Toward Collaborative Autonomous Driving: Simulation Platform and End-to-End System

Genjia Liu, Yue Hu, Chenxin Xu, Weibo Mao, Junhao Ge, Zhengxiang Huang, Yifan Lu, Yinda Xu, Junkai Xia, Yafei Wang, *Member, IEEE*, and Siheng Chen, *Senior Member, IEEE*

***Abstract*—Vehicle-to-everything-aided autonomous driving (V2X-AD) has a huge potential to provide a safer driving solution. Despite extensive research in transportation and communication to support V2X-AD, the actual utilization of these infrastructures and communication resources in enhancing driving performances remains largely unexplored. This highlights the necessity of collaborative autonomous driving; that is, a machine learning approach that optimizes the information sharing strategy to improve the driving performance of each vehicle. This effort necessitates two key foundations: a platform capable of generating data to facilitate the training and testing of V2X-AD, and a comprehensive system that integrates full driving-related functionalities with mechanisms for information sharing. From the platform perspective, we present *V2Xverse*, a comprehensive simulation platform for collaborative autonomous driving. This platform provides a complete pipeline for collaborative driving: multi-agent driving dataset generation scheme, codebase for deploying full-stack collaborative driving systems, closed-loop driving performance evaluation with scenario customization. From the system perspective, we introduce *CoDriving*, a novel end-to-end collaborative driving system that properly integrates V2X communication over the entire autonomous pipeline, promoting driving with shared perceptual information. The core idea is a novel driving-oriented communication strategy, that is, selectively complementing the driving-critical regions in single-view using sparse yet informative perceptual cues. Leveraging this strategy, CoDriving improves driving performance while optimizing communication efficiency. We make comprehensive benchmarks with V2Xverse, analyzing both modular performance and closed-loop driving performance. Experimental results show that CoDriving: i) significantly improves the driving score by 62.49% and drastically reduces the pedestrian collision rate by 53.50% compared to the SOTA end-to-end driving method, and ii) achieves sustaining driving performance superiority over dynamic constraint communication conditions.**

***Index Terms*—Collaborative autonomous driving, end-to-end autonomous driving, V2X communication.**

I. Introduction

VEHICLE-TO-EVERYTHING-COMMUNICATION-AIDED autonomous driving (V2X-AD) targets to improve driving performance by enabling vehicles and their surroundings, such as roadside units, and pedestrians equipped with smart devices, to exchange complementary information. V2X-AD can significantly address the inherent limitations of single-agent autonomous driving [1], [2], [3], [4], [5], [6], like restricted visibility, the unpredictability of other road users, and the development of secure paths. Through the exchange of data, V2X-AD equips individual autonomous vehicles with an enriched perception of their surroundings [7], [8], [9], [10], [11], [12], enabling them to see beyond obstructions and promptly identify smaller, fast-moving entities. This enhanced environmental awareness enables more accurate forecasting and effective route planning, leading to faster responses in emergencies. This capability significantly contributes to accident prevention, promoting safer and more reliable autonomous driving [13], [14].

To enable V2X-AD, previous works have made efforts from various aspects. From the transportation perspective, intelligent infrastructures equipped with roadside units [15], [16], [17], [18], [19], [20] have been strategically deployed along roads. These units provide complementary viewpoints to aid vehicles in perceiving the driving environment. From the communication perspective, customized V2X communication standards and protocols [21], [22], [23], [24], [25], [26], [27], [28] have emerged to facilitate reliable and real-time data exchange among vehicles and infrastructures. Supported by intelligent roadside units and advanced communication protocols, the development of V2X-AD is grounded on a robust foundation.

Recently, the nascent field of collaborative perception has emerged, aiming to address challenges in multi-agent systems from a perceptual standpoint. This approach focuses on enhancing the perceptual capabilities of each agent by facilitating the exchange of complementary perceptual information among them. Specifically, in the context of V2X-AD, the agents are conceptualized as vehicles and roadside units integrated within

Received 12 April 2024; revised 30 October 2024; accepted 8 April 2025. Date of publication 28 April 2025; date of current version 3 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62450162 and Grant 62171276, and in part by the Science and Technology Commission of Shanghai Municipal under Grant 21511100900. Recommended for acceptance by C. Wolf. *(Corresponding author: Siheng Chen.)*

Genjia Liu, Yue Hu, Chenxin Xu, Weibo Mao, Junhao Ge, Zhengxiang Huang, Yifan Lu, Yinda Xu, and Junkai Xia are with the Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: LGJ1zed@sjtu.edu.cn; 18671129361@sjtu.edu.cn; xcxwakaka@sjtu.edu.cn; kirino.mao@sjtu.edu.cn; cancaries@sjtu.edu.cn; huangzhengxiang@sjtu.edu.cn; yifan_lu@sjtu.edu.cn; yinda_xu@sjtu.edu.cn; xiajunkai@sjtu.edu.cn).

Yafei Wang is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wyfjlu@sjtu.edu.cn).

Siheng Chen is with the School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: sihengc@sjtu.edu.cn).

Our code is available at https://github.com/CollaborativePerception/V2Xverse.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3560327, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3560327

infrastructure systems. These entities collaborate by sharing perceptual data, enabling each vehicle to accurately identify all surrounding foreground objects. To achieve this, a bunch of collaborative perception methods have been proposed to address a serious of critical challenges, including the trade-off between perception performance and communication costs [7], [8], [9], [10], [11], [12], [29], [30], [31], [32], [33], [34], [35], [36], [37], robustness to pose error [31], communication latency [38], [39], and the heterogeneous issue [40].

Despite the encouraging progress on collaborative perception, a fundamental inherent limitation remains rarely explored, that is, these works focus solely on optimizing module-level perception capability, it is still unknown how exactly system-level driving capability can be enhanced. To fill this gap, we aim to expand collaborative perception to cover holistic driving capabilities, beginning with perception and extending through essential modules such as planning and control. We define this effort as *collaborative autonomous driving*. Distinct from the concept of V2X-AD, a comprehensive engineering solution that includes infrastructure setup, communication systems, and system optimization among other aspects, collaborative autonomous driving focuses on a machine learning strategy to improve each agent's system-level driving performance through information sharing among multiple agents. To this end, we focus on two essential components in developing collaborative autonomous driving: the closed-loop driving platform and the end-to-end driving system. First, a comprehensive platform that encompasses the full collaborative autonomous driving process, providing training data and supporting driving system evaluation. While a real-world platform is a direct approach, its high cost and safety concerns necessitate the development of a simulation platform as a practical alternative. Second, we need an end-to-end system that integrates information sharing with related driving functions. Such a system allows for a thorough examination of how information sharing benefits the driving capabilities.

From the platform perspective, this paper presents *V2Xverse*, a comprehensive simulation platform for collaborative autonomous driving. The key feature of our proposed V2Xverse is to enable both offline benchmark generation for driving-related subtasks and online closed-loop evaluation for driving performances in diverse scenarios, comprehensively supporting the development of collaborative autonomous driving systems. To create V2X-AD scenarios, V2Xverse outfits multiple vehicles with full driving capabilities and strategically places intelligent infrastructures. This setup offers complementary viewpoints for vehicle perception and facilitates communication between vehicles and infrastructures. To support the subsequent developments of collaborative autonomous driving methods, V2Xverse provides a full set of driving signals and annotations for system training, and it also provides diverse safety-critical scenarios for closed-loop driving evaluation, covering a convincing test length in virtual towns. V2Xverse offers three distinct advantages compared to existing platforms. First, V2Xverse promotes multi-agent simulation while the previous driving platforms [1], [41], [42], [43] only support single-agent driving simulation. Second, V2Xverse promotes full driving functions simulation while previous collaborative perception platforms [10], [44] only support functions related to the perception module. Third, V2Xverse supports comprehensive V2X-AD scenarios, including diverse sensor equipment, model integration, and flexible scenario customization; see a summary in Table I. Through these features, V2Xverse enables the development of diverse collaborative autonomous driving systems.

From the system perspective, we develop *CoDriving*, a novel end-to-end collaborative autonomous driving system that leverages perceptual information sharing to improve driving performance. Contrasting with previous V2X communication strategies in collaborative perception [7], [8], [11], which focus on optimizing perception abilities, our system introduces a novel *driving-oriented communication strategy*. The core idea is to optimize communication content based on the feedback from the driving plan. This mechanism allows each intelligent vehicle to identify driving-critical areas and selectively request collaborative messages to enhance perceptual information in these areas. To achieve this, we develop CoDriving with two functionalities: i) end-to-end autonomous driving, offering full driving capabilities; ii) driving-oriented collaboration, which leverages a novel driving request map that assigns higher scores to the spatial regions near the planned driving waypoints. This request map enables the selection of sparse perceptual features in the driving-critical regions, which are then employed to improve individual driving capabilities. CoDriving offers two distinct advantages. First, CoDriving leverages information sharing to upgrade the whole driving system, improving perception, planning, and driving performance, while previous collaborative perception approaches [11], [12] solely optimize the perception performance. Second, CoDriving promotes adaptability to different communication conditions and demonstrates generalizability to different modalities, while the previous collaborative driving method [45] can only work at a predefined ample communication bandwidth and cater to LiDAR input. To validate the effectiveness of our proposed V2Xverse platform and CoDriving system, we undertake three key evaluations. First, we conduct system-level evaluations in terms of perception, planning, and closed-loop driving performance, which benchmarks earlier collaborative perception methods by integrating them into the whole driving system, showcasing the adaptability and extensibility of our V2Xverse platform. Second, we perform an additional evaluation of perception performance, which compares our CoDriving against previous collaborative perception methods using established collaborative perception benchmarks, including DAIR-V2X [35], V2V4Real [46], OPV2V [10], V2XSIM2.0 [44], and TUMTraf-V2X [20]. Third, we conduct robustness assessments, involving four types of practical issues: communication bandwidth limitation, communication latency, pose error, and effectiveness under heterogeneous setting. These evaluations provide a comprehensive understanding of the strengths and advancements offered by our V2Xverse platform and CoDriving system. Comprehensive experimental results show that CoDriving improves the driving score by 62.49% and drastically reduces the pedestrian collision rate by 53.50% compared to the SOTA single-agent end-to-end driving method.

TABLE I
COMPARISON OF V2XVERSE WITH EXISTING CARLA-BASED CLOSED-LOOP DRIVING PLATFORMS/BENCHMARKS

CARLA-based Benchmarks		Single Agent				Multiple Agents		
		Leadboard v1 [41]	**NoCrash** [42]	**Longest6** [2]	**DOS** [43]	**OpenCDA** [47]	**AutoCastSim** [45]	**V2Xverse (Ours)**
Traffic	**Test Routes**	100	25	36	100	1	3	67
	Average Length	1.7km	N/A	1.5km	N/A	2.8km	N/A	412m
	Scenario Types	21	3	7	4	2	3	24
	Vehicles	✓	✓	✓	✓	✓	✓	✓
	Pedestrians	✓	✓	✓	✓	–	–	✓
Intelligent Agents	**Number**	1	1	1	1	≥2	≥2	≥2
	Roadside Units	–	–	–	–	–	–	✓
	Lidar	✓	–	✓	✓	✓	✓	✓
	Camera	✓	✓	✓	✓	✓	–	✓
Communication	**V2V**	–	–	–	–	✓	✓	✓
	V2I	–	–	–	–	✓	–	✓
Robustness Evaluation	**Latency**	–	–	–	–	–	–	✓
	Pose Error	–	–	–	–	–	–	✓
Online Driving Performance	**Completion**	✓	✓	✓	✓	✓	✓	✓
	Safety	✓	✓	✓	✓	✓	✓	✓
	Efficiency	–	–	–	–	✓	✓	✓
Modular Performance	**Perception**	–	–	–	–	✓	–	✓
	Planning	–	–	–	–	–	–	✓

To sum up, our contributions are:

• We propose V2Xverse, a comprehensive V2X-aided autonomous driving simulation platform. This platform enables the development of collaborative driving systems by supporting offline benchmark generation for driving-related subtasks and online closed-loop driving performance evaluation in diverse scenarios.

• We propose CoDriving, a novel end-to-end collaborative autonomous driving system, which improves driving performance by sharing driving-critical information. • We conduct comprehensive experiments and validate that: i) V2X communication-enabled information sharing significantly outperforms single-agent end-to-end autonomous driving systems, and ii) CoDriving achieves superior performance-bandwidth trade-off in system-level evaluations.

The rest of the paper is organized as follows: in Section II, we review existing works related to V2X-AD. In Section III, we introduce our simulation platform V2Xverse, from platform construction to benchmark generation. In Section IV, we introduce our collaborative driving system CoDriving, including the end-to-end driving pipeline and driving-oriented collaboration strategy. In Section V, we present an offline evaluation on existing widely-used collaborative perception benchmarks, validating the effectiveness of CoDriving compared to previous collaborative perception methods in terms of perception capability. In Section VI, we present system-level evaluations in V2Xverse, validating the effectiveness of CoDriving in terms of perception, waypoints planning and closed-loop driving capacities. Finally, we draw the conclusion of this paper in Section VII.

