vehicles

MDPI

Article

Motion-Aware Geometric Context Adaptation for Streaming 3D Reconstruction of Intelligent Rail Vehicles in Low-Parallax Scenes

Peng Jiang [1], Fuyuan Wang [2,*], Zhiwei Chen [2,*] and Wenbo Pan [3,4,*]

[1] Hunan Engineering Research Center of Special Robot Control Technology and Equipment in Complex Environment, Hunan Vocational Institute of Technology, Xiangtan 411104, China; jiangpeng741@126.com
[2] CRRC Zhuzhou Institute Co., Ltd., 169 Shidai Road, Zhuzhou 412001, China
[3] School of Mechanical Engineering and Mechanics, Xiangtan University, Xiangtan 411105, China
[4] TENGEN INTELLIGENCE INSTITUTE, 899 Xianyuehuan Road, Zhuzhou 412000, China
* Correspondence: wangfy4@csrzic.com (F.W.); im.zwei.chen@gmail.com (Z.C.); panwb@mail.ustc.edu.cn (W.P.)

Abstract

Recent context-aware streaming 3D reconstruction frameworks provide a promising solution for online vehicle perception by maintaining anchor references, local pose windows, and trajectory memory. However, directly applying such frameworks to intelligent rail vehicles remains challenging because rail transit scenes are dominated by long straight motion, low-parallax visual observations, repetitive trackside structures, weak textures, and illumination variations. These characteristics may cause redundant context accumulation, unstable frame registration, and gradual trajectory drift. To address this problem, this paper proposes a motion-aware geometric context adaptation method for streaming 3D reconstruction of intelligent rail vehicles in low-parallax scenes. Instead of requiring task-specific large-scale retraining, the proposed method adapts the inference-stage geometric context using scale-normalized visual motion cues, including scale-normalized translational displacement, turning tendency, and inter-frame viewpoint variation. A motion-aware keyframe selection strategy suppresses redundant low-parallax frames while preserving geometrically informative observations in curved or pose-changing segments. An adaptive local pose reference window further regulates recent visual context to improve frame registration consistency. Experiments on rail transit sequences and the Oxford Spires dataset show that the proposed method achieves lower trajectory error than LingBot-Map and VIPE, while reducing redundant keyframe storage and preserving the qualitative continuity of rail-related structures. The method provides a practical motion-aware streaming 3D perception solution for rail transit inspection and digital infrastructure management.

Keywords: intelligent rail vehicles; streaming 3D reconstruction; motion-aware perception; low-parallax visual degeneration; deep learning

Check for updates

Academic Editors: Lie Yang and Xiangkun He

Received: 12 June 2026
Revised: 8 July 2026
Accepted: 15 July 2026
Published: 20 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Intelligent rail vehicles support rail inspection, digital maintenance, and safe operation. Their onboard cameras continuously capture rails, tunnels, platforms, and trackside facilities during operation [1]. Reconstructing 3D structures from these image streams is valuable for rail digital twins, clearance inspection, asset localization, and infrastructure condition assessment. Compared with offline reconstruction, streaming 3D reconstruction is more

suitable for onboard perception because it processes frames causally and incrementally without requiring the full sequence in advance [2].

Recent advances in visual reconstruction and learning-based geometry have greatly improved camera pose estimation, depth prediction, dense reconstruction, and online mapping. Classical structure-from-motion (SfM), simultaneous localization and mapping (SLAM), and visual–inertial methods provide effective geometric pipelines [3–5], while deep reconstruction models, neural scene representations, and Gaussian-based SLAM systems further improve dense geometry estimation and mapping quality [6–8]. Context-aware streaming reconstruction frameworks also show that structured historical contexts, such as anchor references, local pose windows, and compact trajectory memory, can improve long-sequence inference under bounded memory [2,7,9]. However, these methods are mainly developed for general visual scenes and do not explicitly address the constrained motion and low-parallax degeneration of rail transit environments.

Rail transit scenes differ substantially from general on-road driving scenarios. Rail vehicles often move along long straight tracks, producing consecutive frames with limited viewpoint changes. Such low-parallax observations weaken multi-view geometric constraints and make depth and pose estimation sensitive to small errors. Meanwhile, repetitive rails, sleepers, tunnel walls, cables, poles, and signs introduce ambiguous visual correspondences, while illumination changes, weak textures, reflections, and motion blur further increase reconstruction difficulty [4,10]. As a result, generic context accumulation may preserve many redundant frames in straight sections while still failing to provide stable geometric references in curved or pose-changing segments.

The key insight of this paper is that rail vehicle motion can serve as a useful prior for regulating streaming geometric context. Straight sections usually contain redundant low-parallax observations, whereas curved or pose-changing segments provide more informative viewpoint variation. Therefore, instead of treating all incoming frames uniformly, the reconstruction context should be adapted according to vehicle motion states. Based on this observation, this paper proposes a motion-aware geometric context adaptation method for streaming 3D reconstruction of intelligent rail vehicles in low-parallax scenes. The proposed method uses scale-normalized translational displacement, turning tendency, and inter-frame viewpoint variation to guide keyframe selection and local pose reference windows during inference, improving frame registration stability without task-specific large-scale retraining.

Classical visual odometry and SLAM systems commonly select keyframes using fixed translation, rotation, parallax, tracking-quality, or covisibility thresholds. Therefore, generic keyframe gating and local window adjustment are not claimed as methodological novelties in this work. The contribution instead lies in formulating rail-specific, scale-normalized motion cues and integrating them into the inference time geometric context management of a pretrained streaming reconstruction backbone.

The main contributions of this paper are as follows:

- A motion-aware geometric context adaptation framework is proposed for streaming 3D reconstruction of intelligent rail vehicles under low-parallax visual degeneration.
- A motion-aware keyframe selection strategy is designed to preserve geometrically informative frames and suppress redundant low-parallax observations.
- An adaptive local pose reference window is introduced to improve frame registration stability and reduce unnecessary context accumulation.
- An application-oriented evaluation protocol is adopted for rail transit scenarios, considering reconstruction stability, keyframe compression, trajectory consistency, and rail–structure continuity.

2. Related Work

Visual 3D reconstruction has been widely studied in computer vision, robotics, and intelligent vehicle perception [2,11,12]. Classical pipelines mainly include structure-from-motion, SLAM, visual–inertial odometry, and multi-view stereo. SfM estimates camera poses and sparse structures from unordered images, while SLAM and visual–inertial systems process sequential observations for online localization and mapping [3,10,13]. MVS further recovers dense geometry when camera poses are available [14]. These methods are geometrically interpretable, but they usually depend on reliable feature matching, sufficient parallax, and optimization-based pose refinement. Learning-based SLAM methods, such as DROID-SLAM, improve tracking and dense correspondence through neural optimization [15]. Neural scene representations, including NeRF and 3D Gaussian Splatting, have advanced dense reconstruction and rendering [16,17], and their SLAM extensions demonstrate the potential of neural and Gaussian representations for online mapping [18–20].

Recent 3D foundation models further change the paradigm of visual geometry estimation. DUSt3R directly regresses dense 3D point maps from unposed images [21], while VGGT infers camera parameters, depth maps, point maps, and 3D tracks in a feed-forward manner [6]. InfiniteVGGT extends such visual geometry inference to endless streams [7]. However, many feed-forward reconstruction methods are still designed for offline or short-sequence settings. In vehicle-mounted applications, input frames arrive continuously and must be processed causally. Recent context-aware streaming reconstruction frameworks improve long-sequence inference by organizing historical observations into anchor references, local pose windows, and compact trajectory memory [2,7]. Nevertheless, generic streaming context management does not explicitly consider the constrained motion and low-parallax degeneration of rail transit scenes.

Rail transit localization and mapping have traditionally relied on ground infrastructure or multi-sensor fusion. Balise-based query/response systems provide point-based positioning through ground beacons, but they depend on trackside equipment and cannot offer continuous self-localization when infrastructure fails [22–24]. LiDAR–inertial odometry, visual–inertial systems, and LiDAR–vision–GNSS-IMU fusion have been explored for onboard mapping and localization [25–27]. These methods can improve accuracy, but they may suffer from cumulative drift, GNSS unavailability, high sensor cost, or degraded performance in tunnels and repetitive feature-sparse environments. Backend graph optimization, ICP constraints, absolute pose constraints, and landmark constraints have been introduced to reduce drift in long degraded rail scenarios [1,28].

Different from infrastructure-dependent or heavy multi-sensor solutions, camera-based streaming 3D reconstruction offers a lightweight perception option for intelligent rail vehicles. However, two gaps remain in existing studies. First, current context-aware streaming reconstruction methods usually adopt generic context management policies and rarely exploit the constrained motion pattern of rail vehicles. In particular, long straight rail sections often generate redundant low-parallax frames, whereas curved sections, turnouts, and station areas provide more informative viewpoint changes. Second, existing railway localization and mapping systems mainly rely on infrastructure-assisted positioning, LiDAR, IMU, GNSS, or multi-sensor fusion, while motion-aware context regulation for a monocular visual-only learned streaming reconstruction backbone has received limited attention. These gaps motivate the proposed rail-specific inference time context adaptation strategy based on relative, scale-normalized motion cues.

