CMLGF-LIO: A Cross-Modal Local-Global Fusion Framework for Robust LiDAR-Inertial Odometry

Xin Yang [1], Li Yan [1], Hong Xie [1]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan, China - xinyang@whu.edu.cn, lyan@sgg.whu.edu.cn, hxie@sgg.whu.edu.cn

Keywords: Deep Learning, LiDAR-Inertial Odometry, Cross-Modal Fusion, Local-Global Interaction, Robust Localization.

Abstract

Accurate and robust localization is essential for autonomous vehicles and mobile robots operating in complex, dynamic environments. However, existing learning-based LiDAR-inertial odometry (LIO) methods typically rely on simplistic weighted fusion or purely global attention, which often fail to fully exploit cross-modal complementarity. In this paper, we propose CMLGF-LIO, a cross-modal local-global fusion framework that enhances both the accuracy and robustness of LIO. At the local level, we design a Local Split-Attention (LSA) module that injects IMU-derived motion priors into local LiDAR feature groups and adaptively allocates attention weights, suppressing redundant information while preserving discriminative local geometry for fine-grained fusion. At the global level, we introduce a Global MLP-Mixer (GMM) module that aligns LiDAR and IMU token sequences and models global cross-modal interactions using an MLP-Mixer backbone. Experiments demonstrate that CMLGF-LIO is more robust than learning-based baselines under challenging conditions, and ablation studies validate the effectiveness of the proposed local-global fusion strategy.

1. Introduction

Odometry serves as a cornerstone for robotic navigation and autonomous driving, providing fundamental ego-motion estimation for downstream localization and mapping. Early odometry systems typically relied on a single sensing modality, such as LiDAR odometry (LO), visual odometry (VO), or inertial odometry (IO) (Lee et al., 2024; Fan et al., 2025; Dhaigude and Kulkarni, 2025). However, single-modality solutions suffer from intrinsic limitations when deployed in complex real-world environments. Compared with cameras, LiDAR provides more stable geometric measurements, yet LiDAR-only odometry is prone to performance degradation in scenes with weak or ambiguous geometric constraints, such as long corridors, open areas, and repetitive structures, where insufficient observability can induce drift (Lin et al., 2024). In contrast, an inertial measurement unit (IMU) offers high-rate motion information but suffers from noise and bias accumulation. Consequently, multi-sensor fusion, particularly LIO, has emerged as a mainstream paradigm for robust and accurate state estimation (Chen et al., 2024; Gao et al., 2025; Hoang and Kim, 2026; Yang et al., 2026).

Existing LIO approaches can be broadly categorized into model-based methods and learning-based methods. Model-based LIO systems generally follow a standard pipeline comprising feature extraction, data association, and state estimation, where the state is inferred using filtering or nonlinear optimization. For instance, FAST-LIO2 (Xu et al., 2022) adopts a tightly-coupled iterative extended Kalman filter (IEKF) to directly register point clouds to a local map, achieving competitive performance across multiple datasets. LIO-SAM (Shan et al., 2020) formulates LIO as a factor-graph optimization problem and jointly optimizes LiDAR, IMU, and potentially additional measurements such as global navigation satellite system (GNSS). Despite their success, model-based methods often rely heavily on handcrafted geometric representations and precise cross-modal data association. In challenging scenarios such as environmental degradation, increased sensor noise, and dynamic objects, failures in feature extraction and matching can propagate through the estimation pipeline, severely degrading pose accuracy and potentially causing tracking failure.

In recent years, deep learning has introduced an alternative paradigm for LIO by enabling data-driven learning of robust feature representations, thereby reducing reliance on explicit modeling and manual parameter tuning. Current learning-based LIO methods can be broadly divided into two primary types: simple weighted fusion and attention-based fusion. The first family typically employs separate feature encoders for LiDAR and IMU, followed by late fusion via concatenation or element-wise weighting. For example, Son et al. (2021) employ a convolutional neural network (CNN) to process LiDAR data and a long short-term memory (LSTM) network (Yu et al., 2019) to process IMU sequences, but their fusion is mainly limited to feature concatenation before pose regression, which constrains deep cross-modal interaction. DeepLIO (Iwaszczuk and Roth, 2021) projects point clouds into a 2D representation and introduces a soft fusion mechanism that learns modality weights for element-wise feature reweighting, demonstrating the potential of global fusion. UnDeepLIO (Tu and Xie, 2022) leverages IMU preintegration as an initial pose prior and learns residual corrections in an unsupervised manner. Nevertheless, these approaches commonly perform fusion at the global feature-vector or channel-wise level and often overlook that LiDAR measurements exhibit spatially varying reliability across local regions. As a result, locally corrupted or uninformative features may dominate the aggregated representation, leading to biased pose estimation and significant errors.