II. RELATED WORK

End-to-end autonomous driving. Recently, learning-based end-to-end autonomous driving has emerged as an active topic, which directly maps environment information into control signals thus conceptually avoiding the cascading error of complex modular design. Recent works mainly fall into two categories: reinforcement learning (RL) and imitation learning (IL). Reinforcement learning for end-to-end autonomous driving involves training an autonomous vehicle to navigate and control itself by interacting with the environment, it offers the potential for systems that can adapt to diverse and complex driving scenarios. [48], [49] first map the vision environment into the latent embedding space with auxiliary semantic supervision, and then train the RL agent using the latent representation. WOR [50] builds a RL agent based on a forward model that assumes the world on rails, and then conducts policy distillation to obtain the final agent. Roach [51] utilizes privileged environmental information to train a RL expert agent that maps bird's-eye view (BEV) images to continuous actions. Imitation learning targets to clone the behavior of an expert agent by fitting the recorded driving data. Mainstream methods mainly improve driving performance from four aspects, [1], [3], [6], [43], [52] leverage transformer architecture to learn better representations, [3], [4], [5], [52], [53] involve auxiliary tasks to assist in driving task learning, [1], [2], [3] integrate representations from multi-modalities to exploit their complementary advantages for autonomous driving, and [5], [54] adopt policy distillation strategy. CIL [55] and CILRS [56] stand as early baselines for IL. CIL [55] introduces a conditional architecture that uses different branches for different driving commands, and CILRS [56] extends the CIL model with deeper network and an additional speed prediction head. LBC [54] applies a two-stage imitation learning. A cheating model with access to privileged information first clones the behavior of the expert, which is then imitated by a final model using sensor data. [53] builds scene representations with a predicted online map and occupancy field, which is then followed by a motion planning module. TCP [5] conducts predictions for both

future trajectory and control signal, where the trajectory branch provides guidance for the control prediction, and an intermediate distillation is conducted for better behavior clone. LAV [4] augments the dataset by learning driving strategy from all vehicles around in addition to the ego vehicle. DriveAdapter [57] explores to directly utilize the powerful RL teacher model for planning by adopting it to the predicted BEV segmentation. UniAD [52] introduces a planning-oriented autonomous driving framework, unifying tracking, mapping, motion prediction, occupancy prediction, and planning in a single system, where the key component is a query-based design to connect all modules.

However, single-agent driving systems [2], [3], [4], [5], [50], [51], [52] inevitably encounter limitations in distant or occluded areas which may lead to catastrophic failures. Several approaches have emerged to tackle this issue. For instance, ReasonNet [43] predicts the future evolution of scenes and infers potential occupancy by modeling the interaction between objects and the environment, thereby enhancing perception performance under occlusion. However, it fails to predict the motion of objects that have never appeared in the field of view, as well as invisible objects that do not interact with the environment. Coopernaut [45] introduces a visually-cooperative driving system, where the vehicle aggregates voxel representations from collaborators to improve driving decisions. Coopernaut takes a step towards end-to-end collaborative autonomous driving, but it can only handle a fixed communication bandwidth, and its strongly coupled framework limits the extension in other AD functionalities. To overcome these limitations, we design an end-to-end full-featured collaborative autonomous driving system that ensures adaptability to any communication bandwidth.

V2X communication: V2X communication serves as the fundamental basis of V2X-AD, as it enables information sharing among agents. There are two kinds of mainstream V2X communication technologies: dedicated short-range communication (DSRC) and cellular-based communication. DSRC [25] achieves low end-to-end latency in a short communication range. Cellular communication [26], [27], [28] supports high mobility communication. Through the achieved progress in V2X communication, the communication bandwidth in a realistic communication system is always constrained. To better make use of the precise communication resources, we design a strategy to transmit compact driving features.

Collaborative perception: Collaborative perception [7], [8], [9], [10], [11], [12], [29], [30], [31], [32], [33], [34], [35], [36], [37] is an emerging application of V2X-communication-aided systems, which promotes the crucial perception module of autonomous driving through complementary perceptual information sharing. Several high-quality platforms have emerged [10], [11], [30], [35], which simulate collaborative perception scenarios and provide diverse perception annotations to aid in the development of collaborative perception systems. Collaborative perception systems have made remarkable progress, with CoCa3D [12] achieving nearly complete perception. However, collaborative perception mainly focused on improving perception performance, failed to achieve the ultimate driving goal of V2X-AD due to the lack of other necessary driving functions. In this work, we fill this gap by introducing a novel end-to-end collaborative driving system, achieving complete V2X-AD functionalities.

CARLA-based autonomous driving benchmark: CARLA-based autonomous driving benchmarks [2], [41], [42], [43], [45], [47] involve the creation of comprehensive test scenarios based on the NHTSA pre-crash typology and performance metrics within the CARLA simulator, providing a standardized environment for assessing the capabilities of autonomous systems. NoCrash [42] serves as an early benchmark, it collects training data from CARLA Town01 under 4 specific weather conditions and performs generalization testing with other towns and weather conditions. Meanwhile, the CARLA simulator provides official evaluation leaderboards [41] (v1, v2). Leaderboard v1 involves 100 secret test routes with an average length of 1.7 km. Leaderboard v2 is released in 2023, it is a more challenging version equipped with updated maps and scenarios. Longest6 Benchmark [2] reduces and balances the test routes by choosing the 6 longest routes from each of the 6 available towns. DOS Benchmark [43] is dedicated to validating driving performance in collision scenarios, involving 100 cases of 4 specifically designed scenarios. In addition to the aforementioned extensively discussed single-agent benchmarks, [45], [47] explore the feasibility of multi-agent driving evaluation. OpenCDA [47] serves as a prototype cooperative driving automation platform that offers platooning scenario for benchmark testing. AutoCastSim [45] also supports multi-agent autonomous driving simulation and offers three traffic scenarios with challenging vision occlusion. However, existing multi-agent benchmarks are limited to a small number of scenes and routes, and lack discussion on system-level performance of end-to-end driving pipelines, as well as the impact of practical issues on driving performance. In this work, V2Xverse fills this gap, enabling both offline/online evaluation of collaborative driving systems in various driving scenarios.

III. V2Xverse: V2X-Aided Fully Autonomous Driving Simulation Platform

This section presents V2Xverse, a comprehensive simulation platform for V2X-communication-aided autonomous driving, which supports both data generation as well as closed-loop evaluation for collaborative driving systems. Notably, closed-loop driving requires vehicles to drive continuously with real-time feedback from the environment, and closed-loop evaluation is critical in the assessment of driving systems since i) it more accurately simulates the dynamic conditions of real-world driving, thus addressing the limitations of open-loop evaluations that only document static scenarios and may not effectively highlight the strengths of various driving systems; and ii) it avoids the high expenses and safety risks associated with real-world tests, enabling quick prototyping and testing. This approach promotes the quick iteration of ideas and offers affordable access to a wide range of driving scenarios.

We introduce the platform from four aspects: platform construction, offline benchmark generation, online closed-loop evaluation, and comparisons with previous platforms.

A. Platform Construction

V2Xverse is built based on CARLA simulator [58], it extends capabilities by supporting the deployment of closed-loop V2X-aided autonomous driving in challenging scenarios. To evaluate autonomous driving proficiency, we consider the navigation task in realistic safety-critical traffic scenarios and deploy the complete pipeline for collaborative driving.

Navigation task in safety-critical scenarios: The navigation task requires autonomous vehicles to complete the predefined routes while dealing with high densities of dynamic interactive agents. The driving scenes include static road layouts, static objects and dynamic traffic agents. To assess driving safety more comprehensively, we customize varying safety-critical and challenging scenarios. From the traffic perspective, we turn off traffic lights to create more dynamic and interactive driving scenarios. From the agent perspective, we set up some reckless pedestrians that ignore traffic rules and surrounding vehicles. We also include more large vehicles and roadside obstacles for diverse simulations. Additionally, we design several trigger-based scenarios extended from CARLA Autonomous Driving Leaderboard setting [41], e.g. pedestrians suddenly appearing from occluded regions to cross the road, unexpected agents from occluded buildings, and unprotected turns at intersections, see Fig. 2. See Appendix A.1, available online for more details in scenario generation.

Deployment of V2X collaboration: To simulate V2X situations, we deploy multiple collaborative vehicles and roadside units (RSU) within effective collaboration distance, and build real-time communication between them. The vehicles start from adjacent positions and follow the same route. Meanwhile, RSUs are strategically positioned on the roadside alongside vehicles to ensure a comprehensive view of the traffic conditions, and we ensure that each vehicle has at least one roadside unit for valid collaboration at each moment. Vehicles are equipped with complete driving pipeline, including localization, perception, planning, and control, while RSUs are solely equipped with a perception module. Both vehicles and RSUs are equipped with a communication module and multi-modality sensors including four RGB cameras and a LiDAR. All the driving-related signals, including pose, sensor data, waypoints, and speed, can be exchanged among agents through communication. We demonstrate the deployment of V2X collaboration in Appendix A.2, available online.

B. Offline Benchmark Generation

V2Xverse platform enables offline benchmark generation by providing all the necessary driving annotations, which support the full-stack training of collaborative driving systems. We apply an automatic labeling algorithm to generate data, where multiple autonomous vehicles concurrently drive following a widely-adopted expert policy [1], [2], and their driving behaviors serve as ground-truth labels for imitation learning [54]. The benchmark we provide is fully annotated, including 288 k synchronous images from four views, 72 k LiDAR point clouds, 482 k 2D/3D bounding boxes of vehicles, pedestrians, way-points, speeds, steering angles and brake signals, see Fig. 1 . Statistically, we split the dataset samples: 26322 for training, 5502 for validating, 4306 for testing, spanning 8 towns and 108 routes with a frame rate of 5 FPS. See Appendix A.2, available online for more details in data description.

C. Online Closed-Loop Evaluation

Unlike offline evaluations [10], [35], [44], [46], [52], [59], which lack environmental interaction and adaptive evolution, online driving evaluation offers real-time interactive insights into system performance, allowing for the identification of safety risks and algorithm improvement in realistic environment [41], [42]. V2Xverse enables online closed-loop evaluation by supporting system deployment and driving performance evaluation. We deploy the collaborative autonomous driving system on vehicles and RSUs in CARLA. The collaborative system transforms sensor observations and communicated messages into driving actions, and we assess the driving performance in the navigation task. We set 67 evaluation routes that cover a wide range of driving situations, including three levels of speed and five levels of urgency. As we expect both safe and efficient driving, a bunch of metrics, including the driving score (DS), route completion ratio (RC), infraction score (IS), pedestrian/vehicle/layout collision rate, following [41], and mean speed are used for comprehensive driving performance evaluation.

D. Comparison With Previous Simulation Platforms

A vast array of significant works [2], [41], [42], [43], [45], [47] have been proposed to build autonomous driving platforms and benchmarks based on CARLA. Compared with these platforms, our V2Xverse has the following advantages: i) V2Xverse supports communication aided multi-intelligent-agent simulation. In comparison to mainstream single-intelligent-agent driving benchmarks [2], [41], [42], [43], V2Xverse extends functionality by supporting Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication, and enables the evaluation of system robustness against practical challenges, such as communication latency and pose errors. ii) V2Xverse provides various yet customizable safety-critical traffic scenarios (24 types), sufficient test routes (67), and realistic urban driving environment with pedestrians and vehicles, while existing multi-intelligent-agent simulation platforms [45], [47] caters to vehicle-only and limited scenarios and routes. iii) V2Xverse enables the deployment and comprehensive evaluation for collaborative autonomous driving systems. V2Xverse provides various driving signals and annotations in V2X-AD situations, which support the V2X-based closed-loop driving task and modular tasks including both perception and planning; while [45] only caters to driving performance; see the detailed comparison with previous platforms in Table I.

IV. CoDriving: End-to-End Collaborative Autonomous Driving System

This section introduces our proposed CoDriving, a novel end-to-end collaborative autonomous driving system, which promotes driving performance through information sharing.

