

Research Challenges and Progress in the End-to-End V2X Cooperative Autonomous Driving Competition

Ruiyang Hao¹, Haibao Yu^{2,1,*}, Jiaru Zhong¹, Chuanye Wang¹, Jiahao Wang¹,
Yiming Kan³, Wenxian Yang¹, Siqi Fan¹, Huilin Yin³, Jianing Qiu⁴, Yao Mu^{2,5},
Jiankai Sun⁶, Li Chen^{2,7}, Walter Zimmer⁸, Dandan Zhang⁹, Shanghang Zhang¹⁰,
Mac Schwager⁶, Ping Luo^{2,*}, Zaiqing Nie^{1,*}

¹ Tsinghua University, ² Hong Kong University, ³ Tongji University,

⁴ Chinese University of Hong Kong, ⁵ Shanghai Jiao Tong University, ⁶ Stanford University,

⁷OpenDriveLab, ⁸Technical University of Munich, ⁹Imperial College London,

¹⁰Peking University

Abstract

With the rapid advancement of autonomous driving technology, vehicle-to-everything (V2X) communication has emerged as a key enabler for extending perception range and enhancing driving safety by providing visibility beyond the line of sight. However, integrating multi-source sensor data from both ego-vehicles and infrastructure under real-world constraints, such as limited communication bandwidth and dynamic environments, presents significant technical challenges. To facilitate research in this area, we organized the End-to-End Autonomous Driving through V2X Cooperation Challenge, which features two tracks: cooperative temporal perception and cooperative end-to-end planning. Built on the UniV2X framework and the V2X-Seq-SPD dataset, the challenge attracted participation from over 30 teams worldwide and established a unified benchmark for evaluating cooperative driving systems. This paper describes the design and outcomes of the challenge, highlights key research problems including bandwidth-aware fusion, robust multi-agent planning, and heterogeneous sensor integration, and analyzes emerging technical trends among top-performing solutions. By addressing practical constraints in communication and data fusion, the challenge contributes to the development of scalable and reliable V2X-cooperative autonomous driving systems.

1. Introduction

Autonomous driving has witnessed rapid advancements in recent years, driven by the progress of perception [8, 11], planning [37, 38], and end-to-end [12, 36, 41, 54, 66]

technologies. However, the prevailing paradigm of single-vehicle autonomy, which relies solely on onboard sensors and processing units, is inherently limited by its constrained field of view, susceptibility to occlusions, and lack of awareness of occluded or distant objects [3, 81]. These limitations pose significant challenges in complex urban environments, where safety-critical decision-making demands a more comprehensive understanding of the surrounding traffic context. In particular, scenarios involving intersections, occluded crosswalks, or multi-lane merges often expose the limitations of local perception and lead to suboptimal or unsafe maneuvers.

To address these constraints, vehicle-to-everything (V2X) cooperation has emerged as a promising paradigm [79, 88]. By enabling ego-vehicles to exchange real-time sensory and state information with roadside infrastructure and nearby agents, V2X cooperation extends perception beyond the line of sight and supports more informed and robust perception and final planning performance [57, 58]. The integration of cooperative perception and cooperative planning is thus becoming a pivotal frontier in the development of scalable and safe embodied intelligence systems for autonomous driving.

Despite the growing body of research on V2X-enabled systems, developing deployable and generalizable algorithms for cooperative driving remains challenging. Real-world constraints such as limited communication bandwidth [18], latency, and heterogeneous sensor configurations [80] complicate the design of end-to-end solutions. Moreover, robust fusion of multi-view, multi-agent data [48, 82] for downstream planning under dynamic scenarios is still an open research problem. These challenges are further compounded by the asynchronous nature of inter-agent communication, variable sensor quality across nodes, and

*Corresponding author.

the lack of standardized protocols for representation and fusion.

To promote research in this direction, we organized the **first End-to-End Autonomous Driving through V2X Cooperation Challenge** as part of the Multi-Agent Embodied Intelligent Systems (MEIS) Workshop @ CVPR 2025 (More details in this [link](#)). The challenge aims to benchmark and advance the state-of-the-art in V2X-enhanced driving agents through two complementary tracks: (1) Cooperative Temporal Perception, focusing on multi-agent detection and tracking; and (2) Cooperative End-to-End Planning, targeting V2X-aware sensor-to-action learning. Built upon the open-source UniV2X framework [87] and V2X-Seq-SPD dataset [86], this challenge provides a reproducible platform for evaluating cooperative perception and planning systems in real-world urban driving scenarios.

This paper presents a comprehensive summary of the competition design, research challenges, participant solutions, and key findings. Specifically, we (i) outline the motivation and structure of the challenge, (ii) identify critical research issues emerging from participant submissions, (iii) analyze the technical trends and progress demonstrated, and (iv) discuss future directions for cooperative multi-agent autonomous driving systems.

2. Background

2.1. Related Benchmarks and Challenges

Over the past decade, a variety of datasets and benchmarks have been proposed to evaluate the perception and planning capabilities of autonomous driving systems. Notable examples include nuScenes [5], Waymo Open Dataset [61], Argoverse [10], nuplan-based dataset [6, 23, 34, 55], and the CARLA-based dataset [15, 24, 40], which focus on object detection, motion prediction, and planning under the single-agent paradigm. While these benchmarks have significantly contributed to the development of perception, decision-making and end-to-end pipelines, they largely neglect the potential of inter-agent cooperation and V2X communication [49, 84, 94], which are essential for overcoming occlusion and limited sensor range in congested urban environments. These limitations hinder the modeling of realistic traffic scenes involving multi-agent interactions and limited visibility, such as those found at intersections, curved roads, or occluded pedestrian zones.

Several recent efforts, such as DAIR-V2X [84], V2X-Sim [49], TUMTraf [94], V2X-Real [73], V2v4Real [78], RCooper[33], Griffin[64] and V2XSet [77], have introduced datasets and tasks tailored for cooperative perception. These datasets incorporate multi-view inputs from vehicles and roadside infrastructure, enabling exploration of early and intermediate sensor fusion methods to enhance 3D detection and tracking performance. However, most of these

benchmarks remain focused on perception tasks, with relatively limited emphasis on downstream planning [86]. In particular, few existing datasets provide a unified setting where both perception and planning tasks are evaluated with the same data and scenario structure.

