Advancing V2X Cooperative Perception for Autonomous Driving: Challenges, Developments, and Future Directions

Walter Zimmer[1,13], Ross Greer[2], Arpita Vats[3], Rahul Raja[4]
Max Ronecker[5], Chuheng Wei[6], Haibao Yu[7,8], Rui Song[13,17]
Xingcheng Zhou[1], Jiajie Zhang[1], Lars Ullrich[9], Akshay Gopalkrishnan[10]
Stephany Berrio Perez[11], Zewei Zhou[13], Tianhui Cai[13], Yifan Liu[13]
Haoxuan Ma[13], