GM3DMOT: ONLINE 3D GRAPH-MAMBA MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Online LiDAR 3D multi-object tracking requires reliable association over sparse, unordered, and dynamically changing detections. Existing learning-based geometry-aware methods enrich association representations through sparse graph propagation or dense inter-object interactions. However, sparse graphs capture limited context, while dense interactions may introduce interference from weakly related objects. We propose GM3DMOT, an online LiDAR 3D MOT framework for identity-aware local and frame-level context modeling. Its Geometry-Conditioned Object Graph-Mamba Encoder comprises two complementary branches. The Same-Class Local Interaction Branch uses a same-class BEV graph with edge-gated message passing to model local ambiguity among nearby identity competitors. The Score-Degree Mamba Scan Branch combines detection confidence with same-class graph in-degree to serialize unordered detections for broader frame-level propagation. Scan-path geometric cues encode BEV distance and class consistency between consecutive detections in the serialized sequence. Before association decoding, a lightweight motion-conditioned relation refinement introduces short-term detection–track consistency. On nuScenes, GM3DMOT achieves 73.9% AMOTA with 138 identity switches on validation and 70.3% AMOTA with 198 identity switches on test.

1 INTRODUCTION

Online 3D multi-object tracking (MOT) is a fundamental perception task for autonomous driving, aiming to localize surrounding objects across consecutive frames while preserving consistent identities. Tracking-by-detection remains widely used in online LiDAR 3D MOT because it can leverage strong off-the-shelf detectors while retaining a modular association stage (Yin et al., 2021; Ding et al., 2023; Kim et al., 2025). Because LiDAR detections lack the rich appearance cues available in images, identity discrimination relies primarily on geometry, motion, and detection confidence. Nearby objects of the same class often exhibit similar sizes and motion patterns. Occlusion, missed detections, and confidence fluctuations further increase association ambiguity and can lead to identity switches.

Classical trackers associate detections with tracks using handcrafted motion models and geometric matching (Weng et al., 2020a; Yin et al., 2021). Although efficient, their fixed assumptions can struggle with local identity ambiguity among nearby same-class objects. Learning-based graph and Transformer trackers instead learn object or relation representations (Weng et al., 2020b; Ding et al., 2023; Kim et al., 2022; 2025).

Many representative designs emphasize one of two interaction patterns. Sparse graph propagation preserves relation selectivity by restricting information exchange to a local neighborhood, but its contextual coverage is limited. Dense inter-object interaction provides broader context, but may introduce interference from weakly related objects. Extending contextual coverage while preserving association-relevant local structure therefore remains challenging in online LiDAR 3D MOT.

Selective state space models such as Mamba (Gu & Dao, 2024) avoid dense pairwise interaction through input-dependent state updates, while graph state-space models (Wang et al., 2024; Behrouz & Hashemi, 2024) combine local message passing with node prioritization and selective scanning. Online 3D association introduces a distinct challenge: unordered current detections must be organized using task-relevant cues for identity association. We therefore use detection confidence and same-class

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

local structure for serialization, while explicitly encoding geometric relations between consecutive scan tokens, since adjacency in the serialized sequence does not necessarily imply spatial proximity.

We address this problem with GM3DMOT. Its Geometry-Conditioned Object Graph-Mamba Encoder contains two complementary branches. The Same-Class Local Interaction Branch constructs a same-class BEV graph and uses edge-gated message passing to model local ambiguity among nearby identity competitors. The Score-Degree Mamba Scan Branch combines detection confidence with directed same-class graph in-degree to serialize current detections for broader frame-level propagation.

Scan-path geometric cues encode BEV relations between consecutive detections in the scan sequence. Before association decoding, a lightweight motion-conditioned relation refinement adds short-term position and velocity consistency between detections and propagated tracks.

On nuScenes, GM3DMOT achieves 73.9% AMOTA with 138 IDS on validation and 70.3% AMOTA with 198 IDS on test.

Our contributions are summarized as follows:

- We propose GM3DMOT, an online LiDAR 3D MOT framework for identity-aware association that jointly addresses local same-class ambiguity and broader frame-level context.
- A Geometry-Conditioned Object Graph-Mamba Encoder combines a Same-Class Local Interaction Branch with a Score-Degree Mamba Scan Branch enhanced by scan-path geometric cues. GM3DMOT further applies a lightweight motion-conditioned relation refinement before association decoding.
- Extensive experiments on nuScenes demonstrate strong tracking performance, particularly in reducing identity switches. GM3DMOT achieves 138 IDS on validation and 198 IDS on test. Controlled ablations support the complementary roles of its major components.

2 Related Work

Geometry-Aware Association and Object Relation Modeling. Learning-based 3D trackers enhance association representations through geometric relation modeling. GNN3DMOT (Weng et al., 2020b) learns object interactions with a graph neural network, while OGR3MOT (Zaech et al., 2022) constructs online tracking graphs. 3DMOTFormer (Ding et al., 2023) formulates association as edge classification on a detection–track bipartite graph using an Edge-Augmented Graph Transformer. PolarMOT (Kim et al., 2022) encodes pairwise geometric relations in localized polar coordinates, while ShaSTA (Sadjadpour et al., 2024) learns shape and spatiotemporal affinities between tracks and detections. More recently, GRAE-3DMOT (Kim et al., 2025) combines detection-level spatial relations, detection–track spatiotemporal relations, and distance-aware feature fusion, while DSC-Track (Zhang et al., 2026) exploits spatiotemporal cue consistency through Transformer-based interaction. Recent 3D MOT methods also explore complementary paradigms, including the unified tracking framework MCTrack (Wang et al., 2025), neural-enhanced belief-propagation tracking in NEBP (Liang & Meyer, 2024), optimized Poisson multi-Bernoulli filtering in OptiPMB (Ding et al., 2025), and LiDAR-based two-stage Transformer tracking in LiDAR MOT-DETR (Teye et al., 2025). These works illustrate the diversity of recent approaches to 3D multi-object tracking.

State Space Models for Graphs and Tracking. Graph state-space models have explored different forms of graph serialization and selective propagation. Wang et al. (2024) use graph-centric node prioritization, whereas Behrouz & Hashemi (2024) combine neighborhood tokenization, ordering, local encoding, and selective SSM processing. Related state-space models have also been studied for unordered 3D data (Han et al., 2024; Zhang et al., 2025) and tracking (Xiao et al., 2024; Yan et al., 2025; Tian et al., 2025). The graph-oriented works primarily address generic graph representation learning, while the tracking methods mainly model temporal evolution across trajectories or frames. GM3DMOT instead focuses on organizing variable-size current-frame detections for online 3D association using detection reliability, same-class structure, and scan-path geometry.

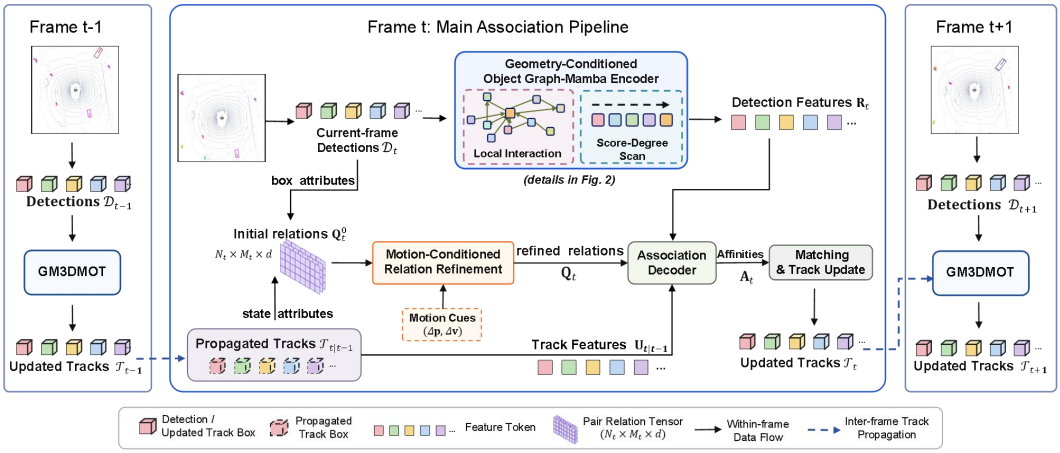

Figure 1: Overall framework of GM3DMOT. Current detections are encoded by the Geometry-Conditioned Object Graph-Mamba Encoder, while motion-conditioned refinement augments detection–track relations before affinity decoding, online matching, and track update. Updated tracks are propagated to the next frame.