To address these limitations, attention-based fusion approaches leverage more sophisticated cross-modal interactions and adaptive weighting to achieve finer-grained multimodal fusion. TransFusionOdom (Sun et al., 2023) highlights the limitations of naive weighting and constructs an end-to-end Transformer-based fusion framework, where multi-head attention is used to

model both intra-modal and inter-modal interactions. To improve robustness under multi-sensor degradation, A2DO (Lai et al., 2025) adopts multi-scale encoding and attention mechanisms to dynamically mitigate the impact of sensor failures, and further enhances generalization via large-scale simulation pretraining. Although attention mechanisms are powerful, their computational overhead and heavy reliance on large-scale training data may hinder real-time deployment on resource-constrained platforms and limit cross-domain generalization.

In summary, existing learning-based LIO methods still face a twofold challenge in fusion design: insufficient local cross-modal interaction and limited global consistency modeling, thereby hindering full exploitation of the complementarity between LiDAR geometry and IMU motion priors. To this end, we propose CMLGF-LIO, a cross-modal local-global fusion framework for robust learning-based LIO. The main contributions of this work are as follows:

(1) We develop an LSA module that jointly encodes local LiDAR geometric structure and IMU motion priors, which learns to allocate attention weights in a confidence-aware manner to enable fine-grained cross-modal interaction and adaptive compensation while suppressing feature redundancy.

(2) We design a GMM module based on an MLP-Mixer backbone to explicitly align and globally integrate multimodal features, effectively capturing long-range cross-modal dependencies while maintaining computational efficiency.

(3) Experiments on the KITTI dataset demonstrate that, under challenging conditions such as aggressive motion and sensor degradation, CMLGF-LIO achieves superior localization accuracy and robustness compared to representative LIO baselines.

2. Method

2.1 Overall Architecture

Figure 1 presents the overall architecture of the proposed CMLGF-LIO framework. At time step t, the inputs consist of a LiDAR point cloud $p_t = \left\{ \mathrm{P}_{t,i} \in \mathbb{R}^3 \right\}_{i=1}^{N_t}$ and synchronized IMU measurements $\mathrm{I}_t = \left\{ \left[\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z, \boldsymbol{\omega}_x, \boldsymbol{\omega}_y, \boldsymbol{\omega}_z \right]^{\mathrm{T}} \in \mathbb{R}^6 \right\}_{j=1}^{M_t}$, where $\mathbf{a}$ and $\boldsymbol{\omega}$ denote the triaxial specific force and angular velocity, respectively. Given these multimodal measurements, the network estimates the relative six-degree-of-freedom (6-DoF) motion between consecutive frames, denoted as $\mathbf{T}_{t-1\to t} \in SE(3)$. The rotational and translational components are parameterized by an Euler-angle vector $\boldsymbol{\varphi}_t \in \mathbb{R}^3$ and a translation vector $\boldsymbol{\rho}_t \in \mathbb{R}^3$, respectively.

LiDAR feature extraction. Following UnDeepLIO, the raw point cloud is projected into a vertex map and a normal map, which preserve local geometric structure while enabling efficient 2D convolutional processing. These two representations are then fed into a ResNet-18 encoder (He et al., 2016) to extract the LiDAR feature $\mathbf{F}_L$.

IMU feature extraction. For the inertial branch, we follow TransFusionOdom to construct temporal feature maps from the normalized IMU sequence. An LSTM network together with fully connected layers is employed to encode motion-related temporal cues, yielding an inertial feature $\mathbf{F}_I$ that complements the LiDAR geometric representation.

Cross-modal local-global fusion. To promote effective interaction between LiDAR geometry and inertial motion cues, we design a local-global fusion module consisting of an LSA unit and a GMM Module. LSA performs confidence-aware local cross-modal recalibration within grouped channel subspaces, while GMM further models global dependencies and aggregates long-range cross-modal interactions. The resulting fused representation is denoted as $\mathbf{F}_{fusion}$. Detailed descriptions of these two modules are provided in Sections 2.2 and 2.3, respectively.