The End-to-End V2X Cooperation Challenge addresses this gap by integrating cooperative perception and planning tasks into a two-track benchmark framework. It builds on the open-source UniV2X system [87] and the V2X-Seq-SPD dataset [86], which jointly support detection, tracking, and motion planning based on multi-agent sensor inputs. By standardizing the task input/output formats and providing an end-to-end development pipeline, the challenge enables participants to explore perception-to-planning integration under realistic multi-view sensing conditions. The use of distinct sensing viewpoints and calibration setups naturally reflects challenges in real-world cooperative driving deployments. This joint benchmark structure promotes a more comprehensive understanding of algorithm performance in multi-agent urban environments.

2.2. UniV2X Framework and Dataset

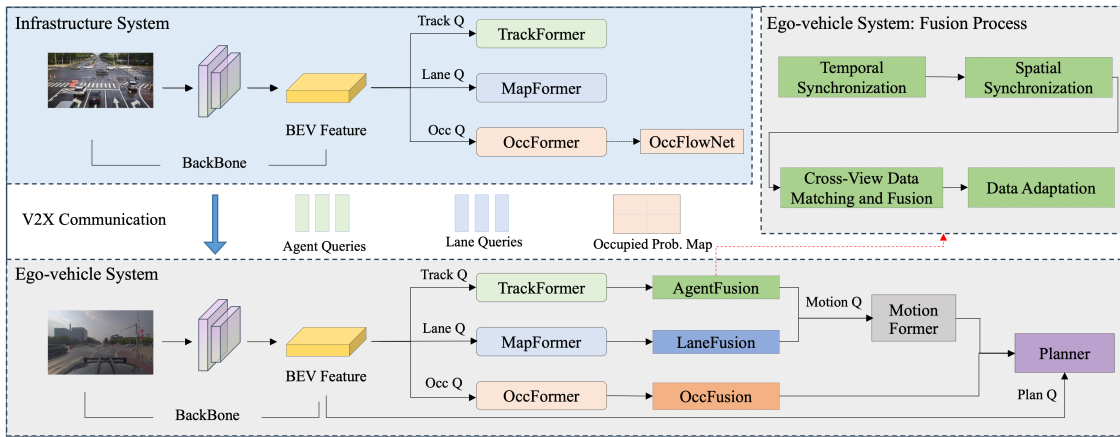
The challenge is built upon the open-source UniV2X framework [87], which serves as the first unified end-to-end pipeline for cooperative autonomous driving. UniV2X integrates multiple key modules—cooperative perception, intermediate representation learning, occupancy forecasting, and planning—into a cohesive architecture. It supports both vehicle-side and infrastructure-side sensing, facilitating multi-view feature alignment and fusion through a hybrid sparse-dense transmission protocol. This allows for efficient message passing while mitigating the communication burden common in dense feature maps, particularly in bird’s-eye-view (BEV) frameworks.

The underlying dataset, V2X-Seq-SPD [86], provides synchronized and calibrated sensor recordings from ego vehicles and roadside units (RSUs), including front-view images, LiDAR point clouds (converted to BEV), and semantic commands. Ground-truth labels for 3D object detection, tracking, and future trajectories are included, allowing evaluation across both perception and planning tasks. The dataset reflects diverse urban driving scenarios with dynamic traffic flow, intersections, and occlusions—thus capturing key challenges faced by V2X systems.

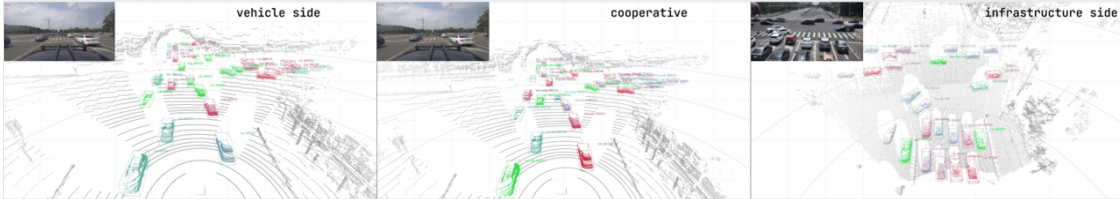
UniV2X serves as the official baseline for both tracks of the competition. In Track 1, it provides a fully sparse 3D detection and tracking solution with anchor-guided query fusion. In Track 2, it offers a modular sensor-to-planning pipeline that leverages query-based adapters to dynamically route fused features into planning heads. These designs provide participants with a strong starting point and encourage innovation in overcoming current bottlenecks.

Table 1. Comparison of autonomous driving datasets by data source, held competitions, task description, V2X support, end-to-end (E2E) support. *Abbreviations:* V2X = V2X model support, E2E = End-to-End driving model support, Det = Detection, Trk = Tracking, MPre = Motion Prediction, Pla = Planning (Open-loop), CL = Closed-loop evaluation

Dataset	Reality	Competition	Task description	V2X	E2E
nuScenes [5]	Real	CVPRW19, ICRAW20, ICRAW21	Det,Trk,MPre,Pla	✗	✓
Waymo [61]	Real	WOD20-25	Det,Trk,MPre,Pla	✗	✓
Argoverse [10]	Real	CVPRW22, CVPRW23, CVPRW25	Det,Trk,MPre,Pla	✗	✓
CARLA [20]	Sim	CVPRW19, NIPSW20-22, CVPRW24	Det,Trk,MPre,Pla,CL	✗	✓
NAVSIM [23]	Real	CVPRW24, CVPRW25, ICCVW25	Det,Trk,MPre,Pla,CL	✗	✓
DAIR-V2X [84]	Real	AIR-Apollo23	Det	✓	✗
TUMTraf [94]	Real	ICCVW25	Det	✓	✗
V2v4Real [78]	Real	-	Det	✓	✗
V2X-Sim [49]	Sim	-	Det,Trk	✓	✗
V2X-Seq [86]	Real	CVPRW25 (Ours)	Det,Trk,Pla	✓	✓



(a) Challenge Baseline: UniV2X Architecture [87].



(b) An example of V2X-Seq-SPD Dataset [86]

Figure 1. Challenge Baseline UniV2X [87] and V2X-Seq-SPD Dataset [86]

3. Challenge Design

3.1. Task Setup and Evaluation Metrics

The challenge comprises two complementary tracks designed to evaluate different aspects of V2X cooperative autonomous driving: Cooperative Temporal Perception and Cooperative End-to-End Planning.

1) Track 1: Cooperative Temporal Perception This track focuses on cooperative 3D detection and multi-object tracking in urban scenarios involving ego vehicles and roadside infrastructure. Each participant receives a stream of

synchronized multi-agent sensor data, including front-view camera images from both ego vehicles and roadside units (RSUs), along with camera calibration parameters, vehicle ego states, and high-level command information. These inputs are drawn from realistic driving sequences, featuring intersections, dynamic obstacles, and partial observability across viewpoints.

