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

Anonymous ICCV DriveX Workshop submission

Paper ID 14

Abstract

With the rapid advancement of autonomous driving technology, vehicle-to-everything (V2X) communication has become a key enabler for enhancing perception range and driving safety by extending 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—poses significant technical challenges. To accelerate progress in this domain, we organized the End-to-End Autonomous Driving through V2X Cooperation Challenge, featuring two tracks: cooperative temporal perception and cooperative end-to-end planning. Built upon the UniV2X framework and the V2X-Seq-SPD dataset, the challenge attracted over 30 teams worldwide and provided a unified benchmark for evaluating cooperative driving systems. This paper presents the design and outcomes of the challenge, identifies key research challenges—such as bandwidth-aware fusion, robust multi-agent planning, and heterogeneous sensor integration—and analyzes emerging technical trends among top-performing solutions. By confronting realistic communication and fusion constraints, the challenge advances 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 [4, 7], planning [18, 19], and end-to-end [8, 17, 21, 25, 32] 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 [1, 40]. 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 [38, 47]. 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 [28, 29]. 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 [10], latency, and heterogeneous sensor configurations [39] complicate the design of end-to-end solutions. Moreover, robust fusion of multi-view, multi-agent data [22, 41] 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 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. 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 [46] and V2X-Seq-SPD dataset [45], 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 [2], Waymo Open Dataset [30], Argoverse [6], nuplan-based dataset [3, 12, 16, 26], and the CARLA-based dataset [9, 13, 20], 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 [23, 43, 52], 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 [43], V2X-Sim [23], TUMTraf [52], V2X-Real [34], V2v4Real [37], RCooper[15], Griffin[31] and V2XSet [36], 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 [45]. 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 [46] and the V2X-Seq-SPD dataset [45], 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 [46], 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 [45], 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.

3. Challenge Design

3.1. Task Setup and Evaluation Metrics

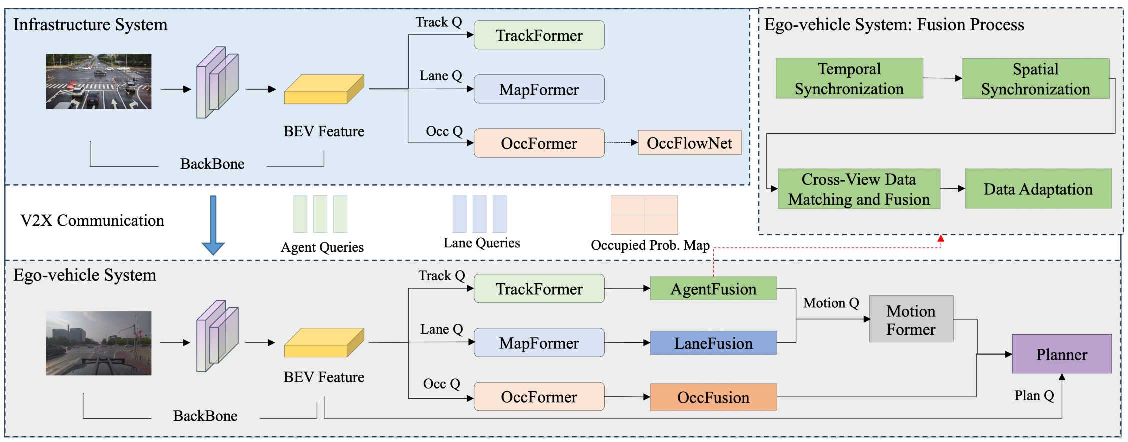
The challenge comprises two complementary tracks designed to evaluate different aspects of V2X cooperative autonomous driving: Cooperative Temporal Perception and Cooperative End-to-End Planning.

