PC-Track: Prior-Conditioned 3D Multi-Object Tracking with Cross-Level Fusion for Autonomous Driving

Jiajie Zhang[1*], Xiangzhong Liu[2], and Alois C. Knoll[1]

***Abstract*—Commercial 3D sensors and vehicle-to-everything (V2X) modules have been widely adopted in autonomous driving systems; however, their outputs are typically available as sparse, post-processed object lists, which are difficult to integrate into mainstream vision-centric 3D multi-object tracking (MOT) pipelines. In this paper, we propose PC-Track, an end-to-end 3D MOT framework based on cross-level fusion (CLF) that incorporates object lists from external black-box sensors as structured priors within a Transformer architecture and jointly fuses them with multi-view image features. To enhance robustness against noisy or delayed priors in real-world deployment, we introduce a prior-conditioned, fusion-aware association mechanism that unifies motion priors and appearance interactions, enabling stable track continuation under a tracking-by-query paradigm. Extensive experiments on the nuScenes benchmark with real and synthetic object lists demonstrate improvements over the CLF tracking baseline and state-of-the-art vision-based 3D trackers.**

***Index Terms*—Autonomous Driving, Smart Sensors, 3D Multi-Object Tracking, Cross-level Sensor Fusion, Denoising Fusion**

I. INTRODUCTION

Accurate and robust 3D multi-object tracking (MOT) is a fundamental component of perception for autonomous driving and intelligent transportation systems. In practical deployments, perception systems are increasingly equipped with heterogeneous sensing and communication modules, including multi-view cameras that provide dense raw observations and external sources such as commercial 3D sensors or vehicle-to-everything (V2X) modules. Notably, due to hardware constraints, privacy policies, and standardization requirements, the outputs from these external modules are often only accessible as sparse post-processed object lists rather than raw data or intermediate features [1]. This real-world constraint poses a critical challenge: how to effectively integrate high-level object list priors with vision-centric systems for accurate and highly efficient end-to-end 3D MOT.

Conventional fusion strategies for tracking typically operate within a single abstraction level [2]. Track-level late fusion is easy to deploy but often suffers from information loss since it relies on highly abstracted tracks and discards dense scene evidence. In contrast, cross-level fusion (CLF), as shown in Fig. 1, has recently emerged as a promising direction that aligns heterogeneous modalities across abstraction levels by integrating object lists as structured priors into Transformer-based architectures, enabling deep interactions between external high-level hypotheses and low-level visual features [3]. While CLF improves cross-modal alignment, applying it to 3D MOT introduces additional challenges beyond object recognition, since tracking performance depends not only on per-frame localization accuracy but also on temporally consistent identity maintenance. In particular, object lists from external black-box sensors may be noisy or delayed in real deployment, which can destabilize query updates and amplify identity switches, making reliable association a key bottleneck for tracking based on cross-level fusion.

Motivated by these observations and aiming to address the aforementioned limitations, we propose a **PC-Track** (Prior-Conditioned 3D Multi-Object Tracking) framework by integrating the CLF strategy with the prior-conditioned tracking-by-query paradigm, thereby enhancing vision-centric perception systems for autonomous driving. Specifically, our contributions in this work are summarized as follows:

1) We develop an end-to-end tracking framework with cross-level fusion that incorporates sparse external object lists as structured priors into a Transformer architecture, and jointly fuses them with dense multi-view image features under the tracking-by-query formulation for 3D MOT through query denoising and attention guidance.
2) To improve robustness to noisy or delayed priors in real-world deployment, we revisit data association under prior-conditioned representations and introduce a fusion-aware association mechanism that explicitly integrates motion state priors with appearance interactions, thereby training the model to be resilient to the uncertainty of external object list inputs.
3) Extensive experiments on the nuScenes [4] benchmark with both real and synthetic object lists demonstrate that our method consistently outperforms state-of-the-art vision-centric baselines, and further indicate its potential to narrow the performance gap between purely vision-based trackers and LiDAR-centric systems.

[1] Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, Germany.
[2] Chair of Data Processing, Technical University of Munich, Germany.
[*] Corresponding author: jiajie.zhang@tum.de.

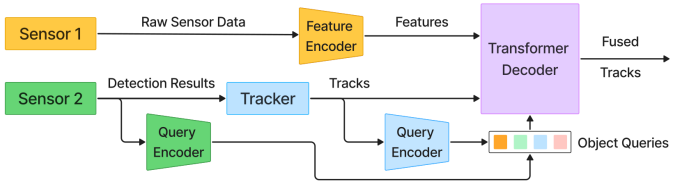

Fig. 1: Illustration of the cross-level fusion strategy for 3D multi-object tracking in autonomous driving.