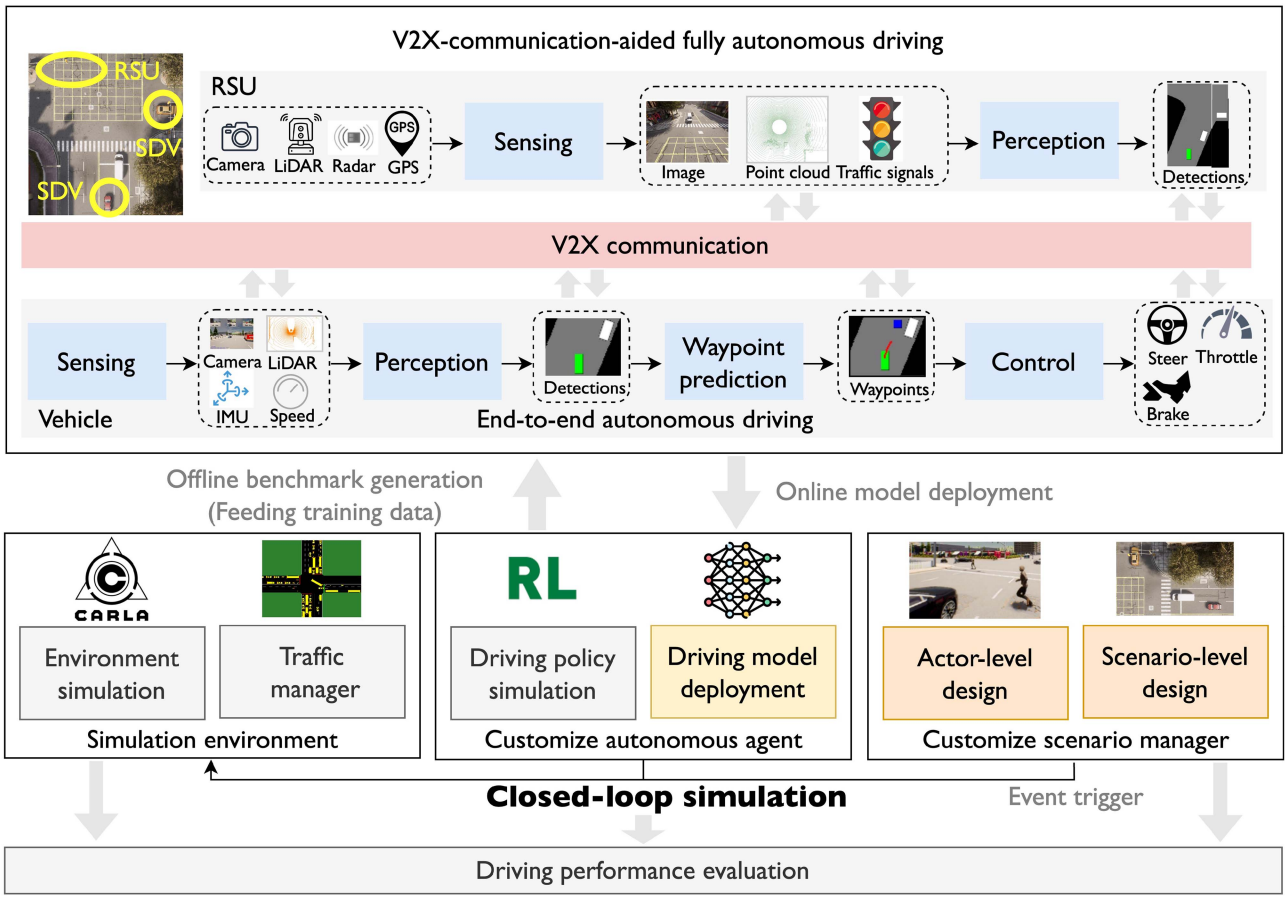


Fig. 1. Platform overview. V2Xverse simulates the complete V2X-AD driving pipeline, incorporating various driving functionalities and delivering extensive driving annotations. It facilitates both the offline benchmark generation and online closed-loop driving performance evaluation.

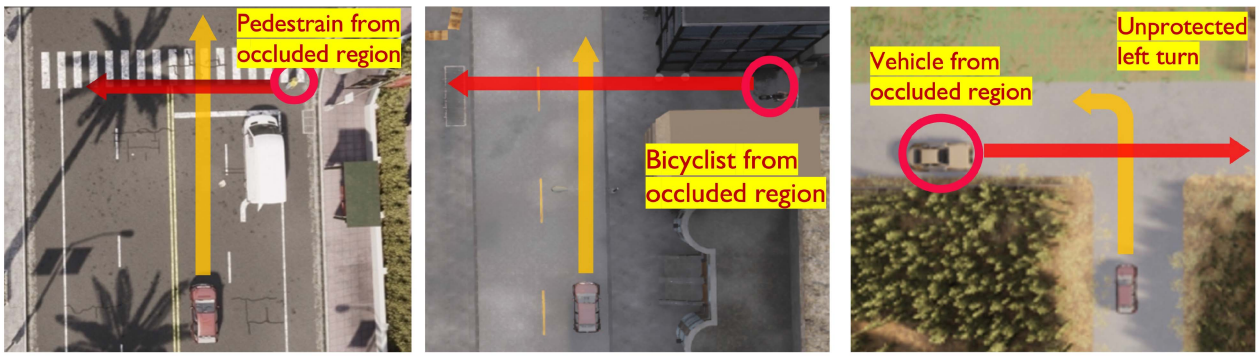


Fig. 2. Safety-critical scenarios caused by occlusion, including reckless pedestrians from occluded vehicles, unexpected bicyclists from occluded buildings, and vehicles rapidly coming from occluded road infrastructures.

A. Problem Formulation

Consider N collaborative agents in the V2X scenario, where agents refer to vehicles and roadside units. We focus on the navigation task, where each vehicle is assigned a specific destination. Given the observations of the N agents as $\{\mathcal{X}_i\}_{i=1}^N$, the overall objective of collaborative autonomous driving is to maximize the driving performance of each vehicle by exchanging information among all vehicles and roadside units within a communication budget B; that is,

$$\max_{\theta,\mathcal{P}} \sum_{i=1}^{N} d\left(\Phi_\theta\left(\mathcal{X}_i, \mathcal{D}_i, \{\mathcal{P}_{j\to i}\}_{j=1}^N\right)\right), \text{s.t.} \sum_{j\neq i} |\mathcal{P}_{j\to i}| \leq B \tag{1}$$

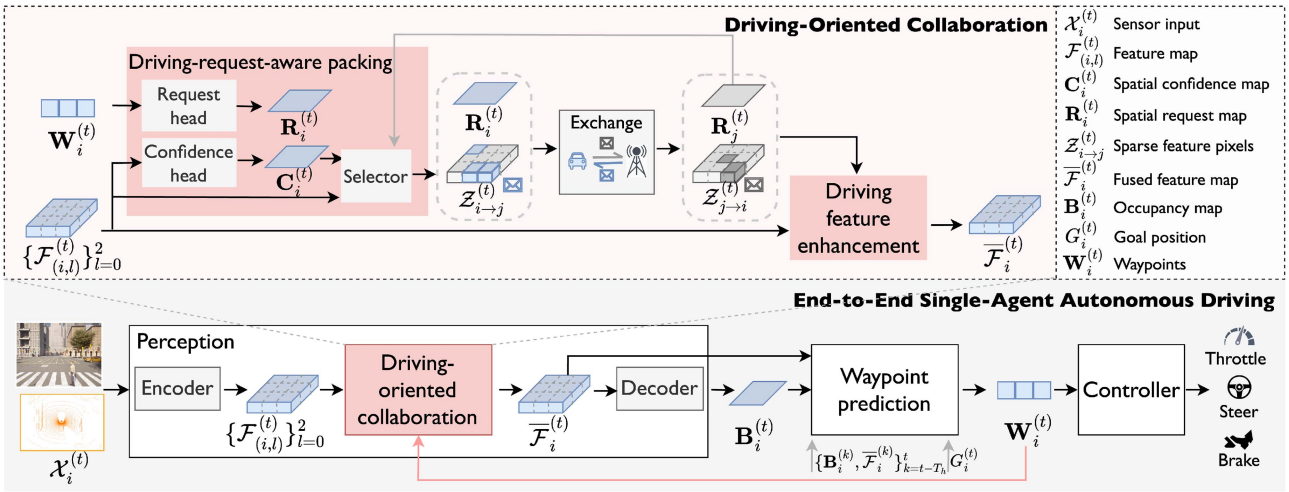


Fig. 3. System overview. CoDriving comprises two components: end-to-end single-agent autonomous driving, which transforms the sensor inputs into driving actions, and driving-oriented collaboration, which enhances the single-agent features by aggregating the driving-critical perceptual features shared through communication. The benefits propagate from the perception module to the entire driving pipeline, enhancing all driving signals.

where $d(\cdot)$ is the driving performance metric, $\Phi_\theta(\cdot)$ is the end-to-end autonomous driving network with trainable parameter θ, and $\mathcal{X}_i, \mathcal{D}_i$ are the observation and destination of the ith agent respectively, and $\mathcal{P}_{j\to i}$ is the message transmitted from the jth agent to the ith agent. Note that the driving performance of roadside units is evaluated as zero.

To optimize the trade-off between driving performance and communication cost, we present CoDriving, which comprises two components; see Fig. 3. First, an end-to-end imitation learning-based autonomous driving is introduced in Section IV-B, which offers full driving capabilities, including perception, waypoints planning, and driving control. Second, a novel driving-oriented collaboration is introduced in Section IV-C, which first leverages a driving-request-aware communication strategy to select sparse driving-critical perceptual information for sharing, and then leverages driving feature enhancement to boost the individual driving capabilities with the received collaborative messages.

B. End-to-End Single-Agent Autonomous Driving

The end-to-end single-agent autonomous driving network learns to output driving actions based on inputs from different modalities. To achieve this, we integrate the driving-needed modular components in a unified system, including a perception module, a waypoint predictor, and a controller. We use the bird's-eye-view (BEV) representation, as it provides a unified global coordinate system, avoiding complex coordinate transformations and better supporting cross-agent collaboration.

Perception: The perception module adapts to inputs from two modalities: RGB images and 3D point clouds, and detects surrounding objects with category and regressed bounding box. For the ith agent, given its input $\mathcal{X}_i^{(t)}$ at timestamp t, we leverage an encoder to extract BEV feature following [40], [60] ; that is, $\Phi_{\text{enc}}(\cdot) = \Phi_{\text{tcv}}(\Phi_{\text{cv}}(\cdot))$, where $\Phi_{\text{cv}}/\Phi_{\text{tcv}}$ represents convolution/transposed-convolution layers. Specifically, the intermediate features from the 3 rd/7th/12th block of Φ_{cv} are denoted as $\{\mathcal{F}_{(i,l)}^{(t)} \in \mathbb{R}^{\frac{X}{2^l}\times\frac{Y}{2^l}\times 2^l D}\}_{l=0}^{2} = \Phi_{\text{cv}}(\mathcal{X}_i^{(t)})$. Then, the final output of BEV encoder is obtained as $\mathcal{F}_i^{(t)} = \Phi_{\text{scv}}(\text{concat}(\Phi_{\text{tcv}}(\{\mathcal{F}_{(i,l)}^{(t)}\}_{l=0}^{2}))) \in \mathbb{R}^{X\times Y\times D}$, where X, Y, D are its height, weight and channel, $\text{concat}(\cdot)$ is feature concatenation, and $\Phi_{\text{scv}}(\cdot)$ is a convolution block to shrink feature dimension. For image input, we additionally extract its 2D convolutional feature and transform it to BEV with a warping function [61], after which the encoder $\Phi_{\text{enc}}(\cdot)$ extracts features in BEV. For the 3D point cloud, we initialize the input by discretizing 3D points as a BEV map [60]. In the collaboration phase in Section IV-C, agents will share the critical regions within BEV features $\{\mathcal{F}_{(i,l)}^{(t)}\}_{l=0}^{2}$.

The detection decoder $\Phi_{\text{dec}}(\cdot)$ comprises a classification head $\Phi_{\text{cls}}(\cdot)$ and a regression head $\Phi_{\text{reg}}(\cdot)$. Given the BEV feature $\mathcal{F}_i^{(t)}$, $\Phi_{\text{cls}}(\cdot)$ generates the object probability heatmap in the view of the ith agent by $\mathbf{S}_i^{(t)} = \Phi_{\text{cls}}(\mathcal{F}_i^{(t)}) \in \mathbb{R}^{X\times Y\times K}$, where K denotes the classes of objects. $\Phi_{\text{reg}}(\cdot)$ generates a dense bounding box regression map by $\mathcal{O}_i^{(t)} = \Phi_{\text{reg}}(\mathcal{F}_i^{(t)}) \in \mathbb{R}^{X\times Y\times 8^{\frown}K}$, where each location of $\mathcal{O}_i^{(t)}$ represents the predicted K classes of 3D box $(x, y, z, h, w, l, \cos\alpha, \sin\alpha)$, denoting position residual, size, and angle. We apply non-maximum suppression (NMS) to generate sparse 3D detections, and subsequently rasterize them into a binary BEV occupancy map $\mathbf{B}_i^{(t)} \in \{0,1\}^{X\times Y}$. The occupancy map $\mathbf{B}_i^{(t)}$ and BEV feature $\mathcal{F}_i^{(t)}$ then serve as input to the planning module. In this manner, the integration of structured occupancy data and detailed perceptual features offers complementary insights for waypoints planning.