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

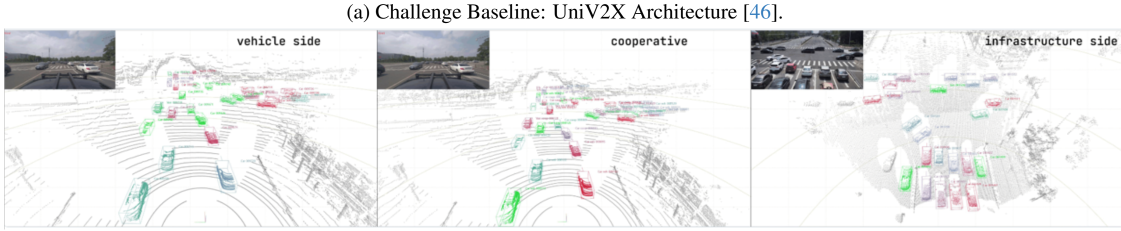
ICCV DriveX Workshop #14

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 [2]	Real	CVPRW19, ICRAW20, ICRAW21	Det,Trk,MPre,Pla	✗	✓
Waymo [30]	Real	WOD20-25	Det,Trk,MPre,Pla	✗	✓
Argoverse [6]	Real	CVPRW22, CVPRW23, CVPRW25	Det,Trk,MPre,Pla	✗	✓
CARLA [11]	Sim	CVPRW19, NIPSW20-22, CVPRW24	Det,Trk,MPre,Pla,CL	✗	✓
NAVSIM [12]	Real	CVPRW24, CVPRW25, ICCVW25	Det,Trk,MPre,Pla,CL	✗	✓
DAIR-V2X [43]	Real	AIR-Apollo23	Det	✓	✗
TUMTraf [52]	Real	ICCVW25	Det	✓	✗
V2v4Real [37]	Real	–	Det	✓	✗
V2X-Sim [23]	Sim	–	Det,Trk	✓	✗
V2X-Seq [45]	***Real***	***CVPRW25 (Ours)***	***Det,Trk,Pla***	✓	✓



(a) Challenge Baseline: UniV2X Architecture [46].



(b) An example of V2X-Seq-SPD Dataset [45]

Figure 1. Challenge Baseline UniV2X [46] and V2X-Seq-SPD Dataset [45]

(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 viewpoints—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 mAP + 0.5 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 quality. A min-max normalization is applied based on predefined reference ranges, and the final score is computed as a weighted sum: $0.5 \times$ normalized L2 Error + $0.25 \times$ normalized Collision Rate + $0.25 \times$ 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 [5, 44]. More recent methods employ sparse query-based methods and transformer for cooperative representations embedding and fusion [14, 33, 51]. 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 underexplored.

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 [35, 49]. 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 [24], varying latency, and intermittent connectivity [27]. However, most existing cooperative driving methods assume idealized or fixed-delay channels [42, 46]. 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 [48, 50]. 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 (Tsinghua University)

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 (Tongji University)

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

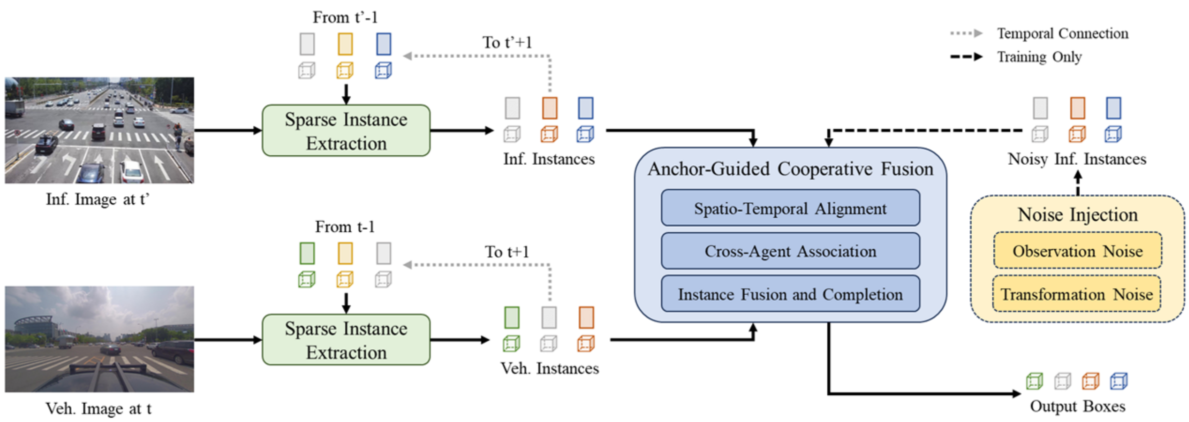


Figure 2. Architecture of **SparseCoop**, the top-ranked solution in **Track 1: Cooperative Temporal Perception**. The method adopts a fully sparse cooperative detection and tracking pipeline, where each object is represented by an *anchor-aided instance query* containing structured geometric attributes (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.