3 METHOD

3.1 OVERALL FRAMEWORK

GM3DMOT follows an online tracking-by-detection formulation and focuses on learning association representations that combine same-class local structure with broader frame-level object context. At frame t, the tracker associates the current detections $\mathcal{D}_t = \{\mathbf{d}_i^t\}_{i=1}^{N_t}$ with the propagated tracks $\mathcal{T}_{t|t-1} = \{\boldsymbol{\tau}_j^{t|t-1}\}_{j=1}^{M_t}$, as illustrated in Figure 1. Each detection contains its 3D center, dimensions, orientation, class, confidence, and velocity.

Each detection is projected by $\phi_{\text{det}}(\cdot)$ into an initial object token, forming $\mathbf{H}_t \in \mathbb{R}^{N_t \times d}$, which is encoded into detection features $\mathbf{R}_t$. Each propagated track contains an extrapolated motion state and a learned feature, with track features stacked as $\mathbf{U}_{t|t-1} \in \mathbb{R}^{M_t \times d}$. For each candidate detection–track pair, an initial relation $\mathbf{Q}_t^0$ is refined with short-term motion cues to obtain $\mathbf{Q}_t$. The association decoder then predicts pairwise affinities from $\mathbf{R}_t$, $\mathbf{U}_{t|t-1}$, and $\mathbf{Q}_t$, followed by online matching and track update.

3.2 GEOMETRY-CONDITIONED OBJECT GRAPH-MAMBA ENCODER

The encoder combines association-relevant same-class local structure with broader frame-level context. As shown in Figure 2, a Same-Class Local Interaction Branch models nearby identity competitors. In parallel, a Score-Degree Mamba Scan Branch organizes broader propagation using detection reliability and same-class graph structure. Their outputs are restored to the original detection order and fused residually with the input representation. The graph and scan order are reconstructed for each frame; the Mamba branch processes current detections rather than a sequence of track histories. Temporal information enters association through propagated track features and detection–track relations.

Same-Class Local Interaction Branch. In the BEV plane, GM3DMOT selects same-class neighbors for each detection and constructs a directed same-class neighborhood graph:

$$\begin{aligned} \mathcal{E}_t^{\rightarrow} &= \left\{(i,j) \mid c_i^t = c_j^t,\ j \in \text{kNN}_k(i)\right\}, \\ d_i^{\text{in}} &= \sum_{j=1}^{N_t} \mathbb{I}\left[(j,i) \in \mathcal{E}_t^{\rightarrow}\right]. \end{aligned} \tag{1}$$

Here, c_i^t denotes the class of detection i, and $\text{kNN}_k(i)$ denotes its k nearest same-class neighbors measured by BEV center distance. Self-neighbors are excluded; if fewer than $k+1$ same-class

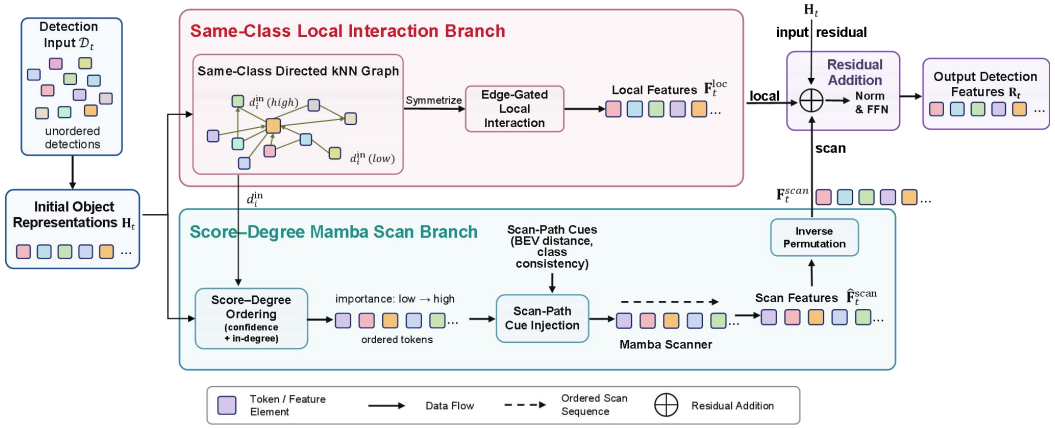

Figure 2: Geometry-Conditioned Object Graph-Mamba Encoder. Shared object representations enter the local interaction and score-degree scan branches; the scan output is inverse-permuted before residual addition.

detections are present, all remaining same-class detections are used as neighbors (Appendix C). The directed in-degree d_i^{in} measures how frequently detection i is selected as a neighbor by other same-class detections, characterizing its local topological prominence.

For local message passing, we augment the directed edge set with its reverse edges:

$$\mathcal{E}_t = \mathcal{E}_t^{\rightarrow} \cup \{(j,i) \mid (i,j) \in \mathcal{E}_t^{\rightarrow}\}. \tag{2}$$

The resulting edge set $\mathcal{E}_t$ enables bidirectional interaction between nearby same-class detections, whereas d_i^{in} remains computed on $\mathcal{E}_t^{\rightarrow}$ and is later used for scan ordering.

Local aggregation follows an edge-gated residual graph-convolution formulation (Bresson & Laurent, 2017). Let $\mathcal{N}_t(i) = \{j \mid (j,i) \in \mathcal{E}_t\}$ denote the incoming neighbors of detection i. For each edge $(j,i) \in \mathcal{E}_t$, the relative position, dimensions, and orientation of the connected detections are encoded as a geometric edge feature $\mathbf{e}_{ji}^t$. The source token, target token, and edge feature are jointly used to predict an element-wise edge gate:

$$\boldsymbol{\gamma}_{ji}^t = \sigma\left(\psi\left(\mathbf{h}_j^t, \mathbf{h}_i^t, \mathbf{e}_{ji}^t\right)\right) \in \mathbb{R}^d. \tag{3}$$

The local representation is then obtained through channel-wise gated aggregation:

$$\mathbf{o}_i^{t,\text{loc}} = \mathbf{W}_{\text{self}}\mathbf{h}_i^t + \frac{\sum_{j \in \mathcal{N}_t(i)} \boldsymbol{\gamma}_{ji}^t \odot \left(\mathbf{W}_{\text{loc}}\mathbf{h}_j^t\right)}{\sum_{j \in \mathcal{N}_t(i)} \boldsymbol{\gamma}_{ji}^t + \epsilon}, \tag{4}$$

where the division is performed element-wise and ϵ prevents numerical instability. The outputs are stacked into

$$\mathbf{F}_t^{\text{loc}} = \left[\mathbf{o}_1^{t,\text{loc}}, \ldots, \mathbf{o}_{N_t}^{t,\text{loc}}\right]^{\top} \in \mathbb{R}^{N_t \times d}. \tag{5}$$

Local message passing primarily focuses on nearby same-class detections. On validation, 91.1% of graph edges lie within 20 m, while only 8.9% exceed 20 m. For these longer edges, the mean gate and relative message contribution drop to 0.338 and 0.317, respectively, versus 0.632/1.000 within 0–5 m. Thus, occasional long-range kNN connections receive substantially weaker propagation. Detailed edge-distance and gating statistics are reported in the supplementary material.

Score-Degree Mamba Scan Branch. Current-frame detections form an unordered set, whereas recurrent state propagation in Mamba is order-sensitive. The scan order is therefore designed not merely to impose a sequence on the detections, but to organize frame-level propagation according to detection-side cues relevant to identity association. GM3DMOT defines an association-oriented

detection-side importance score that combines detection confidence with directed same-class graph in-degree:

$$s_i^{\mathrm{sd}} = \alpha \,\mathrm{RankNorm}(s_i^t) + (1-\alpha)\,\mathrm{RankNorm}(d_i^{\mathrm{in}}), \tag{6}$$

where s_i^t denotes the confidence of detection i, $\mathrm{RankNorm}(\cdot)$ denotes frame-wise ascending-rank normalization to $[0,1]$, and α balances observation reliability and local topological prominence. The ordering score is computed entirely from detection-side cues and is used to organize context before downstream detection–track matching.

Unlike generic topology-based prioritization (Wang et al., 2024), score-degree ordering combines observation reliability with local same-class graph structure. High in-degree indicates that a detection is repeatedly selected by nearby same-class neighbors, making it topologically prominent within a local set of potential identity competitors.