Waypoints planning: The planning module takes a sequence of historical perceptual information and the future goal position as inputs, and outputs the trajectories that are progressing

towards the goal, serving as a planner. Given T_h+1 frames of historical BEV occupancy maps together with BEV features $\{\mathbf{B}_i^{(k)}, \mathcal{F}_i^{(t)}\}_{k=t-T_h}^{t}$ and the goal position $G_i^{(t)}$ at timestamp t, the waypoints planning network $\Phi_{\text{way}}(\cdot)$ generates the waypoints by $\mathbf{W}_i^{(t)} = \Phi_{\text{way}}(\{\mathbf{B}_i^{(k)}, \mathcal{F}_i^{(t)}\}_{k=t-T_h}^{t}, G_i^{(t)}) \in \mathbb{R}^{T_f \times 2}$, where $\mathbf{W}_i^{(t)}$ represents the predicted trajectories of future T_f timestamps, represented with x, y coordinates. Here, we adopt MotionNet [62] to extract spatial-temporal features from the sequence of BEV features and occupancy maps, and MLPs to encode goal position embedding.

Controller: The controller obtains executable driving actions from the predicted waypoints, including steer, throttle, and brake. Two PID controllers are utilized, one for lateral control and one for longitudinal control. The longitudinal controller uses the position vectors between consecutive waypoints, while the lateral controller considers the orientation. Our controller configuration follows [1], [2], [54].

C. Driving-Oriented Collaboration

The end-to-end single-agent autonomous driving network unifies the driving-needed functions and achieves end-to-end autonomous driving. Driving-oriented collaboration leverages the collaboration capability of V2X communication to tackle the single agent's inevitable limited visibility issue through information sharing.

In this work, we present a novel driving-oriented collaboration scheme to optimize both driving performance and communication efficiency. This scheme includes i) driving-request-aware communication, we optimize communication efficiency by exchanging spatially sparse yet driving-critical BEV perceptual features by a driving request head; and ii) driving feature enhancement, we enhance the perceptual features of each agent with received messages. The enhanced BEV feature further serves to drive collaboration benefits throughout the entire system.

Driving-request-aware communication: The driving-request-aware communication targets to pack the complementary perceptual information in driving-critical areas into a compact message. The core idea is to explore the spatial heterogeneity of driving requests and perceptual information. The intuition is that: i) for driving requests, missing information at locations near the planned driving waypoints would cause catastrophic accidents, requesting and obtaining information at these locations could improve driving performance; and ii) for perceptual information, sending information at foreground areas helps recover the miss-detected objects due to the limited view, and background areas could be omitted to save precious bandwidth. Note that incorporating driving requests into communication is a novel design, diverging from previous communication strategies that specifically cater to the module-level perception utility. This new driving-request-aware communication allows for the optimization of valuable communication resources toward enhancing the ultimate system-level driving utility.

To achieve this, we enable a novel driving request map and a perceptual confidence map for each agent. The driving request map is implemented with a heatmap to highlight the regions around the planned driving route, where each element is negatively correlated with the distance to the waypoints predicted based on single agent's observation. The perceptual confidence map is implemented with the object probability heatmap, where each element reflects the confidence of that spatial area containing objects. Let $\mathbf{W}_i^{(t)}$ be the planned waypoints based on the single agent's observation, $\mathbf{S}_i^{(t)}$ be the object probability heatmap. Then, the confidence map $\mathbf{C}_i^{(t)} \in \mathbb{R}^{X \times Y}$ is obtained with its elements given by $\mathbf{C}_i^{(t)}(x,y) = \max(\mathbf{S}_i^{(t)}(x,y))$, and the request map is given by $\mathbf{R}_i^{(t)} = \Phi_{\text{req}}(\mathbf{W}_i^{(t)})$, where $\Phi_{\text{req}}(\cdot)$ is the request head, and we model the negative correlation with a Gaussian distribution with standard variance σ, which is widely used in the detection task to depict the importance decay pattern. The (x,y)th element of request map is $\mathbf{R}_i^{(t)}(x,y) = \exp(-\frac{(x-W_x)^2+(y-W_y)^2}{2\sigma^2})$, where (W_x, W_y) is the nearest waypoint. Confidence map $\mathbf{C}_i^{(t)}$ and request map $\mathbf{R}_i^{(t)}$ reflect foreground and waypoints probability from a bird's-eye view, and agents decide where to communicate based on these maps, offering spatially sparse yet driving-critical supportive features.

According to the driving request map and the perceptual confidence map, we further propose a BEV-based binary selection matrix to reflect where to communicate and then sample from BEV features $\{\mathcal{F}_{(i,l)}^{(t)}\}_{l=0}^{2}$. Let $\mathbf{M}_{j\to i}^{(t)} \in \{0,1\}^{X\times Y}$ be a binary selection matrix to determine the spatial areas of messages sent from agent j to i. Each element in $\mathbf{M}_{j\to i}^{(t)}$ determines whether the feature at the corresponding location should be selected to send. Thus, this sparse mask $\mathbf{M}_{j\to i}^{(t)}$ contributes to transmitting informative regions of the feature map, which are spatially sparse and critical for driving. We obtain the selection matrix $\mathbf{M}_{j\to i}^{(t)}$ by solving a constrained optimization problem conditioned on the jth agent's confidence map $\mathbf{C}_j^{(t)}$, the ith agent's request map $\mathbf{R}_i^{(t)}$ and the bandwidth limit B, as described in (2)

$$\max_{\mathbf{M}_{j\to i}^{(t)}} \mathbf{M}_{j\to i}^{(t)} \odot \mathbf{C}_j^{(t)} \odot \mathbf{R}_i^{(t)},$$
$$\text{s.t.} |\mathbf{M}_{j\to i}^{(t)}| \le b, \mathbf{M}_{j\to i}^{(t)} \in \{0,1\}^{X\times Y}, \tag{2}$$

where $b = B/(\Sigma_{l=0}^{2} 2^{-l} D)$, with its denominator derived from the ratio of total number of elements in $\{\mathcal{F}_{(i,l)}^{(t)}\}_{l=0}^{2}$ to the covered feature area. $\odot$ is element-wise multiplication. By introducing the selection matrix $\mathbf{M}_{j\to i}^{(t)}$, (2) is actually a proxy of (1). The objective of (2) ensures the priority of transmitting the perceptually informative features that are located within driving-critical areas. Fortunately, even with hard constraints and non-differentiability of binary variables, the optimization problem (2) has a closed-form solution that satisfies the constraint. This solution is obtained by selecting those spatial regions whose corresponding elements rank top-b in $\mathbf{C}_j^{(t)} \odot \mathbf{R}_i^{(t)}$. The detailed steps of selection function are: i) arrange the elements in the matrix $\mathbf{C}_j^{(t)} \odot \mathbf{R}_i^{(t)}$ in descending order; ii) given the communication budget constrain, decide the total number b of communication regions; iii) set the spatial regions of $\mathbf{M}_{j\to i}^{(t)}$, where elements rank top-b in $\mathbf{C}_j^{(t)} \odot \mathbf{R}_i^{(t)}$ as 1 and 0 verses.

The selected sparse feature map is then obtained as $\mathcal{Z}_{j\to i}^{(t)} = \{\mathcal{F}_{(j,l)}^{(t)} \odot \mathbf{M}_{j\to i}^{(t,l)}\}_{l=0}^{2}$, where $\mathbf{M}_{j\to i}^{(t,0)} = \mathbf{M}_{j\to i}^{(t)}$ and $\mathbf{M}_{j\to i}^{(t,l)}$ selects the regions of $\mathcal{F}_{(j,l)}^{(t)}$ with corresponding resolution. We obtain $\mathbf{M}_{j\to i}^{(t,l)}$ from $\mathbf{M}_{j\to i}^{(t,l-1)}$ by max-pooling to capture the same critical regions across different feature layers. Overall, the message packs the selected feature map to provide supportive information and the driving request map to provide spatial priors to request driving-critical complementary information, this is, $\mathcal{P}_{i\to j}^{(t)} = (\mathcal{Z}_{i\to j}^{(t)}, \mathbf{R}_i^{(t)})$. Only selected features and their indices are packed, significantly reducing communication costs while retaining sufficient information for perception and driving.

It is worth noting that our perception and waypoints planning model are interconnected via the request map $\mathbf{R}_i^{(t)}$, which feeds the planning outcomes back to influence the front-end perception, establishing a feedback loop. As a result, the perception and planning modules collaboratively optimize the overarching driving objective, fostering synergy across AD system modules.

Driving feature enhancement: We enhance the feature of each agent by aggregating the received messages. Then, based on the augmented features, the driving-related signals across the entire system can be enhanced. Specifically, we utilize pixel-level Scaled Dot-Product Attention [10], [63] to fuse feature vectors from different agents within aligned locations. The non-parametric attention mechanism is computationally efficient in aggregating features, and it preserves robustness against disturbance by weighting feature vectors from each agent with scaled dot-product similarity.

At timestamp t, the ith agent receives messages $\{\mathcal{Z}_{j\to i}^{(t)}\}_{j\in\mathcal{N}_i}$ from other agents, where $\mathcal{N}_i$ denotes the set of collaborators whose detection range overlap with agent i. We also include the ego feature map in fusion and denote $\mathcal{Z}_{i\to i}^{(t)} = \{\mathcal{F}_{(i,l)}^{(t)}\}_{l=0}^{2}$. The fused intermediate feature is obtained as $\overline{\mathcal{F}}_{(i,l)}^{(t)} = \text{softmax}(\mathcal{F}_{(i,l)}^{(t)}\mathcal{F}_{\mathcal{N}_i}^{T}/\sqrt{2^l D})\mathcal{F}_{\mathcal{N}_i}$, where $\mathcal{F}_{\mathcal{N}_i} = [\mathcal{F}_{(j,l)}^{(t)}]_{j\in\mathcal{N}_i}$ is the stack of neighbor features. The collaborative perception feature is then obtained as $\overline{\mathcal{F}}_i^{(t)} = \Phi_{\text{scv}}(\text{concat}(\Phi_{\text{tcv}}(\{\overline{\mathcal{F}}_{(i,l)}^{(t)}\}_{l=0}^{2})))$, fulfilling the role of perception feature $\mathcal{F}_i^{(t)} = \Phi_{\text{enc}}(\mathcal{X}_i^{(t)})$ in single-agent network in Section IV-B.

The collaborative perception feature $\overline{\mathcal{F}}_i^{(t)}$ propagates benefits across the entire system, enhancing all driving signals. Following the pipeline in Section IV-B, we similarly derive an object probability heatmap as $\overline{\mathbf{S}}_i^{(t)} = \Phi_{\text{cls}}(\overline{\mathcal{F}}_i^{(t)})$ and a bounding box regression map as $\overline{\mathbf{O}}_i^{(t)} = \Phi_{\text{reg}}(\overline{\mathcal{F}}_i^{(t)})$. Subsequently, the collaborative BEV occupancy map $\overline{\mathbf{B}}_i^{(t)}$ is generated using $\overline{\mathbf{S}}_i^{(t)}$ and $\overline{\mathbf{O}}_i^{(t)}$ via NMS. During the waypoints planning phase, $\overline{\mathbf{B}}_i^{(t)}, \overline{\mathcal{F}}_i^{(t)}$ substitute $\mathbf{B}_i^{(t)}, \mathcal{F}_i^{(t)}$ in Section IV-B, and are utilized to generate the improved waypoints via $\overline{\mathbf{W}}_i^{(t)} = \Phi_{\text{way}}(\{\overline{\mathbf{B}}_i^{(k)}, \overline{\mathcal{F}}_i^{(t)}\}_{k=t-T_h}^{t}, G_i^{(t)})$. Then, the controller generates the final driving actions based on $\overline{\mathbf{W}}_i^{(t)}$.

To sum up, CoDriving first enhances single-agent detection capabilities by leveraging shared BEV features among collaborators. Specifically, these shared BEV features, drawn from various perspectives, synergize to create a more precise and complete representation of the driving-relevant environment. Second, this enhanced BEV representation results in collaborative detections, which can better reflect the ground truth objects. Such a detailed grasp of the surroundings would simplify the perceiving of dynamics, facilitating the formulation of safer driving plans. Third, the enhanced detection capabilities inherent in collaborative autonomous driving yield more accurate occupancy maps. Fourth, the enhanced BEV representations and occupancy maps serve as critical inputs for the waypoints planning network. With these improved inputs, the network can then generate more precise waypoints plannings. Finally, this improvement directly contributes to safer driving maneuvers, especially in scenarios where objects are obscured in individual view.

D. Losses and Training

To train the overall system, we supervise two tasks: 3D detection and waypoints planning. The overall loss is

$$L = L_{\text{cls}}\left(\overline{\mathbf{C}}_i^{(t)}, \widehat{\mathbf{C}}_i^{(t)}\right) + L_{\text{reg}}\left(\overline{\mathcal{O}}_i^{(t)}, \widehat{\mathcal{O}}_i^{(t)}\right) + L_{\text{way}}\left(\overline{\mathbf{W}}_i^{(t)}, \widehat{\mathbf{W}}_i^{(t)}\right)$$

where $\{\widehat{\mathbf{C}}_i^{(t)}, \widehat{\mathcal{O}}_i^{(t)}\}$ and $\widehat{\mathbf{W}}_i^{(t)}$ represent ground-truth objects and waypoints for the ith agent. $L_{\text{cls}} + L_{\text{reg}}$ and L_{way} are the centerpoint detection loss [64] and planning loss [65]. Specifically, we implement L_{cls} with pixel-wise Gaussian focal loss and $L_{\text{reg}}, L_{\text{way}}$ with weighted L1 loss.