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 highlight several key directions for advancing the field of cooperative autonomous driving under V2X settings. To bridge the gap between benchmark success and real-world deployment, future research should address the following critical aspects:

6.1. Realistic V2X Communication Modeling

Most current solutions assume ideal or simplified communication channels, with constant message delivery and no packet loss. However, real-world V2X systems often operate under non-deterministic network conditions, including variable latency, intermittent connectivity, and data dropouts due to interference or congestion. Future benchmarks and algorithms should incorporate communication-aware learning by simulating:

- Packet loss models based on empirical wireless studies,
- Delay-aware fusion mechanisms, where agents reason with stale or missing messages,
- Redundancy-aware protocols, that prioritize critical information under constraints.

This would enable the design of robust agents that adapt their behavior not only based on perceptual uncertainty, but also on the reliability of the communication channel.

6.2. Bandwidth-Adaptive and Task-Aware Fusion

While sparse fusion strategies showed promise in this challenge, a crucial open question is how to dynamically adjust fusion strategies based on both bandwidth availability and task requirements. Future systems may benefit from:

- Information-theoretic fusion policies, which select features with maximal utility for the downstream task,
- Hierarchical encoding schemes, allowing agents to transmit coarse-to-fine updates depending on the link condition,
- Task-specific prioritization, where planning-critical cues (e.g., dynamic agents near intersections) are communicated more aggressively than static context.

These adaptive fusion mechanisms would support grace-

ICCV DriveX Workshop #14

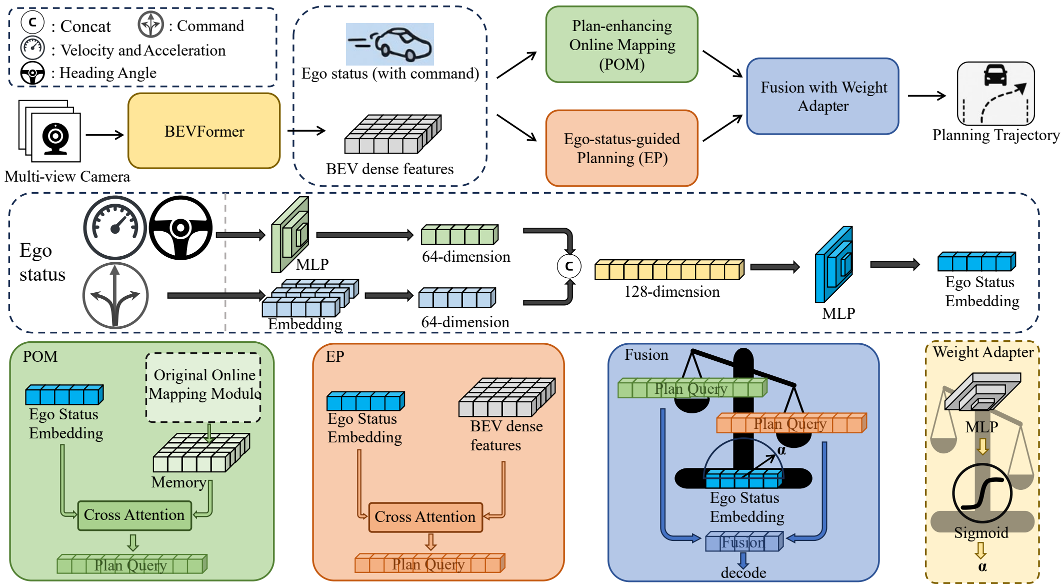


Figure 3. Architecture of **MAP**, the top-ranked solution in **Track 2: Cooperative End-to-End Planning**. This planning-centric framework explicitly incorporates semantic map information into trajectory generation. It consists of two query-generation branches: the *Plan-enhancing Online Mapping (POM)* 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.

ful degradation and efficient resource utilization in large-scale V2X deployments.