Detections are sorted in ascending order of s_i^{sd} to determine the scan permutation $\boldsymbol{\pi}_t = (\pi_1, \ldots, \pi_{N_t})$, satisfying $s_{\pi_1}^{\mathrm{sd}} \le \cdots \le s_{\pi_{N_t}}^{\mathrm{sd}}$. Detections with lower combined score-degree values are placed earlier, whereas those with higher values are placed later. During the unidirectional scan, later tokens receive broader preceding context. We choose this low-to-high direction so that detections supported by higher confidence and local topological prominence are processed after a broader set of frame-level observations. Mamba's input-dependent state updates then modulate how preceding evidence is propagated. The benefit of this direction is evaluated against Reverse S-D in Section 4.4.

Once the scan permutation $\boldsymbol{\pi}_t$ is determined, adjacent tokens in the resulting sequence are not necessarily spatially neighboring objects. To encode this mismatch, GM3DMOT constructs scan-path geometric cues from the BEV distance and class consistency between consecutive detections in the ordered sequence. For $l = 2, \ldots, N_t$, let $\delta_l^t = \|\mathbf{x}_{\pi_l}^t - \mathbf{x}_{\pi_{l-1}}^t\|_2$ denote the corresponding BEV center distance. The scan-path cue is defined as

$$\mathbf{g}_l^t = \phi_{\mathrm{path}}\left(\left[\mathrm{RBF}(\delta_l^t),\ \mathbb{I}\left[c_{\pi_l}^t = c_{\pi_{l-1}}^t\right]\right]\right), \tag{7}$$

where $\mathrm{RBF}(\cdot)$ denotes radial-basis distance encoding and $\mathbf{x}_i^t$ is the BEV center of detection i. The cue is injected as $\widetilde{\mathbf{h}}_{\pi_l}^t = \mathbf{h}_{\pi_l}^t + \eta \mathbf{g}_l^t$, with $\mathbf{g}_1^t = \mathbf{0}$, where η controls the injection strength relative to the original object token.

The geometry-enhanced ordered sequence $[\widetilde{\mathbf{h}}_{\pi_1}^t, \ldots, \widetilde{\mathbf{h}}_{\pi_{N_t}}^t]^\top$ is processed by a unidirectional Mamba module (Gu & Dao, 2024), producing $\widehat{\mathbf{F}}_t^{\mathrm{scan}}$ in scan order. The outputs are restored to the original detection order through inverse permutation:

$$\mathbf{F}_t^{\mathrm{scan}} = \mathrm{InvPerm}\left(\widehat{\mathbf{F}}_t^{\mathrm{scan}}, \boldsymbol{\pi}_t\right).$$

Finally, the input representations, local interaction features, and inverse-permuted scan features are combined through residual addition:

$$\mathbf{R}_t = \mathrm{FFN}\left(\mathrm{LN}\left(\mathbf{H}_t + \mathbf{F}_t^{\mathrm{loc}} + \mathbf{F}_t^{\mathrm{scan}}\right)\right). \tag{8}$$

Layer normalization and the feed-forward network are applied row-wise.

3.3 Motion-Conditioned Relation Refinement

The encoder above organizes context among current-frame detections, whereas detection–track association additionally depends on temporal consistency between each detection and its candidate track. GM3DMOT therefore constructs a base relation representation for each candidate detection–track pair. The relation encodes differences in position, dimensions, orientation, confidence, and elapsed time since the track was last updated:

$$\mathbf{q}_{ij}^{t,0} = \phi_{\mathrm{rel}}\left(\Delta\mathbf{b}_{ij}^t, \Delta t_j^t\right).$$

The pairwise relations are stacked into $\mathbf{Q}_t^0 \in \mathbb{R}^{N_t \times M_t \times d}$, where $\Delta\mathbf{b}_{ij}^t$ collects the pairwise differences in position, dimensions, orientation, and confidence between detection i and propagated track j; the orientation component is computed by direct subtraction, $\theta_i^t - \theta_j^{t|t-1}$. The term Δt_j^t denotes the elapsed time since track j was last updated.

The base relation does not explicitly encode whether the spatial residual is consistent with relative motion. We therefore construct a short-term motion cue from BEV position and velocity residuals:

$$\mathbf{m}_{ij}^{t} = \left[\Delta\mathbf{p}_{ij,\mathrm{BEV}}^{t}, \Delta\mathbf{v}_{ij,\mathrm{BEV}}^{t}\right], \tag{9}$$

Here, $\Delta\mathbf{p}_{ij,\mathrm{BEV}}^{t}$ is the BEV residual between the detection center and the track center predicted from its extrapolated motion state. $\Delta\mathbf{v}_{ij,\mathrm{BEV}}^{t}$ denotes the BEV velocity difference scaled by the frame interval Δt, i.e., $(\mathbf{v}_{i,\mathrm{BEV}}^{t} - \mathbf{v}_{j,\mathrm{BEV}}^{t})\Delta t$. Together, the two cues indicate whether a position offset is consistent with the object's motion. A lightweight mapping refines the base relation:

$$\mathbf{q}_{ij}^{t} = \mathbf{q}_{ij}^{t,0} + \lambda_m \zeta\left(\mathbf{m}_{ij}^{t}\right), \tag{10}$$

where λ_m controls the motion contribution. The refined relations are stacked into $\mathbf{Q}_t \in \mathbb{R}^{N_t \times M_t \times d}$ for association decoding.

3.4 Training Objective and Online Inference

To isolate the proposed representation learning from downstream association and track management, GM3DMOT uses the three-layer association decoder and online matching procedure of GRAE-3DMOT (Kim et al., 2025). Given $\mathbf{R}_t$, $\mathbf{U}_{t|t-1}$, and $\mathbf{Q}_t$, the decoder predicts detection–track affinities and is supervised at all three layers using sigmoid focal loss (Lin et al., 2017) with ground-truth identity labels. A detection–track pair is labeled positive when the two share the same ground-truth identity. During training, track states are maintained through an online rollout using the model-predicted associations and velocity-based extrapolation, rather than being constructed from ground-truth trajectories; ground-truth identities are used only for association supervision.

At inference, the sigmoid final-layer affinity and the exponentially decaying BEV-distance term are equally weighted (0.5/0.5). Class-inconsistent pairs are masked, and the remaining pairs are matched using the adopted two-stage LAPJV procedure (Jonker & Volgenant, 1987). The full scoring formula and track-management details are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

We evaluate GM3DMOT on nuScenes (Caesar et al., 2020) using CenterPoint (Yin et al., 2021) detections. NuScenes contains 700/150/150 training/validation/test scenes, with 2 Hz keyframes and seven tracked classes. We report AMOTA, AMOTP, MOTA, IDS, and FRAG. Published baseline results are taken from the corresponding papers. Our validation results, controlled GRAE-3DMOT comparison, and main-text ablations use identical CenterPoint detections and report three-run means, with checkpoints selected by validation AMOTA. The best full-model validation checkpoint is used for the single official test-set submission. Implementation details are provided in Appendix B.

4.2 Main Results and Controlled Comparison

Benchmark Results. Among the methods compared in Table 1, GM3DMOT achieves the highest validation AMOTA (0.739) and the lowest IDS (138). On test, it ties DSC-Track for the highest AMOTA (0.703) and obtains the lowest IDS (198). Compared with GRAE-3DMOT, GM3DMOT improves AMOTA by 0.002 and reduces IDS from 150 to 138 on validation; on test, it improves AMOTA by 0.001 and reduces IDS from 219 to 198. Relative to the published GRAE-3DMOT results, GM3DMOT reduces IDS but increases FRAG from 315 to 335 on validation and from 366 to 402 on test. Test-set FN also increases (Table 5), while MOTA decreases slightly on both splits, from 0.636 to 0.633 on validation and from 0.574 to 0.573 on test. Relative to DSC-Track, GM3DMOT reduces IDS from 298 to 138 on validation and from 301 to 198 on test.

Controlled GRAE-3DMOT Comparison. Under identical CenterPoint detections and the same three-run protocol, GM3DMOT improves AMOTA from 0.734 to 0.739 and reduces IDS from 158 to 138 (Table 1). It also improves AMOTP and MOTA and reduces FRAG from 346 to 335. Across three runs, it obtains 0.7389 ± 0.0007 AMOTA and 138.0 ± 2.65 IDS. Section 4.5 further reports IDS normalized by covered ground-truth transitions. The controlled reproduction protocol and multi-run stability results are detailed in Appendices D and E, respectively.

