2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

M^{3}CAD: Towards Generic Cooperative Autonomous Driving Benchmark

Morui Zhu[1], Yongqi Zhu[1], Yihao Zhu[1], Qi Chen[2], Deyuan Qu[2], Song Fu[1], Qing Yang[1†]

***Abstract*— We introduce M^{3}CAD, a comprehensive benchmark designed to advance research in generic cooperative autonomous driving. M^{3}CAD comprises 204 sequences with 30,000 frames. Each sequence includes data from multiple vehicles and different types of sensors, e.g., LiDAR point clouds, RGB images, and GPS/IMU, supporting a variety of autonomous driving tasks, including object detection and tracking, mapping, motion forecasting, occupancy prediction, and path planning. This rich multimodal setup enables M^{3}CAD to support both single-vehicle and multi-vehicle cooperative autonomous driving research. To the best of our knowledge, M^{3}CAD is the most complete benchmark specifically designed for cooperative, multi-task autonomous driving research. To test its effectiveness, we use M^{3}CAD to evaluate both state-of-the-art single-vehicle and cooperative driving solutions, setting baseline performance results. Since most existing cooperative perception methods focus on merging features but often ignore network bandwidth requirements, we propose a new multi-level fusion approach which adaptively balances communication efficiency and perception accuracy based on the current network conditions. We release M^{3}CAD, along with the baseline models and evaluation results, to support the development of robust cooperative autonomous driving systems. All resources will be made publicly available on our project webpage.**

I. INTRODUCTION

Cooperative autonomous driving (CAD) refers to the paradigm where multiple autonomous vehicles communicate and coordinate with each other to enhance driving efficiency and safety. To advance research in this domain, there is an urgent need for a comprehensive benchmark that enables evaluation and comparison of CAD algorithms and solutions. To fill this gap, we introduce a benchmark with the following key features: (1) It provides scenarios involving multiple vehicles, focusing on cases where they can collaborate with each other. (2) It supports research on a variety of cooperative driving tasks, e.g., cooperative perception, collaborative mapping, joint motion forecasting, and coordinated path planning. (3) It offers rich diversity, including different sensor types, driving environments, and ego vehicle trajectories.

A. Limitations of Prior Works

Recent advancements in autonomous driving have demonstrated the effectiveness of end-to-end planning frameworks, e.g., UniAD [1], VAD [2], and Drive-VM [3]. Meanwhile, CAD has gained traction, with most efforts focusing on perception [4], [5]. Comprehensive study of cooperative autonomy, however, is still limited by several key factors. First, existing real-world datasets such as KITTI [6], nuScenes [7], and DAIR-V2X [8] are either designed for single-vehicle settings or constrained by limited sensor setups and small-scale collaborations. For instance, the V2V4Real [9] dataset involves only two vehicles, each with just two cameras and one LiDAR. Such limitations restrict their scalability and hinder comprehensive research on cooperative autonomous driving across multiple tasks. Second, existing cooperative perception methods mostly focus on bird's-eye-view (BEV) feature fusion [10], [11], where dense feature maps are transmitted and aligned across vehicles. While effective, this strategy incurs high communication costs and potential redundancy, limiting the applicability under bandwidth constraints. Third, existing autonomous driving datasets [12], [13] created in simulation do not provide a clear pathway for transferring to real-world benchmarks. As a result, current cooperative methods validated solely in simulation cannot be reliably assessed for their effectiveness in real-world scenarios. This sim-to-real gap not only hinders fair evaluation but also limits the practical deployment of cooperative solutions.

B. Contributions

M^{3}CAD. To overcome these issues, we introduce a novel benchmark designed specifically to support research in **M**ulti-vehicle, **M**ulti-task, and **M**ulti-modality **C**ooperative **A**utonomous **D**riving (**M^{3}CAD**). By leveraging the advanced rendering capabilities of the recently-released Unreal Engine 5 (UE5) in CARLA [14], we can simulate realistic multi-vehicle interactions within diverse driving scenarios. M^{3}CAD comprises 204 sequences in total, offering over 30k frames and more than 267k annotated instances, along with the ground truth data for multiple autonomous driving tasks. Each sequence includes 10-60 collaborative vehicles with their precise location and trajectory information at each timestamp, as well as the map and occupancy details. This comprehensive dataset supports a wide range of autonomous driving tasks, including object detection and tracking, mapping, motion forecasting, occupancy, and path planning, addressing the limitations in the single-vehicle end-to-end benchmarks (e.g., nuScenes [7]) and cooperative non-end-to-end datasets (e.g., OPV2V [12]). To the best of our knowledge, M^{3}CAD is currently the most comprehensive benchmark for both single-vehicle and cooperative autonomous driving research, while supporting more realistic vehicle movements and interactions in complex environments.

Multi-Level Fusion. While M^{3}CAD provides the foundation for evaluating cooperative perception in complex scenarios, existing methods are mostly based on dense BEV feature

[1]University of North Texas
[2]Toyota InfoTech Labs
[†]For correspondence and questions: qing.yang@unt.edu

TABLE I: Detailed comparison of M^{3}CAD with existing benchmarks. Traffic: realistic and interactive traffic such as merging, lane crossing and traffic jam. Human-like: human like non-playable characters (NPCs) instead of rule based simulation. MP: Mapping, MF: Motion Forecasting, OCC: Occupancy Prediction, PP: Path Planning. Since perception tasks such as detection and tracking are standard components across datasets, they are not included in this comparison.