3. Materials and Methods

This section presents the proposed motion-aware geometric context adaptation method. Section 3.1 introduces the overall framework. Section 3.2 describes the data materi-

als and preprocessing. Section 3.3 defines the rail vehicle motion state. Sections 3.4 and 3.5 present the motion-aware keyframe selection and adaptive local pose reference window, respectively. Section 3.6 summarizes the streaming inference pipeline, and Section 3.7 introduces the evaluation metrics.

3.1. System Overview

The overall framework is shown in Figure 1. Given a continuous video stream captured by an onboard camera, the system processes each incoming frame in a causal streaming manner. At time step t, the current image is denoted as I_t. A pretrained streaming 3D reconstruction backbone F_θ first extracts frame-level visual tokens through Frame Attention and then performs cross-frame reasoning through Geometric Context Attention. Based on the current image and the maintained geometric context, the backbone predicts the current depth map and camera pose:

$$(\hat{\boldsymbol{D}}_t, \hat{\boldsymbol{P}}_t) = F_\theta(I_t, C_t), \tag{1}$$

where $\hat{\boldsymbol{D}}_t$ is the predicted depth map, $\hat{\boldsymbol{P}}_t$ is the estimated camera pose, F_θ denotes the pretrained reconstruction backbone with parameters θ, and C_t is the geometric context available at time step t. The camera head produces pose estimates, while the depth head predicts the current depth map. The estimated poses are accumulated into a camera trajectory, and the predicted depth maps are back-projected to generate the reconstructed rail scene.

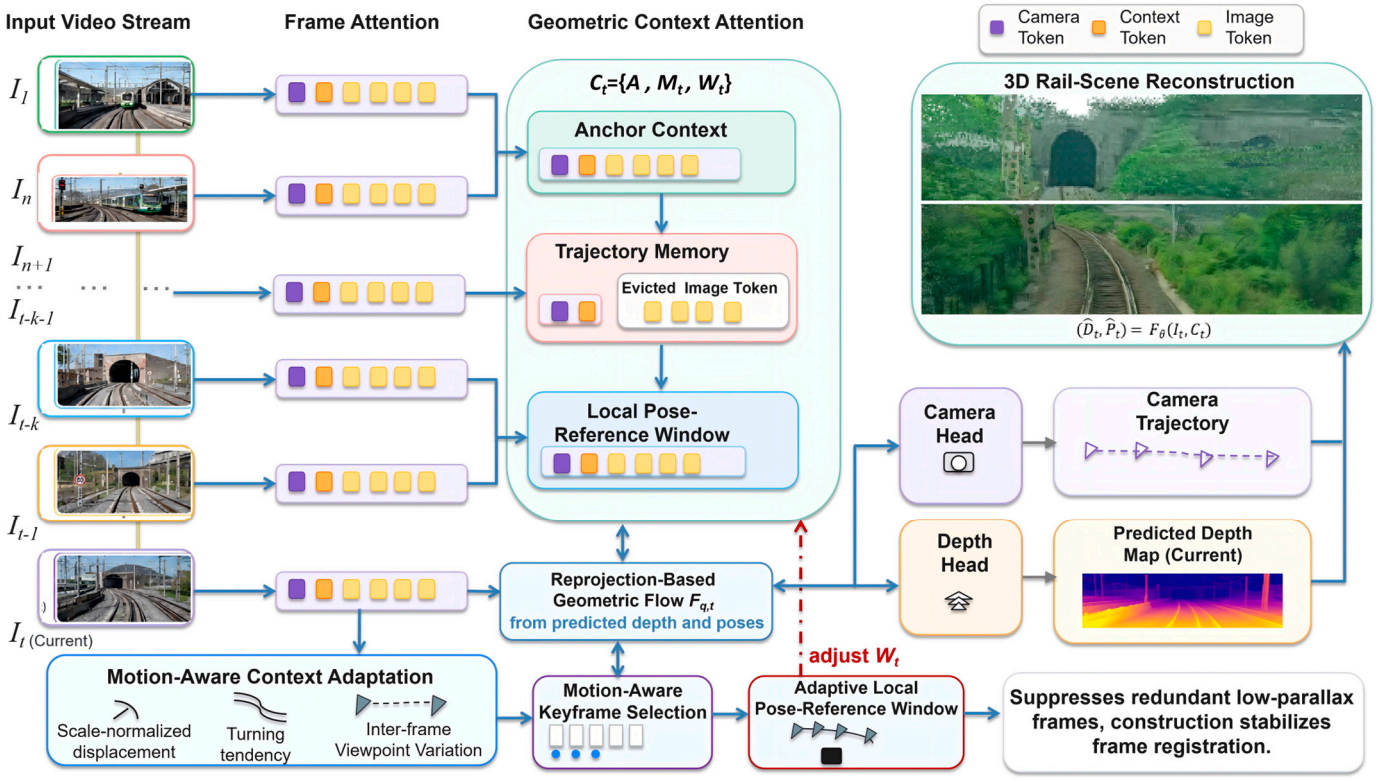


Figure 1. Overall framework of the proposed motion-aware geometric context adaptation method. The pretrained streaming reconstruction backbone processes incoming video frames through Frame Attention and Geometric Context Attention, and predicts pose and depth using task-specific heads. Vehicle motion cues regulate motion-aware keyframe selection and adaptive local pose reference window construction during inference, suppressing redundant low-parallax frames and preserving geometrically informative observations. Solid arrows indicate the main information flow, whereas the red dashed arrow denotes the motion-aware feedback used to update the geometric context. The remaining colors are used only for visual differentiation and carry no additional quantitative meaning.

In this study, the pretrained backbone F_θ is instantiated using LingBot-Map, the released implementation and model corresponding to the Geometric Context Transformer introduced in [2]. We use the officially released long-sequence checkpoint, lingbot-map-long.pt. The pretrained network architecture and weights are kept unchanged, and the proposed method modifies only the inference time context management policy without updating the model parameters.

Following the native LingBot-Map inference configuration, the first $N_s = 8$ input frames are jointly processed as initial scale frames and retained in the key–value (KV) cache before the remaining frames are processed sequentially in causal streaming mode. These frames support the model's internal relative-scale estimation rather than metric-scale initialization. No external navigation measurements or metric-scale calibration are used during this stage; metric scale is introduced only during offline trajectory evaluation through the global Sim(3) alignment described in Section 3.7. In the evaluated native configuration, LingBot-Map further maintains the recent geometric context through a bounded sliding KV cache, retains compact historical trajectory information, and applies an interval-based keyframe retention policy according to the input sequence length.

The geometric context C_t consists of three components: $C_t = \{A, M_t, W_t\}$, where A is the anchor context, M_t is the trajectory memory, and W_t is the local pose reference window. The anchor context provides a stable global reference, the trajectory memory stores compact historical information, and the local pose reference window preserves recent observations with dense visual overlap. Frames not retained in the active context are treated as evicted context. This structured design maintains long-range geometric consistency while avoiding the memory cost of full historical attention.

This work proposes a motion-aware adaptive strategy for geometric context. In rail transit scenes, straight sections often produce redundant low-parallax frames, whereas curved or pose-changing segments provide more informative geometric constraints. To exploit this characteristic, we explicitly incorporate the rail vehicle motion state m_t, which aggregates scale-normalized visual motion cues, including translational displacement, turning tendency, and inter-frame viewpoint variation (detailed in Section 3.3). These motion cues guide two inference-stage operations: motion-aware keyframe selection and adaptive local pose reference window construction. The former suppresses redundant low-parallax frames and determines whether the current frame should be retained as a reference, while the latter adjusts the size and composition of W_t to improve frame registration stability. Since only the inference time context update process is adapted, the proposed method operates on the pretrained LingBot-Map backbone without task-specific retraining or modification of the pretrained network parameters.

3.2. Materials and Data Acquisition

The experimental materials consist of rail transit image sequences, GNSS–IMU reference trajectories, and the Oxford Spires dataset for auxiliary validation. The rail transit data were collected using a forward-facing camera mounted on a rail vehicle. The recorded scenes include straight track sections, curved sections, tunnel entrances, platforms, poles, cables, signs, and other trackside facilities. These scenes contain typical visual challenges for rail perception, such as low-parallax observations, repetitive structures, weak textures, illumination variations, and motion blur.

Figure 2 shows the onboard data acquisition platforms used for rail transit data collection. A forward-facing camera was mounted on the rail vehicle to capture continuous image sequences, while a GNSS–IMU integrated navigation system provided reference trajectory data. In this study, the camera images were used as the input to the proposed reconstruction method, whereas the GNSS–IMU data were used only as a ground-truth

reference for evaluating camera pose estimation and trajectory consistency. During inference, the motion cues used for context adaptation were estimated from the online camera poses predicted by the streaming reconstruction backbone. Specifically, heading variation and inter-frame viewpoint variation were computed from consecutive estimated camera poses, and the turning tendency was approximated from the relative pose change between adjacent frames.