Pose inference. In the final estimation stage, the fused feature of the current frame is combined with the previous hidden state and fed into an LSTM to capture temporal dependencies and motion continuity. We employ a two-layer LSTM with a hidden dimension of 1024. The network then regresses the rotation and translation through a fully connected layer:

$$\left(\boldsymbol{\varphi}_t, \boldsymbol{\rho}_t, \mathbf{h}_t\right) = \mathrm{FC}\left(\mathrm{LSTM}\left(\mathbf{F}_{fusion}^t, \mathbf{h}_{t-1}\right)\right) \tag{1}$$

where $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$ denote the hidden states at time steps t and $t-1$, respectively, and $\mathrm{FC}(\cdot)$ denotes a fully connected layer.

2.2 Local Split-Attention Module

To enable fine-grained cross-modal fusion while preserving intra-modal structural consistency, we introduce an LSA module. As shown in Figure 2, the term “local” refers to channel groups obtained by splitting feature channels, rather than spatial neighborhoods. This design allows the network to adaptively emphasize or suppress different feature subspaces based on the current cross-modal context, which is particularly beneficial when the reliability of LiDAR geometry and IMU motion cues varies under sensor degradation or aggressive motion.

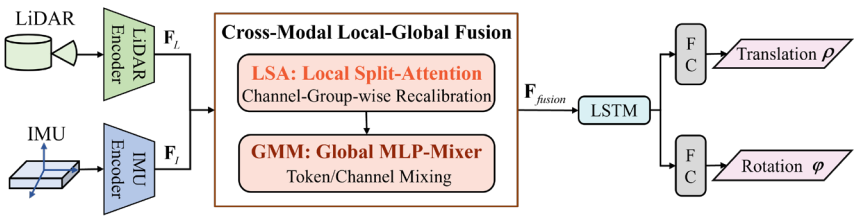

Figure 1. Overall framework of CMLGF-LIO.

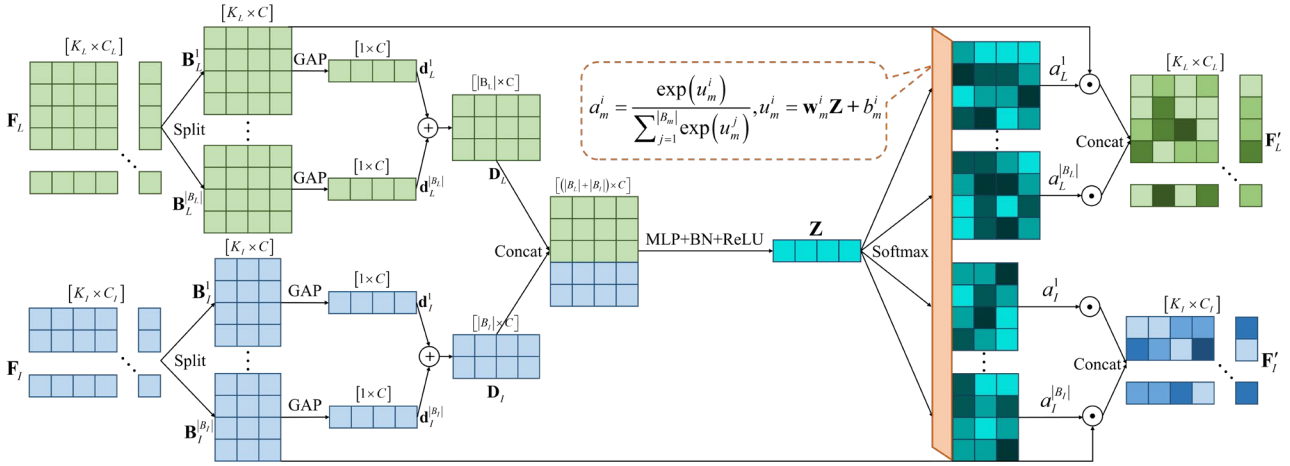

Figure 2. Detailed architecture of the proposed LSA module. The LiDAR and IMU features are first split into channel groups, and group-wise descriptors are then used to generate cross-modal attention weights for local feature recalibration.