			Diverse Scenarios			Diverse Cooperation		Multiple Tasks			
Benchmarks	**3D Labels**	**Source**	**Night / Rain**	**Human-like**	**Pedestrian**	**Coop. Vehicles**	**Coop. Range (m)**	**MP**	**MF**	**OCC**	**PP**
nuScenes [7]	1.4M	real	✓/✓	✓	✓	N/A	N/A	✓	✓	✓	✓
OPV2V [12]	232K	sim	×/×	×	×	2–7	120	✓	×	×	×
V2X-Sim [13]	26.2K	sim	×/×	×	×	2–5	70	×	×	×	×
V2V4Real [9]	240K	real	×/×	✓	✓	2	200	×	✓	×	×
V2X-Seq [15]	10.45K	real	✓/×	✓	✓	2–4	280	✓	✓	×	×
TUMTraf V2X [16]	29.38K	real	✓/×	✓	✓	2–4	200	✓	✓	×	×
V2X-Real [17]	1.2M	real	×/×	✓	✓	4	–	×	✓	×	×
DAIR-V2X [8]	464K	real	✓/✓	✓	✓	2	200	×	×	×	×
V2XSet [18]	230K	sim	×/×	✓	✓	2-7	280	×	×	×	×
V2X-Radar [19]	350K	real	✓/✓	✓	✓	2	200	×	×	×	×
WHALES [20]	2.01M	sim	×/×	×	✓	8.4	200	×	×	×	×
M^{3}CAD (ours)	**267K**	**sim**	✓/✓	✓	✓	**10-60**	**200**	✓	✓	✓	✓

fusion [10], [11], resulting in high communication costs and poor scalability. To close this gap, we propose a multi-level fusion method that adaptively balances communication efficiency and perception accuracy. Specifically, our framework explores three complementary strategies: *BEV Feature Fusion*, which fuses dense feature maps to provide spatial information; *Query Fusion*, which fuses compact, trajectory-aware features that preserve temporal and sequence information; and *Reference Points Fusion*, which shares only sparse spatial to directly guide attention. By dynamically selecting the appropriate fusion level according to network conditions and system requirements, our framework enables scalable and efficient cooperative perception, maintaining strong accuracy while significantly reducing bandwidth consumption.

Transfer to Real-World Benchmark. While multi-level fusion works well in simulations, it's unclear how it performs on real-world data. To explore this, we align M^{3}CAD with the nuScenes dataset and the cross-domain performance. Our results show that UniAD [1], when pre-trained on M^{3}CAD, can be effectively fine-tuned with only 10% of nuScenes data, leading to significant performance improvements in real-world settings.

Extensive Evaluations. Finally, we conduct comprehensive experiments to validate our framework. These include: (1) systematic benchmarking across multiple cooperative driving tasks, (2) analysis of communication bandwidth requirements to assess the efficiency of the proposed multi-level fusion, and (3) robustness tests under ego-pose and sensor noises. These evaluations demonstrate the effectiveness, robustness, and generalizability of our M^{3}CAD benchmark.

II. Related Works

Benchmarks for A multi-Task Autonomous Driving. Multi-task autonomous driving framework jointly optimizes multiple modules for individual tasks while prioritizing the ultimate planning objective, benefiting from enhanced safety and interpretability [21]. To support multi-task research, the nuScenes benchmark [7] is proposed and is recognized as the most valuable resource in autonomous driving research, due to its comprehensive and high-quality data that encompasses a wide range of multi-task driving scenarios. Several end-to-end autonomous driving solutions have been evaluated using the nuScenes dataset, e.g., DiffStack [22], MP3 [23], P3 [24], ST-P3 [25], and VAD [2]. Among them, a notable example is UniAD [1], a unified framework that seamlessly integrates perception, prediction, and planning into a single architecture. While multi-task end-to-end autonomous driving frameworks have shown promising performance by jointly optimizing all modules, there remains a significant gap in exploring collaborative strategies among multiple vehicles for autonomous driving tasks.

Benchmarks for Cooperative Perception. Despite the progress in end-to-end autonomous driving, multi-vehicle collaboration has primarily focused on cooperative perception, leading this field underexplored. To support this direction, several benchmarks have been introduced, including OPV2V [12], V2V4Real [9], V2X-Seq [15], TUMTraf-V2X [16], DAIR-V2X-C [8], V2X-Set [18], V2X-Radar [19], and WHALES [20]. However, these datasets are largely centered around perception tasks (particularly object detection), which limits their utility for more comprehensive autonomous driving challenges such as object tracking, motion forecasting, and ultimately, path planning. Table I presents a comparative summary of existing benchmarks and our proposed M^{3}CAD, highlighting the differences in scale of dataset, diversity of scenarios, cooperation settings, and types of supported autonomous driving tasks.