6.3. Generalization Across Heterogeneous Agents and Scenarios

Real-world autonomous systems will inevitably involve heterogeneous participants, including vehicles and infrastructure units with varying sensor types, fields of view, and computing capabilities. Robust fusion and planning under such heterogeneous conditions remain largely unsolved. Future research should explore:

- Calibration-agnostic fusion frameworks, resilient to partial or inaccurate sensor alignment,
- Meta-learning or domain adaptation techniques, to generalize across sensor configurations, cities, and deployment environments,
- Scalable fusion topologies, that support dynamic participation (e.g., vehicles entering/leaving the scene).

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

6.4. Interpretability, Safety, and Standardization

For cooperative systems to be adopted in safety-critical applications such as autonomous driving, interpretability and verifiability become essential. Future work should focus on:

- Transparent fusion architectures, that expose which agents and observations contributed to decisions,
- Uncertainty quantification in cooperative predictions and plans,
- Conformance to communication and safety standards, such as SAE J2735 or ETSI ITS-G5.

Moreover, establishing open benchmarks and evaluation protocols for interpretability and fault tolerance will further accelerate research translation into practice.

6.5. Community Building and Ecosystem Development

The field of V2X cooperative driving remains fragmented across perception, networking, and control communities. To build a coherent and impactful research direction, we encourage:

- Continued development of open-source toolkits, such as UniV2X, that support full-stack experimentation,
- Expansion of datasets to include adverse conditions (e.g., night, rain, sensor failure),
- Organization of long-term multi-institutional benchmarks, fostering reproducibility, collaboration, and community-wide progress.
- Increased support for V2X-specific challenges and competitions, as most existing benchmarks (see Table 1) remain focused on single-agent autonomy.

Through sustained infrastructure and shared challenge platforms, we can drive the maturation of V2X research from academic prototypes to robust, real-world systems.

7. Conclusion

This paper presented a comprehensive overview of the End-to-End Autonomous Driving through V2X Cooperation Challenge, held as part of the MEIS Workshop @ CVPR 2025. The challenge was designed to advance the state of cooperative autonomous driving by 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 on the open-source UniV2X framework and the V2X-Seq-SPD dataset.

Through participation from over 30 teams worldwide, the challenge revealed both significant progress and critical bottlenecks in V2X-enabled driving systems. Top solutions leveraged sparse, query-based fusion, modular architectures, and temporal reasoning to achieve strong results in both perception and planning. At the same time, key research challenges were identified in areas such as communication-aware fusion, robust planning under partial observability, and heterogeneous agent generalization.

The insights from this challenge underscore the importance of designing V2X systems that are not only accurate and efficient, but also adaptive, interpretable, and deployable under real-world constraints. By promoting open benchmarks, reproducible baselines, and community collaboration, this initiative aims to bridge the gap between academic research and practical deployment of cooperative autonomous driving technologies.

Future editions of the challenge will expand in scope and complexity, incorporating richer sensor setups, more realistic communication models, and diversified driving scenarios. We invite the broader research community to join in this effort to build safe, scalable, and intelligent multi-agent driving systems for the cities of tomorrow.

References

[1] 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. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3

[3] 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. 2

[4] 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. 1

[5] 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. 4

[6] 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. 2, 3

[7] 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. 1

[8] 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. 1

[9] 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. 2

[10] 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. 1

[11] Contributors. Carla autonomous driving leaderboard, 2024. 3

[12] 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. 2, 3

[13] 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. 2

[14] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Quest: Query stream for practical cooperative perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18436–18442. IEEE, 2024. 4

[15] 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. 2

[16] 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. 2

[17] 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. 1

[18] 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. 1

[19] 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. 1

[20] 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. 2

[21] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 1

[22] 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. 1

[23] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2, 3

[24] 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. 5

[25] 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. 1

[26] 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. 2

[27] 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. 5

[28] 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. 1

[29] 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. 1

[30] 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. 2, 3

[31] 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. 2

[32] 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. 1

[33] 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. 4

[34] 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 largs-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, pages 455–470. Springer, 2024. 2

[35] 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. 5

[36] 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. 2

[37] 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. 2, 3

[38] 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. 1

[39] Xincao 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. 1

[40] 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. 1

[41] 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. 1

[42] 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. 5

[43] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui 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. 2, 3

[44] 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. 4

[45] 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. 1, 2, 3

[46] 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. 1, 2, 3, 5

[47] 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. 1

[48] 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. 5

[49] 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. 5

[50] 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. 5

[51] 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. 4

[52] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22668–22677, 2024. 2, 3