Let $\mathbf{F}_m \in \mathbb{R}^{K_m \times C_m}$ denote the input feature of modality $m \in \{L, I\}$, where K_m is the token dimension and C_m is the channel dimension. We split $\mathbf{F}_m$ along the channel dimension into $|B_m|$ groups with a fixed group size C, yielding $\mathbf{F}_m = \left[B_m^1, \cdots, B_m^{|B_m|}\right]$ with $\mathbf{B}_m^i \in \mathbb{R}^{K_m \times C}$. If C_m is not divisible by C, we pad the remaining channels of the last group to ensure that all groups have the same size C. For each group $\mathbf{B}_m^i$, we compute a compact descriptor using global average pooling (GAP) along the token dimension:

$$\mathbf{d}_m^i = \mathrm{GAP}\left(\mathbf{B}_m^i\right) = \frac{1}{K}\sum_{k=1}^{K_m} \mathbf{B}_m^i(k,:) \tag{2}$$

Stacking all group descriptors yields a modality-level descriptor matrix $\mathbf{D}_m = \left[\mathbf{d}_m^1; \cdots; \mathbf{d}_m^{|B_m|}\right] \in \mathbb{R}^{|B_m| \times C}$. To model cross-modal correlations, the LiDAR and IMU descriptor matrices are concatenated along the group dimension and projected into a shared latent representation:

$$\mathbf{Z} = \mathrm{ReLU}\left(\mathbf{W}_Z \mathrm{vec}\left(\left[\mathbf{D}_L; \mathbf{D}_I\right]\right) + \mathbf{b}_Z\right) \tag{3}$$

where $\mathrm{vec}(\cdot)$ denotes vectorization, and the projection is implemented by a lightweight MLP. Based on $\mathbf{Z}$, LSA predicts a scalar importance score for each local group of modality m:

$$a_m^i = \frac{\exp\left(u_m^i\right)}{\sum_{j=1}^{|B_m|} \exp\left(u_m^j\right)}, u_m^i = \mathbf{w}_m^i \mathbf{Z} + b_m^i \tag{4}$$

where a_m^i reflects the relative importance of the i-th channel group under the current cross-modal context, and the Softmax is applied over the group index within each modality. The original group feature is then recalibrated by its corresponding importance coefficient, which is broadcast along both token and channel dimensions:

$$\mathbf{F}_m' = \mathrm{Group}\left(\mathrm{Concat}\left(a_m^1 \odot \mathbf{B}_m^1, \ldots, a_m^{|B_m|} \odot \mathbf{B}_m^{|B_m|}\right)\right) \tag{5}$$

where $\odot$ denotes element-wise scaling with broadcasting. $\mathrm{Concat}(\cdot)$ restores the original channel order, and $\mathrm{Group}(\cdot)$ removes the zero-padded channels of the last group when necessary. Consequently, $\mathbf{F}_m' \in \mathbb{R}^{K_m \times C_m}$ has the same dimensionality as $\mathbf{F}_m$.

Through this channel-group-wise recalibration, LSA suppresses unreliable or redundant feature subspaces while enhancing informative ones, producing refined modality features $\mathbf{F}_L'$ and $\mathbf{F}_I'$ that are better aligned for the subsequent global fusion stage.

2.3 Global MLP-Mixer Module

The proposed LSA module enhances cross-modal interaction within local channel groups. However, due to its localized aggregation, LSA alone does not explicitly propagate information across the full multimodal token sequence. To further capture long-range dependencies and build a globally coherent fused representation, we introduce a GMM module. As illustrated in Figure 3, GMM adopts a pre-normalized residual all-MLP design and models global interactions through sequential token mixing and channel mixing. Compared with attention-based global fusion, the MLP-Mixer design provides a simple and effective global interaction mechanism without explicit pairwise attention computation.

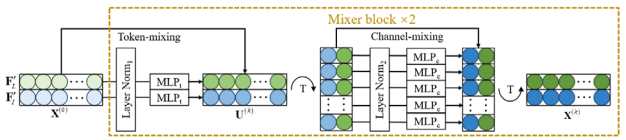

Figure 3. Detailed architecture of the proposed GMM module. The input multimodal representation is first processed by token mixing to capture global inter-token dependencies and then by channel mixing to refine channel-wise responses.