II. RELATED WORK

A. Taxonomy of Sensor Fusion for the Object Tracking Task in Autonomous Driving

1) Conventional Strategies: Conventional sensor fusion strategies can be broadly categorized into data-level fusion and track-level fusion [1]. Data-level fusion integrates sensor information before tracking and includes early fusion on raw signals, intermediate (feature-level) fusion on processed representations (e.g., point clouds or feature maps), and late fusion on object-level outputs, such as object lists [2]. While early and intermediate fusion can preserve dense information, they are often costly and difficult to deploy due to heterogeneous sensor formats or privacy policies. Late fusion is ideal to implement, but typically makes tracking heavily dependent on the quality of per-sensor detections. In contrast, track-level fusion directly merges processed tracks from multiple sources, which is efficient and modular; yet it may suffer from reduced estimation accuracy since raw evidence is discarded [5].

2) Cross-Level Fusion: Cross-level fusion (CLF) has recently emerged as a new fashion in sensor fusion for autonomous driving, aiming to address the practical challenge of efficiently aligning heterogeneous sensor modalities during real-world deployment. Traditional fusion approaches normally operate on information representations at the same abstraction level [2]. However, in practice, the data formats accessible from different sensors can differ substantially, e.g., raw images from cameras versus object lists commonly produced by external black-box sensors or V2X modules. Cross-level fusion tackles this discrepancy by aligning data from different abstraction levels within a dedicated end-to-end mechanism, thereby enabling dense raw observations and sparse high-level prior object state information to complement each other and achieve implicit fusion [3].

B. Paradigm of Multi-Object Tracking in Autonomous Driving

1) Tracking by Detection: In the field of 3D MOT, tracking-by-detection is a classic and widely adopted paradigm. Under this paradigm, detection and tracking are two sequential yet independent processes: the detection outputs (e.g., bounding boxes, categories, and confidence scores) are associated over time to match current-frame detections with historical tracks, and are further combined with motion models for state update and track management [6]. Data association is typically formulated using cost matrices or matching algorithms, such as Kalman-filter-based motion prediction gating [7], and bipartite optimal matching [8] exemplified by the Hungarian methods. In scenarios involving occlusion, interaction, or visually similar targets, ReID or appearance embeddings are often incorporated to enhance identity consistency [9]. However, the tracking performance of such frameworks is largely dependent on detection accuracy and is not end-to-end controllable.

2) Tracking by Query: In recent years, the tracking-by-query paradigm has emerged as a promising alternative. Unlike the conventional sequential pipeline of tracking-by-detection, tracking-by-query leverages Transformer queries as object representations, unifying detection and tracking within a shared prediction framework and attention-based interaction mechanism. This paradigm revolves around two key frameworks:

First, temporal query propagation is employed to carry tracking information across time, e.g., by feeding queries corresponding to historical tracks into the Transformer decoder of the current frame, or by fusing historical features via a memory bank or temporal attention. DETR-style [10] query-based models [11], [12] adopt query propagation throughout video sequences to enforce cross-frame consistency. Second, set matching and identity binding are introduced to align prediction slots with target instances using strategies such as Hungarian matching, and to utilize the resulting alignment for track continuation. For instance, approaches for 3D and multi-camera scenarios, such as PF-Track [13] and DQTrack [14], incorporate 3D-to-2D projection relationships or decoupled queries into multi-view perception to maintain identities.

Under the tracking-by-query paradigm, detection and tracking share representations and feature extraction paths in an end-to-end manner, thereby reducing the reliance on heuristic post-hoc data association. However, both training and inference are more susceptible to issues such as query–target switching, query drifting under occlusion, and unstable attention focusing. Recent studies have therefore incorporated denoising-based training [15] and spatial attention modulation (e.g., SMCA-like location-prior guidance [16]) to improve convergence stability and the quality of feature alignment.

III. METHODOLOGY

Motivated by the above observations, we develop an end-to-end 3D MOT framework across modalities of different abstraction levels by combining cross-level fusion with tracking-by-query. Referring to the design in [3], we first inject the sparse object lists from black-box 3D sensors/V2X modules as structured priors into the Transformer decoder based on the PETRv2 [18] architecture, and fuse them with image features extracted from cameras for multimodal fusion. Specifically, CLF achieves effective cross-level alignment by integrating (i) denoising-like prior queries and (ii) location-prior-guided spatial attention modulation, as shown in Fig. 2.