III. The M^{3}CAD Benchmark

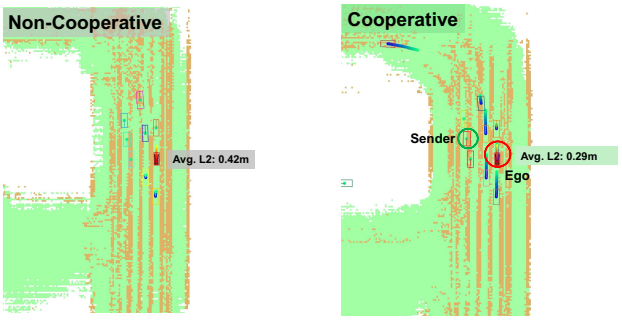
M^{3}CAD is designed to serve as a versatile platform to facilitate various cooperative driving tasks. To achieve this goal, we deployed multiple vehicles in the Town01, Town02, Town03, Town04, and Town10HD maps in CARLA. Each vehicle is equipped with six 1920×1080-pixel (110° FOV) cameras, a 64-beam LiDAR, and GPS/IMU sensors to collect

multi-modality sensor data during daytime, nighttime, and various weather conditions.

Annotation. We annotate the collected data based on the nuScenes format [7], enabling research on multi-vehicle, multi-task and multi-modality autonomous driving. To support research on cooperative driving tasks, we provide in M^{3}CAD comprehensive annotations for object detection, tracking, motion prediction, occupancy forecasting, and path planning. All objects located within the $102.4m \times 102.4m$ BEV range centered on the ego vehicle are annotated, including attributes: location, speed, bounding box center, and extent (length, width, height). To support effective collaboration between vehicles, M^{3}CAD also provides transformation matrices that align the ego vehicle with any collaborating vehicles, enabling accurate information fusion.

For object detection and tracking tasks, only objects within the BEV range are considered to reflect realistic constraints. The dataset includes motion histories and future trajectories for vehicles, as well as binary BEV occupancy maps generated from LiDAR point clouds to support occupancy forecasting. Vehicle trajectories are also recorded to assist the planning task. Additionally, the semantic map contains multiple layers, e.g., drivable areas, lane dividers, and road dividers, all aligned with the CARLA global coordinate system and structured according to the nuScenes specification.

Dataset Split and Multiple Tasks. The M^{3}CAD dataset is divided into training, validation, and test subsets using a 70/15/15 split. The rich data generated by these vehicles allows for the exploration of various tasks. These tasks include object detection, object tracking, mapping, motion forecasting, occupancy prediction, and, critically, path planning. Fig. 1 shows the results of UniAD's different tasks performed on the M^{3}CAD dataset.



(a) Non-Cooperative (b) Cooperative

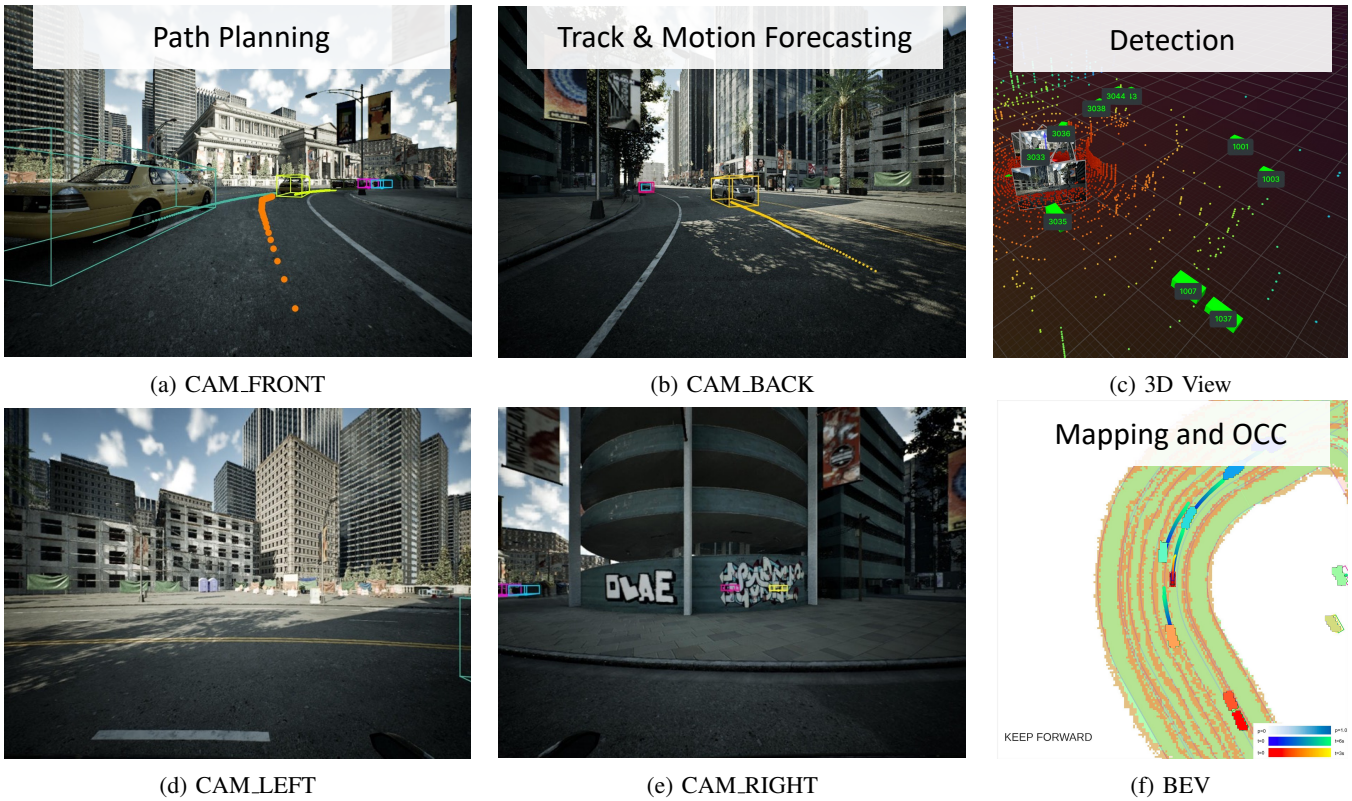
Fig. 2: Qualitative comparison between cooperative and non-cooperative path planning. All visualizations are shown in the BEV, with each vehicle depicted as a uniquely colored box. The blue-green curves represent the predicted trajectories of vehicles over the next 6 seconds, showing only the top-1 predicted trajectories here. The red-yellow curves present the planned paths of the ego vehicle within a 3-second window. In the underlying semantic map, green areas indicate drivable regions and yellow dots mark lane dividers.