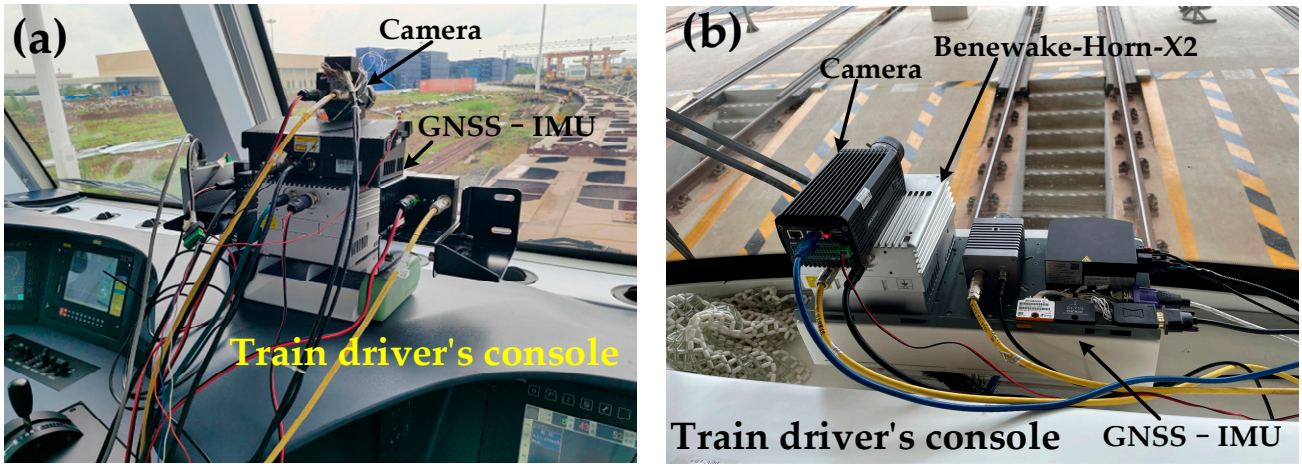


Figure 2. Onboard data acquisition platforms used for rail transit data collection: (**a**) locomotive-based platform; (**b**) urban rail platform. A forward-facing camera was used to capture image sequences. A GNSS–IMU integrated navigation system provided reference trajectory data in open-air sections, while post-processed navigation results were used as reference data in underground or GNSS-degraded sections.

The Oxford Spires dataset was additionally used for auxiliary validation under general long-sequence visual reconstruction conditions [29]. This dataset contains complex indoor–outdoor transitions, illumination changes, and challenging visual regions, which help verify the general behavior of the proposed inference-stage context adaptation strategy beyond rail transit scenes. However, the rail transit sequences were used as the primary evaluation data because they better represent constrained motion, repetitive structures, and low-parallax degeneration.

Before inference, the image sequences were temporally ordered and synchronized with the GNSS–IMU reference records for evaluation. Invalid frames with severe blur, overexposure, or incomplete recording were removed. The remaining images were resized to the input resolution required by the reconstruction backbone. All frames were processed sequentially without using future observations, ensuring a causal streaming evaluation setting.

3.3. Motion State Representation

The proposed framework dynamically adapts the geometric context according to the online motion state estimated from consecutive camera poses. During inference, the pose-derived motion state m_t is computed entirely from the camera poses predicted by the streaming reconstruction backbone, rather than from the GNSS-IMU reference trajectory, which is reserved strictly for evaluation. The keyframe selection score additionally uses a reprojection-based geometric flow cue derived from the predicted depth and poses, as described in Section 3.4. No external optical flow estimator is introduced. It should be noted that the predicted poses from a monocular streaming reconstruction backbone may not have a globally consistent metric scale. Therefore, the proposed method does not treat the translational term as an absolute physical velocity. Instead, it uses scale-normalized visual

motion cues to regulate context updating. These cues describe relative motion patterns within the current sequence and are used for keyframe selection and local reference window adaptation, rather than for safety-critical vehicle localization.

At time step t, the estimated camera pose is defined as $\hat{P}_t = \{\hat{R}_t, \hat{p}_t\}$, where $\hat{R}_t$ represents the camera rotation matrix and $\hat{p}_t$ denotes the camera position coordinates. The current vehicle motion state is characterized by a three-dimensional vector:

$$m_t = [\tilde{d}_t, \rho_t, \Delta\theta_t], \tag{2}$$

where $\tilde{d}_t$ denotes the scale-normalized inter-frame translational displacement, ρ_t represents the relative turning tendency, and $\Delta\theta_t$ signifies the inter-frame viewpoint variation. Together, these cues quantify the geometric informativeness of the current frame. In straight low-parallax sections, ρ_t and $\Delta\theta_t$ are usually small, even though the forward translational displacement may be large. Such forward-dominant motion often produces limited lateral parallax, making adjacent frames visually redundant. Conversely, in curved tracks or during rapid pose transitions, ρ_t and $\Delta\theta_t$ noticeably increase, indicating that the incoming frame introduces more informative geometric constraints for scene reconstruction.

The scale-normalized translational displacement is computed from consecutive camera position estimates as follows:

$$\tilde{d}_t = \|\hat{p}_t - \hat{p}_{t-1}\|_2 / (median(j \in \Omega_t)(\|\hat{p}_j - \hat{p}_{j-1}\|_2) + \varepsilon), \tag{3}$$

where Ω_t denotes a sliding temporal window containing up to the most recent $N_\Omega = 16$ inter–frame translational displacements, and ε is a small positive constant for numerical stability. This formulation converts frame-to-frame translation into a relative motion-intensity cue within the current sequence, thereby reducing the influence of monocular scale ambiguity. During initialization, before 16 displacement measurements are available, all currently available inter-frame displacements are used for normalization. Once the window is fully populated, only the most recent 16 displacement measurements are maintained.

The turning tendency is estimated from the heading change between adjacent frames:

$$\rho_t = |\Delta\psi_t|, \tag{4}$$

where $\Delta\psi_t$ denotes the estimated heading change derived from consecutive camera orientations. Unlike metric curvature, this turning tendency does not rely on absolute traveled distance. A larger ρ_t indicates that the vehicle is undergoing a more pronounced turning or orientation-changing motion, which usually corresponds to more informative viewpoint variation.

The inter-frame viewpoint variation is derived from the relative rotation between two consecutive camera orientations:

$$\Delta\theta_t = \arccos((trace(\hat{R}_{t-1}^T \hat{R}_t) - 1)/2), \tag{5}$$

where $trace(\cdot)$ denotes the matrix trace operator. This metric provides a direct angular measurement of the orientation shift between adjacent frames. Compared to utilizing translational displacement alone, $\Delta\theta_t$ captures variations in the viewing direction more sensitively, which is crucial for identifying geometrically informative keyframes in inherently forward-moving, low-parallax rail scenarios.

The derived motion state m_t subsequently governs two downstream modules: In Section 3.4, it guides the motion-aware keyframe selection mechanism to filter out redundant low-parallax frames. In Section 3.5, it dynamically modulates the adaptive local pose

reference window to maintain stable and accurate frame registration across diverse rail motion profiles.

3.4. Motion-Aware Keyframe Selection

In streaming 3D reconstruction, the keyframe selection strategy directly affects both reconstruction stability and context efficiency. In this work, a keyframe refers to a frame retained in the streaming geometric context or KV cache as a future reference. Non-keyframes are still processed to produce current depth and pose predictions, but they are not persistently stored as reference frames. Standard interval-based strategies tend to retain redundant frames in straight, low-parallax rail sections, whereas purely appearance-based methods may become unstable when repetitive track structures or weak textures cause unreliable visual matching. To mitigate these issues, we propose a motion-aware keyframe selection strategy that jointly considers reprojection-based geometric flow and scale-normalized motion cues. The objective is to preserve geometrically informative frames while systematically suppressing redundant low-parallax observations.

Unlike classical keyframe selection policies based on fixed translation or rotation thresholds, the proposed criterion does not rely on a single unnormalized motion magnitude. Instead, it combines scale-normalized displacement with reprojection-based geometric flow, turning tendency, viewpoint variation, and a redundancy-aware interaction term. This formulation is intended to distinguish redundant low-parallax motion from geometrically informative viewpoint changes in rail transit sequences.