The primary task is to detect vehicles of the merged “Car” category in 3D space and associate consistent tracking IDs across time, leveraging both temporal information and cross-agent collaboration. The design emphasizes the need for participants to model how complementary view-

points—e.g., an RSU’s top-down view and the ego vehicle’s forward-facing camera—can be fused over time to disambiguate occluded or partially visible objects.

To evaluate performance, we employ two widely used metrics in cooperative perception benchmarks: mean Average Precision (mAP), which measures spatial detection accuracy, and Average Multi-Object Tracking Accuracy (AMOTA), which captures temporal consistency of object identities. The final evaluation score is computed as the unweighted average of the two ($0.5 \text{ mAP} + 0.5 \text{ AMOTA}$), allowing fair comparison between detection and tracking capabilities.

This task encourages the design of fusion algorithms capable of aligning features from spatially distinct viewpoints and maintaining identity consistency across frames, even under object occlusion, motion blur, or disjoint agent fields of view. It also offers a platform to evaluate temporal modeling techniques such as query-based memory propagation, agent-aware attention, and cross-frame association strategies. Ultimately, this track aims to advance the robustness and scalability of cooperative perception systems deployed in real-world driving environments.

2) Track 2: Cooperative End-to-End Planning This track aims to evaluate complete sensor-to-planning pipelines that generate future motion trajectories based on fused perception from multiple agents. Participants are tasked with predicting a sequence of future waypoints over a 5-second horizon, using the same input modalities as in Track 1, including ego and infrastructure camera images, calibration data, command signals, and current ego vehicle states.

Unlike modular approaches that decouple perception and planning, this track encourages joint reasoning across the full autonomous driving stack, from raw sensor input to trajectory-level output. The data spans a variety of challenging urban situations—such as intersection negotiation, overtaking, and lane turning—requiring the agent to anticipate dynamic scene evolution and react safely under partial observability.

Performance is assessed using three complementary metrics:

- L2 Error, which measures the Euclidean distance between predicted and ground-truth waypoints, reflecting trajectory accuracy;
- Collision Rate, which quantifies how often the predicted trajectory intersects with other traffic participants;
- Off-road Rate, which measures deviation from the drivable area and thus reflects constraint violation or poor lane adherence.

To obtain a comprehensive evaluation, each metric is averaged at three future timestamps (2.5s, 3.5s, 4.5s), balancing short-term responsiveness and long-term planning qual-

ity. A min-max normalization is applied based on predefined reference ranges, and the final score is computed as a weighted sum: $0.5 \times \text{normalized L2 Error} + 0.25 \times \text{normalized Collision Rate} + 0.25 \times \text{normalized Off-road Rate}$.

This track emphasizes planning robustness in complex multi-agent scenes, and highlights the importance of integrating spatial-temporal reasoning, intent understanding, and safety guarantees into the learning process. It offers a testbed for evaluating architectures such as transformer-based fusion planners, modular policy networks, and multi-head decoding strategies under realistic traffic conditions.

3.2. Participation

Over 30 teams registered, with 5 finalists achieving ranked results. Participants came from academic institutions and industry research labs across China, Japan, the Middle East, the United States, and Europe. Most teams adopted the open-source UniV2X baseline as a foundation, developing innovative fusion architectures and planning strategies on top of it. To recognize outstanding solutions, the challenge organizers awarded monetary prizes to the top-ranked teams in each track. The diversity in approaches—from sparse query-based perception pipelines to modular planning frameworks—reflects the richness and complexity of the V2X cooperation landscape.

4. Research Challenges

The V2X Cooperation Challenge was intentionally designed to reflect real-world difficulties in cooperative autonomous driving. Through analysis of participant submissions and related work, several core research challenges emerged, spanning multi-agent fusion, communication efficiency, planning robustness, and realistic deployment modeling. These challenges reveal both the current limitations of existing solutions and promising directions for future research.

Multi-Agent Sensor Fusion under Bandwidth Constraints. A fundamental challenge lies in effectively aggregating heterogeneous sensor inputs from ego vehicles and infrastructure, particularly under tight communication budgets. Naïvely transmitting dense feature maps from multiple viewpoints (e.g., bird’s-eye view or BEV) quickly exhausts bandwidth and leads to latency bottlenecks [9, 85]. More recent methods employ sparse query-based methods and transformer for cooperative representations embedding and fusion [25, 69, 92, 93]. This necessitates the development of sparse, information-aware representations that can preserve critical scene understanding while minimizing message size.

Top-performing teams in Track 1 adopted query-based attention fusion mechanisms, such as anchor-guided sparse

queries and cooperative instance denoising, to mitigate these issues. However, challenges remain in dynamically selecting which information to transmit, how to encode uncertainty from partial observations, and how to align features from spatially and temporally misaligned views. Efficient and adaptive feature compression strategies, potentially guided by learned importance scores, are still under-explored.

Robust Planning in Dynamic and Complex Environments Track 2 highlighted the difficulty of producing reliable motion plans in highly dynamic, multi-agent urban scenes. When relying on fused perception from multiple sources, temporal inconsistency, latency-induced misalignment, and partial observability can significantly degrade planning performance [74, 90]. Ego agents must reason not only about static obstacles and drivable regions, but also about the future intentions and potential interactions of nearby vehicles.

Moreover, the planning module must cope with command diversity (e.g., turns, stops, merges) and structural uncertainty in intersections or occluded traffic elements. These issues call for more robust multi-modal trajectory prediction, tighter integration of intent inference, and online failure recovery mechanisms in planning architectures.

Communication-Aware System Design and Modeling Realistic V2X deployment is subject to a range of networking imperfections, including packet loss [53], varying latency, and intermittent connectivity [56]. However, most existing cooperative driving methods assume idealized or fixed-delay channels [83, 87]. The challenge dataset incorporates limited communication constraints (e.g., message size limits), but further progress depends on building systems that are explicitly aware of and adaptive to the communication channel.

Few teams explored bandwidth-adaptive fusion strategies or uncertainty-aware planning under degraded connectivity. Future systems can reason about when, what, and how to communicate, potentially leveraging learned policies or information-theoretic objectives. Modeling the trade-off between perception gain and communication cost remains an open research question, especially when agents must operate asynchronously or with partial participation.

Generalization and Transfer under Domain Shift Although the dataset provides consistent sensor configurations, real-world deployments often involve heterogeneous sensor suites, diverse camera placements, and varying calibration quality [89, 91]. Designing fusion and planning models that generalize across these variations remains challenging. Furthermore, reliance on known object models or

tightly coupled training scenarios can hinder transferability to new domains.