Let $\mathbf{F}'_L \in \mathbb{R}^{S_L \times C_L}$ and $\mathbf{F}'_I \in \mathbb{R}^{S_I \times C_I}$ denote the LSA-refined features from the LiDAR and IMU branches, respectively, where S_L and S_I are the numbers of tokens in the two modalities, and C_L and C_I are their corresponding channel dimensions. To embed the two modalities into a unified feature space, we first project them to a shared channel dimension C by modality-specific linear projections. The initial multimodal token representation is then constructed by concatenating the projected features along the token dimension:

$$\mathbf{X}^{(0)} = \left[\mathrm{Proj}_L\left(\mathbf{F}'_L\right); \mathrm{Proj}_I\left(\mathbf{F}'_I\right)\right] \in \mathbb{R}^{(S_L+S_I)\times C} \tag{6}$$

GMM is implemented by stacking two identical Mixer blocks. For the k-th block $(k=1,2)$, token mixing is first performed to enable information exchange across the full token set:

$$\mathbf{U}^{(k)} = \mathbf{X}^{(k-1)} + \left(\mathrm{MLP}_{\mathrm{t}}\left(\mathrm{LN}_1\left(\mathbf{X}^{(k-1)}\right)^{\mathrm{T}}\right)\right)^{\mathrm{T}} \tag{7}$$

where $\mathrm{LN}_1(\cdot)$ denotes Layer Normalization, and $\mathrm{MLP}_{\mathrm{t}}$ is a two-layer perceptron with GELU activation applied along the token dimension. This step allows each token to aggregate global context from the entire multimodal sequence, thereby explicitly modeling long-range inter-token dependencies. After global token interaction, channel mixing is applied to refine the feature responses within each token:

$$\mathbf{X}^{(k)} = \mathbf{U}^{(k)} + \mathrm{MLP}_{\mathrm{c}}\left(\mathrm{LN}_2\left(\mathbf{U}^{(k)}\right)\right) \tag{8}$$

This operation further enhances inter-channel dependency modeling after global context propagation, leading to more expressive fused token representations. By stacking two such Mixer blocks, the proposed GMM progressively transforms the locally refined multimodal input $\mathbf{X}^{(0)}$ into a globally integrated representation $\mathbf{X}^{(2)}$, which is forwarded to the pose inference stage described in Section 2.1.

In summary, LSA and GMM play complementary roles in the proposed local-global fusion framework. LSA focuses on local cross-modal recalibration within grouped feature subspaces, whereas GMM performs global dependency modeling through sequential token mixing and channel mixing. Their combination yields a multimodal representation that is both locally

discriminative and globally coherent, which is beneficial for robust odometry estimation.

2.4 Loss Function

We train the proposed CMLGF-LIO using a mean squared error (MSE) objective defined on the relative pose increment between consecutive frames. For a training sequence of length T, the loss is computed as:

$$Loss_{\mathrm{pose}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\beta\left\|\hat{\boldsymbol{\varphi}}_t - \boldsymbol{\varphi}_t\right\|_2^2 + \left\|\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t\right\|_2^2\right) \tag{9}$$

where $(\boldsymbol{\varphi}_t, \boldsymbol{\rho}_t)$ and $(\hat{\boldsymbol{\varphi}}_t, \hat{\boldsymbol{\rho}}_t)$ denote the ground-truth and predicted rotation and translation increments between frames $t-1$ and t, respectively. The scalar β balances the rotation and translation terms due to their different numerical scales and units, and we set $\beta = 100$ in all experiments.

3. Experiments

3.1 Datasets and Evaluation Metrics

We evaluate CMLGF-LIO on the KITTI Odometry benchmark, which is widely used for assessing odometry and SLAM methods in autonomous driving scenarios and provides synchronized LiDAR and IMU measurements across diverse urban, suburban, and highway environments. The dataset contains 22 sequences in total, among which sequences 00-10 provide ground-truth trajectories. Following common practice, we use sequences 00, 01, 02, 04, 05, 06, 07, and 08 for training, and reserve 09 and 10 for testing. Sequence 03 is excluded because the corresponding IMU measurements are unavailable, making the LiDAR-IMU fusion setting inconsistent. We further hold out 10% of the training data as a validation set for model selection and hyperparameter tuning.