One of the prominent features of M^{3}CAD is its support for multi-vehicle collaboration, enabling cooperative path planning to enhance driving safety. Although cooperative object detection has been extensively studied [26], [12], [10], [27], [28], [29], other tasks such as cooperative mapping, motion forecasting, occupancy prediction, and planning remain much less explored. Recent studies have explored cooperative motion planning [30], [31]; however, a comprehensive understanding of how different tasks interact and depend on each other is still missing. As shown in Fig. 2, with the assistance from the sender vehicle through a BEV fusion method, the ego vehicle can obtain a better perception of the surrounding environment, allowing it to follow a trajectory much closer to the ground truth (with a L2 error of $0.29m$). Without M^{3}CAD, it would be extremely difficult, if not impossible, to systematically evaluate the benefits and trade-offs of different tasks within cooperative autonomous driving settings.

IV. Multi-level Fusion

To better understand how cooperative perception improves the ego vehicle's performance, we now take a closer look at what information can be shared between CAVs and how this sharing happens. Unlike existing approaches, which are commonly divided into high-level, intermediate-level [26], [32], [33], and low-level sharing [34], we propose a unified multi-level information sharing framework to support both high- and intermediate-level cooperative perception. As shown in Fig. 3, the input to our framework can be multi-modal sensor data. In this work, we instantiate it with camera inputs, which will be processed by various perception modules e.g., BEVFormer [35] or MOTR [36], to obtain different types of data: BEV features, query features, and reference points. These outputs are then passed to the prediction model to generate the final path planning results.

Relying on the BEVFormer, BEV features are extracted to represent the spatial and semantic layout of the environment from a bird's-eye view. These features can be shared among vehicles to realize cooperative perception [10], [11]. Although effective, it suffers from high communication cost, limiting its real-world applications. To overcome this issue, we explore another way to represent the information, known as query features [37], [38]. Modern transformer-based perception modules, e.g., UniAD-TrackFormer [1], can generate detailed track-level outputs, including queries, reference points, object indices, confidence scores, and predicted boxes. Among these, queries are the most informative while remaining compact. They include not just spatial information, but also capture important details about time, motion, and object identity, making them especially useful for tracking tasks. Queries from different vehicles can be fused to realize cooperative perception, which uses much less bandwidth than fusing BEV features. To further reduce bandwidth requirements, we suggest that vehicles share only reference point information to achieve cooperative perception.



(a) CAM_FRONT (b) CAM_BACK (c) 3D View

(d) CAM_LEFT (e) CAM_RIGHT (f) BEV

Fig. 1: Illustrations of various autonomous driving tasks using the M^{3}CAD dataset. (a) Demonstrates the path planning (PP) results, where the ego vehicle's predicted trajectory is represented by a dotted line. (b) Shows object tracking (OT) and motion forecasting (MF) results where dotted lines represent predicted trajectories of other vehicles. (c) Presents object detection (OD) results in 3D space. (f) Depicts mapping (MP) and occupancy prediction (OCC) results.

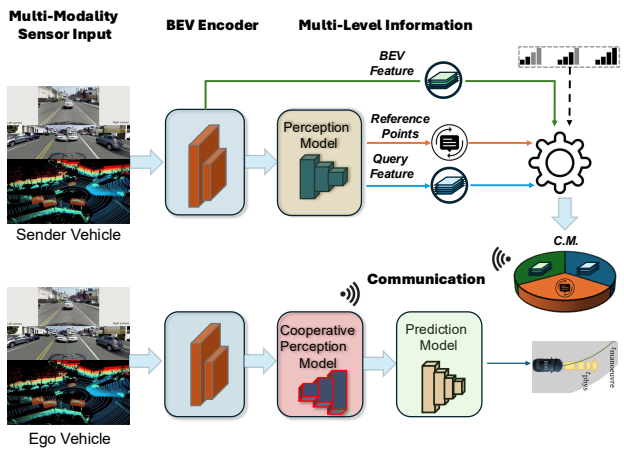


Fig. 3: Multi-level cooperative perception on CAVs. Sender vehicles generate BEV features, queries, and reference points, which are selectively packaged into a cooperative message (C.M.) based on bandwidth and task requirements. The ego vehicle receives these messages and performs multi-level fusion to achieve cooperative perception.