Some participants addressed this by employing modular architectures with adaptable feature backbones, but the issue of domain robustness under limited supervision persists. Robustness to weather, lighting, and sensor degradation was not evaluated in this challenge but constitutes a necessary extension for real-world readiness.

5. Progress and Analysis

The competition attracted a diverse set of participants from academia and industry, contributing a broad spectrum of approaches across cooperative perception, feature fusion, and planning architectures. While implementations varied in complexity and formulation, a number of converging trends emerged. In particular, the most effective solutions reflect a growing shift toward modular, interpretable, and task-centric designs that emphasize structured information flow between agents and system components.

This section introduces the top-performing solutions from each track of the challenge. These methods represent state-of-the-art approaches in cooperative 3D perception and end-to-end planning with V2X input, and demonstrate the effectiveness of structured representations and adaptive fusion strategies.

5.1. Track 1 Top Method: SparseCoop

Wang et al. from Tsinghua University proposed SparseCoop, a fully sparse, instance-centric cooperative perception framework (Fig. 2) designed to simultaneously address the communication and computational bottlenecks of traditional dense BEV-based approaches and the challenges of newer sparse, query-based methods, including their insufficiently expressive query representations for handling real-world scenarios and their inherent training instability.

At its core, SparseCoop introduces the concept of the anchor-aided instance query, where each object is represented by a rich feature vector coupled with an explicit anchor box. The anchor includes structured geometric and motion attributes—namely the object’s 3D position, dimensions, velocity, and yaw. This representation enables precise, physically grounded fusion across agents with different viewpoints and asynchronous observations.

To address the training instability common in sparse query systems, SparseCoop incorporates a cooperative instance denoising task. During training, noise is deliberately added to ground-truth objects in the form of “Observation Noise” and “Transformation Noise”. The model is then supervised to recover clean object states, which generates a robust and abundant stream of positive training signals. This design improves convergence speed and accuracy.

SparseCoop achieves state-of-the-art detection and tracking performance, demonstrating strong robustness to

viewpoint diversity, temporal misalignment, and perception noise under the V2X-Seq-SPD benchmark.

5.2. Track 2 Top Method: MAP

The MAP framework (Fig. 3), proposed by Kan et al. from Tongji University, emerged from a critical reevaluation of the role of perception in end-to-end autonomous driving. While many recent approaches favor minimal input paradigms that rely solely on ego history, MAP challenges this trend by demonstrating that explicitly and effectively utilizing semantic map information can substantially enhance planning robustness.

At its core, MAP transforms semantic segmentation from a passive supervision target into a direct planning input. It introduces a two-branch query generation pipeline: The Ego-status-guided Planning (EP) module leverages the current ego state for trajectory planning, while the other extracts map-guided priors through a Plan-enhancing Online Mapping (POM) module. The resulting semantic-aware and ego-status-driven queries are then fused via a learned Weight Adapter, which adaptively predicts a fusion scalar α based on the current driving context.

This adaptive weighting mechanism allows the planner to rely more on ego information in simple scenes, and to prioritize semantic priors in complex or ambiguous scenarios, leading to context-sensitive and reliable decision-making. Importantly, MAP achieves strong performance without stacked modules such as tracking or occupancy prediction.

On the DAIR-V2X-Seq-SPD benchmark, MAP improves the overall normalized score by 44.5% over the UniV2X baseline and ranks first on the planning leaderboard, showing competitive results across all sub-metrics, including L2 error and off-road rate.

6. Future Directions

The challenge results and observed limitations across both tracks reveal a clear gap between current benchmark performance and the requirements for real-world deployment of cooperative autonomous driving under V2X settings. To close this gap, future research should address the following interconnected directions, progressing from foundational communication challenges to system-level adoption.

6.1. Realistic V2X Communication Modeling

Most current solutions assume ideal or simplified communication channels with perfect message delivery [77, 83]. In practice, V2X networks suffer from variable latency, intermittent connectivity, and packet drops caused by interference or congestion [19, 59]. Future benchmarks and algorithms should embed communication-aware learning by:

- Simulating packet loss models grounded in empirical wireless measurements [32],

- Incorporating delay-aware fusion mechanisms that reason with stale or missing data [1, 76],
- Designing redundancy-aware protocols or modules to prioritize safety-critical and planning-oriented information under bandwidth constraints [39].

These modules would enable robust agents that adapt to both perceptual uncertainty and communication reliability.

6.2. Bandwidth-Adaptive and Task-Aware Fusion

Fusion strategies must adapt to both bandwidth availability and task requirements to ensure reliable performance at scale [85]. However, most existing approaches remain static and fail to capture the dynamic trade-offs between efficiency and accuracy. Beyond sparse fusion, future feature fusion modules may benefit from:

- Information-theoretic feature selection to maximize task utility per transmitted bit [14, 52],
- Hierarchical encoding schemes for coarse-to-fine updates based on link conditions [46, 50],
- Task-driven prioritization, sending safety-critical and planning-critical cues (e.g., nearby dynamic agents) more aggressively than static context.

These adaptive fusion mechanisms would support graceful degradation and efficient resource utilization in large-scale V2X deployments.

6.3. Generalization Across Heterogeneous Agents and Scenarios

Real-world deployments will involve heterogeneous vehicles and infrastructure with varying sensors, fields of view, and computational capabilities [65, 70, 89]. Robust fusion and planning under such diversity can be supported by:

- Calibration-agnostic fusion mechanisms or frameworks tolerant to partial or inaccurate alignment [27],
- Meta-learning or domain adaptation methods to generalize across sensor setups, cities, and conditions [43, 45],
- Scalable fusion topologies that support dynamic participation as agents enter or leave the scene [62].

Addressing these challenges will significantly improve the deployability of cooperative driving systems across different geographies and manufacturers.