In real-world deployment, external object lists may be noisy or delayed, which can destabilize query updates and exacerbate identity switches. To address this issue, we introduce a fusion-aware association mechanism that explicitly models both motion and appearance priors under prior-conditioned representations. Rather than relying on a single association

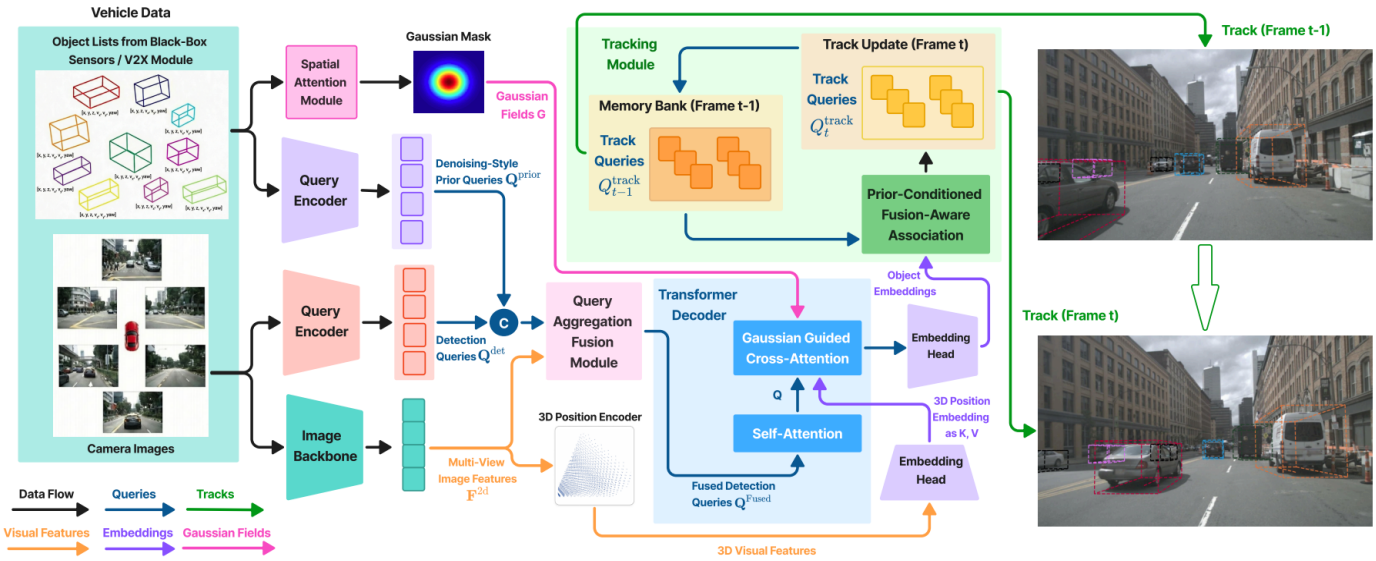

Fig. 2: Overview of the proposed PC-Track framework for 3D multi-object tracking in autonomous driving via cross-level fusion. External object lists from black-box sensors/V2X modules are encoded as structured priors and fused with multi-view image features through cross-level query aggregation and Gaussian-guided attention via a Transformer decoder. The fused representations are propagated through a memory-aware tracking module with fusion-aware association, where motion priors and appearance interactions jointly maintain identity consistency across frames. Colors of the output 3D bounding boxes are randomly assigned from the evaluation color pool [17] according to object IDs.

cue, the mechanism is designed to exploit their complementary roles: motion consistency provides stable geometric grounding, while appearance interactions preserve identity information under viewpoint changes and representation shifts, jointly contributing to robust tracking continuity.

A. Cross-Level Modality Alignment and Fusion

We consider cross-level fusion with two inputs at time t: multi-view images $\mathbf{I} = \{I_n \in \mathbb{R}^{3\times H_I \times W_I}\}_{n=1}^{N}$ and an external object list $\mathbf{O} = \{o_m\}_{m=1}^{M}$. Each object entry is a compact state hypothesis

$$o_m = \big(\mathbf{p}_m, \mathbf{s}_m, r_m, \mathbf{v}_m, c_m, \boldsymbol{\sigma}_m\big), \tag{1}$$

where $\mathbf{p}_m \in \mathbb{R}^3$ denotes 3D position, $\mathbf{s}_m \in \mathbb{R}^3$ for size, $r_m \in \mathbb{R}$ for yaw, $\mathbf{v}_m \in \mathbb{R}^2$ for planar velocity, $c_m \in \mathcal{C}$ for category, and $\boldsymbol{\sigma}_m$ represents estimation uncertainty.

Images are first processed by a backbone to obtain multi-view feature maps $\mathbf{F}^{2d} = \{F_n^{2d} \in \mathbb{R}^{C\times H_F \times W_F}\}_{n=1}^{N}$, which are then filled into a unified 3D embedding space via camera-aware projection and positional encoding. The resulting 3D embeddings serve as keys and values in the cross-attention module of the Transformer decoder based on the PETRv2 [18] architecture. Our cross-level fusion can be viewed as a mapping $\Psi : (\mathbf{I}, \mathbf{O}) \to \mathbf{H}$, producing decoder representations $\mathbf{H}$ used for 3D detection and subsequent tracking task.