Method	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
Validation set					
OGR3MOT (Zaech et al., 2022)	0.693	0.627	0.602	262	332
PolarMOT† (Kim et al., 2022)	0.711	–	–	213	332
CenterPoint (Yin et al., 2021)	0.665	0.567	0.562	562	424
3DMOTFormer (Ding et al., 2023)	0.712	0.515	0.607	341	436
DSC-Track (Zhang et al., 2026)	0.732	0.498	0.625	298	–
GRAE-3DMOT (Kim et al., 2025)	0.737	0.488	**0.636**	150	**315**
GM3DMOT (Ours)	**0.739**	**0.486**	0.633	**138**	335
Test set					
OGR3MOT (Zaech et al., 2022)	0.656	0.620	0.554	288	371
PolarMOT† (Kim et al., 2022)	0.664	0.566	0.561	242	**332**
CenterPoint (Yin et al., 2021)	0.638	0.555	0.537	760	529
3DMOTFormer (Ding et al., 2023)	0.682	0.496	0.556	438	529
DSC-Track (Zhang et al., 2026)	**0.703**	0.476	**0.575**	301	–
GRAE-3DMOT (Kim et al., 2025)	0.702	0.475	0.574	219	366
GM3DMOT (Ours)	**0.703**	**0.474**	0.573	**198**	402
Controlled comparison, identical CenterPoint detections and three-run protocol					
GRAE-3DMOT (local repro.)	0.734	0.495	0.632	158	346
GM3DMOT (Ours)	**0.739**	**0.486**	**0.633**	**138**	**335**

Table 1: Representative nuScenes results with CenterPoint detections. † denotes an offline method. The controlled block reruns GRAE-3DMOT under our three-run protocol. Full comparisons and official test statistics are in Appendix A.

Variant	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
Full GM3DMOT	**0.739**	**0.486**	0.633	**138**	**335**
Dense Self-Attention	0.736	0.490	0.632	158	363
GRU Scan	0.735	0.491	0.632	162	359
w/o Local Interaction	0.735	0.492	0.633	151	348
w/o Mamba Branch	0.733	0.495	0.632	172	366
w/o Geo. Scan Cue	0.737	0.491	**0.636**	164	363
w/o Motion Ref.	0.737	0.490	0.634	164	359

Table 2: Component ablations and controlled context-operator comparisons on the nuScenes validation set. The dense self-attention and GRU variants replace the Mamba context operator while retaining the remaining applicable encoder design. The w/o Mamba Branch variant also removes the dependent scan-path geometric cue.

Cross-Detector Transfer. In inference-only transfer, one CenterPoint-trained checkpoint per tracker is reused without retraining or detector-specific tuning. Relative to the controlled GRAE-3DMOT reproduction, GM3DMOT improves AMOTA from 0.478 to 0.483 and reduces IDS from 388 to 331 with BEVFormer v2 detections; with StreamPETR, AMOTA improves from 0.517 to 0.520 and IDS decreases from 372 to 350. Full results are reported in Appendix G.

4.3 Local and Frame-Level Context Analysis

We analyze the contributions of local structure, frame-level propagation, and scan-path geometry through component and context-operator ablations.

Component Ablation. Removing local interaction increases IDS from 138 to 151, while removing the Mamba branch together with its dependent scan-path cue reduces AMOTA to 0.733 and increases IDS to 172. Removing the scan-path cue alone yields 0.737 AMOTA and 164 IDS, whereas removing the Mamba branch yields 0.733 AMOTA and 172 IDS, indicating an additional contribution from frame-level propagation beyond the scan-path cue. Removing motion-conditioned refinement also yields 0.737 AMOTA and 164 IDS. MOTA varies non-monotonically across these variants, whereas AMOTA and IDS show clearer degradation. Taken together, these ablations support complementary

roles for local same-class interaction, broader frame-level propagation, scan-path geometry, and short-term detection–track motion consistency.

Context Operator Comparison. Replacing Mamba with dense self-attention or GRU yields 0.736/158 and 0.735/162 AMOTA/IDS, respectively, versus 0.739/138 for the full model, indicating that Mamba achieves higher mean AMOTA and lower mean IDS than both alternatives under this controlled setting. The GRU follows the same score-degree recurrent scan, while self-attention performs dense interaction without recurrent scan dependence. Self-attention is faster (15.63 vs. 18.01 ms/frame), whereas GM3DMOT uses less peak memory (605 vs. 637 MB). Tracker-only efficiency and complexity analyses are reported in Appendices I–J.

4.4 Scan Ordering Analysis

Because the Mamba branch is order-sensitive, we compare serialization strategies that vary the ordering cue and scan direction.

Strategy	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
Random	0.735	0.493	**0.634**	156	356
Bi-Random	0.736	0.489	0.632	153	**332**
Score	0.737	0.491	0.633	142	340
Degree	0.737	0.494	**0.634**	153	341
Reverse S-D	0.736	0.490	0.633	149	340
S-D	**0.739**	**0.486**	0.633	**138**	335

Table 3: Ablation on Mamba scan ordering strategies. S-D denotes score-degree ordering; Bi-Random and Reverse S-D denote bidirectional-random and reversed score-degree ordering, respectively.

Table 3 shows that combining confidence and in-degree achieves the lowest IDS (138) and best AMOTA/AMOTP among the tested orderings, compared with 142 IDS for Score-only and 153 for Degree-only. MOTA remains within a narrow range of 0.632–0.634 across the tested orderings and does not follow the AMOTA/IDS trend. The IDS gain over Score-only is consistent across all three matched runs, with reductions of 4, 3, and 5 identity switches, respectively. On the validation detections, 60.8% of detections belong to class-frame groups with non-uniform raw in-degree. A tie-aware analysis further shows that Score-Degree changes the scan position of 43.6% of detections relative to Score-only, while reversing the same criterion increases IDS from 138 to 149. Additional ordering analyses are provided in Appendix F.

Ordering Robustness. We additionally test input permutation and small confidence perturbations. Random input permutation yields 0.738 AMOTA and 141 IDS, compared with 0.739/138 without perturbation. At inference, adding Gaussian noise with $\sigma = 0.010$ to detection confidence—distinct from the training-time ordering regularizer—changes 22.6% of scan positions and yields 0.737 AMOTA and 147 IDS.

4.5 Identity-Switch Analysis

Transition-Coverage Analysis. We normalize identity switches by identity-maintenance opportunities, where a valid ground-truth transition is defined as the same identity appearing in two consecutive evaluable keyframes. Using the same three matched runs as the controlled comparison, GM3DMOT covers 87,420 of 95,427 valid transitions (91.61%), compared with 87,180 (91.36%) for the controlled GRAE-3DMOT reproduction. The corresponding IDS rate decreases from 1.812 to 1.579 per 1,000 covered transitions, showing that the IDS reduction persists after accounting for transition coverage. The reference baseline for this analysis is the controlled GRAE-3DMOT reproduction in Section 4.2, under the same matched-run protocol.

Density-Stratified Analysis. To examine local same-class ambiguity, we define the density of ground-truth object i at keyframe t as $\rho_i^t = \sum_{j \neq i} \mathbb{I}[c_j^t = c_i^t, \|\mathbf{p}_j^t - \mathbf{p}_i^t\|_2 < 10 \text{ m}]$. Transitions are grouped as sparse ($\rho = 0$), medium ($\rho = 1$), or dense ($\rho \geq 2$), and IDS is reported per 1,000 ground-truth transitions to account for different group sizes.

Density	#Trans.	Full		w/o Local	w/o Mamba
		IDS	Rate↓	Rate↓ (Δ)	Rate↓ (Δ)
Sparse	30,896	18	0.583	0.615 (+0.032)	0.712 (+0.129)
Medium	21,888	14	0.640	0.777 (+0.137)	0.914 (+0.274)
Dense	42,643	106	2.486	2.697 (+0.211)	3.049 (+0.563)

Table 4: Density-stratified IDS on nuScenes validation, averaged over three runs. Rates are IDS per 1,000 ground-truth transitions, and Δ denotes the increase over the full model within each group.

Table 4 shows that the full-model IDS rate reaches 2.486 in dense regions, more than four times the 0.583 rate in sparse regions, showing that denser same-class neighborhoods are associated with greater identity ambiguity. Removing local interaction raises the IDS rate by 0.032, 0.137, and 0.211 from sparse to dense regions, respectively. The absolute IDS-rate gap therefore increases with same-class density.

Removing the Mamba branch and its dependent scan-path cue causes a larger degradation at every density level, including a 0.129 increase in sparse regions compared with 0.032 without local interaction.