For each incoming frame I_t, the system computes a dynamic keyframe score S_t by combining the reprojection-based geometric flow cue with the pose-derived motion cues defined in Section 3.3:

$$S_t = \alpha_1 F_{q,t} + \alpha_2 \rho_t + \alpha_3 \Delta\theta_t + \alpha_4 \eta_t, \tag{6}$$

where S_t denotes the keyframe score of the current frame, and $F_{q,t}$ denotes the reprojection-based geometric flow cue between the current frame I_t and the most recently retained keyframe I_q. Instead of being estimated by an external optical flow network, $F_{q,t}$ is computed directly from the depth and camera poses predicted by the pretrained reconstruction backbone. Specifically, valid pixels in the current predicted depth map are back-projected into 3D, transformed into the coordinate system of the keyframe using the predicted relative pose, and then reprojected onto the keyframe image plane. $F_{q,t}$ is calculated as the mean Euclidean displacement between the original and reprojected pixel coordinates over the valid pixels. Before evaluating S_t the computed $F_{q,t}$, ρ_t, $\Delta\theta_t$, and η_t are independently normalized to comparable ranges. For notational simplicity, the same symbols are retained for their normalized values in Equation (6). The terms ρ_t and $\Delta\theta_t$ represent the relative turning tendency and inter-frame viewpoint variation, respectively, while η_t denotes a redundancy-aware motion factor based on the scale-normalized translational displacement. The hyperparameters α_1, α_2, α_3, and α_4 control the relative contributions of reprojection-based geometric flow, turning tendency, viewpoint variation, and redundancy-aware motion, respectively.

Distinct from general driving environments, a large forward displacement in rail scenes does not inherently imply high geometric informativeness. In long straight sections, the rail vehicle may move forward along the tracks, yet the viewing direction remains almost unchanged, yielding limited lateral parallax. Consequently, the translational cue is not directly used as an absolute velocity to increase the keyframe score. Instead, it is formulated as a scale-normalized redundancy-aware factor:

$$\eta_t = \tilde{d}_t \cdot (\rho_t + \Delta\theta_t). \tag{7}$$

In Equation (7), ρ_t and $\Delta\theta_t$ denote the original pose-derived quantities defined in Equations (4) and (5). After η_t is computed, it is normalized together with the other score components before Equation (6) is evaluated. By implication, the scale-normalized translational displacement contributes to the keyframe score mainly when accompanied by turning tendency or inter-frame viewpoint variation. When the vehicle traverses straight tracks with minimal ρ_t and $\Delta\theta_t$, η_t remains suppressed even if the forward displacement is relatively large, effectively preventing the over-insertion of redundant straight-section frames into the active context.

The current frame is formally designated as a keyframe if its score exceeds a predefined threshold:

$$I_t \in \mathcal{K}, \text{ if } S_t > \tau_k, \tag{8}$$

where $\mathcal{K}$ denotes the active keyframe set, and τ_k is the selection threshold. If this condition is not satisfied, the frame is still forwarded through the pipeline to estimate its depth and pose, but it is not inserted into the keyframe set or retained as an active local reference. In this way, redundant low-parallax frames are processed for current-frame reconstruction while being prevented from increasing the reference context in subsequent inference.

This selection mechanism offers two distinctive advantages tailored to rail transit environments. First, along extensive straight segments, it minimizes unnecessary keyframe insertions, preventing computational and memory bloat within the geometric context. Second, across curved tracks, turnouts, station areas, or sudden pose-changing segments, the elevated turning tendency and orientation variations naturally elevate the keyframe score. This allows the system to adaptively increase the density of reference frames where complex geometry demands tighter registration constraints. Consequently, the resulting geometric context remains highly compact, structurally adaptive, and explicitly aligned with the motion characteristics of rail vehicles.

3.5. Adaptive Local Pose Reference Window

Following the keyframe selection phase, the retained reference frames are utilized to construct the local pose reference window. In general streaming reconstruction frameworks, camera reference context is usually managed by generic causal or fixed-window rules, without explicitly considering rail vehicle motion states. However, a static window footprint is ill-suited for the unique spatial dynamics of rail transit scenes. In long straight sections, adjacent frames offer limited lateral parallax, necessitating a prolonged temporal window to accumulate sufficient geometric constraints and stabilize camera pose registration. It should be noted that a larger window in straight sections does not mean inserting more new keyframes; instead, it reuses a longer range of already selected reference frames to stabilize registration. Conversely, curved tracks, turnouts, station areas, or rapid pose-changing segments inherently present much tighter geometric bounds, where a highly compact window suffices for accurate localization while mitigating unnecessary context accumulation. To exploit this operational variance, we introduce an adaptive local pose-reference window whose capacity is dynamically regulated by the vehicle motion state estimated in Section 3.3.

Let W_t denote the local pose-reference window at time step t. It comprises the most recent k_t reference frames selected from the active keyframe set $\mathcal{K}$:

$$W_t = \{the\ kt\ most\ recent\ keyframes\ in\ \mathcal{K}\ before\ time\ step\ t\}, \tag{9}$$

where k_t represents the adaptive window size at time step t. Specifically,k_t denotes the number of retained keyframes used as local pose references, rather than the number of consecutive input frames. Instead of employing a constant window length, k_t is modulated

as a function of the turning tendency and viewpoint variation of the rail vehicle. The continuous formulation is defined as follows:

$$k_t = k_{min} + (k_{max} - k_{min}) \cdot 2[1 - \sigma(\beta_1 \rho_t + \beta_2 \Delta\theta_t)], \tag{10}$$

where k_{min} and k_{max} denote the lower and upper bounds of the window size, respectively. The function $\sigma(\cdot)$ represents the standard sigmoid operator. Since ρ_t and $\Delta\theta_t$ are normalized non-negative motion cues, their sum is used as a normalized motion variation cue for adaptive window regulation. The factor $2[1 - \sigma(\beta_1 \rho_t + \beta_2 \Delta\theta_t)]$ decreases as the motion variation increases. The computed window size is rounded to an integer and constrained within the predefined range $[k_{min}, k_{max}]$ before constructing the local pose reference window. This formulation is consistent with the motion characteristics of rail transit. When the vehicle traverses a straight, low-parallax track segment, both ρ_t and $\Delta\theta_t$ remain small, resulting in a larger local pose reference window. In this scenario, the system reuses a longer range of already selected reference frames to provide extended temporal support for stable frame-to-frame registration. When the vehicle enters a curved track or undergoes a pronounced viewpoint transition, ρ_t and $\Delta\theta_t$ increase, reducing the local pose reference window accordingly. Consequently, the system relies more on immediate recent observations in geometrically informative pose-changing segments.

To reduce the risk of drift-induced context updates, the proposed strategy adopts bounded context regulation. The local pose reference window is always constrained within the predefined range $[k_{min}, k_{max}]$, and a minimum number of recent reference keyframes is preserved even when the estimated motion state indicates rapid pose changes. In addition, the anchor context is maintained throughout streaming inference, preventing temporary pose noise from completely removing stable historical references. Therefore, occasional errors in the online visual pose estimates are prevented from causing unbounded window compression or complete removal of the reference context, although they may still influence subsequent context updates. It should be noted that this mechanism forms a causal one-step feedback process rather than an instantaneous recursive loop. At time step t, the backbone first predicts the current pose and depth from C_t, after which the estimated motion cues are used to construct C_{t+1} for subsequent inference. The predefined window bounds limit the magnitude of context changes, while persistent anchor retention prevents temporary pose errors from completely removing the historical reference. These constraints reduce the risk of aggressive context updates, although they do not provide a formal guarantee against error propagation under severe or persistent pose estimation failures.

The adaptive window directly shapes the local geometric context fed into the Geometric Context Attention module. Compared to a rigid, fixed-size baseline, the proposed adaptive strategy improves context efficiency by preventing reference saturation in geometrically rich segments, while providing robust temporal support in visually degenerate straight sections. Consequently, the local pose reference window conforms to the motion characteristics of rail vehicles and supports more accurate and stable streaming 3D reconstruction. The overall motion-aware context update mechanism, including keyframe selection, adaptive local pose reference window construction, and bounded context regulation, is illustrated in Figure 3.