We evaluate localization accuracy using the standard KITTI odometry metrics: the relative translational error t_{rel} (%) and the relative rotational error r_{rel} (°/100m). t_{rel} measures the average translational drift normalized by the traveled distance and is reported as a percentage, while r_{rel} measures the average rotational drift per 100 meters. Following the KITTI evaluation protocol, errors are computed over multiple sub-trajectories with path lengths ranging from 100m to 800m at 100m intervals. The reported values are averaged over all valid sub-sequences.

3.2 Implementation Details

Our network is implemented in PyTorch and trained on a workstation equipped with an NVIDIA RTX 4090 GPU. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 16, and the model is trained for 120 epochs. We employ a stepwise learning rate decay schedule with an initial learning rate of 1×10^{-4}. Specifically, the learning rate is halved every 30 epochs to facilitate stable convergence in the later training stages. For LiDAR inputs, raw point clouds are projected into a vertex image and a normal image before being fed to the ResNet-18 encoder. The LiDAR projection resolution is set to $H = 64$ and $W = 720$. In the LSA module, the channel group size is fixed to $C = 128$, which provides a practical trade-off between representational capacity and computational efficiency and ensures consistent granularity for channel-group interactions across modalities.

3.3 Localization Performance Comparison

As reported in Table 1, we compare CMLGF-LIO with representative model-based and learning-based baselines, including FAST-LIO2, LO-Net (Li et al., 2019), and DeepLIO. Overall, learning-based methods tend to outperform conventional model-based approaches on the KITTI benchmark under our experimental setting. Meanwhile, CMLGF-LIO achieves more stable and competitive localization performance on most evaluated sequences. In particular, on the test sequences 09 and 10, CMLGF-LIO attains an average t_{rel} and r_{rel} of 1.16% and 0.42 °/100m, respectively, which clearly

outperform those of FAST-LIO2 (5.31%, 1.38°/100m) and LO-Net (1.59%, 0.76°/100m), indicating stronger cross-sequence generalization and reduced accumulated drift.

A closer comparison suggests that FAST-LIO2 relies primarily on geometric constraints and the quality of point-to-map registration. Its performance degrades more noticeably when geometric observability is reduced, dynamic disturbances occur, or motion patterns change. In contrast, learning-based methods such as LO-Net and DeepLIO can learn more robust representations and fusion strategies in a data-driven manner, resulting in lower overall error levels. However, they may still suffer from limited global consistency and inadequate modeling of long-range dependencies across sequences. For instance, DeepLIO exhibits relatively high t_{rel} on several sequences, suggesting that fusion strategies dominated by global or coarse-grained interactions can lead to pronounced drift accumulation under degraded conditions. By comparison, CMLGF-LIO integrates a cross-modal local-global fusion strategy that simultaneously enhances local feature reliability and enforces global consistency, thereby achieving lower t_{rel} and more competitive r_{rel} on most sequences. These results validate the effectiveness and robustness of the proposed framework in complex driving environments.

Method		00*	01*	02*	04*	05*	06*	07*	08*	09	10	Mean on 09 and 10
FAST-LIO2	t_{rel}	7.79	10.12	8.04	5.46	1.97	4.28	1.01	3.56	6.89	3.72	5.31
	r_{rel}	1.05	2.75	1.61	2.18	0.93	1.47	0.71	1.03	1.60	1.15	1.38
LO-Net	t_{rel}	1.47	1.36	1.52	**0.51**	1.04	0.71	1.70	2.12	1.37	1.80	1.59
	r_{rel}	0.72	0.47	0.71	0.65	0.69	0.50	0.89	0.77	0.58	0.93	0.76
DeepLIO	t_{rel}	1.60	5.90	1.96	3.70	1.24	1.97	1.92	2.34	4.40	4.00	4.20
	r_{rel}	**0.38**	**0.19**	**0.23**	**0.12**	0.21	**0.14**	0.32	**0.34**	**0.21**	0.51	**0.36**
w/o LSA	t_{rel}	2.23	4.10	3.45	2.76	1.39	1.84	1.21	1.68	3.02	3.50	3.26
	r_{rel}	0.96	1.11	0.84	0.60	0.55	0.91	0.56	0.78	0.82	1.09	0.96
w/o GMM	t_{rel}	1.77	3.96	2.03	1.74	0.90	1.38	1.00	1.75	2.31	2.15	2.23
	r_{rel}	0.69	0.84	1.15	0.97	0.63	0.75	0.88	1.36	0.64	0.97	0.81
CMLGF-LIO	t_{rel}	**0.86**	**0.71**	**0.95**	0.57	**0.52**	**0.38**	**0.70**	**1.35**	**0.82**	1.49	**1.16**
	r_{rel}	0.45	0.28	0.42	0.30	**0.19**	0.26	**0.31**	0.66	0.36	**0.48**	0.42

Table 1. Localization results based on KITTI dataset in terms of relative translational error t_{rel} (%) and relative rotational error r_{rel} (°/100m). The best results are highlighted in bold, and the second-best results are underscored. The symbol ‘*’ indicates training sequences.