A. BEV Feature Fusion (BFF)

In this paradigm, each vehicle transmits its own BEV feature maps to the ego vehicle for alignment and fusion. Assuming vehicle v_j is selected for cooperation, it needs to send its BEV feature $\mathcal{F}j$ to the ego vehicle, along with its pose $\mathcal{P}j$ and location $\mathcal{L}j$ information. Based on the affine transformation function $T_j = \Psi(\mathcal{P}j, \mathcal{L}j)$, the transformation matrix T_j can be obtained. Then, the ego vehicle transforms the received feature map to align with its perspective, resulting in a transformed feature map $\mathcal{F}j^{'}$. The transformed $\mathcal{F}j^{'}$ is then fused into the ego's as follows:

$$\mathcal{F}^* = \Phi(\mathcal{F}e \parallel \mathcal{F}j^{'}) \in \mathbb{R}^{H \times W} \quad (1)$$

where $\mathcal{F}e$ is the ego vehicle's BEV feature, H and W are the height and width of the feature map, and $\Phi(\cdot)$ denotes the fusion function. A variety of methods can implement the function Φ, e.g., those in F-Cooper [26], Attentive Fusion [12], CoBEVT [10], V2VNet [27], Where2Comm [39], V2VAM [40], V2X-ViT [18], CoAlign [41], and SiCP [28]. Within the M^{3}CAD benchmarks, several above-mentioned fusion strategies are provided, offering the flexibility of selecting or extending them with new fusion methods.

B. Query Feature Fusion (QFF)

While BFF is effective, it requires sending a lot of data. To reduce this cost, we propose query feature fusion (QFF), which uses smaller query features to achieve cooperative perception. Different perception modules produce queries with different meanings, e.g., object queries generated by DETR [42] represent potential objects in the scene, while track queries produce by UniAD-TrackFormer represent

tracked objects across time. Here, we use track query as an example to illustrate how cooperative perception can be achieved on the tracking task.

In UniAD-TrackFormer, each tracked object is represented by a track query, which includes a learnable embedding, reference points, a class score, and a bounding box. These queries are passed from frame to frame using query interactions, helping the model keep track of the same object over time. To enable track query fusion, we first align the sender's track instances into the ego's coordinate system, using extrinsic calibration information. Next, we combine the matching query embeddings using a Multi-Layer Perceptron (MLP):

$$\tilde{q}_i = \text{MLP}\big([q_{i,\text{ego}} \,\|\, q_{j,\text{sender}\to ego}]\big), \tag{1}$$

where $[\cdot\|\cdot]$ denotes concatenation. Finally, the fused queries $\tilde{q}_i$ are passed into the decoder on the ego vehicle to update its tracking results.

Note that the QFF mechanism can also be used for other perception tasks by fusing queries generated from their corresponding transformer-based modules. QFF greatly reduces communication compared to BFF; however, query embeddings are still high-dimensional, making them costly to transmit over current networks. To further cut down bandwidth requirements, while keeping important spatial information, we introduce reference points fusion.

C. Reference Point Fusion (RPF)

The key idea of RPF is is to move from sharing large amounts of dense feature data to sharing smaller, more meaningful, high-level information. The reference points from UniAD-TrackFormer show where potential tracked objects might be. By using reference points shared from other vehicles, the ego vehicle can improve its own tracking performance. Specifically, the ego vehicle combines its own reference points with those received from senders. Formally, let $\mathcal{R}^t_{ego}$ and $\mathcal{R}^t_{sender\to ego}$ denote the ego's and sender's reference point sets at frame t. The fused set is obtained as

$$\tilde{\mathcal{R}}^t = \mathcal{R}^t_{ego} \cup \left(\mathcal{R}^t_{sender\to ego} \setminus \mathcal{R}^t_{ego}\right).$$

This ensures ego preserves all self-derived reference points while enlarging its search space (within BEV) with more priors contributed by the sender. By transmitting only sparse information reference point fusion thus yields the minimal communication cost, making it highly practical for current vehicular networks. By combining BFF, QFF, and RPF, we build a progressive multi-level framework that adaptively balances perception accuracy and communication cost.

V. Experiments

To test how well our multi-level fusion framework works, we evaluate it on multiple cooperative perception tasks using the M^{3}CAD dataset.

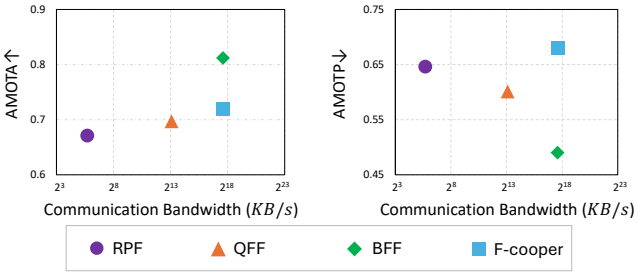


Fig. 4: Comparison of cooperative tracking performance of different multi-level fusion methods and the F-Cooper method. It shows how much communication bandwidth each method needs, using a log_2 scale. QFF requires approximately 9,063 KB/s, BFF requires 200,000 KB/s, while RPF requires only 53 KB/s. These values are based on *float32* precision and a 5 FPS transmission rate.