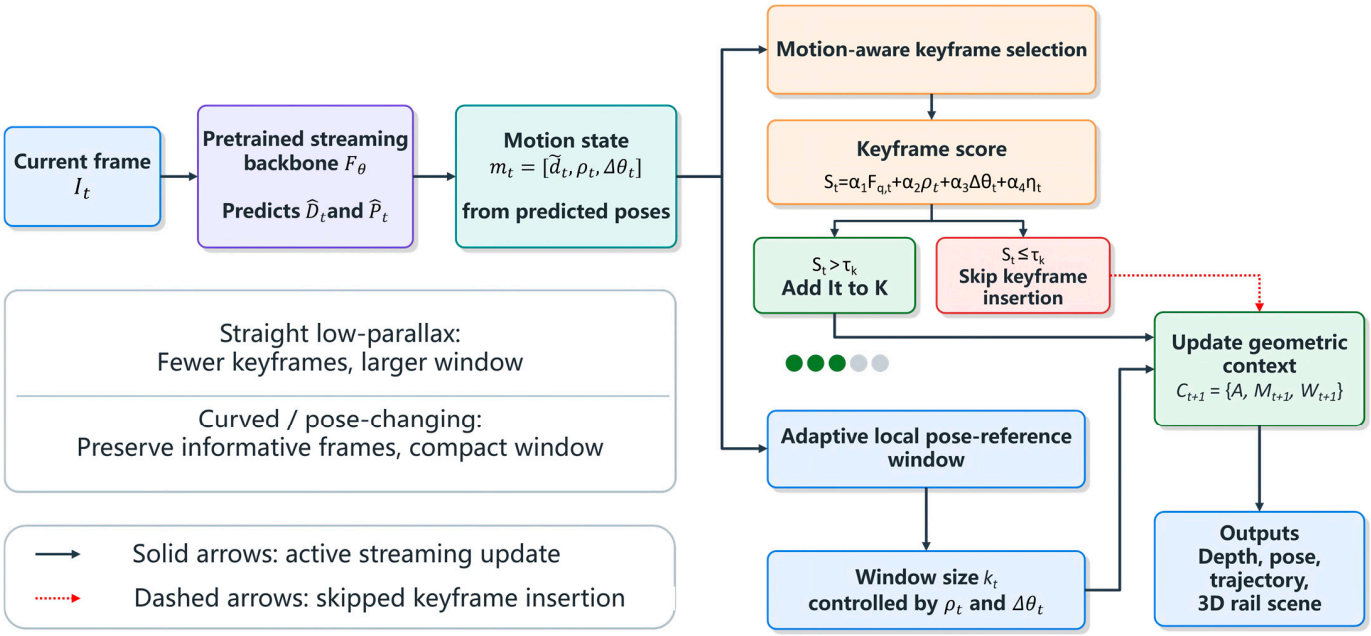


Figure 3. Motion-aware context update mechanism during streaming inference. The current frame is processed by the pretrained streaming backbone to predict depth and camera pose. The predicted pose is used to estimate the motion state, which guides motion-aware keyframe selection and adaptive local pose reference window construction.

3.6. Streaming Inference Pipeline

The proposed framework operates in a strictly causal streaming manner, processing the current frame I_t at each time step t using only the maintained geometric context C_t. The inference procedure consists of six primary steps. First, the pretrained backbone processes I_t and C_t to predict the current depth map D_t and camera pose P_t. Second, the scale-normalized motion state m_t is estimated from consecutive predicted poses, including scale-normalized translational displacement, turning tendency, and inter-frame viewpoint variation. Third, the motion-aware keyframe selection module computes a score S_t; if $S_t > \tau_k$, I_t is added to the active keyframe set $\mathcal{K}$ as a future reference. Fourth, the local pose reference window W_t is adaptively updated based on m_t to balance temporal support and efficiency. These components are then integrated to synthesize the geometric context C_{t+1} for the next iteration. Finally, predicted poses and depth maps are accumulated to reconstruct the camera trajectory and the global 3D rail scene. The overall procedure is summarized as follows:

- Step 1: Input the current frame I_t and geometric context C_t into the pretrained backbone.
- Step 2: Predict the current depth map $\hat{\boldsymbol{D}}_t$ and camera pose $\hat{\boldsymbol{P}}_t$.
- Step 3: Estimate the motion state m_t from consecutive predicted poses.
- Step 4: Compute the keyframe score S_t and update the active keyframe set $\mathcal{K}$.
- Step 5: Adapt the local pose reference window and obtain W_{t+1} according to the bounded motion-aware update rule.
- Step 6: Update the geometric context C_{t+1} and output the reconstructed depth, pose, trajectory, and 3D rail structure.

3.7. Evaluation Metrics

The evaluation framework focuses on trajectory accuracy, context efficiency, runtime performance, and visual reconstruction stability. All metrics are selected to be directly measurable from the predicted poses, selected keyframes, runtime logs, and final reconstruction results.

First, camera trajectory accuracy is evaluated against the GNSS–IMU reference trajectory. After temporal association, each estimated monocular trajectory is aligned with the corresponding metric reference trajectory using a global Sim(3) similarity transformation estimated by the Umeyama method, including one global scale factor, rotation, and translation. The Absolute Trajectory Error (ATE) is then reported as the Root Mean Square Error (RMSE):

$$ATE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\hat{p}_i^{\text{aligned}} - p_{ref,i}\|_2^2}, \tag{11}$$

where N denotes the number of temporally associated frames, $\hat{p}_i^{\text{aligned}}$ is the estimated camera position after global Sim(3) alignment, and $p_{ref,i}$ is the corresponding metric reference position obtained from the GNSS–IMU or post-processed navigation trajectory. The same alignment protocol is applied independently to all evaluated methods, without frame-wise or segment-wise rescaling. The reference trajectory is used only for offline evaluation.

Second, keyframe efficiency is quantified by the keyframe ratio:

$$R_{kf} = N_{kf}/N, \tag{12}$$

where N_{kf} is the count of selected keyframes and N is the total number of input frames. A lower R_{kf} indicates that the system effectively filters out more redundant frames, thereby maintaining a leaner active context.

Third, runtime efficiency is evaluated using the average inference throughput in frames per second under identical input and hardware settings. Context management efficiency is characterized by the keyframe ratio and the size of the local pose reference window.

Finally, reconstruction stability is assessed via qualitative visual inspection of the reconstructed rail scene. We specifically examine the geometric continuity of critical rail infrastructures, including rails, tunnel boundaries, platforms, catenary poles, and trackside structures. This qualitative evaluation is adopted because complete, sequence-aligned, high-fidelity 3D ground-truth models are not available for the collected large-scale rail transit sequences. Therefore, dense-reconstruction metrics such as point-to-point error, Chamfer distance, and F-score are not reported in this study, because using incomplete or non-uniform 3D references may lead to misleading quantitative conclusions. Instead, this study mainly reports trajectory accuracy, keyframe efficiency, runtime performance, and qualitative reconstruction stability, while quantitative dense-reconstruction evaluation will be further investigated when complete reference models become available.

4. Results

4.1. Experimental Setup

The proposed method was evaluated on five sequences, including two rail station-yard sequences, two rail mainline sequences, and one sequence from the Oxford Spires dataset [29]. The two additional rail sequences were included to evaluate whether the observed improvements remain consistent across different recordings of the same scene types. For comparison, LingBot-Map, the released implementation of the Geometric Context Transformer [2], and VIPE [12] were included as baseline methods. The LingBot-Map baseline represents the pretrained LingBot-Map framework with the officially released lingbot-map-long.pt checkpoint operating under its native inference and context management policy. It therefore serves as a controlled same-backbone baseline for isolating the contribution of the proposed inference time motion-aware context adaptation. VIPE is included as an independent optimization-based visual pose estimation method.

The same image sequences, temporally associated frames, and evaluation protocol were used for all methods. For trajectory evaluation, each estimated trajectory was in-

dependently aligned with the corresponding metric reference trajectory using a single global Sim(3) similarity transformation estimated by the Umeyama method, including one global scale factor, rotation, and translation. No frame-wise or segment-wise rescaling was applied. For the rail transit data, GNSS–IMU measurements or post-processed navigation results were used as metric reference trajectories only for offline evaluation and were not involved in the inference process. Under fixed inputs, model weights, and inference configurations, the evaluated inference pipeline contains no intentional stochastic sampling. The evaluation focused on visual reconstruction quality, trajectory consistency, and Absolute Trajectory Error (ATE). The principal tunable parameters introduced by the proposed method are summarized in Table 1.

Table 1. Principal hyperparameter settings of the proposed motion-aware context-adaptation method.

Parameter	Description	Value
α_1	Reprojection-based geometric flow weight	0.40
α_2	Turning tendency weight	0.20
α_3	Viewpoint-variation weight	0.25
α_4	Redundancy-aware motion weight	0.15
τ_k	Keyframe selection threshold	0.50
k_{min}	Minimum local pose reference window	16
k_{max}	Maximum local pose reference window	64

4.2. Qualitative Streaming Reconstruction Results

Figure 4 shows the qualitative reconstruction results of the proposed method under the monocular visual-only streaming setting used throughout this study. Given a continuous video stream, the method estimates camera poses incrementally and reconstructs the surrounding rail scene in a streaming manner. The reconstructed point cloud preserves the main spatial structures of the rail environment, including rails, trackside facilities, and surrounding infrastructure. The camera trajectory is also visually consistent with the direction of vehicle motion, indicating coherent pose estimation in this qualitative example.

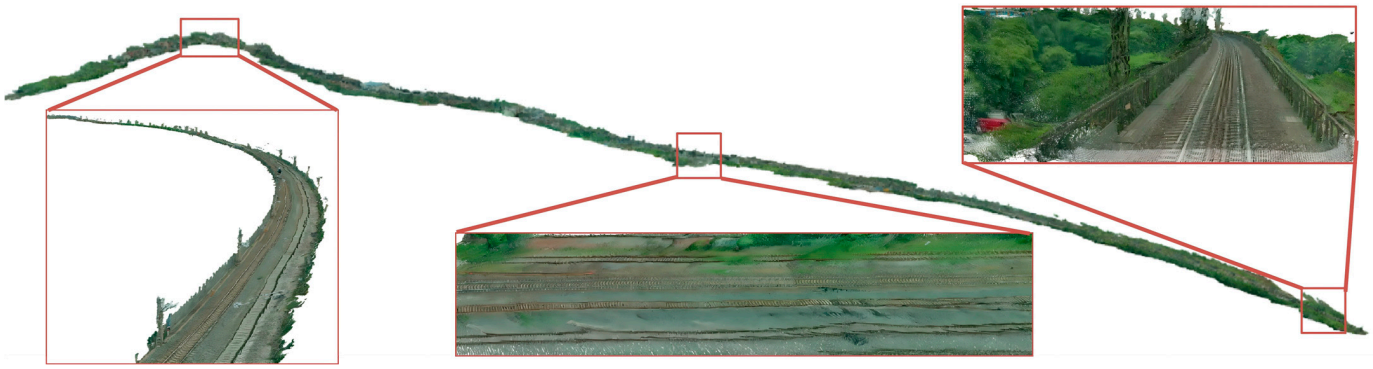


Figure 4. Qualitative rail scene reconstruction result under the monocular visual-only streaming setting.