3.4 Ablation Study

To quantify the contributions of the proposed LSA and GMM modules, we conduct component-level ablation studies on the KITTI dataset, as summarized in Table 1 and visualized in Figure 4. Specifically, we remove the LSA module (w/o LSA) and the GMM module (w/o GMM), respectively, while keeping the rest of the network unchanged. As shown in Table 1, removing LSA increases the mean errors on sequences 09 and 10 to 3.26% and 0.96°/100m, respectively, indicating that the network loses the ability to adaptively enhance reliable local cross-modal cues. Removing GMM also leads to a noticeable performance drop, highlighting the importance of global dependency modeling and global consistency constraints for controlling drift over long trajectories. Figure 4 provides qualitative trajectory comparisons on sequences 09 and 10. The ablated variants deviate more from the ground truth, especially in segments with faster motion or weaker geometric constraints, whereas the full CMLGF-LIO trajectory remains consistently closer to the ground truth. Overall, these results demonstrate that the joint use of LSA and GMM enables the network to preserve local feature reliability while enforcing global consistency, thereby achieving more robust and accurate localization under challenging motion and sensor degradation conditions.

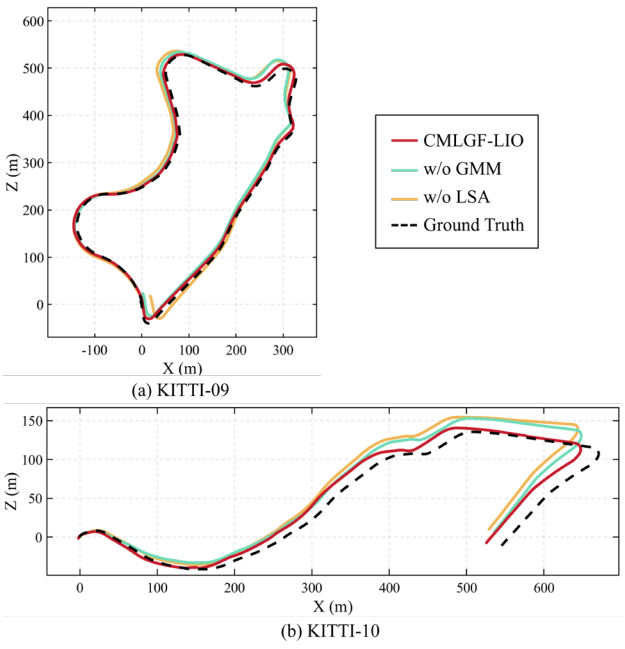

Figure 4. Trajectory comparison of the full model and ablated variants on KITTI sequences 09 and 10.

4. Conclusion

In this study, we propose CMLGF-LIO, a learning-based LIO framework that targets accurate and robust localization in complex, dynamic environments by explicitly improving cross-modal fusion. Specifically, the LSA module performs channel-group-wise recalibration to enhance fine-grained cross-modal interaction and suppress unreliable feature subspaces, thereby improving robustness under locally degraded sensing conditions. Moreover, the GMM module models long-range dependencies by alternately mixing information along token and channel dimensions, strengthening global consistency of the fused representation and mitigating drift accumulation. Extensive experiments on the KITTI dataset demonstrate that CMLGF-LIO achieves consistently competitive performance relative to representative model-based and learning-based baselines, and that the proposed local-global fusion mechanism contributes substantially to improved accuracy and generalization on challenging test sequences. Future work will investigate extending the proposed local-global fusion strategy to additional modalities, such as cameras or radar, while maintaining real-time performance.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant number 42394061], the Natural Science Foundation of Wuhan [Grant number 2024040701010028], and the Zhejiang Province "Vanguard" and "Geese Leading" Research and Development Plan [Grant number 2025C01073].