A. Multiple Tasks for Cooperative Perception and Prediction

We compare the performance of several perception and prediction tasks using no fusion, early fusion [26], and our proposed multi-level fusion. The tasks include tracking, mapping, motion forecasting, occupancy prediction, and planning. The detailed results for all methods are shown in Table II, showcasing the different performances of our multi-level fusion strategies. Since BFF sends the most data, it achieves the best or second-best results in almost every task, which explains why most existing cooperative perception methods are BFF-based. Compared to no fusion, however, BFF's improvement is significant, showing how important cooperative perception really is. Interestingly, we found that RPF also performs very well, i.e., its L2 planning error is only $8cm$ larger ($0.300m$ vs. $0.221m$) than QFF, which is not a too big difference. Notably, RPF performs best in mapping tasks, likely because reference points provide precise spatial information, which is especially important for tasks that need high accuracy like mapping.

B. Network Bandwidth Requirements

While the above results demonstrate the benefits of cooperative perception across different tasks, real-world deployment also needs to take communication costs into account. To explore this, we compare how well different information fusion strategies perform in cooperative tracking, based on how much network bandwidth they require. As illustrated in the Fig. 4, each fusion method has its own strengths. RPF works best in low-bandwidth situations where it transmits minimal amount of data, about $1/3800$ of what BFF needs. BFF gives the best tracking accuracy but uses the most bandwidth, making it ideal when performance matters more than bandwidth usage. QFF strikes a balance, offering strong tracking performance while keeping communication costs reasonable.

C. Benchmarking on Real-World Dataset

To demonstrate that models trained on M^{3}CAD can transfer effectively to real-world datasets, we evaluate the well-

TABLE II: Comparison of multi-level cooperative methods performance on various autonomous driving tasks.

	Tracking			Mapping (%)		Motion Forecasting (m)		Occupancy Prediction (%)		Planning (m)
Method	AMOTA↑	AMOTP↓	Recall↑	IoU-Lane↑	IoU-Road↑	ADE↓	FDE↓	IoU-n↑	IoU-f↑	L2↓
No fusion	0.21	0.66	0.48	49.6	94.0	0.36	0.38	76.2	57.5	0.43
F-cooper	0.720	0.680	0.816	-	-	-	-	-	-	-
RPF	0.671	0.684	0.758	**58.3**	**96.0**	0.346	0.358	79.5	62.5	0.300
QFF	0.697	0.601	0.835	55.6	95.6	0.3490	0.3627	80.5	**63.8**	**0.221**
BFF	**0.774**	**0.579**	**0.846**	50.67	95.34	**0.2797**	**0.2976**	**80.9**	63.0	0.3109

known multi-task autonomous driving model UniAD on the nuScenes benchmark after pretraining it on M^{3}CAD. We compare three training strategies to test the sim-to-real transfer: (1) UniAD trained directly on 100% of nuScenes data, (2) BEVFormer pretrained and then fine-tuned on 10% of nuScenes data, and (3) UniAD pretrained on 100% of M^{3}CAD and then fine-tuned on 10% of nuScenes data.

As shown in Table III, the M^{3}CAD pretraining approach achieves significant improvements over the baseline BEVFormer approach. Specifically, pretraining on M^{3}CAD reduces the average trajectory error from $1.91m$ to $1.30m$ (a 32% improvement) and lowers the collision rate from 1.3% to 0.57% (a 56% reduction) when fine-tuned with only 10% nuScenes data. These results demonstrate that our synthetic dataset provides effective pretraining that substantially improves performance in data-limited real-world scenarios, validating the quality and transferability of M^{3}CAD.

TABLE III: We evaluate planning performance of UniAD [1] on the nuScenes dataset using different training datasets.

Training datasets	**L2** $\downarrow$ (m)	**Collision** $\downarrow$ (%)
100% nuScenes	1.03	0.31
10% nuScenes	1.91	1.3
100% M^{3}CAD + 10% nuScenes	1.30	0.57

D. Impact of Noise in M^{3}CAD.

To make our M^{3}CAD dataset more realistic, we introduce two types of noise commonly seen in real-world autonomous driving: localization errors and calibration errors. Specifically, we model the ego vehicle's localization errors follow a zero-mean Gaussian distribution. For the translational component, we apply a noise with standard deviations of $\sigma_x = 0.1m$, $\sigma_y = 0.08m$, and $\sigma_z = 0.02m$, which reflects the typical accuracy characteristics of automotive-grade localization systems. For the rotational component, we introduce angular noise with standard deviations of $\sigma_{roll} = 0.2°$, $\sigma_{yaw} = 1.0°$, and $\sigma_{pitch} = 0.2°$. Similarly, we introduce camera calibration errors by adding zero-mean Gaussian noise to the sensor-to-LiDAR transformation. The rotational errors have standard deviations of $\sigma_{roll} = 0.1°$, $\sigma_{pitch} = 0.1°$, and $\sigma_{yaw} = 0.2°$, while the translational errors have standard deviations of $\sigma_x = 0.01m$, $\sigma_y = 0.01m$, and $\sigma_z = 0.02m$. For the camera intrinsics, we perturb the focal lengths (f_x, f_y) by ± 2.0 pixels and the principal point coordinates (c_x, c_y) by ± 1.0 pixel.

As shown in Table IV, under only localization noise, our cooperative perception solution remains robust. The mAP decreases from 0.785 to 0.666 (a 15% drop), which is acceptable given typical GPS accuracy limits. AMOTA falls from 0.774 to 0.664 (a 14% drop), showing some impact on tracking but still maintaining reliable performance for safe operation. Most importantly, planning remains highly resilient, with the L2 error rising only from $0.31m$ to $0.49m$, confirming that the system can still ensure safe navigation under realistic localization uncertainties.