The training pipeline consists of two stages. First, the perception module is trained from scratch with supervision from the detection loss. In the second stage, we freeze the perception module and train the planner for waypoint planning. During this phase, we randomly set the feature selection rate of the matrix $\mathbf{M}_{j\to i}^{(t)}$ within the range of 0 to 1, with the logarithm of the selection rate following a uniform distribution. This strategy simulates the uncertain communication bandwidth and helps the system adapt to varying bandwidth conditions by augmenting the transmitted features. In this way, the system is further optimized to fit the objective in (1). The random selection rate is introduced in the later stages of training to ensure convergence.

E. Discussion

CoDriving has two distinct advantages. First, CoDriving breaks the limitation of single agent perception by sharing BEV features and comprehensively enhances the performance of the entire system, encompassing improvements in perception, planning, and driving behavior. Second, CoDriving achieves driving-oriented communication, it considers the driving request and specifically targets the driving-critical regions, which improves communication efficiency by transmitting spatially compact features.

Compared to existing collaborative perception methods [7], [8], [9], [11], [20], [29], [30], [31], [34], [36], CoDriving shows distinction in both aspects of system purpose and collaborative

information selecting. First, CoDriving aims to extend beyond environmental perception, producing waypoints planning and driving control signals as outputs, while collaborative perception methods only focus on addressing perception tasks. Second, CoDriving optimizes the selection of collaborative information tailored for driving tasks. This driving-centric collaboration is facilitated through a feedback mechanism, wherein outputs from the downstream planning module are utilized to highlight areas of driving requests, and these driving request maps are then relayed back to the front-end perception module. Such a process underscores the collaboration within a unified system. In comparison, collaborative perception methods optimize the selection of shared information tailored for perceptual tasks. Compared to existing end-to-end autonomous driving systems [2], [3], [4], [5], [43], [50], [52], CoDriving enables feature sharing among multiple agents, enhancing the perception and planning ability of each participant. Especially, compared to Coopernaut [45], our system outperforms in three aspects: i) CoDriving demonstrates adaptability to any communication bandwidth by solving the optimization problem (2), while Coopernaut relies on voxel pooling and can only handle a fixed communication bandwidth; ii) CoDriving adapts input modality of either RGB image or point clouds, while Coopernaut is designed for point clouds only; and iii) CoDriving facilitates driving-critical BEV feature sharing, enhancing interpretability, while Coopernaut communicates all the encoded point representations.

V. Perception Evaluation

In this section, we compare the proposed CoDriving with previous state-of-the-art methods in collaborative perception using offline collaborative perception datasets [10], [20], [35], [44], [46], covering both real-world and simulated scenarios. The evaluation includes the trade-off between perception performance and communication cost under both homogeneous and heterogeneous scenarios, and the robustness against two practical issues including communication latency and pose errors. We begin by outlining the task setting, evaluation metrics, and details of the dataset used. Following these, the experimental results are presented. Note that CoDriving is a full-spectrum collaborative autonomous driving system, with capabilities in perception, planning, planning, and control. Here, we retain only the perception module of CoDriving and prune the subsequent modules, allowing it to be trained solely with annotations relevant to perception.

A. Tasks and Metrics

We follow the common collaborative perception task setting, concentrating on collaborative 3D detection tasks. Here, each agent has a specific detection objective: to identify all objects within a predefined spatial area, based on its sensor data and collaborative messages enabled by communication. To evaluate this task, we consider several assessment scenarios: i) the trade-off between performance and communication cost, evaluating collaborative detection capabilities under varying communication bandwidth constraints; ii) heterogeneous scenarios, where agents in the same collaborative scene are equipped with different types of sensors. Here, we randomly assign agents in the scene either LiDAR or camera, resulting in a balanced 1:1 ratio of agents across the different modalities; iii) communication latency issues, evaluating collaborative detection's performance when there are delays in the received collaborative messages; and iv) pose error issues, evaluating performance when the collaborative messages received are incorrectly localized.

Detection performance: Following the collaborative perception methods [10], [11], [20], [44], the detection results are evaluated by 1) Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.30, 0.50. 2) Mean average precision (mAP) in BEV perspective, considering the BEV center distance.

Communication cost: Following the collaborative perception methods [11], [29], the communication cost in the feature-based intermediate collaboration setting is calculated as $\log_2(H \times W \times ||\mathbf{M}||_1 \times C \times 32/8)$, where $\mathbf{M}$ is the selection matrix representing the selected feature to be packed in the messages. Here, 32 represents the float32 data type, and 8 converts bits to bytes.

Communication latency and pose error settings: For communication latency, we follow SyncNet [38], using latency varying from 0 ms to 500 ms. For pose error issues, following setting in CoAlign [31], we use Gaussian noise with a mean of 0 m/0° and standard deviations ranging from 0 m/0° to 0.6 m/0.6°.

B. Datasets

We conduct comprehensive experiments on collaborative 3D object detection tasks using five commonly used datasets, including the real-world datasets, DAIR-V2X [35], V2V4Real [46], TUMTraf-V2X [20], and simulation datasets, OPV2V [10] and V2X-SIM2.0 [44].

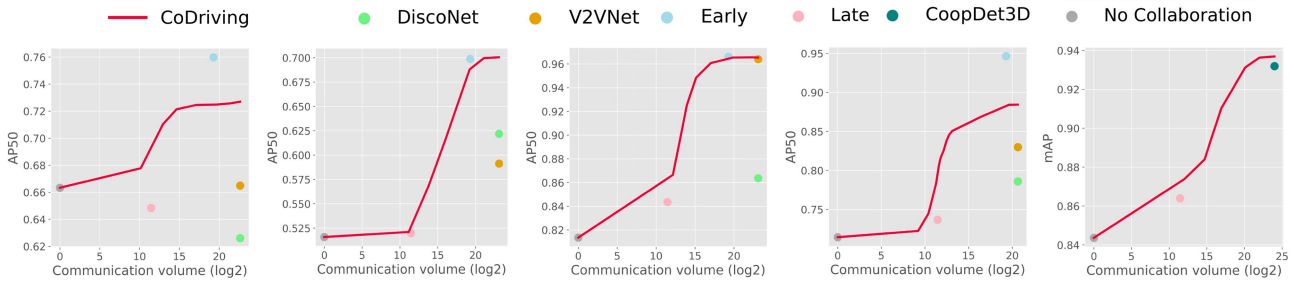
DAIR-V2X: DAIR-V2X [35] is a popular real-world collaborative perception dataset. Each scene contains two agents: a vehicle and a roadside unit. Each agent is equipped with a LiDAR and a camera. The perception range is 204.8 m× 102.4 m.

V2V4Real: V2V4Real [46] is a large real-world vehicle-to-vehicle collaborative perception dataset. It includes a total of 20 K frames of LiDAR point cloud captured by Velodyne VLP-32 sensor and 40 K frames of RGB images captured by two mono cameras (front and rear) with 240 K annotated 3D bounding boxes. The perception range is 280 m× 80m.

TUMTraf-V2X: TUMTraf-V2X [20] is a real-world collaborative perception dataset. It includes sensor data from 4 roadside cameras, 1 roadside LiDAR, 1 vehicle camera, and 1 vehicle LiDAR. The perception range is 150 m× 150m.

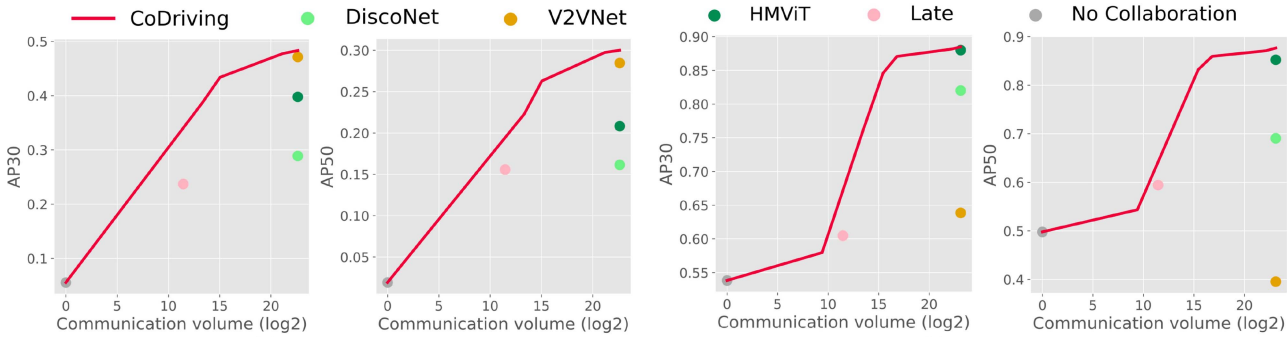
OPV2V: OPV2V [10] is a vehicle-to-vehicle collaborative perception simulation dataset, co-simulated by OpenCDA [47] and CARLA [58]. It includes 12 K frames of 3D point clouds and RGB images with 230 K annotated 3D boxes. The perception range is 280 m× 80m.

V2X-SIM2.0: V2X-SIM2.0 [44] is a vehicle-to-vehicle collaborative perception simulation dataset. Each scene contains a 20-second traffic flow at a certain intersection of three CARLA



(a) DAIR-V2X (b) V2V4Real (c) OPV2V (d) V2XSIM2.0 (e) TUMTraf-V2X

Fig. 4. Benchmark CoDriving and previous collaborative perception methods on commonly used collaborative perception real-world and simulation datasets under homogeneous setting.



(a) DAIR-V2X (b) OPV2V

Fig. 5. Benchmark CoDriving and previous collaborative perception methods on commonly used collaborative perception real-world and simulation datasets under heterogeneous settings.

towns. It includes 47.2 k samples with 10 k frames of 3D point clouds and RGB images. The perception range is 100 m $\times$ 80 m.

C. Quantitative Results

Perception performance and communication cost trade-off under both homogeneous and heterogeneous scenarios: Figs. 4 and 5 compare the proposed CoDriving with previous methods in terms of the trade-off between detection performance and communication bandwidth on real-world datasets DAIR-V2X, V2V4Real, TUMTraf-V2X and simulation datasets OPV2V, V2XSIM2.0 under homogeneous and heterogeneous settings, respectively. Baselines include no collaboration, V2VNet [9], DiscoNet [29], HMViT [66], CoopDet3D [20], early fusion, and late fusion, where agents exchange the detected 3D boxes directly. We see that CoDriving: i) achieves a far-more superior perception-communication trade-off across all the communication bandwidth choices on all the collaborative perception datasets under both homogeneous and heterogeneous scenarios; ii) significantly improves the detection performance, especially under extremely limited communication bandwidth, improving the SOTA performance by 47.58/9.66% on DAIR-V2X and V2XSIM2.0 even when the bandwidth is constrained by a factor of 1 K; iii) outperforms previous SOTA, DiscoNet, with significantly reduced communication cost: 142/654 times less on V2V4Real and V2XSIM2.0, and achieves competitive performance compared to CoopDet3D on TUMTraf-V2X while reducing communication costs by 8 times; and iv) can adapt to varying communication bandwidth while previous methods are limited to specific communication choices. The reason is that CoDriving can adjust the information selection under varying bandwidth limits by solving the constrained optimization problem (2), while previous methods instinctively transmit the complete feature map.

Robustness of offline perception performance against pose error and communication latency issues: We validate the robustness of CoDriving against pose error and communication latency on V2V4Real, DAIR-V2X, OPV2V, and V2XSIM2.0. Figs. 6 and 7 show the detection performances as a function of pose error and latency, respectively. We see: i) while perception performance generally declines with increasing levels of pose error and latency, CoDriving achieves superior performance under all imperfect conditions across all the collaborative perception datasets. Note that early fusion initially outperforms CoDriving under minor pose error perturbations in DARI-V2X and V2XSIM2.0, but its performance declines quickly due to the limited ability to correct communication errors. ii) CoDriving consistently surpasses no collaboration on DAIR-V2X, V2V4Real, V2XSIM2.0 with metric AP30, whereas baselines fail when pose error exceeds 0.4 m and latency exceeds 300 ms.