1) Prior Query Initialization and Cross-Level Query Aggregation: We initialize a set of prior queries from the information given by object lists, which serve as denoising-style [15] queries that explicitly model noisy, ground-truth(GT)-like hypotheses from external black-box sensors, using their 3D centers $\mathbf{P} = \{\mathbf{p}_m\}_{m=1}^{M}$. Specifically,

$$\mathbf{Q}^{\text{prior}} = \text{MLP}\big(\phi_{\text{pos}}^{3d}(\mathbf{P})\big), \tag{2}$$

where $\phi_{\text{pos}}^{3d}(\cdot)$ is a 3D positional encoding and $\text{MLP}(\cdot)$ maps it to the query embedding dimension.

Initial prior queries are then refined by cross-attending to image features:

$$\mathbf{Q}^{\text{prior}\,*} = \text{CrossAttn}\big(\mathbf{Q}^{\text{prior}}, \mathbf{F}^{2d}\big). \tag{3}$$

This step refines noisy object-list hypotheses by letting prior queries attend to spatially aligned visual evidence, effectively correcting perturbed spatial centers through feature-wise verification. Functionally, it resembles denoising-based [15] training, in which the decoder processes GT-like but corrupted queries and is encouraged to map them back to consistent object representations, thereby improving robustness to prior noise under the tracking-by-query formulation.

Following the denoising fusion principle [3], we do not assume all prior queries from external object lists to be reliable; instead, similarity-weighted aggregation enables the model to selectively absorb consistent, denoised priors while suppressing noisy or misaligned ones. This mechanism acts as a soft reliability filter in the embedding space, allowing robust prior integration without enforcing hard selection. We therefore merge the prior-refined queries with detection queries $\mathbf{Q}^{\text{det}}$ from images via similarity-weighted aggregation:

$$\mathbf{Q}^{\text{Fused}} = \alpha\,\mathbf{Q}^{\text{det}} + \beta\,\text{sim}\big(\mathbf{Q}^{\text{det}}, \mathbf{Q}^{\text{prior}\,*}\big)\,\mathbf{Q}^{\text{prior}\,*}, \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is normalized cosine similarity and α, β are learnable scalars controlling the fusion ratio.

2) Shared Location-Prior Spatial Attention Guidance via Object Lists: To complement query aggregation, we explicitly inject spatial priors into cross-attention. Instead of adopting SMCA-style [16] query-specific one-to-one Gaussian masks,

we follow [3] and construct a multi-target shared Gaussian field G by aggregating all object-list hypotheses with view-specific projection. This shared field provides a soft spatial bias over the image plane, aligning attention with expected target regions and enabling location-prior-guided cross-attention.

We incorporate this Gaussian field into the attention logits:

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \frac{\log G}{\sqrt{d_k}}\right)\mathbf{V}, \quad (5)$$

where the $\log(\cdot)$ term prevents peaked Gaussian fields from dominating the attention distribution, and the $\sqrt{d_k}$ normalization follows standard scaled dot-product attention for stability.

B. Tracking via Prior-Conditioned Fusion-Aware Association

A key challenge in the tracking-by-query paradigm with external prior information is that these priors may be noisy or delayed [19], which can destabilize identity association even when query propagation itself remains reliable. To address this issue, we adopt the decoupled query propagation framework of [14] while redefining the association mechanism under prior-conditioned fusion-aware representations.

1) Prior Information as Embeddings: Our tracking module leverages a query-based tracking design that is decoupled from object detection. Specifically, persistent track queries are propagated across frames and matched with the current-frame detections via an explicit association matrix $\mathbf{P}^t$. Given fused detection queries Q^{fused} and propagated track queries Q^{track}, we first predict the 3D states and compute time-compensated tracking centers through velocity-based propagation. We then construct a pairwise motion prior using the Euclidean distance between the detection centers and the propagated tracking centers. This distance is embedded into motion features via a learnable distance encoder.

At time t, the decoder produces fusion-conditioned query embeddings $\tilde{\mathbf{q}}_i^t \in \mathbb{R}^C$, together with object states including BEV center $\mathbf{c}_i^t \in \mathbb{R}^2$, velocity $\mathbf{v}_i^t \in \mathbb{R}^2$, and confidence s_i^t. We maintain a set of track embeddings $\mathbf{u}_j^{t-1} \in \mathbb{R}^C$ and track states $(\mathbf{c}_j^{t-1}, \mathbf{v}_j^{t-1})$ from the previous frame. Following a conditioned association design, we encode two complementary priors – the **motion states** and the **appearances** of objects as embeddings:

$$\mathbf{m}_{ij}^t = \phi\big(\big\|\mathbf{c}_i^t - \hat{\mathbf{c}}_j^t\big\|_2\big), \qquad \mathbf{a}_{ij}^t = \mathbf{o}_i^t \odot \mathbf{u}_j^{t-1}, \quad (6)$$

where $\hat{\mathbf{c}}_j^t = \mathbf{c}_j^{t-1} + \Delta t\, \mathbf{v}_j^{t-1}$ is the velocity-propagated track center, $\phi(\cdot)$ maps a scalar motion discrepancy to a C-dimensional motion prior embedding, and $\mathbf{o}_i^t = f_{\text{obj}}(\tilde{\mathbf{q}}_i^t) \in \mathbb{R}^C$ is the object embeddings from fused detection representations. This formulation explicitly separates *motion consistency* and *appearance interaction* as two embedding-level priors.

2) Prior-Conditioned Fusion-Aware Association Modes: To systematically analyze data association under prior-conditioned representations, we provide a unified interface with three modes: (i) **motion-driven association** (motion prior dominant), where motion-geometric consistency provides the primary matching signal via negative time-compensated distance logits; (ii) **appearance-driven association** (embedding similarity dominant), which predicts association logits from appearance interaction features; and (iii) **unified association**, which combines motion priors and appearance interactions as described above. This design enables controlled ablations and reveals how cross-level priors modulate the relative importance of motion versus appearance cues in data association.

We fuse the two priors in embedding space:

$$\mathbf{f}_{ij}^t = \mathbf{m}_{ij}^t + \mathbf{a}_{ij}^t, \qquad \ell_{ij}^t = g(\mathbf{f}_{ij}^t), \quad (7)$$

where $g(\cdot)$ is a lightweight projection head producing a scalar logit ℓ_{ij}^t for matching detection i to track j. We then obtain a detection-to-track affinity matrix for explicit query association via row-wise normalization:

$$\mathbf{P}^t(i,j) = \mathrm{Softmax}_j\big(\ell_{ij}^t\big). \quad (8)$$

Given this association matrix $\mathbf{P}^t(i,j)$, matched detection–track pairs are obtained via Hungarian assignment, and the corresponding track queries are refreshed using the fused embeddings of the matched detections. Specifically, for each associated pair, the track query inherits the updated appearance embedding and velocity-aware state from the current frame, while unmatched tracks follow the standard survival and birth rules of the baseline tracker [14]. This preserves the original query propagation protocol while allowing our fusion-aware association to modulate only the matching stage.

3) Training Objective: We optimize the model end-to-end with standard detection and tracking supervision, and additionally regularize the association logits to improve the stability of prior-conditioned matching (refer to Subsec. IV-B). The overall loss is:

$$\mathcal{L} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{track}}\mathcal{L}_{\text{track}} + \lambda_{\text{asso}}\mathcal{L}_{\text{asso}}, \quad (9)$$

where $\mathcal{L}_{\text{det}}$ includes classification and 3D box regression for decoder outputs, $\mathcal{L}_{\text{track}}$ supervises temporal consistency of track states, and $\mathcal{L}_{\text{asso}}$ is a cross-entropy loss defined on the association matrix $\mathbf{P}^t$ using the ground-truth identity correspondence between detections and active tracks; λ_{det}, λ_{track} and λ_{asso} denote the loss balancing weights, all of which are set to 1.0. This loss encourages the fusion-conditioned representations to produce stable and discriminative association scores under noisy/delayed priors, while keeping the inference pipeline unchanged from the baseline trackers [14], [19].

IV. Experiment and Evaluation

A. Dataset

Public autonomous driving datasets seldom provide object lists as a standalone modality that is directly plug-and-play for cross-level fusion. Several V2X datasets (e.g., [20], [21]) provide synchronized image frames and 3D annotations in cooperative driving (vehicle–infrastructure) settings; the required object lists can be theoretically constructed by converting the available 3D ground-truth boxes (from heterogeneous vehicle- or infrastructure-side detections/messages) into a unified list format and aligning them to the ego-vehicle coordinate frame. However, such conversion is more involved due to additional calibration chains and frame alignment between heterogeneous

agents. In contrast, the nuScenes dataset [4] is a large-scale public benchmark for autonomous driving and has been widely used for perception tasks such as 3D MOT. It provides synchronized data from six calibrated surround-view cameras, capturing seven movable object categories relevant to MOT with full 360° coverage. nuScenes offers a standardized multi-camera benchmark with well-established tracking protocols, making it well-suited for validating the effectiveness of our proposed method in a reproducible manner.

Building on the above considerations, all experiments in this work are conducted on the nuScenes dataset.

B. Experimental Setup

1) Object List Construction and Data Preparation: To emulate realistic black-box priors, we follow the pseudo object list generation (POLG) strategy of our CLF baseline [19] and apply controlled perturbations to the generated object lists during training and evaluation processes.