Although no quantitative point cloud metric is reported in this visualization; the result demonstrates that the proposed method can produce visually coherent rail scene reconstruction from monocular video streams. This supports the feasibility of applying streaming 3D reconstruction to rail vehicle perception and infrastructure inspection.

4.3. Trajectory Preservation in Rail Transit Scenes

Figure 5 compares the estimated trajectories of LingBot-Map, VIPE, and the proposed method in a rail transit station yard scene. This sequence includes track switches and curved rail segments before entering the station area. Such regions are challenging because visual observations contain repetitive rail structures, frequent geometric transitions, and limited stable texture. Compared with the baseline methods, the proposed method better preserves the trajectory shape through the turnout and station yard region. The trajectory remains more consistent with the rail path, suggesting that motion-aware keyframe selection and adaptive local pose reference window construction help retain informative frames in curved and pose-changing segments.

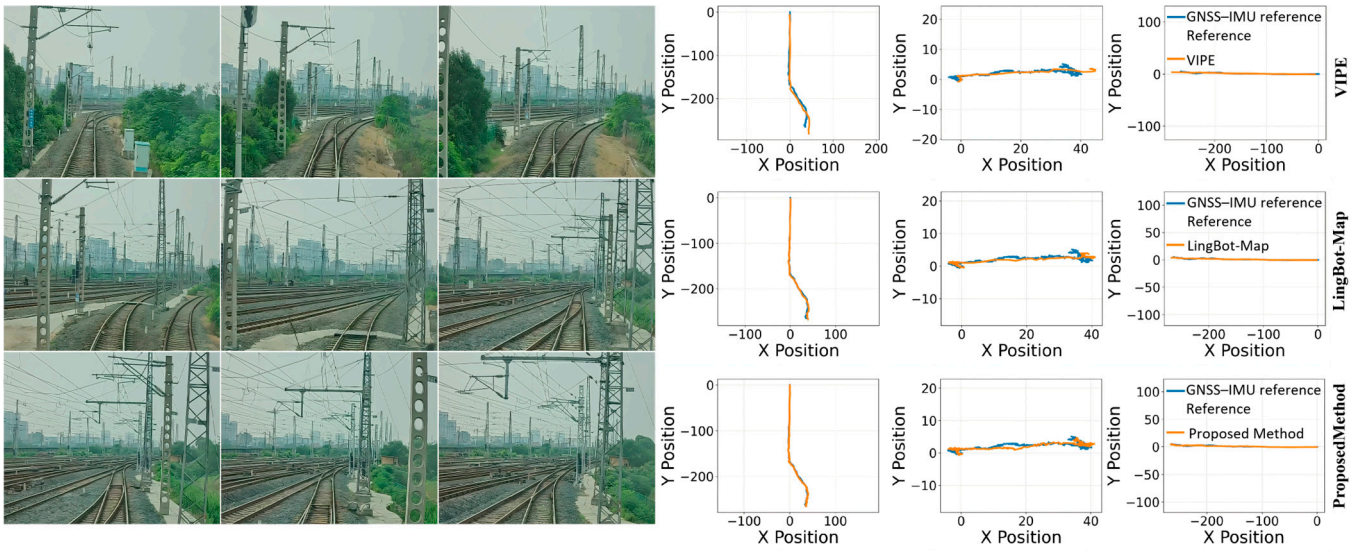


Figure 5. Representative trajectory comparison on a rail transit station yard sequence with turnouts. The estimated trajectories are globally aligned to and plotted together with the GNSS–IMU reference trajectory. For each row, the line colors and styles shown in the legend of the third column apply to all three trajectory projections.

Figure 6 further evaluates the methods on a rail mainline sequence. Compared with station-yard scenes, mainline rail transit data contain longer straight segments and stronger low-parallax degeneration. In this case, adjacent frames often provide redundant forward-looking observations, which may lead to unnecessary context accumulation and gradual trajectory deviation. The proposed method maintains a more stable trajectory along the mainline sequence. This result indicates that suppressing redundant low-parallax frames while preserving suitable temporal context is beneficial for long straight rail scenes.

4.4. Generalization on Oxford-Spires

To further evaluate the general applicability of the proposed context adaptation strategy, experiments were also conducted on the Oxford Spires dataset [29]. Figure 7 shows the trajectory comparison among LingBot-Map, VIPE, and the proposed method. Oxford Spires contains complex indoor–outdoor transitions and challenging dark stair regions, which are different from rail transit scenes but similarly require robust long-sequence trajectory preservation.

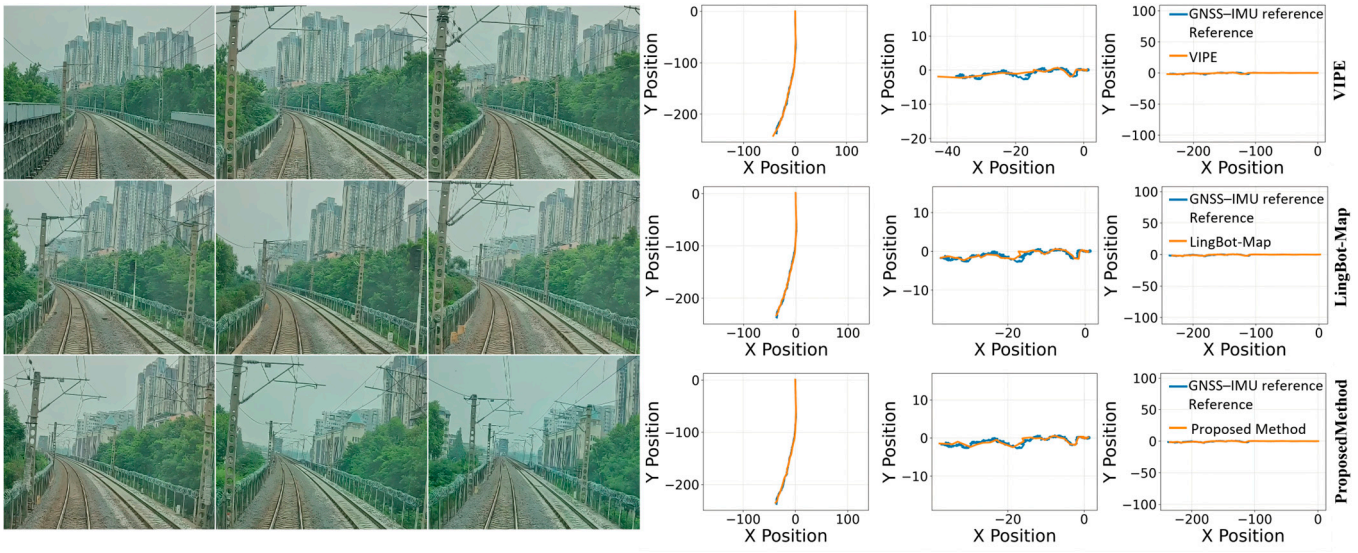


Figure 6. Representative trajectory comparison on a rail mainline sequence. The estimated trajectories are globally aligned to and plotted together with the GNSS–IMU reference trajectory. For each row, the line colors and styles shown in the legend of the third column apply to all three trajectory projections.