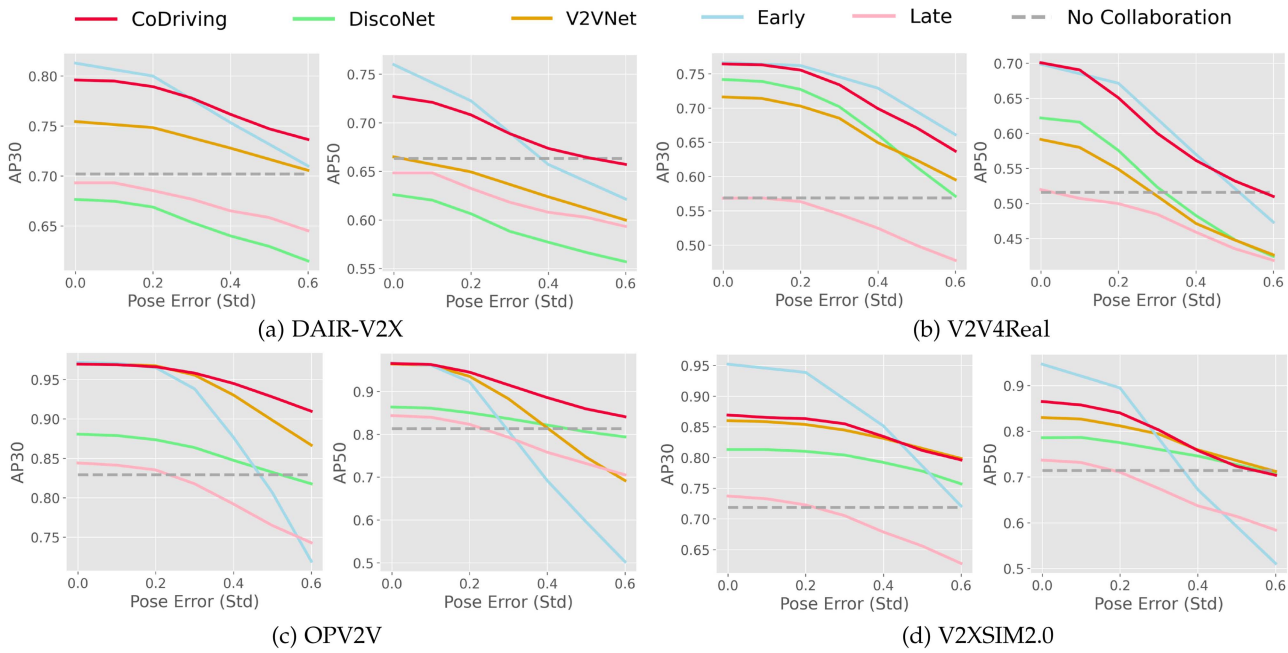


Fig. 6. The robustness to pose error of CoDriving on commonly used collaborative perception real-world and simulation datasets.

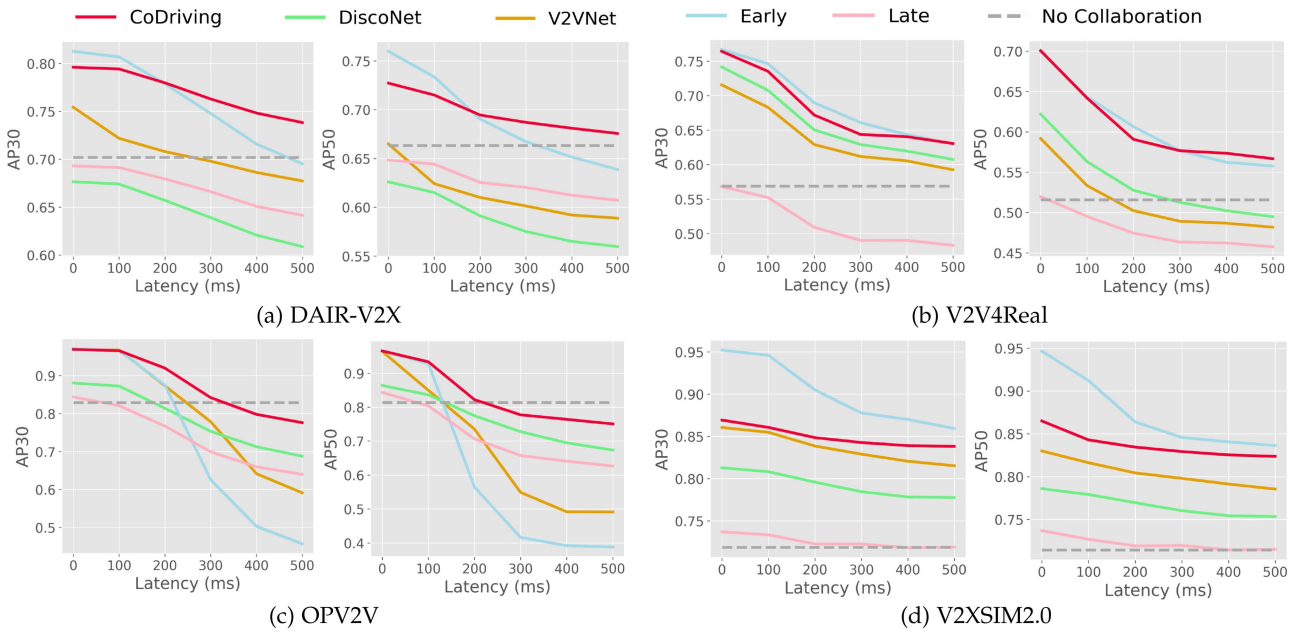


Fig. 7. The robustness to communication latency of CoDriving on commonly used collaborative perception real-world and simulation datasets.

The observed robustness is attributed to: i) the attention fusion helps to filter out the disturbed feature values by assigning lower attention weights; 2) CoDriving fuses features at three resolution level, as described in Section IV-C, which incorporates low-resolution feature fusion that is less sensitive to positional disturbances.

Advantages: Our advantages are twofold. First, compared to previous collaborative perception methods which only address a subset of evaluation scenarios, we conduct a thorough benchmarking of the existing state-of-the-art (SOTA) approaches under a wide range of evaluation settings, including four commonly used datasets, both homogeneous and heterogeneous sensor configurations, and two practical challenges: pose error and communication latency. Our comprehensive evaluation lowers the barriers to researching the advantages of collaboration facilitated by V2X communication from the perceptual perspective. Second, this specialized variant of CoDriving consistently outperforms previous SOTAs in collaborative perception across

these extensive evaluation scenarios. As a result, CoDriving serves as an effective method within the field of collaborative perception research.

VI. System-Level Driving Evaluation

V2Xverse facilitates the system-level evaluation of collaborative autonomous driving under unified data distribution. In this section, we evaluate the online closed-loop driving performance of CoDriving within V2Xverse simulation environment, and utilize the offline benchmark of V2Xverse to evaluate the modular performance.

A. Experimental Setting and Details

Our experiments comprehensively evaluate CoDriving, including three tasks: the online driving task, the perception (3D detection) task, and the waypoints planning task. Two types of input modality are covered in the evaluation, including 3D point cloud and RGB image. We also evaluate the trade-off between performance and communication band-width, and the robustness against two types of practical issues, including communication latency and pose error.

Metrics: We evaluate the closed-loop driving performance with a bunch of driving safety and efficiency metrics introduced in Section III-C, including driving score (DS), route completion ratio (RC), infraction score (IS), pedestrian/vehicle/layout collision rate, and mean speed; see details in Appendix A.2, available online. The 3D detection task is evaluated using the Average Precision (AP) metric for different categories of objects at various IoU thresholds, along with the mean Average Precision (mAP) computed across all categories. The waypoints planning task is evaluated with average displacement error (ADE) and final displacement error (FDE).

Implementation details: In closed-loop evaluation, the ego vehicle can access and communicate with one RSU simultaneously. In the perception and the planning tasks, the ego vehicle can communicate with at most one RSU and one vehicle. Both vehicles and roadside units (RSU) are equipped with cameras and LiDAR. The image resolution is 600 $\times$ 800 and the LiDAR is 64-channel. The detection range is 48m $\times$ 24m, with 36m/12m/12m/12m at front/rear/left/right.

Baselines: We compare our method with representative collaborative methods and single-agent end-to-end driving methods. First, collaborative baselines are categorized into two types: i) collaborative end-to-end driving method, Coopernaut [45], and ii) collaborative perception methods, V2X-ViT [34], Fcooper [67], CoopDet3D [20], perceptual-output-based late collaboration (Late fusion) and raw-inputs-based early collaboration (Early fusion). Note that, to assess the system-level driving performance of each collaborative perception modular method, we integrate their perception components into our driving system, enabling them with the planner and controller of CoDriving, and train the planning modules accordingly. Specifically, we re-implement these collaborative perception methods with PointPillar BEV encoding [60] and centerpoint detection loss [64], as done in CoDriving. Second, for individual-agent end-to-end driving baselines, we identify two groups: 1) state-of-the-art methods, including LAV [4], WOR [50], TCP [5], Transfuser [1], InterFuser [3] and 2) a handcrafted expert autopilot system which has access to the privileged information in the CARLA simulator.

B. Quantitative Results

Closed-loop driving performance evaluation: Table II presents the comparison of our CoDriving with representative state-of-the-art methods in terms of closed-loop driving performance. We see that i) our CoDriving achieves the highest driving score and the highest infraction score while maintaining the second highest speed, providing a safe and efficient driving solution; ii) compared to Coopernaut, the existing end-to-end collaborative driving method, our CoDriving achieves significantly more effective driving behavior with a higher mean speed by **231.52%** improvement and a higher driving score by **822.85%** improvement. In contrast, Coopernaut experiences the lowest route completion due to its overfitted policy; iii) compared with other collaborative perception approaches like Fcooper, V2X-Vit, and CoopDet3D, our method achieves the lowest pedestrian and vehicle collision rate, showing the effectiveness of our collaborative perception method on pedestrian and vehicle detection; iv) compared with TCP, the state-of-the-art end-to-end individual driving framework, CoDriving achieves a more efficient driving behavior with a higher mean speed by **23.98%** improvement and a higher route completion by **48.60%** improvement. Meanwhile, CoDriving ensures safety with a significantly lower pedestrian collision rate by **53.50%**. TCP takes a relatively conservative driving strategy, resulting in a low route completion rate due to exceeding the time limit.

Modular tasks evaluation on different collaboration strategies: Table III compares CoDriving with other collaboration strategies on modular tasks including 3D object detection and waypoints planning. We see that: i) all the collaboration strategies outperform no collaboration on both perception and planning task; ii) our CoDriving achieves the best detection and planning performance.

Driving and modular tasks evaluation on different modalities: Tables IV and V present the comparison of no collaboration and Codriving collaboration under different input modalities: Camera-based (C) and LiDAR-based (L). We evaluate the whole system and report the performance of perception, planning and driving. From Tables IV and V, we see that i) for both modalities, the communication-enabled multi-agent collaboration significantly enhances the driving system, validating the generalizability of collaboration; and ii) collaboration benefits the entire driving system, improving the driving score by 98.43%, the detection mAP30 by 19.30%, and reducing the planning FDE by 3.22% based on LiDAR inputs.

Trade-off between performance and communication bandwidth: To compare communication results straightforwardly and fairly under different communication bandwidths, we do not consider any extra data/feature/model compression. Fig. 8 compares CoDriving with/without (red/orange curve) the proposed driving-request-aware communication with other collaboration strategies, illustrating the trade-off between performance and

TABLE II
CLOSED-LOOP DRIVING PERFORMANCE EVALUATION

	Methods	Driving Score↑	Route Completion↑	Infraction Score↑	Pedestrian Collision↓	Vehicle Collision↓	Layout Collision↓	Mean Speed↑
		Driving Performance						
Single	LAV [4]	38.27	55.07	0.64	2.83	3.97	0.44	2.13
Single	WOR [50]	12.05	22.27	0.60	3.13	5.60	7.16	1.30
Single	TCP [5]	47.48	62.14	0.81	1.57	**0.39**	0.20	2.46
Single	TransFuser [1]	30.23	73.47	0.47	5.12	1.74	0.26	2.63
Single	InterFuser [3]	40.31	85.07	0.47	2.63	2.68	1.02	2.55
Collaborative	No Fusion	33.81	89.74	0.37	2.99	1.71	1.87	2.84
Collaborative	Late Fusion	52.40	90.72	0.57	2.07	1.30	1.75	2.89
Collaborative	Early Fusion	59.12	90.72	0.63	1.41	2.35	1.76	2.89
Collaborative	Fcooper [67]	44.00	90.42	0.46	2.59	1.70	1.94	2.47
Collaborative	V2X-ViT [34]	39.35	91.98	0.42	2.61	4.05	0.13	**3.43**
Collaborative	CoopDet3D [20]	63.04	88.12	0.71	1.67	0.83	1.09	2.76
Collaborative	Coopernaut [45]	8.36	12.38	0.62	1.02	3.45	3.20	0.92
Collaborative	CoDriving (Ours)	**77.15**	**92.34**	**0.82**	**0.73**	0.47	**0.04**	3.05
	Expert	82.58	92.00	0.89	1.40	0.16	0.00	2.53

↑ means the higher the better. ↓ means the lower the better. The bold font denotes the best performance. Expert is a powerful handcrafted agent, presented for reference and not included in the comparison. Our CoDriving ensures the highest driving score with the second highest speed, and at the same time, achieves low collision rates, providing a reliable and efficient driving solution.