Specifically, following the best experimental setting in [19] that balances realism and robustness, we adopt the POLG strategy implemented under the MMDetection3D framework [17] to construct the required synthetic object lists from the nuScenes ground truth (GT), serving as a proxy for detections and tracks produced by external sensors. Concretely, for each object list sampled from GT, we independently perturb its position, velocity, and size with Gaussian noise whose standard deviation is randomly sampled from [0, 0.06]. In addition, we introduce random variations with an object drop rate of 0.2, a false positive rate of 0.1, and a label change rate of 0.3, so as to better emulate real perception outputs from external black-box sensors and make our experiments more realistic [3].

2) Experimental Implementation: Following the baseline tracker [19], we adopt its stage-wise training strategy: Specifically, we initialize our PC-Track from its detection-pretrained model and then fine-tune our entire network end-to-end for 5 epochs using the joint training objective described in Subsection III-B. All experiments are conducted on a single NVIDIA RTX 3090 GPU (24GB VRAM). Due to memory constraints, both training and evaluation are performed with a batch size of 1. V2-99 [22], which achieves the best performance on the same single-GPU setting, is adopted as the backbone for visual feature extraction. The input camera images are resized to the maximum resolution supported by a single RTX 3090, i.e., 800 × 320. During fine-tuning, we set the initial learning rate to 1e-5 and use the AdamW optimizer. For fair comparison, we keep the number of queries and the tracking threshold parameters identical to those used in our baselines [14], [19].

C. Quantitative Evaluation

1) Comparison with Baselines: Table I summarizes the overall performance of representative tracking baselines on the nuScenes validation set. Our PC-Track achieves the best performance among camera-centric tracking-by-query approaches using the same backbone. Our method also outperforms the CLF tracker, in both state estimation quality and association consistency (AMOTA improves from 0.593 to 0.611, a relative gain of over 3%, also with improvements in AMOTP and Recall). Meanwhile, it effectively reduces IDS by approximately 5% (1997→1897). Notably, compared with DQ-Track, which follows a similar tracking paradigm, our method yields a substantial improvement (AMOTA: 0.439→0.611). Although IDS may increase due to more aggressive track births and potential representation distribution shifts introduced by object-list priors, the overall results indicate that the proposed association mechanism enables competitive temporal identity maintenance. We also report a LiDAR-centric tracker (CenterPoint [25]), which has been widely deployed in commercial systems, as a reference upper bound, highlighting that incorporating object lists can significantly narrow the gap between pure-vision tracking systems and LiDAR-dominant perception.

2) Class-Wise Evaluation: Table II further presents the class-wise tracking performance of our approach. The results indicate that the proposed tracking framework performs particularly well on pedestrians and cars, achieving AMOTA scores of 0.830 and 0.790, respectively, together with high recalls. These dominant classes in the dataset involve dense traffic interactions, suggesting that the proposed fusion-aware association effectively stabilizes identity maintenance in crowded scenes. For medium-scale objects such as bicycles and motorcycles, the tracker maintains competitive results (AMOTA of 0.717 and 0.652, respectively). Larger vehicles such as buses and trucks exhibit moderate AMOTA performance (0.597 and 0.518), where the limitation is likely related to long-term occlusion and sparse training samples. Trailers remain the most challenging (AMOTA 0.173), which is consistent with the difficulty of nuScenes due to extreme aspect ratios and rare occurrences. Overall, the class-wise distribution suggests that the advantages of the proposed framework generalize across diverse motion patterns and object scales in the 3D MOT task.

3) Ablation Study for Track Association Modes: Table III investigates the impact of different prior-conditioned association modes of our PC-Track on 3D MOT performance. The results clearly indicate that both motion-driven and appearance-driven associations provide important prior cues for 3D MOT, and that each mode alone can reasonably support the tracking task. Motion-only association offers slightly stronger geometric consistency (AMOTA 0.590), while appearance-only association remains competitive (AMOTA 0.568), suggesting that the learned embeddings still preserve identity information under cross-level fusion despite viewpoint changes and representation shifts. When the two priors are combined, the performance reaches the optimum (AMOTA 0.611), demonstrating a clear complementary effect: motion priors provide stable spatiotemporal grounding, whereas appearance interactions refine identity matching in uncertain scenarios, such as partial occlusions. The improvement across all metrics indicates that robust association in prior-conditioned query-based tracking benefits from jointly modeling motion-geometric consistency and embedding similarity, rather than relying on a single cue.