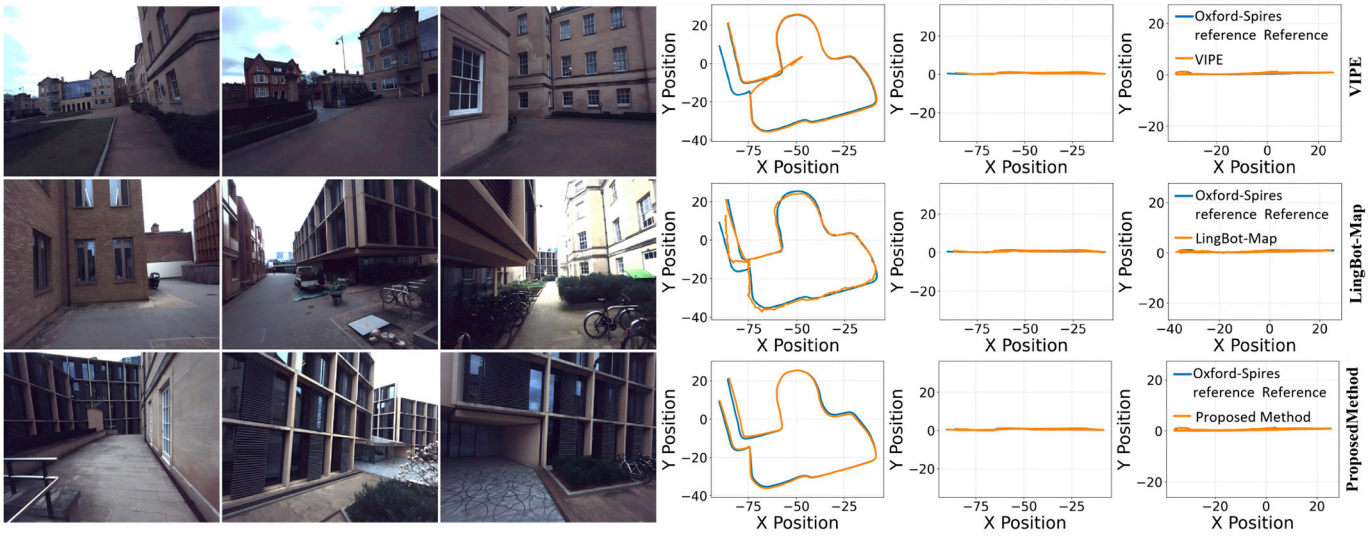


Figure 7. Trajectory comparison on the Oxford–Spires dataset. The top, middle, and bottom rows show the results of LingBot-Map [2], VIPE [12], and the proposed method, respectively. In each row, the estimated trajectory is plotted together with the Oxford–Spires reference trajectory in the XY, XZ, and YZ planes. The line colors and styles shown in the legend of the third column apply to all three trajectory projections.

The proposed method preserves the trajectory more accurately in the complex transition regions, indicating that the motion-aware context adaptation does not only benefit rail transit data but also improves robustness in general long-sequence reconstruction scenarios. This suggests that regulating keyframe insertion and local window construction according to estimated motion states can improve streaming reconstruction stability beyond the specific rail domain.

4.5. Quantitative Trajectory Evaluation

To quantitatively evaluate trajectory accuracy, each estimated trajectory was temporally associated with the corresponding reference trajectory and aligned using a global Sim(3) similarity transformation before computing the ATE RMSE. Table 2 reports the tra-

jectory errors of LingBot-Map [2], VIPE [12], and the proposed method on five evaluation sequences, including two rail station yard sequences, two rail mainline sequences, and one Oxford Spires sequence. The table also reports the average inference throughput measured under the same input and hardware settings.

Table 2. Quantitative comparison of trajectory accuracy and inference throughput on the five evaluated sequences.

Dataset/Sequence	LingBot-Map [2]	VIPE [12]	Proposed Method
Station-yard 1, ATE (m)	6.02 m	9.92 m	2.76 m
Station-yard 2, ATE (m)	5.79 m	9.53 m	2.51 m
Mainline 1, ATE (m)	2.82 m	3.86 m	1.34 m
Mainline 2, ATE (m)	3.16 m	4.29 m	1.55 m
Oxford-Spires, ATE (m)	6.51 m	10.62 m	5.93 m
Mean ATE (m)	4.86 m	7.64 m	2.82 m
Average inference throughput (FPS)	20.26	4.33	20.05

Note: All methods were evaluated using the same five input sequences, an image resolution of 640 × 480, and a single NVIDIA RTX 5090 GPU. Mean ATE denotes the arithmetic mean across the five sequences, and inference throughput is averaged over the same evaluation set.

As shown in Table 2, the proposed method achieves the lowest ATE on each of the five evaluated sequences. Its mean ATE is 2.82 m, compared with 4.86 m for LingBot-Map and 7.64 m for VIPE, corresponding to reductions of approximately 42.0% and 63.1%, respectively. The results on the two newly added rail sequences are consistent with those obtained on the original station yard and mainline sequences, indicating that the observed improvement is not limited to a single recording from each rail scene type.

Under the same input resolution and hardware settings, the proposed method achieves an average inference throughput of 20.05 FPS, compared with 20.26 FPS for LingBot-Map and 4.33 FPS for VIPE. The throughput of the proposed method is therefore comparable to that of the native LingBot-Map baseline, with a difference of approximately 1.04%, while remaining substantially higher than that of VIPE. These results indicate that the proposed motion-aware context adaptation introduces only limited computational overhead while preserving real-time streaming capability.

Figures 5–7 present representative qualitative trajectory comparisons for the station-yard, mainline, and Oxford Spires sequences, respectively, while the complete quantitative results for all five evaluated sequences are reported in Table 2.

4.6. Ablation Study

To evaluate the contribution of the proposed parameterized components, an ablation study was conducted using the average results over all evaluated sequences. Four settings were compared: the baseline method without motion-aware adaptation, Variant A with only motion-aware keyframe selection, Variant B with only adaptive pose reference window construction, and the full proposed method with both components enabled.

As shown in Table 3, the baseline method obtains an average ATE of 4.86 m with an average keyframe ratio of 29.2%. When only motion-aware keyframe selection is enabled, the ATE decreases to 3.09 m, and the average keyframe ratio is reduced to 26.5%. This result indicates that the keyframe score weights and the keyframe selection threshold help suppress redundant low-parallax frames while preserving geometrically informative observations. When only the adaptive pose reference window is enabled, the ATE decreases from 4.86 m to 4.01 m, while the keyframe ratio remains unchanged at 29.2%. This confirms

that the adaptive window mainly improves local reference selection and frame registration consistency rather than changing the number of retained keyframes.

Table 3. Ablation study of the proposed motion-aware context adaptation.

Method	Motion-Aware Keyframe Selection	Adaptive Pose Reference Window	ATE RMSE	Average Keyframe Ratio
Baseline	×	×	4.86 m	29.2%
Variant A	√	×	3.09 m	26.5%
Variant B	×	√	4.01 m	29.2%
Proposed Method	√	√	2.82 m	26.5%

Note: Average keyframe ratio denotes the proportion of frames retained as future reference frames after the initial scale frames. The enabled modules use the same parameter settings as summarized in Table 1. The non-adaptive setting uses a fixed local reference window, whereas the adaptive setting updates the local pose reference window according to the estimated motion state.

The full proposed method achieves the best overall performance, with an average ATE of 2.82 m and an average keyframe ratio of 26.5%. Compared with the baseline, it reduces the average ATE by approximately 42.0% and decreases the keyframe ratio by 2.7 percentage points, corresponding to a relative reduction of approximately 9.2%. These results demonstrate that the two parameterized components are complementary; motion-aware keyframe selection improves context compactness, while the adaptive pose reference window further enhances trajectory consistency by regulating the local reference context according to the estimated motion state.

To further characterize the operational behavior of the adaptive local pose reference window, the rounded window sizes were divided into three regimes. The small-window regime was defined as $k_t \leq k_{min} + 0.2(k_{max} - k_{min})$, the large-window regime was defined as $k_t \geq k_{min} + 0.8(k_{max} - k_{min})$, and the remaining cases were treated as the intermediate-window regime. With $k_{min} = 16$ and $k_{max} = 64$, these thresholds correspond approximately to $k_t \leq 26$ and $k_t \geq 55$, respectively. In the evaluated sequences, the small-window regime occurred in 6.89% of the context updates, the large-window regime occurred in 62.80%, and the intermediate-window regime occurred in 30.31%. The frequent use of the large-window regime is consistent with the long straight low-parallax sections in the rail sequences, where a longer local reference window is beneficial for registration stability. The anchor context was retained throughout all inference steps by design.

5. Discussion

The experimental results show that motion-aware geometric context adaptation improves trajectory preservation in rail transit scenes. Compared with general road-driving or indoor scenes, rail vehicle motion is highly constrained by the track geometry. Long straight sections usually generate forward-dominant motion with limited lateral parallax, while turnouts, station areas, and curved sections introduce more informative viewpoint changes. This observation explains why treating all incoming frames equally is not optimal for rail-oriented streaming reconstruction. By using the estimated motion state to regulate keyframe selection and local pose reference windows, the proposed method suppresses redundant low-parallax frames and preserves more useful geometric observations.

The proposed strategy is complementary to recent context-aware streaming reconstruction frameworks. Geometric context modeling, such as anchor context, local pose reference windows, and trajectory memory, provides an effective basis for long-sequence online reconstruction. However, such context structures are usually designed for general scenes and do not explicitly consider the motion regularity of rail vehicles. In this work, the geometric context is not reconstructed from scratch or retrained with large-scale rail-specific data. Instead, the context update process is adapted during inference according to

scale-normalized translational displacement, turning tendency, and inter-frame viewpoint variation. Moreover, bounded context regulation is adopted to constrain the local reference window and preserve stable historical references, reducing the risk that occasional pose noise causes aggressive context truncation. This makes the method lightweight and suitable for practical deployment where large annotated rail datasets are difficult to obtain.

The quantitative trajectory results further support this interpretation. The proposed method achieves lower ATE than LingBot-Map and VIPE on both rail station yard and mainline sequences. The improvement is more evident in rail scenarios than in Oxford Spires, suggesting that the proposed adaptation is particularly beneficial when the motion pattern is strongly constrained and low-parallax degeneration is frequent. On Oxford Spires, the improvement is smaller but still positive, indicating that the method can also provide robustness in general long-sequence reconstruction with complex transitions.