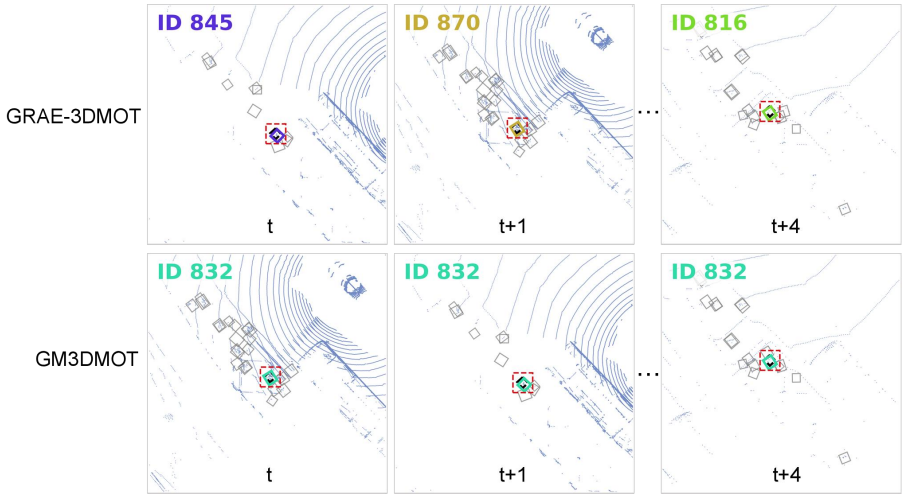

Figure 3: Qualitative comparison on nuScenes. Across frames t, $t+1$, and $t+4$, GRAE-3DMOT changes IDs (845→870→816), whereas GM3DMOT retains ID 832; IDs are method-specific.

5 Conclusion

We presented GM3DMOT, which combines same-class local interaction with score-degree-ordered frame-level propagation for online LiDAR 3D MOT. It achieves 73.9% AMOTA/138 IDS on nuScenes validation and 70.3%/198 on test. Controlled operator and ordering studies support the proposed local–frame-level design and score-degree serialization, while density-stratified analysis shows that the IDS-rate gap caused by removing local interaction increases with same-class density.

Limitations. Evaluation is currently limited to nuScenes; cross-dataset generalization and robustness to larger detector-distribution shifts remain to be studied. Comparisons with additional recurrent or state-space operators would further assess the generality of the proposed serialization. GM3DMOT also has higher tracker-only latency than the dense self-attention variant despite lower peak memory.

Reproducibility Statement

Architecture, training, and inference settings are provided in Appendix B, with graph construction, scan ordering, relation refinement, and online matching details in Appendix C. Controlled baseline reproduction, multi-run stability, cross-detector evaluation, parameter sensitivity, and efficiency analyses are reported in the subsequent appendices.

AI Use Statement

In this work, we used generative AI tools, including ChatGPT and Claude, for manuscript restructuring, language editing, formatting, reference checking, and editorial assistance in presenting the experimental analysis. The research methodology, experimental design, implementation, and reported numerical results were produced and verified by the authors. All AI-assisted text was reviewed against the experimental records, source code, and cited literature. The authors take full responsibility for the final content of the manuscript.

References

Ali Behrouz and Farnoosh Hashemi. Graph Mamba: Towards learning on graphs with state space models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 119–130, 2024.

Nuri Benbarka, Jona Schröder, and Andreas Zell. Score refinement for confidence-based 3D multi-object tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8083–8090, 2021.

Xavier Bresson and Thomas Laurent. Residual gated graph ConvNets. *arXiv:1711.07553*, 2017.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2020.

Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3DMOTFormer: Graph transformer for online 3D multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9784–9794, 2023.

Guanhua Ding, Yuxuan Xia, Runwei Guan, Qinchen Wu, Tao Huang, Weiping Ding, Jinping Sun, and Guoqiang Mao. OptiPMB: Enhancing 3D multi-object tracking with optimized Poisson multi-Bernoulli filtering. *IEEE Transactions on Intelligent Transportation Systems*, 26(12):22312–22328, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3D: Enhancing local features for 3D point cloud analysis via state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4995–5004, 2024.

Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

Aleksandr Kim, Guillem Brasó, Aljoša Ošep, and Laura Leal-Taixé. PolarMOT: How far can geometric relations take us in 3D multi-object tracking? In *European Conference on Computer Vision (ECCV)*, pp. 41–58, 2022.

Hyunseop Kim, Hyo-Jun Lee, Yonguk Lee, Jinu Lee, Hanul Kim, and Yeong Jun Koh. GRAE-3DMOT: Geometry relation-aware encoder for online 3D multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11697–11706, 2025.

Mingchao Liang and Florian Meyer. Neural enhanced belief propagation for multiobject tracking. *IEEE Transactions on Signal Processing*, 72:15–30, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Jianan Liu, Liping Bai, Yuxuan Xia, Tao Huang, Bing Zhu, and Qing-Long Han. GNN-PMB: A simple but effective online 3D multi-object tracker without bells and whistles. *IEEE Transactions on Intelligent Vehicles*, 8(2):1176–1189, 2023.

Ziqi Pang, Zhichao Li, and Naiyan Wang. SimpleTrack: Understanding and rethinking 3D multi-object tracking. In *Computer Vision – ECCV 2022 Workshops*, pp. 680–696, 2022.

Tara Sadjadpour, Jie Li, Rares Ambrus, and Jeannette Bohg. ShaSTA: Modeling shape and spatio-temporal affinities for 3D multi-object tracking. *IEEE Robotics and Automation Letters*, 9(5):4273–4280, 2024.

Martha Teiko Teye, Ori Maoz, and Matthias Rottmann. LiDAR MOT-DETR: A LiDAR-based two-stage Transformer for 3D multiple object tracking. In *British Machine Vision Conference (BMVC)*, 2025.

Shengjing Tian, Yinan Han, Xiantong Zhao, Xuehu Liu, and Qi Lang. MambaTrack3D: A state space model framework for LiDAR-based object tracking under high temporal variation. *arXiv:2511.15077*, 2025.

Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal Tracker: Tracklet never dies. *arXiv:2111.13672*, 2021.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3621–3631, 2023.

Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-Mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv:2402.00789*, 2024.

Xiyang Wang, Shouzheng Qi, Jieyou Zhao, Hangning Zhou, Siyu Zhang, Guoan Wang, Kai Tu, Songlin Guo, Jianbo Zhao, Jian Li, Hailong Qin, and Mu Yang. MCTrack: A unified 3D multi-object tracking framework for autonomous driving. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4551–4558, 2025.

Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10359–10366, 2020a.

Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris Kitani. GNN3DMOT: Graph neural network for 3D multi-object tracking with 2D-3D multi-feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6499–6508, 2020b.

Changcheng Xiao, Qiong Cao, Zhigang Luo, and Long Lan. MambaTrack: A simple baseline for multiple object tracking with state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4082–4091, 2024.

Zhuohao Yan, Shaoquan Feng, Xingxing Li, Yuxuan Zhou, Chunxi Xia, and Shengyu Li. S3MOT: Monocular 3D object tracking with selective state space model. *arXiv:2504.18068*, 2025.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17830–17839, 2023.

Table 5: Full nuScenes test-set comparison with CenterPoint detections, including additional baselines and official test statistics. References for methods shared with Table 1 are provided there; additional baselines are cited in the accompanying text. † denotes offline methods, and ‡ denotes methods using 10 Hz input data.

Method	AMOTA↑	AMOTP↓	MOTA↑	MT↑	ML↓	TP↑	FP↓	FN↓	IDS↓	FRAG↓
OGR3MOT	0.656	0.620	0.554	5278	2094	95264	17877	24013	288	371
PolarMOT†	0.664	0.566	0.561	5701	1686	97909	17856	21414	242	**332**
CenterPoint	0.638	0.555	0.537	5584	1681	95877	18612	22928	760	529
CBMOT	0.649	0.592	0.545	5319	1966	94916	**16469**	24092	557	450
SimpleTrack‡	0.668	0.550	0.566	5476	1780	95539	17514	23451	575	591
ImmortalTracker	0.677	0.599	0.572	5565	1669	97584	18012	21661	320	477
GNN-PMB	0.678	0.560	0.563	5698	**1622**	97274	17071	21521	770	431
3DMOTFormer	0.682	0.496	0.556	5466	1896	95790	18322	23337	438	529
DSC-Track	**0.703**	0.476	**0.575**	–	–	–	–	–	301	–
GRAE-3DMOT	0.702	0.475	0.574	5785	1781	**99023**	19343	**20323**	219	366
GM3DMOT (Ours)	**0.703**	**0.474**	0.573	**5933**	1796	98625	19407	20742	**198**	402

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11784–11793, 2021.