6.4. Air-Ground Collaboration

The growth of the low-altitude economy [67] will introduce drones as additional sensing and communication agents for autonomous driving [26, 30, 35, 64]. By acting as “free viewpoints,” drones can enrich situational awareness in dense traffic and occluded intersections. At the same time, their deployment raises issues of sensor vibration, communication overhead, and limited endurance. Air-ground cooperative systems can be advanced by:

- Addressing sensor vibration artifacts in drone-mounted

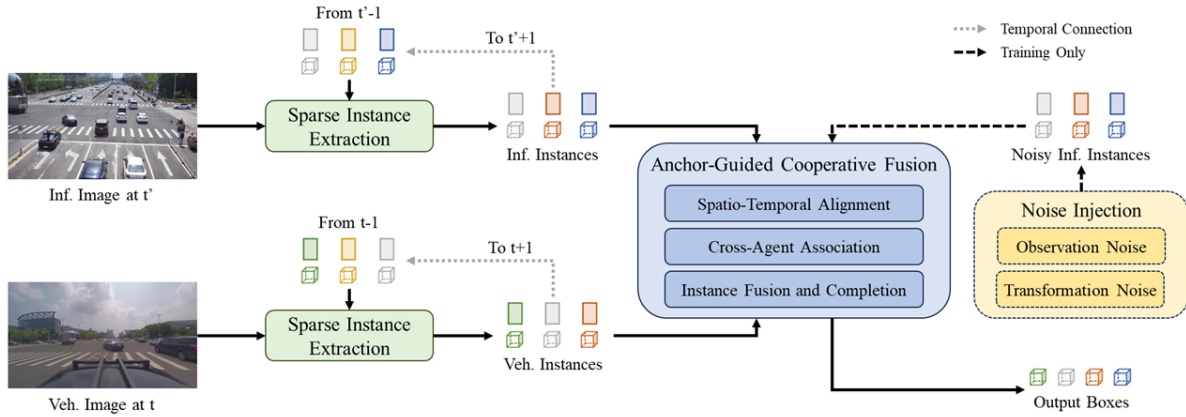


Figure 2. Architecture of **SparseCoop**, the top-ranked solution in **Track 1: Cooperative Temporal Perception**. The method adopts a fully sparse cooperative detection and tracking pipeline, where each object is represented by an *anchor-aided instance query* containing structured geometric attributes (position, size, velocity, orientation) and semantic features. Cross-agent fusion is performed directly at the object level without relying on intermediate BEV features. A *cooperative instance denoising task* is applied during training to inject noise into ground-truth anchors and improve convergence robustness through reconstruction supervision.

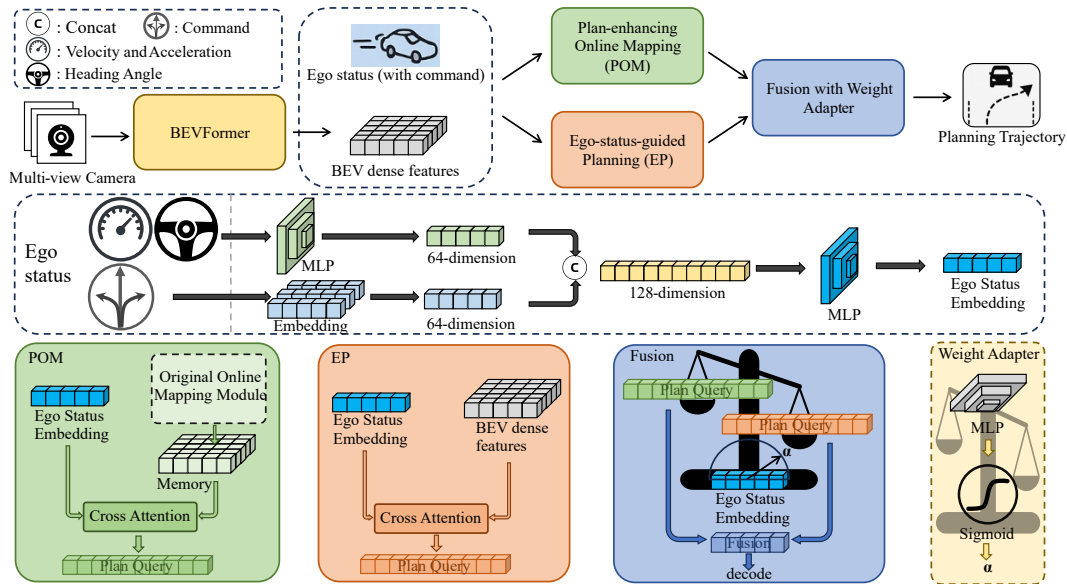


Figure 3. Architecture of **MAP**, the top-ranked solution in **Track 2: Cooperative End-to-End Planning**. This planning-centric framework explicitly incorporates semantic map information into trajectory generation. It consists of two query-generation branches: the *Plan-enhancing Online Mapping (POM)* module extracts semantic priors from segmentation outputs, while the *Ego-status-guided Planning (EP)* module models motion dynamics. A learned adapter fuses the two planning queries with weights conditioned on ego state, enabling context-aware trajectory generation under varying traffic complexity.

perception, which can degrade image quality and downstream detection and planning performance [4],

- Developing bandwidth-aware fusion algorithms for transmitting aerial data efficiently,
- Coordinating multiple drones under battery and flight-time constraints for persistent coverage [13, 17].

Leveraging aerial viewpoints in coordination with ground vehicles could unlock richer situational awareness

and enable safer, large-scale deployment of cooperative driving systems.

6.5. Interpretability, Safety, and Standardization

For cooperative systems to be adopted in safety-critical applications such as autonomous driving, interpretability, verifiability and standardization become essential [2, 44, 87]. Progress can be made through:

- Transparent fusion architectures that expose the contribution of each agent and observation [44, 82],
- Uncertainty quantification in cooperative perception and planning outputs [7, 47],
- Conformance to communication and safety standards such as SAE J2735 [60] or ETSI ITS-G5 [63].

These directions are critical to ensuring that cooperative driving systems are not only performant but also trustworthy and certifiable for real-world use.

6.6. Language as Communication Medium

Recent advances in Large Vision-Language Models [42, 68, 75] have opened the possibility of using natural language as a medium for V2X communication [16, 21, 28, 29, 31, 51, 71, 72]. Language-based communication promises transparency, efficiency, and interoperability, but its deployment in safety-critical driving scenarios remains largely unexplored. Future research should address:

- Developing structured and unambiguous protocols for language-based V2X exchanges to avoid ambiguity [28, 71],
- Combining natural language with traditional feature-level or state-level communication in hybrid pipelines,
- Studying robustness to multilingual, noisy, or adversarial language inputs in cooperative driving,
- Exploring decision-level negotiation mechanisms that go beyond perception sharing [22].

Addressing these challenges would transform natural language from a promising idea into a practical medium for cooperative autonomous driving.

6.7. Community and Ecosystem Development

Progress in V2X cooperative driving will be accelerated by cohesive community efforts and shared infrastructure. Key steps include:

- Continuing development of open-source toolkits, such as UniV2X, for full-stack experimentation,
- Expanding datasets to cover adverse conditions (e.g., night, rain, sensor failures),
- Establishing long-term multi-institutional benchmarks to ensure reproducibility and collaboration,
- Organizing V2X-specific competitions to stimulate innovation beyond single-agent autonomy.
- Considering broader ethical and policy implications of multi-agent cooperation, including inter-manufacturer trust and data governance among automotive original equipment manufacturers (OEMs), as well as data security and regulatory compliance.