The visual reconstruction results also demonstrate the practical value of the method. In the rail scene reconstruction visualization, the reconstructed point cloud preserves the main spatial structures of the scene, and the estimated camera trajectory remains consistent with the vehicle motion direction. Although complete high-fidelity 3D ground truth is difficult to obtain for rail transit environments, visual continuity of rails, platforms, and trackside facilities provides useful application-level evidence for infrastructure inspection and digital maintenance.

Several limitations remain, mainly due to the monocular visual-only nature of the proposed streaming reconstruction framework. First, the current evaluation mainly focuses on trajectory accuracy, context efficiency, and visual reconstruction quality, while dense 3D reconstruction metrics are not reported due to the lack of complete ground-truth point clouds. This evaluation focus is consistent with the practical constraints of long-distance streaming rail perception, where complete global 3D reference models are usually unavailable and the online system must prioritize trajectory stability, compact context maintenance, and continuous structural perception. Second, the motion-aware parameters and thresholds are empirically selected, and a more systematic parameter sensitivity analysis could further strengthen the evaluation. Third, the proposed motion-aware adaptation relies on online visual pose estimates and therefore inherits the limitations of monocular streaming reconstruction. In particular, the translational cue used in this work is not treated as an absolute metric velocity, but as a sequence-normalized relative motion cue for context regulation. Therefore, the method is designed to improve inference-stage context management rather than to replace safety-critical localization modules. Severe illumination changes, image blur, reflective surfaces, highly textureless tunnel regions, or drift in the visual backbone may still affect pose prediction and context updating. In practical rail applications, IMU, LiDAR, or GNSS–IMU measurements can be integrated to provide a more reliable metric motion prior and further reduce the risk of drift-induced context updates. Further parameter sensitivity analysis, quantitative dense-reconstruction evaluation with complete reference models, multi-sensor extension, and comparison with railway-specific SLAM methods under matched sensor configurations will be conducted in future work.

Overall, the results indicate that rail vehicle motion provides an effective prior for regulating streaming geometric context. The proposed method offers a lightweight and practical extension to context-aware streaming 3D reconstruction, improving trajectory consistency in low-parallax rail transit scenes in the evaluated visual-only setting, without introducing additional navigation inputs or test-time parameter optimization.

6. Conclusions

This paper proposed a motion-aware geometric context adaptation method for streaming 3D reconstruction of intelligent rail vehicles in low-parallax scenes. The method

uses online estimated motion states to guide keyframe selection and adaptive local pose-reference window construction, improving context efficiency without task-specific large-scale retraining.

Experiments on rail transit sequences and Oxford-Spires showed that the proposed method achieves lower trajectory error than LingBot-Map and VIPE, with an average ATE of 2.82 m. Visual results also demonstrate improved camera-trajectory consistency relative to the evaluated baselines and coherent rail scene reconstruction in station yard, mainline, and complex transition scenarios. These results indicate that rail vehicle motion is a useful prior for practical online 3D reconstruction, supporting applications such as rail inspection, infrastructure perception, and digital maintenance.

Author Contributions: Conceptualization, P.J. and W.P.; methodology, P.J., Z.C. and W.P.; software, P.J. and Z.C.; validation, P.J. and F.W.; formal analysis, P.J. and F.W.; investigation, P.J. and F.W.; resources, W.P.; data curation, P.J. and F.W.; writing—original draft preparation, P.J.; writing—review and editing, Z.C., F.W. and W.P.; visualization, P.J. and F.W.; supervision, W.P.; project administration, W.P.; funding acquisition, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Hunan Provincial Natural Science Foundation of China under Grant Nos. 2025JJ50726 and 2025JJ50735.

Data Availability Statement: The public dataset analyzed in this study is openly available from the source cited in the manuscript. The rail transit datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Fuyuan Wang and Zhiwei Chen were employed by the company CRRC Zhuzhou Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, C.; Pan, W.; Yuan, X.; Huang, W.; Yuan, C.; Wang, Q.; Wang, F. High-precision map construction in degraded long tunnel environments of urban subways. *Remote Sens.* **2024**, *16*, 809. [CrossRef]
2. Chen, L.Z.; Gao, J.; Chen, Y.; Cheng, K.L.; Sun, Y.; Hu, L.; Xue, N.; Zhu, X.; Shen, Y.; Yao, Y.; et al. Geometric context transformer for streaming 3d reconstruction. *arXiv* **2026**, arXiv:2604.14141. [CrossRef]
3. Schonberger, J.L.; Frahm, J.-M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
4. Campos, C.; Elvira, R.; Rodriguez, J.J.G.; Montiel, J.M.M.; Tardos, J.D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
5. Fan, X.; Pan, W. Elevated high-precision mapping and localization technology for periodic inspections. *Sci. Rep.* **2025**, *15*, 4810. [CrossRef] [PubMed]
6. 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, Nashville, TN, USA, 10–17 June 2025. [CrossRef]
7. Yuan, S.; Yang, Y.; Yang, X.; Zhang, X.; Zhao, Z.; Zhang, L.; Zhang, Z. InfiniteVGGT: Visual Geometry Grounded Transformer for Endless Streams. *arXiv* **2026**, arXiv:2601.02281. [CrossRef]
8. Pan, W.; Chen, Z.; Fan, X. Gaussian Splatting SLAM for Enhanced Monocular Vehicle Sensor Localization and Roadside Scene Reconstruction. *IEEE Sens. Lett.* **2025**, *9*, 1–4. [CrossRef]
9. Huang, H.; Li, L.; Cheng, H.; Yeung, S.-K. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024. [CrossRef]
10. Qin, T.; Li, P.; Shen, S. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
11. Pan, W.; Fan, X.; Li, H.; He, K. Long-Range Perception System for Road Boundaries and Objects Detection in Trains. *Remote Sens.* **2023**, *15*, 3473. [CrossRef]

12. 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* **2025**, arXiv:2508.10934. [CrossRef]
13. Temmen, T.; Debougnoux, J.; Li, L.; Krautwig, B.; Brinkmann, T.; Eisenbarth, M.; Andert, J. 3D Environment Generation from Sparse Inputs for Automated Driving Function Development. *Vehicles* **2026**, *8*, 47. [CrossRef]
14. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1362–1376. [CrossRef] [PubMed]
15. Teed, Z.; Deng, J. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16558–16569. [CrossRef]
16. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
17. Kerbl, B.; Kopanas, G.; Leimkuehler, T.; Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [CrossRef]
18. Fan, X.; Chen, Z.; Pan, W. A monocular Gaussian splatting SLAM framework for cultural heritage digitization. *npj Herit. Sci.* **2026**. [CrossRef]
19. Keetha, N.; Karhade, J.; Jatavallabhula, K.M.; Yang, G.; Scherer, S.; Ramanan, D.; Luiten, J. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 21357–21366. [CrossRef]
20. Matsuki, H.; Murai, R.; Kelly, P.H.J.; Davison, A.J. Gaussian splatting slam. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024. [CrossRef]
21. 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, Seattle, WA, USA, 16–22 June 2024. [CrossRef]
22. Wu, Y.; Weng, J.; Tang, Z.; Li, X.; Deng, R.H. Vulnerabilities, attacks, and countermeasures in balise-based train control systems. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 814–823. [CrossRef]
23. Elbahhar, F.; Marc, H. Advanced train positioning/communication system. In *Modern Railway Engineering*; IntechOpen: London, UK, 2018; pp. 107–129. [CrossRef]
24. Fikejz, J.; Kavička, A. RegioRail—GNSS train-positioning system for automatic indications of crisis traffic situations on regional rail lines. *Appl. Sci.* **2022**, *12*, 5797. [CrossRef]
25. Wang, Y.; Song, W.; Wang, Y.; Dai, X.; Lou, Y. Metroloc: Metro vehicle mapping and localization with lidar-camera-inertial integration. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 1441–1453. [CrossRef]
26. Wang, Y.; Song, W.; Lou, Y.; Zhang, Y.; Huang, F.; Tu, Z.; Liang, Q. Rail vehicle localization and mapping with LiDAR-vision-inertial-GNSS fusion. *IEEE Robot. Autom. Lett.* **2022**, *7*, 9818–9825. [CrossRef]
27. Wang, Y.; Song, W.; Lou, Y.; Huang, F.; Tu, Z.; Zhang, S. Simultaneous localization of rail vehicles and mapping of environment with multiple LiDARs. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8186–8193. [CrossRef]
28. Cai, Y.; Qian, W.; Dong, J.; Zhao, J.; Wang, K.; Shen, T. A LiDAR–Inertial SLAM method based on virtual inertial navigation system. *Electronics* **2023**, *12*, 2639. [CrossRef]
29. Tao, Y.; Muñoz-Bañón, M.Á.; Zhang, L.; Wang, J.; Fu, L.F.T.; Fallon, M. The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods. *Int. J. Robot. Res.* **2026**, *45*, 839–857. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.