Sensor calibration drift shows similar impact patterns, as shown in table IV, with mAP decreasing to 0.643 and AMOTA to 0.628, representing an 18% and 19% degradation, respectively. Interestingly, calibration errors have a more pronounced effect on perception tasks than localization errors, as evidenced by the slightly larger performance drops. However, the planning module demonstrates consistent robustness, with L2 error increasing to only $0.39m$, confirming that our multi-level fusion strategy provides sufficient redundancy to maintain safe autonomous driving capabilities, despite sensor calibration uncertainties.

When both localization and calibration errors exist, our system experiences the expected cumulative effect with mAP dropping to 0.609 and planning L2 error reaching $0.49m$. These results show that although individual error sources cause some performance drop, our cooperative framework still performs at an acceptable level. The relatively small degradation under challenging conditions confirms the effectiveness of our approach for practical use in a real-world setting.

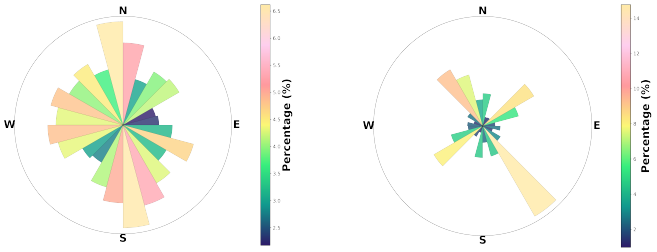
E. Importance of Camera Data

While the previous experiments demonstrate the robustness of cooperative perception under various noise conditions, an equally important question arises: how critical is environmental perception data for autonomous driving? Prior research [43], [44] suggest that when datasets mainly contain simple ego-vehicle trajectories, path planning can rely solely on internal states, e.g., velocity, acceleration, and steering angle, without using camera or LiDAR data. This happens because existing real-world datasets, e.g., nuScenes, often include oversimplified trajectories, where the ego vehicle mostly drives in straight lines for easier data collection and synchronization. We argue that this conclusion is misleading, as real-world driving is far more complex. As shown in Fig. 5, M^{3}CAD contains more diverse ego-vehicle trajectories, making sensor data essential for understanding the surrounding environment and achieving safe path planning.

We compare UniAD's performance on M^{3}CAD with Ego-MLP [43], a planning method that relies mainly on internal

TABLE IV: Comparison of cooperative perception and planning performance on M^{3}CAD with different types of noise. Type 1 noise: ego localization errors, Type 2 noise: sensor calibration drifts.

Noise Type	Object Detection and Tracking				Mapping (%)		Motion Forecasting (m)			Occupancy Prediction (%)				Plan (m, %)	
	mAP↑	AMOTA↑	AMOTP↓	Recall↑	IoU-Lane↑	IoU-Road↑	ADE↓	FDE↓	MR↓	IoU-n↑	IoU-f↑	VPQ-n↑	VPQ-f↑	L2↓	Col.↓
Baseline (No noise)	0.785	0.774	0.579	0.846	50.67	95.34	0.2797	0.2976	0.0001	80.9	63.0	76.9	61.4	0.3109	0.02
Type 1 noise	0.666	0.664	0.803	0.749	47.46	94.24	0.3754	0.4097	0.0004	76.2	54.4	71.9	51.6	0.4906	0.13
Type 2 noise	0.643	0.628	0.846	0.709	48.25	94.48	0.3661	0.3982	0.0007	75.4	52.0	70.5	48.2	0.3910	0.08
Type 1+2 noise	0.609	0.623	0.888	0.710	47.41	94.22	0.4270	0.4579	0.0005	75.0	50.4	70.2	45.7	0.4914	0.0014



(a) M^{3}CAD (b) nuScenes

Fig. 5: **Directional distribution of ego-vehicle trajectories in M^{3}CAD and nuScenes.** The polar plots show the normalized distribution of ego-vehicle trajectory segments across movement directions, with wedge length and color intensity indicating the percentage of trajectories in each 15° angular bin. Trajectory counts are normalized across both datasets to ensure fair comparison. (a) M^{3}CAD exhibits a balanced distribution across directions, reflecting comprehensive scenario coverage in simulation. (b) nuScenes displays a strong concentration in specific directions, suggesting a dataset bias toward straight-driving behaviors with fewer turning maneuvers. The broader directional coverage in M^{3}CAD supports more diverse training conditions for autonomous driving models.

states, e.g., speed, acceleration, steering, rather than perception data from cameras and LiDAR. As shown in Table V, Ego-MLP performs similarly to UniAD on nuScenes dataset ($0.35m$ vs. $0.46m$ on average L2 error), but its performance drops sharply on M^{3}CAD dataset ($2.04m$ vs. $0.46m$ on average L2 error). This $4.4\times$ gap highlights that when trajectories show realistic and complex behaviors such as lane changes, turns, and multi-vehicle interactions, effective planning depends heavily on rich perception data. These results confirm that M^{3}CAD better captures the complexity of real-world autonomous driving.

TABLE V: Path planning results of different solutions on nuScenes and M^{3}CAD benchmarks. † Results on nuScenes are reported from [43].