4) Control Study for Real and Synthetic Object Lists: Table IV further evaluates the robustness of PC-Track un-

Model	Modality	Backbone	Image Resolution	AMOTA ↑	AMOTP ↓	MOTA ↑	Recall ↑	IDS ↓
PF-Track [13]	Camera	V2-99	800 x 320	0.408	1.343	0.376	0.507	**166**
Focal-PETR* [23]	Camera	V2-99	800 x 320	0.421	1.260	0.368	0.519	2739
DQTrack [14]	Camera	V2-99	800 x 320	0.439	1.286	0.381	0.556	1445
StreamPETR* [24]	Camera	V2-99	800 x 320	0.522	1.127	0.457	0.612	1321
CLF-DQTrack [19]	Camera + Object List	V2-99	800 x 320	0.593	1.100	0.547	0.680	1997
PC-Track (Ours)	Camera + Object List	V2-99	800 x 320	**0.611**	**1.075**	**0.551**	**0.685**	1897
CenterPoint [25]	LiDAR	VoxelNet	–	0.637	0.606	–	–	760

TABLE I: Comparison of 3D multi-object tracking performance across different trackers on the nuScenes [4] validation set. * denotes integrating CenterTrack [26] with the corresponding detectors for tracking.

Object Class	AMOTA ↑	AMOTP ↓	MOTA ↑	Recall ↑
pedestrian	0.830	0.784	0.752	0.872
bicycle	0.717	0.887	0.669	0.777
motorcycle	0.652	1.022	0.604	0.709
car	0.790	0.862	0.707	0.844
bus	0.597	1.229	0.551	0.690
truck	0.518	1.157	0.419	0.625
trailer	0.173	1.587	0.155	0.280

TABLE II: Class-wise evaluation results of the proposed PC-Track framework on the nuScenes [4] validation set.

Association Mode		Tracking Metrics			
motion-driven	appearance-driven	AMOTA↑	AMOTP↓	MOTA↑	Recall↑
✓		0.590	1.076	0.524	0.678
	✓	0.568	1.119	0.521	0.665
✓	✓	0.611	1.075	0.551	0.685

TABLE III: Ablation study of the proposed approach PC-Track under different prior-conditioned association modes (refer to Subsection III-B) on the nuScenes [4] validation set.

der different sources of external object lists, including lists converted from LiDAR tracker outputs (CenterPoint [25] and PointPillars [27]) and GT-derived synthetic lists constructed via POLG [3]. Across all sources, PC-Track consistently outperforms the CLF tracking baseline [19], indicating stronger robustness to variations in prior quality and distribution. The results suggest that our PC-Track can more reliably exploit external cues under heterogeneous and imperfect prior inputs.

5) Real-Time Efficiency: Following the configurations in Subsec. IV-B, Tables I and IV, we measure the end-to-end inference speed on a single RTX 3090 GPU. When using pseudo object lists generated from nuScenes-GT via the POLG strategy [3], which introduces noise, object drop, false positives, and label perturbations, our PC-Track achieves an average inference speed of 5.21 FPS, higher than 5.17 FPS of CLF-DQTrack [19] under the same setting. When using the object lists obtained from the CenterPoint [25] tracker, the inference speed increases to 6.16 FPS. We attribute the speed gap to the higher noise level and false positives in the pseudo object lists, which increase the number of effective prior queries and the matching workload during fusion-aware association. Nevertheless, compared with the baselines, our method demonstrates greater potential to meet the real-time requirements of autonomous driving tasks when combined with subsequent engineering-level optimizations.

	CLF-DQTrack [19]		PC-Track (Ours)	
Object List Source	AMOTA ↑	AMOTP ↓	AMOTA ↑	AMOTP ↓
CenterPoint [25]	0.505	1.212	**0.549**	**1.125**
PointPillars [27]	0.395	1.392	**0.421**	**1.350**
nuScenes-GT by POLG [3]	0.593	1.100	**0.611**	**1.075**

TABLE IV: Comparison between our CLF baseline and the proposed PC-Track approach under different object list sources, evaluated on the nuScenes [4] validation set.

D. Qualitative Analysis

Fig. 3 demonstrates a qualitative comparison of 3D multi-object tracking performance between the proposed PC-Track framework and the baseline across three representative driving scenarios, highlighting the key improvements introduced by our proposed approach.

First, under adverse weather conditions with rain and visual blur, our method successfully tracks a truck whose only rear is visible and whose appearance is visually similar to the sky (top row), whereas the baseline fails to maintain a valid track for this object. Second, in dense scenes, PC-Track not only achieves accurate localizations but also effectively reduces tracking hallucinations exhibited by the baseline, such as the false pedestrian detection shown on the right side of the middle-row baseline result. Finally, the bottom row shows images captured by a rear-view camera, where objects on the left sides are partially occluded. In this challenging scenario, PC-Track demonstrates accurate and consistent tracking by jointly leveraging motion priors and appearance interactions under the proposed prior-conditioned formulation.