References

Chen, Z., Xu, Y., Yuan, S., Xie, L., 2024. IG-LIO: An incremental gicp-based tightly-coupled lidar-inertial odometry. IEEE Robotics and Automation Letters, 9(2), 1883-1890. doi.org/10.1109/LRA.2024.3349915.

Dhaigude, V., Kulkarni, A., 2025. A Comprehensive Review of Sensor Fusion Techniques in SLAM: Trends, Challenges and Future Directions. In 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), IEEE, 1-8. doi.org/10.1109/ICAET63349.2025.10932171.

Fan, Z., Zhang, L., Wang, X., Shen, Y., Deng, F., 2025. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: A systematic review. Artificial Intelligence Review, 58(6), 174. doi.org/10.1007/s10462-025-11187-w.

Gao, Y., Zhao, L., Zhang, B., 2025. ININ-LIO: Informer-based neural inertial network aided lidar-inertial odometry. IEEE Transactions on Instrumentation and Measurement, 74, 1-10. doi.org/10.1109/TIM.2025.3546399.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778. doi.org/10.1109/CVPR.2016.90.

Hoang, Q.H., Kim, G.W., 2026. External Disturbances Compensation for LiDAR-Inertial Odometry Under Vibration Conditions on Quadruped Robot. IEEE Robotics and Automation Letters, 11(4), 4713-4720. doi.org/10.1109/LRA.2026.3665313.

Iwaszczuk, D., Roth, S., 2021. DeepLIO: deep lidar inertial sensor fusion for odometry estimation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 1, 47-54. doi.org/10.5194/isprs-annals-V-1-2021-47-2021.

Lai, H., Chen, Q., Zhang, J., Pu, J., 2025. A2DO: Adaptive Anti-Degradation Odometry with Deep Multi-Sensor Fusion for Autonomous Navigation. 2025 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 8589-8595. doi.org/10.1109/ICRA55743.2025.11128429.

Lee, D., Jung, M., Yang, W., Kim, A., 2024. Lidar odometry survey: recent advancements and remaining challenges. Intelligent Service Robotics, 17(2), 95-118. doi.org/10.1007/s11370-024-00515-8.

Li, Q., Chen, S., Wang, C., Li, X., Wen, C., Cheng, M., Li, J., 2019. LO-Net: Deep real-time lidar odometry. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 8473-8482. doi.org/10.1109/CVPR.2019.00867.

Lin, X., Yang, X., Yao, W., Wang, X., Ma, X., Ma, X., Ma, B., 2024. Graph-based adaptive weighted fusion SLAM using multimodal data in complex underground spaces. ISPRS Journal of Photogrammetry and Remote Sensing, 217, 101-119. doi.org/10.1016/j.isprsjprs.2024.08.007.

Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D., 2020. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), 5135-5142. doi.org/10.1109/IROS45743.2020.9341176.

Son, H., Lee, B., Sung, S., 2021. Synthetic deep neural network design for LiDAR-inertial odometry based on CNN and LSTM. International Journal of Control, Automation and Systems, 19(8), 2859-2868. doi.org/10.1007/s12555-020-0443-2.

Sun, L., Ding, G., Qiu, Y., Yoshiyasu, Y., Kanehiro, F., 2023. TransFusionOdom: Transformer-based LiDAR-inertial fusion odometry estimation. IEEE Sensors Journal, 23(18), 22064-22079. doi.org/10.1109/JSEN.2023.3302401.

Tu, Y., Xie, J., 2022. UnDeepLIO: Unsupervised deep lidar-inertial odometry. Asian Conference on Pattern Recognition. Cham: Springer International Publishing, 189-202. doi.org/10.1007/978-3-031-02444-3_14.

Xu, W., Cai, Y., He, D., Lin, J., Zhang, F., 2022. FAST-LIO2: Fast direct lidar-inertial odometry. IEEE Transactions on Robotics, 38(4), 2053-2073. doi.org/10.1109/TRO.2022.3141876.

Yang, X., Yan, L., Xie, H., Lin, X., Zhu, L., Feng, D., Fei, L., 2026. IaNDT-SLAM: intensity-augmented NDT for robust LiDAR-inertial SLAM in unstructured environments. Geo-Spatial Information Science, 1-15. doi.org/10.1080/10095020.2026.2642558.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270. doi.org/10.1162/neco_a_01199.