Benchmarks	Methods	Avg. L2 (m) ↓				Avg. Col. (%) ↓			
		1s	2s	3s	Avg	1s	2s	3s	Avg
nuScenes†	Ego-MLP	**0.15**	**0.32**	**0.59**	**0.35**	**0.00**	0.27	0.85	**0.37**
	UniAD	0.20	0.42	0.75	0.46	0.02	**0.25**	**0.84**	**0.37**
M^{3}CAD	Ego-MLP	1.02	2.04	3.06	2.04	0.00	0.09	0.30	0.13
	UniAD	**0.34**	**0.46**	**0.58**	**0.46**	**0.00**	**0.00**	**0.07**	**0.02**

VI. Conclusions

In this paper, we present M^{3}CAD, the first comprehensive benchmark designed for cooperative autonomous driving that supports multi-vehicle, multi-task, and multi-modality evaluation. Building upon this benchmark, we further introduce a multi-level fusion method that adaptively balances communication efficiency and perception accuracy by supporting BEV feature fusion, query fusion, and reference point fusion. Experiments demonstrate that our approach not only advances cooperative perception within simulation but also transfers effectively to real-world datasets such as nuScenes, achieving strong improvements in data-limited settings. We believe the M^{3}CAD benchmark, together with the proposed multi-level fusion method and benchmark evaluations, will help drive further research in cooperative autonomous driving and support its transition into real-world applications.

References

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.

[3] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.

[4] S. Hong, Y. Liu, Z. Li, S. Li, and Y. He, "Multi-agent collaborative perception via motion-aware robust communication network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 301–15 310.

[5] J. Zhang, K. Yang, Y. Wang, H. Wang, P. Sun, and L. Song, "Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 575–12 584.

[6] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *arXiv preprint arXiv:2109.13410*, 2021.

[7] 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.

[8] T. U. Institue for AI Industry Research (AIR), "Vehicle-infrastructure collaborative autonomous driving: Dair-v2x dataset," 2021.

[9] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 712–13 722.

[10] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[11] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[12] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

[13] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[14] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[15] 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.

[16] 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.

[17] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng, L. Jin, M. Lei, Z. Ma, Z. He, H. Ma, Y. Yuan, Y. Zhao, and J. Ma, "V2x-real: A largs-scale dataset for vehicle-to-everything cooperative perception," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 455–470.

[18] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 107–124.

[19] X. Huang, J. Wang, Q. Xia, S. Chen, B. Yang, C. Wang, and C. Wen, "V2x-r: Cooperative lidar-4d radar fusion for 3d object detection with denoising diffusion," *arXiv preprint arXiv:2411.08402*, 2024.

[20] Y. Wang, S. Chen, Z. Song, and S. Zhou, "Whales: A multi-agent scheduling dataset for enhanced cooperation in autonomous driving," 2025. [Online]. Available: https://arxiv.org/abs/2411.13340

[21] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] P. Karkus, B. Ivanovic, S. Mannor, and M. Pavone, "Diffstack: A differentiable and modular control stack for autonomous vehicles," in *Conference on robot learning*. PMLR, 2023, pp. 2170–2180.

[23] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[24] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 414–430.

[25] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549.

[26] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *2019 ACM/IEEE Symposium on Edge Computing (SEC)*, p. 88–100.

[27] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 605–621.

[28] D. Qu, Q. Chen, T. Bai, H. Lu, H. Fan, H. Zhang, S. Fu, and Q. Yang, "Sicp: Simultaneous individual and cooperative perception for 3d object detection in connected and automated vehicles," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8905–8912.

[29] D. Qu, Q. Chen, Y. Zhu, Y. Zhu, S. S. Avedisov, S. Fu, and Q. Yang, "Head: A bandwidth-efficient cooperative perception approach for heterogeneous connected and autonomous vehicles," *arXiv preprint arXiv:2408.15428*, 2024.

[30] X. Zhang, Z. Zhou, Z. Wang, Y. Ji, Y. Huang, and H. Chen, "Co-mtp: A cooperative trajectory prediction framework with multi-temporal fusion for autonomous driving," 2025. [Online]. Available: https://arxiv.org/abs/2502.16589

[31] Z. Wang, Y. Wang, Z. Wu, H. Ma, Z. Li, H. Qiu, and J. Li, "Cmp: Cooperative motion prediction with multi-agent communication," *IEEE Robotics and Automation Letters*, 2025.

[32] Z. Wang, S. Fan, X. Huo, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, "Emiff: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 388–16 394.

[33] S. Huang, J. Zhang, Y. Li, and C. Feng, "Actformer: Scalable collaborative perception via active queries," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 716–14 723.

[34] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524.

[35] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers."

[36] 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.

[37] H. Yu, W. Yang, J. Zhong, Z. Yang, S. Fan, P. Luo, and Z. Nie, "End-to-end autonomous driving through v2x cooperation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9598–9606.

[38] Z. Wang, S. Xu, X. Zhuang, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, "Coopdetr: A unified cooperative perception framework for 3d detection via object query," *arXiv preprint arXiv:2502.19313*, 2025.

[39] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022.

[40] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 4, pp. 2650–2660, 2023.

[41] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4812–4818.

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

[43] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is ego status all you need for open-loop end-to-end autonomous driving?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 864–14 873.

[44] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang, "Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes," 2023. [Online]. Available: https://arxiv.org/abs/2305.10430