The above evidence further highlights the robustness of PC-Track across diverse and challenging driving scenarios.

V. Conclusion and Future Work

In this work, we introduce PC-Track, a prior-conditioned 3D multi-object tracking framework with cross-level fusion for enhancing vision-centric perception systems in autonomous driving. The proposed framework incorporates object lists from external black-box sensors or V2X modules as structured high-level priors into a Transformer architecture, and performs joint fusion with dense multi-view image features under a tracking-by-query paradigm via query denoising and explicit attention guidance. To further address the instability of external priors in real-world deployment, we introduce a prior-conditioned, fusion-aware association mechanism that combines motion priors with appearance interactions to ensure stable track continuation. Experiments on the nuScenes [4] benchmark with both real and synthetic object lists demonstrate that our method achieves significant improvements in tracking quality over the strong cross-level fusion tracking

Fig. 3: Qualitative comparison of 3D multi-object tracking results between the baseline (left column) and the proposed PC-Track (right column) on the nuScenes [4] validation set across three representative driving scenarios. Top: Accurate tracking of a large truck under rainy, visually blurred conditions. Middle: Reduced false positives in dense pedestrian scenes. Bottom: Continuous and geometrically accurate tracking of partially occluded objects (truck, car, trailer). Colors of the 3D bounding boxes are randomly assigned from the evaluation color pool [17] according to object IDs.

baseline and state-of-the-art vision-centric trackers. Future work will investigate how to fully exploit temporal information as a stronger conditioned prior to better handle the uncertainty of object lists in practice, and extend the proposed framework to real-world cooperative perception datasets with multi-agent calibration uncertainty and asynchronous communication.

REFERENCES

[1] C. Wei, Z. Qin, Z. Zhang, G. Wu, and M. J. Barth, "Integrating multi-modal sensors: a review of fusion techniques for intelligent vehicles," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025, pp. 1817–1824.

[2] H. Qian, M. Wang, M. Zhu, and H. Wang, "A review of multi-sensor fusion in autonomous driving," *Sensors*, vol. 25, no. 19, p. 6033, 2025.

[3] X. Liu, J. Zhang, and H. Shen, "Cross-level sensor fusion with object lists via transformer for 3d object detection," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025, pp. 236–241.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[5] B. Fan, Y. Jia, and S. Zhou, "Sensor selection for multi-level collaborative perception with covariance intersection," in *2025 IEEE 101st Vehicular Technology Conference (VTC2025)*. IEEE, 2025, pp. 1–6.

[6] Z. Peng, W. Liu, Z. Ning, Q. Zhao, and J. Hu, "3d multi-object tracking in autonomous driving: A survey," in *2024 36th Chinese Control and Decision Conference (CCDC)*. IEEE, 2024, pp. 4964–4971.

[7] H.-K. Chiu, C.-Y. Wang, M.-H. Chen, and S. F. Smith, "Probabilistic 3d multi-object cooperative tracking for autonomous driving via differentiable multi-sensor kalman filter," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 18 458–18 464.

[8] S. Sun, C. Shi, C. Wang, Q. Zhou, R. Sun, B. Xiao, Y. Ding, and G. Xi, "Intra-frame graph structure and inter-frame bipartite graph matching with reid-based occlusion resilience for point cloud multi-object tracking," *Electronics*, vol. 13, no. 15, p. 2968, 2024.

[9] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "Strongsort: Make deepsort great again," *IEEE Transactions on Multimedia*, vol. 25, pp. 8725–8737, 2023.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, and A. Kirillov, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[11] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[12] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking with transformer," in *European conference on computer vision*. Springer, 2022, pp. 659–675.

[13] Z. Pang, J. Li, P. Tokmakov, D. Chen, S. Zagoruyko, and Y.-X. Wang, "Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 928–17 938.

[14] Y. Li, Z. Yu, J. Philion, A. Anandkumar, S. Fidler, J. Jia, and J. Alvarez, "End-to-end 3d tracking with decoupled queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 302–18 311.

[15] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.

[16] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of detr with spatially modulated co-attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3621–3630.

[17] M. Contributors, "Mmdetection3d: Openmmlab next-generation platform for general 3d object detection," 2020.

[18] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3262–3272.

[19] X. Liu, X. Wang, and H. Shen, "Cross-level fusion: Integrating object lists with raw sensor data for 3d object tracking," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 10 124–10 130.

[20] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.

[21] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 22 668–22 677.

[22] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 906–13 915.

[23] S. Wang, X. Jiang, and Y. Li, "Focal-petr: Embracing foreground for efficient multi-camera 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1481–1489, 2023.

[24] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3621–3631.

[25] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[26] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European conference on computer vision*. Springer, 2020, pp. 474–490.

[27] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.