TABLE III
MODULAR TASKS PERFORMANCE EVALUATION ON DIFFERENT COLLABORATION STRATEGIES IN TOWN5 TEST SET

Collaborative methods	Detection								Waypoints Planning	
	Vehicle			Cyclist		Pedestrian		Mean		
	AP30↑	AP50↑	AP70↑	AP30↑	AP50↑	AP30↑	AP50↑	mAP30↑	ADE↓	FDE↓
No Fusion	0.89	0.84	0.73	0.40	0.30	0.41	0.24	0.57	0.636	1.460
Late Fusion	0.88	0.86	0.81	0.43	0.38	0.45	0.27	0.59	0.631	1.454
Fcooper [67]	0.93	0.82	0.68	0.44	0.29	0.56	0.33	0.64	0.627	1.446
V2X-ViT [34]	0.93	**0.91**	**0.84**	0.50	0.36	0.41	0.12	0.61	0.629	1.447
CoopDet3D [20]	0.93	0.90	0.81	0.48	**0.41**	0.53	0.31	0.65	0.623	1.439
CoDriving (Ours)	**0.94**	**0.91**	0.83	**0.52**	**0.41**	**0.58**	**0.35**	**0.68**	**0.619**	**1.413**

On perception/planning, CoDriving outperforms other collaboration strategies.

TABLE IV
ABLATION STUDY FOR MULTI-MODALITY (C: CAMERA/L: LIDAR) AND COLLABORATION (×/✓) IN DRIVING PERFORMANCE

	Collaboration	Modality	Driving Score↑	Route Completion↑	Infraction Score↑	Pedestrian Collision↓	Vehicle Collision↓	Layout Collision↓	Mean Speed↑
Driving	×	C	38.88	92.13	0.40	4.21	3.51	1.23	**3.12**
		L	33.81	89.74	0.37	2.99	1.71	1.87	2.84
	✓	C	45.06	**95.28**	0.46	2.29	2.31	0.23	2.83
		L	**77.15**	92.34	**0.82**	**0.73**	**0.47**	**0.04**	3.05

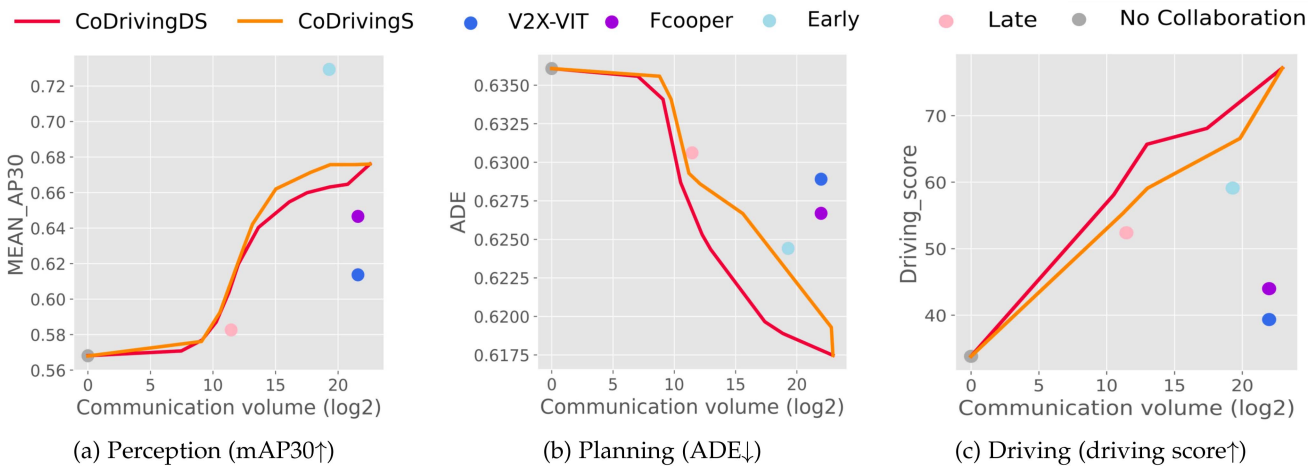
Collaboration improves the driving score by 98%(LiDAR), 18%(camera).

communication bandwidth in the driving task and modular tasks. The absence of driving request map indicates that the communication is determined solely by the confidence map. We see that: i) CoDriving outperforms other intermediate collaboration strategies across all communication bandwidth on all three tasks, reducing the communication cost 90/1176 times less on perception/planning than Fcooper and 630/4389 times less than V2X-VIT. ii) Compared to perceptual-confidence-only communication, the proposed driving-request-aware communication achieves a superior trade-off between performance and communication cost in both planning and driving tasks. In the ultimate closed-loop driving task, the driving-request mechanism achieves an average improvement of 9.6% in driving score across communication volume ranging from 2^{11} to 2^{20}. Note that the driving-request mechanism slightly reduces perception performance, since it prioritizes objects in regions near the predicted

TABLE V
ABLATION STUDY FOR MULTI-MODALITY (C: CAMERA/L: LIDAR) AND COLLABORATION (×/✓) ON PERCEPTION AND PLANNING

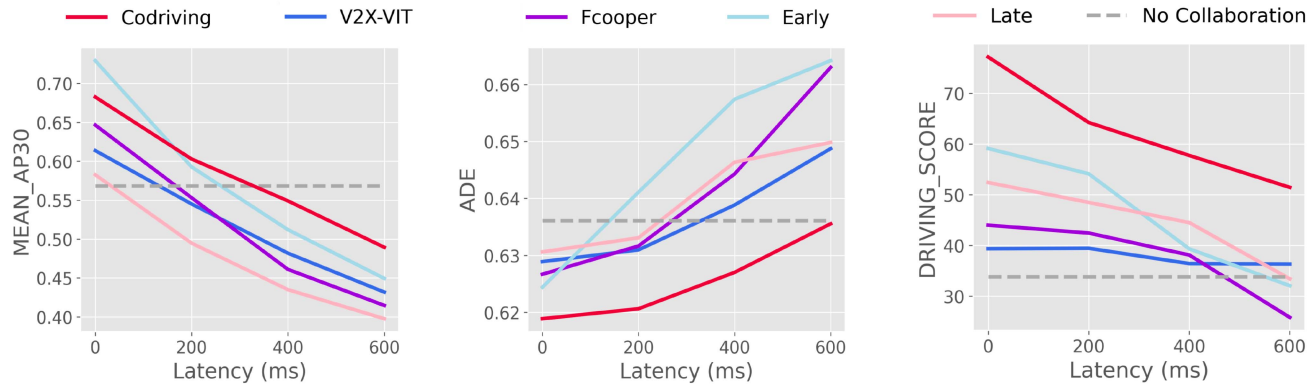
Collaboration	Modality	Detection								Waypoints Planning	
		Vehicle			Cyclist		Pedestrian		Mean		
		AP30↑	AP50↑	AP70↑	AP30↑	AP50↑	AP30↑	AP50↑	mAP30↑	ADE↓	FDE↓
×	C	0.73	0.63	0.51	0.36	0.22	0.27	0.14	0.45	0.706	1.531
	L	0.89	0.84	0.73	0.40	0.30	0.41	0.24	0.57	0.636	1.460
✓	C	0.91	0.86	0.76	**0.52**	0.37	0.43	0.28	0.62	0.634	**1.375**
	L	**0.94**	**0.91**	**0.83**	**0.52**	**0.41**	**0.58**	**0.35**	**0.68**	**0.619**	1.413

Collaboration improves mAP30 by 19.30% (LiDAR), 37.78% (camera) and FDE by 3.22% (LiDAR), 10.19% (camera).



(a) Perception (mAP30↑) (b) Planning (ADE↓) (c) Driving (driving score↑)

Fig. 8. The trade-off between performance and communication bandwidth. CoDrivingDS/S denotes CoDriving with/without driving-request. Compared with previous collaboration strategies, CoDriving with driving-request-aware communication consistently achieves superior performance over varying bandwidths in perception, planning and driving task.

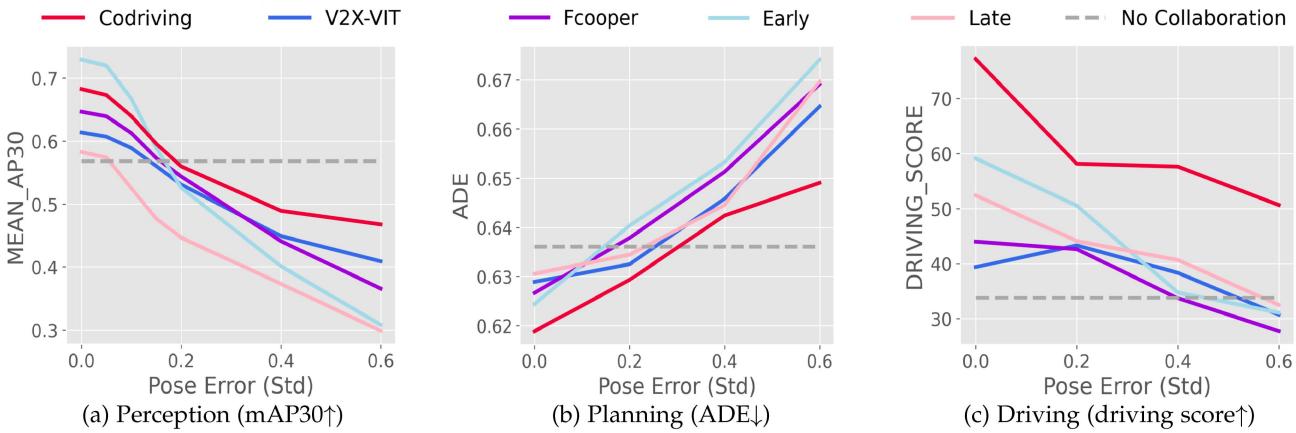


(a) Perception (mAP30↑) (b) Planning (ADE↓) (c) Driving (driving score↑)

Fig. 9. Robustness to latency on V2Xverse. CoDriving consistently outperforms previous collaboration strategies.

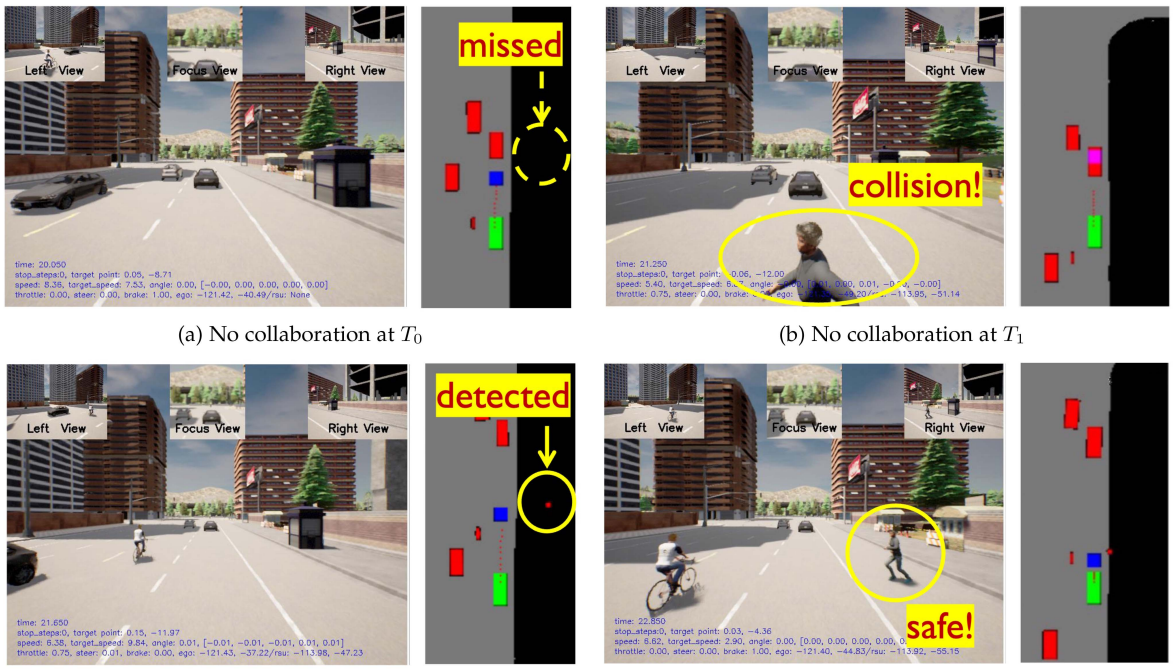
waypoints, and the detection precision for objects in other areas, which constitute a substantial portion, is somewhat compromised. Nevertheless, the driving-request mechanism achieves superior performance in driving and planning by allocating precise communication bandwidth to driving-critical regions, which improves communication efficiency.

Robustness of driving performance against communication latency and pose errors: For communication latency, Fig. 9 presents the comparison of our CoDriving with other collaboration strategies under various communication latency settings. We see that: i) the system-level driving performance and module-level perception, planning performance degrade



(a) Perception (mAP30↑) (b) Planning (ADE↓) (c) Driving (driving score↑)

Fig. 10. Robustness to pose error on V2Xverse. CoDriving consistently outperforms previous collaboration strategies.