Jan-Nico Zaech, Alexander Liniger, Dengxin Dai, Martin Danelljan, and Luc Van Gool. Learnable online graph representations for 3D multi-object tracking. *IEEE Robotics and Automation Letters*, 7(2):5103–5110, 2022.

Tao Zhang, Haobo Yuan, Lu Qi, Jiangning Zhang, Qianyu Zhou, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Point Cloud Mamba: Point cloud learning via state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10121–10130, 2025.

Haonan Zhang, Xinyao Wang, Boxi Wu, Tu Zheng, Wang Yunhua, and Zheng Yang. Delving into dynamic scene cue-consistency for robust 3D multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12484–12492, 2026.

A Full nuScenes Benchmark Comparison

Table 1 in the main text reports a representative subset of methods for space. Tables 5 and 6 report the full comparison, including additional baselines and, for the test set, the official MT, ML, TP, FP, and FN statistics. On the test set, GM3DMOT records 20,742 FN, compared with 20,323 for the published GRAE-3DMOT result (Table 5).

Additional baselines include CBMOT (Benbarka et al., 2021), SimpleTrack (Pang et al., 2022), ImmortalTracker (Wang et al., 2021), and GNN-PMB (Liu et al., 2023).

B Additional Implementation Details

Table 7 summarizes the principal architecture, training, and online inference settings. All ablation variants retain the remaining applicable settings. Because the scan-path geometric cue is defined along the Mamba scan, the w/o Mamba Branch variant removes both the branch and its dependent scan-path cue, as described in Section 4.3.

For all CenterPoint-based ablations and controlled comparisons, the same precomputed detections are used. In the cross-detector experiments (Appendix G), only the input detection results are replaced. The GM3DMOT validation result and the controlled GRAE-3DMOT result in Table 1, together with all main-text ablations, are averaged over the same three independent training runs. Count-based metrics are averaged and then rounded to the nearest integer. For each run, we select the checkpoint with the highest validation AMOTA. Among the three selected full-model checkpoints, the one with the highest validation AMOTA is used for the single official test-set submission reported in Table 1.

Table 6: Full nuScenes validation-set comparison with CenterPoint detections. † denotes offline methods, and ‡ denotes methods using 10 Hz input data.

Method	AMOTA↑	AMOTP↓	MOTA↑	IDS↓	FRAG↓
OGR3MOT	0.693	0.627	0.602	262	332
PolarMOT†	0.711	–	–	213	332
CenterPoint	0.665	0.567	0.562	562	424
CBMOT	0.675	–	–	494	–
SimpleTrack‡	0.696	0.547	0.602	–	403
ImmortalTracker	0.702	–	0.601	–	385
GNN-PMB	0.707	0.560	–	650	345
3DMOTFormer	0.712	0.515	0.607	341	436
DSC-Track	0.732	0.498	0.625	298	–
GRAE-3DMOT	0.737	0.488	**0.636**	150	**315**
GM3DMOT (Ours)	**0.739**	**0.486**	0.633	**138**	335

Table 7: Principal architecture, training, and online inference settings of GM3DMOT.

Setting	Value	Setting	Value
Hidden dimension d	128	Association decoder layers	3
Detection-token dimension	17	Sparse edge-input dimension	9
Same-class kNN size k	12	Score-degree weight α	0.7
Graph-Mamba encoder layers	1	Scan direction	Unidirectional
Score-degree order	Ascending	Inference scan permutations	1
Mamba state dimension	16	Mamba convolution width	4
Mamba expansion factor	1	Graph-Mamba dropout	0.1
Training-time order-noise std.	0.01	Feed-forward expansion factor	2
Distance RBF bases	8	RBF distance range	[0, 50] m
Scan-path MLP hidden dimension	64	Scan-path cue scale η	0.1
Motion-cue dimension	4	Motion refinement weight λ_m	0.1
Motion interval Δt	0.5 s	Velocity clipping range	$[-20, 20]$ m/s
Training epochs	20	Clip length	8 frames
Optimizer	AdamW	Learning rate	2×10^{-4}
Weight decay	2×10^{-3}	Batch size / accumulation	1 / 8 steps
Effective batch size	8	Independent training runs	3
Checkpoint selection	Highest val. AMOTA	Training GPU	NVIDIA RTX 3090 Ti
High-confidence threshold	0.24	New-track threshold	0.12
Maximum track age	12 frames	Detector input	Precomputed

C Additional Method and Inference Details

This section provides implementation-level details omitted from Section 3.2–3.4 because of space constraints.

C.1 Detection Tokens and Sparse Edge Features

Each detection is represented by a 17-dimensional object token:

$$\mathbf{z}_i^t = \left[x_i^t, y_i^t, z_i^t, l_i^t, w_i^t, h_i^t, \theta_i^t, v_{x,i}^t, v_{y,i}^t, s_i^t, \mathrm{onehot}(c_i^t)\right]. \tag{11}$$

The token contains the 3D center, box dimensions, yaw angle, BEV velocity, detection confidence, and seven-class one-hot encoding. A two-layer MLP projects this input to the 128-dimensional initial object representation $\mathbf{h}_i^t$.

Unlike the dense pairwise spatial representation used in GRAE-3DMOT, the same-class local interaction branch computes geometric edge features only for retained sparse graph edges. The scan branch separately constructs geometric cues between consecutive tokens in the ordered sequence.

For a retained directed edge (j, i), where j and i denote the source and target detections, respectively, the 9-dimensional edge input is

$$\mathbf{e}_{ji}^t = \left[\mathbf{p}_j^t - \mathbf{p}_i^t,\ \mathbf{b}_j^t - \mathbf{b}_i^t,\ \sin(\theta_j^t - \theta_i^t),\ \cos(\theta_j^t - \theta_i^t),\ \|\mathbf{p}_{j,\mathrm{BEV}}^t - \mathbf{p}_{i,\mathrm{BEV}}^t\|_2\right]. \tag{12}$$

BEV edge distance	Edge fraction	Mean gate	Relative contribution
0–5 m	24.8%	0.632	1.000
5–10 m	31.7%	0.586	0.842
10–20 m	34.6%	0.497	0.603
$>$ 20 m	8.9%	0.338	0.317

Table 8: Distance distribution and learned gating behavior of same-class kNN edges on the nuScenes validation detections. Relative contribution is normalized to the 0–5 m group.

Here, $\mathbf{p}_i^t \in \mathbb{R}^3$ and $\mathbf{b}_i^t \in \mathbb{R}^3$ denote the detection center and box dimensions, respectively. The sparse edge input is projected to the model dimension and used by the same-class local interaction branch.

C.2 Same-Class Graph Construction

The directed same-class BEV kNN graph is reconstructed independently at each frame. Self-neighbors are excluded. If a category contains fewer than $k+1$ detections, each detection is connected to all other detections of the same category.

The original directed graph is used to compute the in-degree required by score-degree ordering. For local message passing, reverse edges are added and duplicate directed edges are removed. Therefore, the directed graph statistic used for ordering and the symmetrized graph used for local feature propagation are derived from the same same-class neighborhood structure but serve different roles.

The edge-gated local branch predicts channel-wise gates from the source token, target token, and projected edge feature. Incoming messages are normalized by the summed gates, with a small numerical constant added to the denominator. A learnable self projection is then added to the aggregated neighborhood representation.

Edge-Distance and Gating Audit. To examine whether fixed-k same-class neighborhoods introduce substantial long-range interactions in sparse class-frame groups, we measure the BEV distance distribution of constructed graph edges on the nuScenes validation detections and summarize the learned gating behavior by distance.

Overall, 56.5% of constructed edges lie within 10 m and 91.1% within 20 m, while only 8.9% exceed 20 m. The learned gate decreases with edge distance, from 0.632 in the 0–5 m group to 0.338 beyond 20 m. The relative message contribution shows the same trend, dropping from 1.000 to 0.317. These statistics indicate that the fixed-k graph remains predominantly local in practice, while occasional longer edges induced in sparse class-frame groups receive substantially weaker propagation.