Building such community resources and ecosystems will help translate academic advances into robust, deployable cooperative driving systems worldwide.

7. Conclusion

This paper has presented a comprehensive overview of the End-to-End Autonomous Driving through V2X Cooperation Challenge, organized as part of the MEIS Workshop @ CVPR 2025. The challenge was designed to advance the state of cooperative autonomous driving by rigorously evaluating perception and planning systems under realistic multi-agent and communication-constrained conditions. It comprised two tracks: cooperative temporal perception and end-to-end planning, built upon the open-source UniV2X framework [87] and the V2X-Seq-SPD dataset [86].

Participation from numerous teams highlighted both substantial progress and persistent limitations in V2X-enabled driving systems. The strongest submissions adopted sparse, query-based fusion, modular architectures, and temporal reasoning, achieving competitive results in both perception and planning tasks. At the same time, we summarize the critical open challenges, including communication-aware fusion, robust planning under partial observability, and generalization across heterogeneous agents.

These findings and insights from related research, emphasize that the development of cooperative driving systems must go beyond accuracy and efficiency, extending to adaptability, interpretability, and robustness in real-world deployments. The challenge has further underscored the importance of open benchmarks, reproducible baselines, and cross-community collaboration as essential drivers for translating academic innovation into deployable systems.

Looking forward, future editions of the challenge will broaden their scope to incorporate richer sensor modalities, more realistic and dynamic communication models, and increasingly diverse driving environments. By continuing to integrate technical advances with ethical, regulatory, and societal considerations, this initiative seeks to foster the design of safe, scalable, and intelligent multi-agent driving systems for the urban mobility ecosystems of tomorrow.

Acknowledgement

The authors would like to express their sincere gratitude to all participating teams of our challenge for their valuable contributions. In particular, we acknowledge the outstanding efforts of the team led by Ziyi Song and Dr. Sheng Zhou from Tsinghua University, as well as the team led by Dr. Ehsan Javanmardi from the University of Tokyo. We would also like to thank Xiangbo Gao and Dr. Zhengzhong Tu from Texas A&M University for their valuable suggestions on future work. We also gratefully acknowledge the sponsorship provided by the Multimedia Laboratory at the University of Hong Kong and Shanghai Songying Technology Co., Ltd. Furthermore, we sincerely acknowledge the support from the Wuxi Research Institute of Applied Technologies at Tsinghua University under Grant No. 20242001120.

References

- [1] Ahmed N. Ahmed, Siegfried Mercelis, and Ali Anwar. Delawarecol: Delay aware collaborative perception. *IEEE Open Journal of Vehicular Technology*, 6:1164–1177, 2025.
- [2] Moin Ali, Ali Nauman, Muhammad Ali Jamshed, Su Min Kim, and Junsu Kim. Vehicles-to-everything standardization, services and enhancements for intelligent transportation systems. *IEEE Communications Standards Magazine*, 2025.
- [3] Hamidreza Bagheri, Md Noor-A-Rahim, Zilong Liu, Haeyoung Lee, Dirk Pesch, Klaus Moessner, and Pei Xiao. 5g nr-v2x: Toward connected and cooperative autonomous driving. *IEEE Communications Standards Magazine*, 5(1):48–54, 2021.
- [4] Matteo Bertocco, Alessandro Brighente, Gianluca Ciattaglia, Ennio Gambi, Giacomo Peruzzi, Alessandro Pozzebon, and Susanna Spinsante. Malicious drone identification by vibration signature measurement: A radar-based approach. *IEEE Transactions on Instrumentation and Measurement*, 74:8004415, 2025.
- [5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [6] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.
- [7] Kunyang Cai, Ting Qu, Fen Liu, Hong Chen, and Lihua Xie. Cooperative perception with localization uncertainty: A cubature split covariance intersection framework. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):18006–18024, 2024.
- [8] Yingfeng Cai, Tianyu Luan, Hongbo Gao, Hai Wang, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. Yolov4-5d: An effective and efficient object detector for autonomous driving. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.
- [9] Cheng Chang, Jiawei Zhang, Kunpeng Zhang, Wenqin Zhong, Xinyu Peng, Shen Li, and Li Li. Bev-v2x: Cooperative birds-eye-view fusion and grid occupancy prediction via v2x-based data sharing. *IEEE Transactions on Intelligent Vehicles*, 8(11):4498–4514, 2023.
- [10] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.
- [11] Long Chen, Shaobo Lin, Xiankai Lu, Dongpu Cao, Hangbin Wu, Chi Guo, Chun Liu, and Fei-Yue Wang. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3234–3246, 2021.
- [12] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024.
- [13] Wu Chen, Jiayi Zhu, Jijia Liu, and Hongzhi Guo. A fast coordination approach for large-scale drone swarm. *Journal of Network and Computer Applications*, 221:103769, 2024.
- [14] Yihao Chen and Zefang Wang. An effective information theoretic framework for channel pruning. *arXiv preprint arXiv:2408.16772*, 2024.
- [15] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.
- [16] Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*, 2025.
- [17] Omer Chughtai, Nadia Nawaz Qadri, Zeeshan Kaleem, and Chau Yuen. Drone-assisted cooperative routing scheme for seamless connectivity in v2x communication. *IEEE Access*, 12:17369–17381, 2024.
- [18] Joseph Clancy, Darragh Mullins, Brian Deegan, Jonathan Horgan, Enda Ward, Ciarán Eising, Patrick Denny, Edward Jones, and Martin Glavin. Wireless access for v2x communications: Research, challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 26(3):2082–2119, 2024.
- [19] Baldomero Coll-Perales, M Carmen Lucas-Estañ, Takayuki Shimizu, Javier Gozalvez, Takamasa Higuchi, Sergei Avedisov, Onur Altintas, and Miguel Sepulcre. End-to-end v2x latency modeling and analysis in 5g networks. *IEEE Transactions on Vehicular Technology*, 72(4):5094–5109, 2022.
- [20] Contributors. Carla autonomous driving leaderboard, 2024.
- [21] Jiaxun Cui, Chen Tang, Jarrett Holtz, Janice Nguyen, Alessandro G Allievi, Hang Qiu, and Peter Stone. Towards natural language communication for cooperative autonomous driving via self-play. *arXiv preprint arXiv:2505.18334*, 2025.
- [22] Yiming Cui, Shiyu Fang, Peng Hang, and Jian Sun. A vehicle-infrastructure multi-layer cooperative decision-making framework. *arXiv preprint arXiv:2503.16552*, 2025.
- [23] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.
- [24] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.
- [25] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Quest: Query stream for practical cooperative percep-