(a) No collaboration at T_0 (b) No collaboration at T_1

(c) CoDriving at T_0 (d) CoDriving at T_1

Fig. 11. Visualization of a potentially dangerous scenario and detection results. Compared to no collaboration, our CoDriving perceives the actors better with the complement of RSU information and avoids collision.

as latency increases, while CoDriving consistently achieves superior performance; ii) in planning and driving tasks, CoDriving surpasses no collaboration even under extreme asynchrony, while other collaboration strategies deteriorate to the level of no collaboration. For pose errors, Fig. 10 compares CoDriving with other collaboration strategies under various pose error levels. The experimental setting follows Section V-A. We see that: i) the system and modular performance degrade as pose error increases, while CoDriving achieves superior performance consistently; ii) even in an extreme noise of 0.6 Std, CoDriving surpasses no collaboration by 47.06% in driving score. The observed robustness is attributed to the multi-scale attention fusion strategy, which helps reduce the sensitivity to disturbances and filter out noise, as analyzed in Section V-C. This advantage extends from perception to the entire driving system, improving both planning and driving performance. See further ablation studies in Appendix C.1, available online.

C. Qualitative Analysis

Does collaboration benefit driving behavior? Fig. 11 depicts a safety-critical scenario caused by occlusion, a reckless pedestrian abruptly emerges from a telephone booth. From Fig. 11(a)(b), we observe that the single-agent driving system would

encounter a severe pedestrian collision caused by the miss detection. In contrast, from Fig. 11(c)(d), we observe that our CoDriving system avoids a catastrophic collision by braking in advance as it detects the occluded pedestrian based on the complementary information shared by roadside unit through communication. Compared to single-agent driving, CoDriving provides a safer and more reliable autonomous driving solution.

VII. Conclusion and Limitations

This work advances collaborative autonomous driving through both a simulation platform and an end-to-end system. We develop a comprehensive closed-loop V2X fully autonomous driving simulation platform V2Xverse. This platform enables the complete pipeline for developing collaborative autonomous driving systems that target the ultimate driving performance. Meanwhile, V2Xverse maintains adaptability and extensibility to integrate and validate single-functional modules and single-agent driving systems. Furthermore, we propose a novel end-to-end collaborative autonomous driving system, CoDriving, which enhances driving performance while optimizing communication efficiency by sharing driving-critical perceptual information. Comprehensive evaluations on the entire driving system show that CoDriving significantly outperforms single-agent systems across varying communication bandwidths.

This work presents significant potential for future extensions. From a platform perspective, while V2Xverse implements V2X-aided driving scenarios through rule-based simulation, there is room to improve the realism of traffic trajectories and the fidelity of image rendering by incorporating data-driven generative simulation. From a methodological perspective, although our approach leverages low-level planning outputs to enable efficient sharing of perceptual features, an open question remains as to how shared high-level planning intentions could further facilitate planning negotiation. Exploring this direction could maximize the benefits of information exchange, contributing to safer and more efficient autonomous driving.

References

[1] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7073–7083.

[2] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[3] H.-C. Shao, L. Wang, R. Chen, H. Li, and Y. T. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Proc. Conf. Robot Learn.*, 2022, pp. 726–737.

[4] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17201–17210.

[5] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 6119–6132.

[6] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15773–15783.

[7] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4106–4115.

[8] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6876–6883.

[9] 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 *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 605–621.

[10] R. Xu, H. Xiang, X. Xia, X. Han, J. Liu, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Automat.*, 2021, pp. 2583–2589.

[11] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 4874–4886.

[12] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake LiDAR in 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9243–9252.

[13] W. G. Najm et al., "Pre-crash scenario typology for crash avoidance research," Dept. Transportation, Nat. Highway Traffic Saf. Admin., USA, 2007.

[14] T. Wang et al., "Deepaccident: A motion and accident prediction benchmark for V2X autonomous driving," 2023, *arXiv:2304.01168*.

[15] A. R. Khan et al., "DSRC technology in vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) IoT system for intelligent transportation system (ITS): A review," in *Proc. Recent Trends Mechatronics Towards Ind. 4.0*, 2021, pp. 97–106.

[16] C. M. Elias, O. M. Shehata, E. I. Morgan, and C. Stiller, "Emerging of V2X paradigm in the development of a ROS-based cooperative architecture for transportation system agents," in *Proc. IEEE Intell. Veh. Symp.*, 2022, pp. 1303–1308.

[17] N. Bouchemal and S. Kallel, "Testbed of V2X infrastructure for autonomous vehicles," *Ann. Telecommun.*, vol. 76, pp. 731–743, 2021.

[18] Z. Liu, H. Song, H. Tan, H. Hao, and F. Zhao, "Evaluation of the cost of intelligent upgrades of transportation infrastructure for intelligent connected vehicles," *J. Adv. Transp.*, vol. 1, pp. 1–15, 2022.

[19] P. Schörner, M. Conzelmann, T. Fleck, M. R. Zofka, and J. M. Zöllner, "Park my car! automated valet parking with different vehicle automation levels by V2X connected smart infrastructure," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 836–843.

[20] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "TUMTraf V2X cooperative perception dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22668–22677.

[21] S. U. Bhover, A. Tugashetti, and P. Rashinkar, "V2X communication protocol in VANET for co-operative intelligent transportation system," in *Proc. Int. Conf. Innov. Mechanisms Ind. Appl.*, 2017, pp. 602–607.

[22] I. Wahid et al., "Vehicular ad hoc networks routing strategies for intelligent transportation system," *Electronics*, vol. 11, 2022, Art. no. 2298.

[23] X. Gu et al., "Intelligent surface aided D2D-V2X system for low-latency and high-reliability communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 11624–11636, Nov. 2022.

[24] N. Raza, S. Jabbar, J. Han, and K. J. Han, "Social vehicle-to-everything (V2X) communication model for intelligent transportation systems based on 5G scenario," in *Proc. 2nd Int. Conf. Future Netw. Distrib. Syst.*, 2018, pp. 1–8.

[25] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," in *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.

[26] C. Hoymann et al., "LTE release 14 outlook," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 44–49, Jun. 2016.

[27] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tut.*, vol. 23, no. 1, pp. 222–255, First Quarter 2021.

[28] T. Higuchi, M. Giordani, A. Zanella, M. Zorzi, and O. Altintas, "Value-anticipating v2v communications for cooperative perception," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1947–1952.

[29] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29541–29552.

[30] Y. Li et al., "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.

[31] Y. Lu et al., "Robust collaborative 3D object detection in presence of pose errors," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 4812–4818.

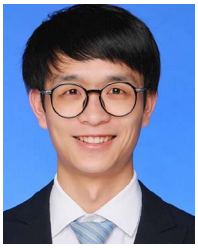
[32] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *Proc. Conf. Robot Learn.*, 2022, pp. 2062–2072.

[33] P. Gao, R. Guo, H. Lu, and H. Zhang, "Regularized graph matching for correspondence identification under uncertainty in collaborative perception," in *Proc. Robot.: Sci. Syst. XVI*, 2020. [Online]. Available: https://par.nsf.gov/servlets/purl/10215121
[34] 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 *Proc. 17th Eur. Conf. Comput. Vis.*, Israel, Springer, 2022, pp. 107–124.
[35] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21329–21338.
[36] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Proc. Conf. Robot Learn.*, 2022, pp. 989–1000.
[37] Y. Hu et al., "Pragmatic communication in multi-agent collaborative perception," 2024, *arXiv:2401.12694*.
[38] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 316–332.
[39] S. Wei et al., "Asynchrony-robust collaborative perception via bird's eye view flow," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 28462–28477.
[40] Y. Lu, Y. Hu, Y. Zhong, D. Wang, S. Chen, and Y. Wang, "An extensible framework for open heterogeneous collaborative perception," 2024, *arXiv:2401.13964*.
[41] C. leaderboard, 2019. [Online]. Available: https://leaderboard.carla.org/leaderboard/
[42] F. Codevilla, A. M. López, V. Koltun, and A. Dosovitskiy, "On offline evaluation of vision-based driving models," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, Springer, 2018, pp. 246–262.
[43] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "Reasonnet: End-to-end driving with temporal and global reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13723–13733.
[44] Y. Li et al., "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10914–10921, 2022.
[45] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17231–17241.
[46] R. Xu et al., "V2V4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2023, pp. 13712–13722.
[47] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "OpenCDA: An open cooperative driving automation framework integrated with co-simulation," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 1155–1162.
[48] M. Toromanoff, É. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," 2019, *arXiv: 1911.10868*.
[49] R. Chekroun, M. Toromanoff, S. Hornauer, and F. Moutarde, "GRI: General reinforced imitation and its application to vision-based autonomous driving," 2021, *arXiv:2111.08575*.
[50] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15590–15599.
[51] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. V. Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 15202–15212.
[52] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.
[53] S. Casas, A. Sadat, and R. Urtasun, "MP3: A unified model to map, perceive, predict and plan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14398–14407.
[54] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," 2019, *arXiv: 1912.12294*.
[55] F. Codevilla, M. Müller, A. Dosovitskiy, A. M. López, and V. Koltun, "End-to-end driving via conditional imitation learning," 2017, *arXiv: 1710.02410*.
[56] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," 2019, *arXiv: 1904.08980*.
[57] X. Jia, Y. Gao, L. Chen, J. Yan, P. L. Liu, and H. Li, "DriveAdapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving," 2023, *arXiv:2308.00398*.
[58] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, "Carla: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.
[59] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern*, 2020, pp. 11618–11628.
[60] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 12689–12697.
[61] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Comput. Vis.–ECCV 2020: 16th Eur. Conf., Glasgow, UK, August 2328, 2020, Part XIV 16*. Springer Int. Publishing, 2020, pp. 194–210.
[62] P. Wu, S. Chen, and D. N. Metaxas, "Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11382–11392.
[63] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.
[64] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv: 1904.07850*.
[65] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6318–6327.
[66] H. Xiang, R. Xu, and J. Ma, "HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 284–295.
[67] Q. Chen, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, 2019, pp. 88–100.
[68] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[69] H. Yu et al., "V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5486–5495.
[70] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8551–8560.
[71] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv: 1802.01436*.

Genjia Liu received the BE degree in information engineering from Shanghai Jiao Tong University, in 2021. He is currently working toward the PhD degree with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University. His research interests include graph machine learning and collaborative autonomous driving.

Yue Hu received the BE and MS degrees in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2017 and 2020, respectively. She is currently working toward the PhD degree with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University since 2021. Her research interests include multi-agent collaboration, communication efficiency, and 3D vision.

Chenxin Xu is working toward the joint PhD degree with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University and with Electrical and Computer Engineering, National University of Singapore since 2019. His research interests include trajectory prediction and multi-agent system. He is the reviewer of some prestigious international journals and conferences, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, CVPR, ICCV, ICML, and NeurIPS.



Weibo Mao received the BE degree from Shanghai Jiao Tong University, in 2021. He was working toward the MS degree when the work was done. His research interests include trajectory prediction and memory intelligence.

Junhao Ge is working toward the BE degree with Cooperative Medianet Innovation Center in Shanghai Jiao Tong University. His current research field includes traffic simulation and collaborative data generation.

Zhengxiang Huang is working toward the BE degree with Computer Science Department, Shanghai Jiao Tong University. His current research field includes collaborative data generation and machine learning system design.

Yifan Lu received the BE degree in computer science from Shanghai Jiao Tong University, in 2022. He is currently working toward the MS degree with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University. His current research field includes 3D scene data simulation and generation.

Yinda Xu received the BS and MS degrees in electrical engineering from Zhejiang University, in 2017 and 2021, respectively. He is currently working toward the PhD degree in computer science with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University since 2023. He also served as an algorithm engineer with the industry of Autonomous Driving (Momenta.ai and Deeproute.ai, respectively) between 2021 and 2023. His research interests include robot learning and embodied AI.

Junkai Xia received the BE degree in information engineering from Shanghai Jiao Tong University, in 2022. He is currently working toward the MS degree with SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University and the Ingénieur Polytechnicien Program at École Polytechnique. His research interests include traffic scenarios generation and collaborative autonomous driving.

Yafei Wang (Member, IEEE) received the BS degree in internal combustion engine from Jilin University, Changchun, China, in 2005, the MS degree in vehicle engineering from Shanghai Jiao Tong University, Shanghai, China, in 2008, and the PhD degree in electrical engineering from The University of Tokyo, Tokyo, Japan, in 2013. From 2008 to 2010, he was with automotive industry for nearly two years. From 2013 to 2016, he was a postdoctoral researcher with the University of Tokyo. He is currently a associate professor with the School of Mechanical Engineering, Shanghai Jiao Tong University. His research interests include state estimation and control for connected and automated vehicles.

Siheng Chen (Senior Member, IEEE) received the doctorate degree from Carnegie Mellon University, in 2016. He is a tenure-track associate professor with Shanghai Jiao Tong University. He was an autonomy engineer with Uber Advanced Technologies Group (ATG), working on self-driving cars. His work on sampling theory of graph data received the 2018 IEEE Signal Processing Society Young Author Best Paper Award. He contributed to the project of scene-aware interaction, receiving MERL President's Award. His research interests include collective intelligence and AI agents.