C.3 Rank Normalization and Scan Ordering

For a frame containing N_t detections, the frame-wise ascending-rank normalization used in score-degree ordering is

$$\mathrm{RankNorm}(z_i) = \frac{\mathrm{rank}^{\uparrow}(z_i) - 1}{\max(N_t - 1, 1)}. \tag{13}$$

In the implemented ranking, equal input values may receive distinct sequential ranks according to the sorting order. We examine the effect of this tie handling using a tie-aware diagnostic in Appendix F. The normalized ranks lie in $[0, 1]$. When only one detection is present, its normalized rank is set to zero.

Confidence and directed same-class graph in-degree are independently rank-normalized before being combined (Eq. 6). Detections are sorted in ascending score-degree order, placing detections with larger combined values later in the unidirectional scan. During training, Gaussian noise with standard deviation 0.01 is added to the combined importance values as a small ordering regularizer. Inference uses deterministic score-degree ordering without test-time permutation averaging.

C.4 Scan-Path Geometric Cue

For consecutive detections in the ordered sequence, the scan-path cue contains an eight-basis RBF encoding of their BEV distance and a binary same-class indicator. The RBF centers are evenly distributed between 0 and 50 meters. These features are processed by a two-layer MLP with a 64-dimensional hidden layer and a 128-dimensional output. The output layer is initialized to zero. The resulting geometric embedding is multiplied by $\eta = 0.1$ before being added to the corresponding ordered object token.

The scan-path cue of the first token is fixed to zero because no preceding scan token exists. After the Mamba operation, inverse permutation restores the original detection order. The local feature, inverse-permuted scan feature, and input representation are combined through residual addition, followed by layer normalization and a feed-forward network with expansion factor two and GELU activation.

C.5 Motion-Conditioned Relation Refinement

For a detection–track candidate pair, the four-dimensional motion cue is

$$\mathbf{m}_{ij}^{t} = \left[\mathbf{p}_{i,\mathrm{BEV}}^{t} - \hat{\mathbf{p}}_{j,\mathrm{BEV}}^{t},\ (\mathbf{v}_{i,\mathrm{BEV}}^{t} - \mathbf{v}_{j,\mathrm{BEV}}^{t})\Delta t\right], \tag{14}$$

where $\hat{\mathbf{p}}_{j,\mathrm{BEV}}^{t}$ is the propagated track center and $\Delta t = 0.5$ s. Detection and track velocities are clipped to $[-20, 20]$ m/s before constructing the cue. The four-dimensional cue is mapped by an MLP with dimensions $4 \to 32 \to 128$ and ReLU activation. The final projection is initialized to zero, and the result is scaled by $\lambda_m = 0.1$ before being added to the base detection–track relation representation.

C.6 Affinity and BEV-Distance Fusion

Let A_{ij}^{t} denote the final decoder affinity logit between detection i and propagated track j, and let $d_{ij}^{t,\mathrm{BEV}}$ denote their BEV center distance. The final matching score is

$$S_{ij}^{t} = 0.5\,\sigma\big(A_{ij}^{t}\big) + 0.5\,\exp\Big(-d_{ij}^{t,\mathrm{BEV}}\Big). \tag{15}$$

Pairs with inconsistent semantic classes are masked before bipartite matching. The fusion retains the learned association evidence while preserving an explicit geometric preference for spatially plausible matches.

C.7 Track Motion-Feature Update

For a matched detection–track pair (i, j), the stored motion feature is updated using the fused matching score:

$$\mathbf{u}_{j}^{t} = \big(1 - S_{ij}^{t}\big)\,\mathbf{u}_{j}^{t|t-1} + S_{ij}^{t}\,\mathbf{u}_{i,\mathrm{det}}^{t}, \tag{16}$$

where S_{ij}^{t} is the fused association score in Eq. 15, $\mathbf{u}_{j}^{t|t-1}$ is the propagated track motion feature, and $\mathbf{u}_{i,\mathrm{det}}^{t}$ is the current detection motion feature. A larger fused matching score therefore assigns greater weight to the current detection representation.

C.8 Two-Stage Matching and Track Management

Detections with confidence above 0.24 enter the first LAPJV matching stage. The remaining low-confidence detections are matched against the tracks left unmatched by the first stage. The LAPJV cost limits are 0.9 and 0.8 for the high- and low-confidence stages, respectively. Class-inconsistent pairs are invalidated before both stages.

Unmatched detections with confidence above 0.12 initialize new tracks. Unmatched active tracks are propagated using their estimated velocity and retained until their age reaches 12 frames. The same matching thresholds and track-management settings are used in all validation, ablation, and cross-detector experiments unless otherwise specified.

D CONTROLLED REPRODUCTION OF GRAE-3DMOT

We rerun the official GRAE-3DMOT implementation using the same CenterPoint detections, three-run training protocol, and validation-based checkpoint-selection criterion as GM3DMOT. The resulting comparison is reported in the controlled block of Table 1; published benchmark results are reported separately in the other blocks of that table.

E MULTI-RUN STABILITY

We evaluate training stability over three independent runs with different random initializations. The data, architecture, optimization settings, online inference settings, and checkpoint-selection criterion remain unchanged. For each run, the checkpoint with the highest validation AMOTA is selected. Standard deviations are sample standard deviations computed over the three runs.

Table 9: Individual validation results and aggregate statistics over three independent training runs with different random initializations.

Run	AMOTA↑	AMOTP (m)↓	MOTA↑	IDS↓	FRAG↓
Run 1	0.7394	0.4782	0.6342	135	340
Run 2	0.7392	0.4867	0.6337	139	339
Run 3	0.7381	0.4921	0.6321	140	326
Mean $\pm$ Std.	0.7389 ± 0.0007	0.4857 ± 0.0070	0.6333 ± 0.0011	138.0 ± 2.65	335.0 ± 7.81

Across the three runs, GM3DMOT obtains 0.7389 ± 0.0007 AMOTA and 138.0 ± 2.65 IDS. The limited variation in AMOTA, MOTA, and IDS indicates stable tracking accuracy and identity-switch behavior across random initializations. The rounded means reproduce the full-model row reported in Table 1.

The context-operator, component, ordering, and density-stratified ablations in Section 4 follow the same three-run protocol. Individual runs are provided here for the full model to expose its training stability without duplicating all main-text ablation tables.

F ADDITIONAL SCAN-ORDERING ANALYSIS

F.1 MATCHED-RUN SCORE VS. SCORE-DEGREE COMPARISON

To examine whether the improvement over confidence-only ordering is consistent across random initializations, we compare Score and Score-Degree ordering under the same three matched training runs.

Table 10: Matched-run IDS comparison between Score-only and Score-Degree (S-D) ordering on the nuScenes validation set. Each row uses the same random initialization for the two ordering strategies.

Run	Score IDS↓	S-D IDS↓	Reduction
Run 1	139	135	4
Run 2	142	139	3
Run 3	145	140	5
Mean	142	138	4

Score-Degree ordering reduces IDS relative to Score-only ordering in all three matched runs, with reductions of 4, 3, and 5 identity switches, respectively. The mean IDS decreases from 142 to 138, consistent with Table 3.

F.2 DIRECTED IN-DEGREE AUDIT

We further examine whether directed same-class in-degree provides structural variation that can affect the serialization beyond detection confidence. The audit uses the same CenterPoint validation detections, $k = 12$, and $\alpha = 0.7$ as the main experiments. To separate raw in-degree variation from implementation-specific rank ties, ordering statistics are additionally computed with a tie-aware diagnostic in which equal raw in-degree values receive the same rank. This diagnostic does not alter the trained model or any reported tracking result.

Table 11: Audit of directed same-class in-degree and its effect on Score-Degree serialization on the nuScenes validation detections.

Statistic	Value
Detections in groups with non-uniform raw in-degree	60.8%
Class-frame groups with non-uniform raw in-degree	27.8%
Tie-aware tokens with changed scan position	43.6%
Class-frame groups with changed within-class ordering	21.4%
Mean normalized rank displacement	6.3%

With $k = 12$, 39.2% of detections occur in class-frame groups containing at most 13 same-class detections, for which every node selects all other same-class nodes and the raw in-degree is therefore uniform. In contrast, the remaining detections occur in groups where the directed graph can provide non-uniform local topology. Overall, 60.8% of detections belong to groups with non-uniform raw in-degree, although such groups account for 27.8% of all class-frame groups, showing that non-uniform degree is concentrated in larger groups.