- tion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18436–18442. IEEE, 2024.
- [26] Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. U2data: A large-scale cooperative perception dataset for swarm uavs autonomous flight. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7600–7608, 2024.
- [27] Fuji Fu, Jinfu Yang, Jiaqi Ma, and Jiahui Zhang. Self-supervised visual odometry based on scene appearance-structure incremental fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- [28] Xiangbo Gao, Keshu Wu, Hao Zhang, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Automated vehicles should be connected with natural language. *arXiv preprint arXiv:2507.01059*, 2025.
- [29] Xiangbo Gao, Yuheng Wu, Rujia Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. Langcoop: Collaborative driving with language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4226–4237, 2025.
- [30] Xiangbo Gao, Yuheng Wu, Fengze Yang, Xuewen Luo, Keshu Wu, Xinghao Chen, Yuping Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. Airv2x: Unified air-ground vehicle-to-everything collaboration. *arXiv preprint arXiv:2506.19283*, 2025.
- [31] Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. Stamp: Scalable task and model-agnostic collaborative perception. *arXiv preprint arXiv:2501.18616*, 2025.
- [32] Kevin Herman Muraro Gulate, João Paulo Javidi Da Costa, José Alfredo Ruiz Vargas, Antonio Santos Da Silva, Giovanni Almeida Santos, Yuming Wang, Christian Alfons Müller, Christoph Lipps, Rafael Timóteo de Sousa Júnior, Walter de Britto Vidal Filho, et al. Integrating cybersecurity in v2x: A review of simulation environments. *IEEE Access*, 2024.
- [33] Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Rcooper: A real-world large-scale dataset for roadside cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22347–22357, 2024.
- [34] Ruiyang Hao, Bowen Jing, Haibao Yu, and Zaiqing Nie. Styledrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. *arXiv preprint arXiv:2506.23982*, 2025.
- [35] Yunhao Hou, Bochao Zou, Min Zhang, Ran Chen, Shangdong Yang, Yanmei Zhang, Junbao Zhuo, Siheng Chen, Jiansheng Chen, and Huimin Ma. Agc-drive: A large-scale dataset for real-world aerial-ground collaboration in driving scenarios. *arXiv preprint arXiv:2506.16371*, 2025.
- [36] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.
- [37] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3903–3913, 2023.
- [38] Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving. *IEEE transactions on neural networks and learning systems*, 35(11): 15222–15236, 2023.
- [39] Yilong Hui, Jie Hu, Nan Cheng, Gaosheng Zhao, Rui Chen, Tom H Luan, and Khalid Aldubaikhy. Rcf: Redundancy-aware collaborative federated learning in vehicular networks. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):5539–5553, 2023.
- [40] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.
- [41] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggong Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023.
- [42] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. *arXiv preprint arXiv:2506.24044*, 2025.
- [43] Xianghao Kong, Wentao Jiang, Jinrang Jia, Yifeng Shi, Runsheng Xu, and Si Liu. Dusa: Decoupled unsupervised sim2real adaptation for vehicle-to-everything collaborative perception. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1943–1954, 2023.
- [44] Mingyue Lei, Zewei Zhou, Hongchen Li, Jiaqi Ma, and Jia Hu. Risk map as middleware: Towards interpretable cooperative end-to-end autonomous driving for risk-aware planning. *arXiv preprint arXiv:2508.07686*, 2025.
- [45] Baolu Li, Jinlong Li, Xinyu Liu, Runsheng Xu, Zhengzhong Tu, Jiacheng Guo, Xiaopeng Li, and Hongkai Yu. V2x-dgw: Domain generalization for multi-agent perception under adverse weather conditions. *arXiv preprint arXiv:2403.11371*, 2024.
- [46] Hanlei Li, Guangyi Zhang, Kequan Zhou, Yunlong Cai, and Guanding Yu. Coarse-to-fine: A dual-phase channel-adaptive method for wireless image transmission. *arXiv preprint arXiv:2412.08211*, 2024.
- [47] Wei Li, Lin Ma, Haoze Chang, Xiangyun He, and Longteng Huang. Efficient collaborative perception with integrated uncertainty estimation via evidence regression. *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- [48] Xiang Li, Junbo Yin, Wei Li, Chengzhong Xu, Ruigang Yang, and Jianbing Shen. Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3215, 2024.
- [49] Yiming Li, Dekun Ma, Ziyang An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent col-

- laborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022.
- [50] Xiaoqing Luo, Juan Wang, Zhancheng Zhang, and Xiao-jun Wu. A full-scale hierarchical encoder-decoder network with cascading edge-prior for infrared and visible image fusion. *Pattern Recognition*, 148:110192, 2024.
- [51] Xuewen Luo, Fengze Yang, Fan Ding, Xiangbo Gao, Shuo Xing, Yang Zhou, Zhengzhong Tu, and Chenxi Liu. V2x-unipool: Unifying multimodal perception and knowledge reasoning for autonomous driving. *arXiv preprint arXiv:2506.02580*, 2025.
- [52] Xi-Ao Ma, Hao Xu, Yi Liu, and Justin Zuopeng Zhang. Class-specific feature selection using fuzzy information-theoretic metrics. *Engineering Applications of Artificial Intelligence*, 136:109035, 2024.
- [53] Nadia Mouawad and Valérian Mannoni. Collective perception messages: New low complexity fusion and v2x connectivity analysis. In *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pages 1–5. IEEE, 2021.
- [54] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024.
- [55] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023.
- [56] Shunli Ren, Zixing Lei, Zi Wang, Mehrdad Dianati, Yafei Wang, Siheng Chen, and Wenjun Zhang. Interruption-aware cooperative perception for v2x communication-aided autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 9(4):4698–4714, 2024.
- [57] Rui Song, Andreas Festag, Abhishek Dinkar Jagtap, Maximilian Bialdyga, Zhiran Yan, Maximilian Otte, Sanath Tiptur Sadashivaiah, and Alois Knoll. First mile: An open innovation lab for infrastructure-assisted cooperative intelligent transportation systems. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1635–1642, 2024.
- [58] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17996–18006, 2024.
- [59] Alireza Souiri, Mani Zarei, Atefeh Hemmati, and Mingliang Gao. A systematic literature review of vehicular connectivity and v2x communications: Technical aspects and new challenges. *International Journal of Communication Systems*, 37(10):e5780, 2024.
- [60] Roy Sumner, Bruce Eisenhart, John Baker, et al. Sae j2735 standard: applying the systems engineering process. Technical report, United States. Department of Transportation. Intelligent Transportation . . . , 2013.
- [61] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.
- [62] Jiayao Tan, Fan Lyu, Linyan Li, Fuyuan Hu, Tingliang Feng, Fenglei Xu, Zhang Zhang, Rui Yao, and Liang Wang. Dynamic v2x perception from road-to-vehicle vision. *IEEE Transactions on Intelligent Vehicles*, 2024.
- [63] Jonas Vogt. A comprehensive overview of the protocols associated with intelligent transportation systems. *arXiv preprint arXiv:2407.12799*, 2024.
- [64] Jiahao Wang, Xiangyu Cao, Jiaru Zhong, Yuner Zhang, Haibao Yu, Lei He, and Shaobing Xu. Griffin: Aerial-ground cooperative detection and tracking dataset and benchmark. *arXiv preprint arXiv:2503.06983*, 2025.
- [65] Sichao Wang, Ming Yuan, Chuang Zhang, Lei He, Qing Xu, and Jianqiang Wang. V2x-dgpe: Addressing domain gaps and pose errors for robust collaborative 3d object detection. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 2074–2080. IEEE, 2025.
- [66] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.
- [67] Yixian Wang, Geng Sun, Zemin Sun, Jiacheng Wang, Jiahui Li, Changyuan Zhao, Jing Wu, Shuang Liang, Minghao Yin, Pengfei Wang, et al. Toward realization of low-altitude economy networks: Core architecture, integrated technologies, and future directions. *arXiv preprint arXiv:2504.21583*, 2025.
- [68] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, et al. Generative ai for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025.
- [69] Zhe Wang, Shaocong Xu, Xucai Zhuang, Tongda Xu, Yan Wang, Jingjing Liu, Yilun Chen, and Ya-Qin Zhang. Coopdetr: A unified cooperative perception framework for 3d detection via object query. *arXiv preprint arXiv:2502.19313*, 2025.
- [70] Chuheng Wei, Ziye Qin, Walter Zimmer, Guoyuan Wu, and Matthew J Barth. Hecofuse: Cross-modal complementary v2x cooperative perception with heterogeneous sensors. *arXiv preprint arXiv:2507.13677*, 2025.
- [71] Keshu Wu, Pei Li, Yang Zhou, Rui Gan, Junwei You, Yang Cheng, Jingwen Zhu, Steven T Parker, Bin Ran, David A Noyce, et al. V2x-llm: Enhancing v2x integration and understanding in connected vehicle corridors. *arXiv preprint arXiv:2503.02239*, 2025.
- [72] Yuchen Xia, Quan Yuan, Guiyang Luo, Xiaoyuan Fu, Yang Li, Xuanhan Zhu, Tianyou Luo, Siheng Chen, and Jinglin Li. One is plenty: A polymorphic feature interpreter for immutable heterogeneous collaborative perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1592–1601, 2025.

- [73] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, et al. V2x-real: a large-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, pages 455–470. Springer, 2024.
- [74] Hao Xiang, Zhaoliang Zheng, Xin Xia, Seth Z Zhao, Letian Gao, Zewei Zhou, Tianhui Cai, Yun Zhang, and Jiaqi Ma. V2x-realo: An open online framework and dataset for cooperative perception in reality. *arXiv preprint arXiv:2503.10034*, 2025.
- [75] Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, et al. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *arXiv preprint arXiv:2412.15206*, 2024.
- [76] Fan Xu, Chen Chen, Haifeng Zheng, and Xinxin Feng. Delay-aware cooperative perception with deep reinforcement learning in vehicular networks. In *2024 9th International Conference on Computer and Communication Systems (ICCCS)*, pages 980–985. IEEE, 2024.
- [77] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022.
- [78] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13712–13722, 2023.
- [79] Runsheng Xu, Chia-Ju Chen, Zhengzhong Tu, and Ming-Hsuan Yang. V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 650–662, 2025.
- [80] Xincan Xu, Kai Liu, Penglin Dai, Ruitao Xie, Jingjing Cao, and Jiangtao Luo. Cooperative sensing and heterogeneous information fusion in vcps: A multi-agent deep reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):4876–4891, 2023.
- [81] Xun Yang, Yunyang Shi, Jiping Xing, and Zhiyuan Liu. Autonomous driving under v2x environment: state-of-the-art survey and challenges. *Intelligent Transportation Infrastructure*, 1:liac020, 2022.
- [82] Sheng Yi, Hao Zhang, and Kai Liu. V2iviewer: Towards efficient collaborative perception via point cloud data fusion and vehicle-to-infrastructure communications. *IEEE Transactions on Network Science and Engineering*, 11(6):6219–6230, 2024.
- [83] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251*, 2024.
- [84] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jiru Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21361–21370, 2022.
- [85] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36:34493–34503, 2023.
- [86] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023.
- [87] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9598–9606, 2025.
- [88] Syed Adnan Yusuf, Arshad Khan, and Riad Souissi. Vehicle-to-everything (v2x) in the autonomous vehicles domain—a technical review of communication, sensor, and ai technologies for road user safety. *Transportation Research Interdisciplinary Perspectives*, 23:100980, 2024.
- [89] Yuanyuan Zha, Wei Shangguan, Junjie Chen, Linguo Chai, Weizhi Qiu, and Antonio M López. Heterogeneous multi-scale cooperative perception for connected autonomous vehicles via v2x interaction. *IEEE Internet of Things Journal*, 2025.
- [90] Lin Zhao, Mikael Nybacka, Maytheewat Aramrattana, Malte Rothhämel, Azra Habibovic, Lars Drugge, and Frank Jiang. Remote driving of road vehicles: A survey of driving feedback, latency, support control, and real applications. *IEEE Transactions on Intelligent Vehicles*, 2024.
- [91] Seth Z Zhao, Hao Xiang, Chenfeng Xu, Xin Xia, Bolei Zhou, and Jiaqi Ma. Coopre: Cooperative pretraining for v2x cooperative perception. *arXiv preprint arXiv:2408.11241*, 2024.
- [92] Jiaru Zhong, Haibao Yu, Tianyi Zhu, Jiahui Xu, Wenxian Yang, Zaiqing Nie, and Chao Sun. Leveraging temporal contexts to enhance vehicle-infrastructure cooperative perception. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 915–922. IEEE, 2024.
- [93] Jiaru Zhong, Jiahao Wang, Jiahui Xu, Xiaofan Li, Zaiqing Nie, and Haibao Yu. Cooptrack: Exploring end-to-end learning for efficient cooperative sequential perception. *arXiv preprint arXiv:2507.19239*, 2025.
- [94] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumor v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22668–22677, 2024.