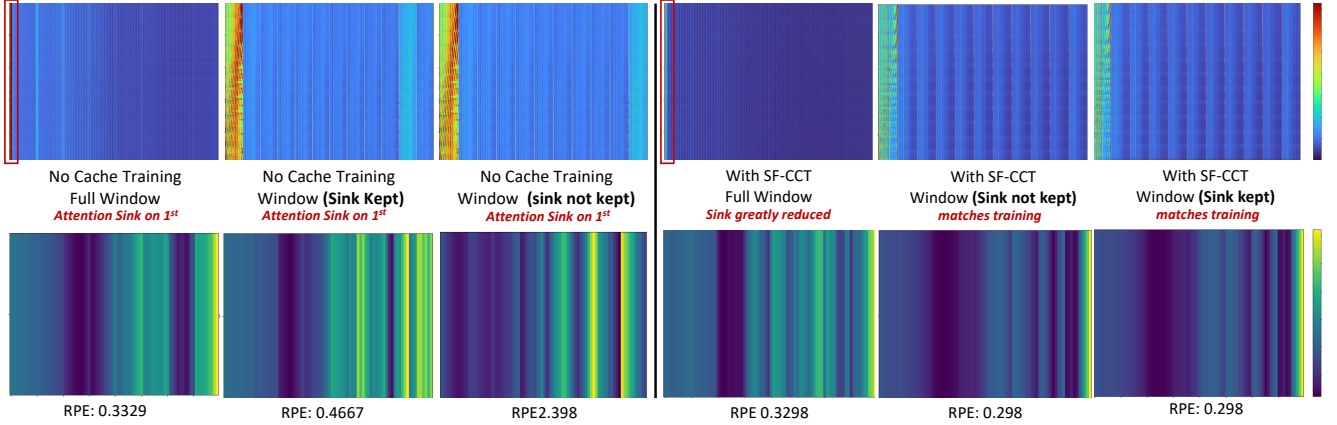


Figure 7. **Cache-consistent training (CCT)**. We show attention maps (top) and Relative Pose Error (RPE) heatmaps (bottom) under different training–inference settings. Without CCT (left), causal inference develops a strong attention sink; windowed inference either amplifies this sink when it is kept or collapses when it is removed. With CCT (right), the sink is strongly suppressed in causal mode and likewise suppressed in both windowed modes, yielding stable and best accuracy. Light blue denotes attention to the keyframe.

should primarily rely on temporally adjacent frames; however, this imbalance causes rapid growth in RPE and unstable long-range predictions. In windowed inference, retaining the sink yields accelerated degradation, whereas removing it leads to collapse, indicating that the baseline is strongly dependent on the initial frame.

With our cache-consistent KV-cache training (CCT), the attention distribution becomes more balanced. The model reduces its reliance on the first frame and allocates relatively more attention to nearby frames, resulting in more stable behavior across both full-window and sliding-window inference. Nonetheless, as sequence length approaches ~ 80 frames, we still observe a gradual shift of attention toward earlier history, consistent with cache saturation effects.

Overall, these visualizations highlight the underlying mechanism of long-sequence degradation: baseline models develop a strong first-frame attraction and long-range bias, while CCT encourages attention patterns that better align with temporal geometric coherence.

9. Long-Sequence Stability Analysis

Table 6 reports long-sequence results on Waymo #520018670 (135 m) and KITTI #03 (561 m). Streaming methods, including baselines and LongStream without refresh, exhibit non-linear error growth as sequence length increases. This suggests that under strictly-online constraints, longer histories or revisiting do not necessarily improve performance and may instead amplify long-horizon effects as the sequence grows. In contrast, LongStream remains stable over long sequences by removing first-frame anchoring and mitigating long-history effects through cache-consistent training with periodic cache refresh.

Method	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
Ours (w/o SW)	0.151	0.192	0.477	1.699	0.347	20.83
Ours	0.151	0.183	0.343	0.723	0.164	3.81

Table 6. ATE (m) as sequence length increases on Waymo #520018670 (Left, 135 m) and KITTI #03 (Right, 561 m). w/o SW denotes the variant without cache refresh and sliding window.

10. Additional Hyperparameter Analysis

In this section, we provide detailed ablation studies on hyperparameters. These experiments were conducted on the vKITTI dataset to validate our design choices.

10.1. Impact of Keyframe Interval

We first examine the sensitivity of the model to the keyframe interval N . As presented in Table 7, setting an extremely short interval such as $N = 1$ degenerates the system into frame-to-frame tracking, leading to rapid error accumulation. Conversely, extending the interval to 15 also degrades performance. This happens because the training chunk is fixed at 22 frames. With such sparse keyframe switches, the model receives too few supervision signals. It cannot reliably learn the switching behaviour.

10.2. Impact of Cache Window Size

We further investigate the influence of the cache window size W . As shown in Table 8, while a window size of 10 is sufficient to maintain context, increasing it to 30 significantly impairs accuracy with the ATE rising to 0.516. This empirical evidence supports our theory of “geometric sat-

Interval 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

Table 7. **Effect of Keyframe Interval.** $N = 10$ yields the best trade-off between drift accumulation and training dynamics.

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

Table 8. **Effect of Cache Window Size.** $W = 10$ prevents geometric saturation while maintaining sufficient context.

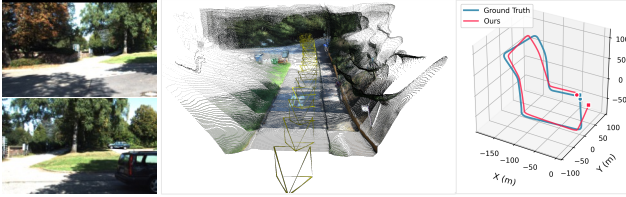


Figure 8. Without loop-closure correction, LongStream shows mild drift when revisiting the same place. Adding online loop-closure cues is a promising direction for improving global consistency.

uration” where an excessively long history cache accumulates outdated features that pollute the attention mechanism. Thus, a window size of 10 is adopted to minimize computational cost while preventing long-term drift.

11. Additional Limitation

As illustrated in Figure 8, LongStream does not perform explicit loop-closure optimization, and therefore does not benefit from the strong trajectory correction achievable in offline global bundle adjustment. While the proposed relative pose formulation and cache-consistent training already provide stable drift behavior over long horizons, incorporating lightweight online loop-closure cues may further improve global consistency, especially in large loops. We leave this as future work.