The tie-aware Score-Degree diagnostic changes the global scan position of 43.6% of detections relative to Score-only ordering, with a mean normalized displacement of 6.3%. Moreover, only 0.3% of all class-frame groups have uniform raw in-degree yet exhibit a changed within-class ordering. The scan positions produced by the implemented Score-Degree ranking and the tie-aware diagnostic differ for only 3.8% of detections. This suggests that the observed ordering changes cannot be explained solely by rank-tie handling.

G CROSS-DETECTOR TRANSFER

We evaluate cross-detector transfer by directly replacing the CenterPoint detections used at inference. For each tracker, one CenterPoint-trained checkpoint is selected using only CenterPoint validation performance and is then reused without retraining, fine-tuning, checkpoint reselection, detector-specific adaptation, or threshold retuning. Thus, unlike the three-run averaged validation results in the main text, the cross-detector results correspond to a single selected checkpoint for each tracker.

Within each target-detector setting, exactly the same detection results are provided to GRAE-3DMOT and GM3DMOT. The association procedure, confidence thresholds, class filtering, and track-management settings remain unchanged. No target-detector ground truth is used for checkpoint selection or hyperparameter tuning. AMOTP is a center-position error in meters, consistent with the main text; values above one are expected here given the noisier detection inputs relative to CenterPoint.

Table 12: Inference-only cross-detector comparison on the nuScenes validation set. Both trackers are trained using CenterPoint detections and directly evaluated with BEVFormer v2 (Yang et al., 2023) and StreamPETR (Wang et al., 2023) detections without retraining or detector-specific tuning.

Inference Detector	Tracker	AMOTA↑	AMOTP (m)↓	MOTA↑	IDS↓	FRAG↓
BEVFormer v2	GRAE-3DMOT (local reproduction)	0.478	1.042	0.429	388	901
BEVFormer v2	GM3DMOT	0.483	1.035	0.433	331	883
StreamPETR	GRAE-3DMOT (local reproduction)	0.517	1.030	0.454	372	778
StreamPETR	GM3DMOT	0.520	1.020	0.453	350	803

With BEVFormer v2 detections, GM3DMOT improves AMOTA from 0.478 to 0.483 and MOTA from 0.429 to 0.433. It also reduces AMOTP from 1.042 to 1.035, IDS from 388 to 331, and FRAG from 901 to 883. The IDS reduction is 57, corresponding to a relative reduction of 14.7%.

With StreamPETR detections, GM3DMOT improves AMOTA from 0.517 to 0.520, reduces AMOTP from 1.030 to 1.020, and reduces IDS from 372 to 350. The IDS reduction is 22, corresponding to a relative reduction of 5.9%. MOTA changes slightly from 0.454 to 0.453, while FRAG increases from 778 to 803.

Overall, GM3DMOT reduces IDS under both alternative detection sources, showing that the observed identity-switch reduction also transfers to BEVFormer v2 and StreamPETR detections.

H Parameter Sensitivity

We vary one parameter at a time while retaining all other default settings. All configurations within each sensitivity study are trained from scratch using the same fixed random initialization and the same checkpoint-selection criterion. The default row corresponds to Run 1 in Table 9 and is therefore a single-run result rather than the three-run mean reported in Table 1.

H.1 Sensitivity to the Neighborhood Size

Table 13: Single-run sensitivity to the same-class neighborhood size k on the nuScenes validation set.

Neighborhood Size k	AMOTA↑	AMOTP (m)↓	MOTA↑	IDS↓	FRAG↓
8	0.736	0.491	0.634	163	326
12 (default)	0.739	0.478	0.634	135	340
16	0.738	0.492	0.636	144	348

The default $k = 12$ achieves the best AMOTA, AMOTP, and IDS. The smaller $k = 8$ configuration obtains the lowest FRAG, whereas $k = 16$ achieves the highest MOTA. These complementary metric trends indicate that $k = 12$ provides the strongest overall balance, potentially avoiding both insufficient neighborhood coverage and the inclusion of more weakly related same-class detections.

H.2 Sensitivity to the Score-Degree Weight

Table 14: Single-run sensitivity to the score-degree balancing coefficient α on the nuScenes validation set.

Score-Degree Weight α	AMOTA↑	AMOTP (m)↓	MOTA↑	IDS↓	FRAG↓
0.5	0.738	0.491	0.633	**135**	**337**
0.7 (default)	**0.739**	**0.478**	**0.634**	**135**	340
0.9	0.737	0.488	0.632	147	342

Among the tested settings, $\alpha = 0.7$ achieves the highest AMOTA and MOTA, the lowest AMOTP, and ties for the lowest IDS. The $\alpha = 0.5$ setting obtains the lowest FRAG, while increasing α to 0.9 raises IDS from 135 to 147.

I Tracker-Only Efficiency

We measure tracker-only efficiency on one NVIDIA GeForce RTX 3090 Ti using the same precomputed CenterPoint detections and batch size one. Detector inference, data loading, result serialization, and official metric computation are excluded. CUDA synchronization is performed immediately before and after each timed region. Each profiling run uses 80 effective warm-up frames followed by 5,939 timed validation frames. Runtime and FPS are reported as mean $\pm$ sample standard deviation over three independent profiling runs. Peak allocated memory is measured using PyTorch CUDA peak-memory statistics.

Table 15: Tracker-only efficiency under identical local hardware and profiling protocols. The dense self-attention variant replaces only the Mamba operator while retaining the remaining GM3DMOT encoder design.

Method	Params (M)↓	Runtime (ms/frame)↓	FPS↑	CUDA Peak Alloc. Memory (MB)↓
GRAE-3DMOT (local reproduction)	**0.60**	16.69 ± 0.24	59.91 ± 0.88	689.03
Dense Self-Attention	0.63	$\mathbf{15.63 \pm 0.54}$	$\mathbf{64.02 \pm 2.17}$	636.66
GM3DMOT	0.63	18.01 ± 0.06	55.53 ± 0.19	**605.29**

A separate 200-frame runtime audit confirms that the dense self-attention variant executes the self-attention branch and scaled dot-product attention operation, while neither the Mamba module nor the legacy dense-input path is invoked. Profiling hooks are disabled during timing, and output serialization is excluded. The total, forward-used, and possibly unused parameter counts of this variant are 0.63M, 0.63M, and 0.00M, respectively.

Dense self-attention achieves lower latency and higher throughput in this short-sequence regime, where GPU-optimized attention kernels are effective. Compared with this variant, GM3DMOT improves AMOTA from 0.736 to 0.739, reduces AMOTP from 0.490 to 0.486, lowers IDS from 158 to 138, and reduces FRAG from 363 to 335. It also provides a modest 4.9% reduction in peak allocated memory, from 636.66 to 605.29 MB, at the cost of 15.2% higher measured runtime and 13.3% lower throughput.

Relative to the local GRAE-3DMOT reproduction, GM3DMOT reduces peak allocated memory from 689.03 to 605.29 MB, corresponding to a 12.2% reduction. Its runtime increases by 7.9%, throughput decreases by 7.3%, and parameter count increases by 0.03M, while retaining tracker-only throughput of 55.53 FPS.

Overall, relative to dense self-attention, GM3DMOT improves all reported tracking metrics and modestly reduces peak allocated memory, with the measured runtime and throughput trade-offs reported above.

J Complexity Analysis

Let N_t denote the number of current detections, d the hidden dimension, and $|\mathcal{E}_t|$ the number of retained same-class graph edges.

Table 16: Principal asymptotic complexities for fixed state dimensions. The cost of same-class kNN construction depends on the neighbor-search implementation.

Operation	Time	Memory
Local message passing	$O(\lvert\mathcal{E}_t\rvert d)$	$O((\lvert\mathcal{E}_t\rvert + N_t)d)$
Score-degree sorting	$O(N_t \log N_t)$	$O(N_t)$
Mamba token mixing	$O(N_t d)$	$O(N_t d)$
Dense self-attention	$O(N_t^2 d)$	$O(N_t^2)$

For fixed k, the number of retained sparse graph edges grows approximately linearly with the number of detections. The sparse local branch and Mamba token mixing therefore scale approximately linearly with N_t, apart from same-class neighbor search and the score-degree sorting step.

However, asymptotic token-mixing complexity does not alone determine practical latency. In the nuScenes tracking setting, each frame contains a relatively short object sequence, and the optimized scaled dot-product attention implementation yields lower measured latency than the Mamba implementation.

The complete GM3DMOT encoder additionally includes same-class neighbor search, score-degree sorting, scan-path cue construction, and local graph propagation. Practical efficiency should therefore be assessed using the controlled runtime, throughput, and peak-memory measurements in Table 15, rather than by considering the Mamba operator in isolation.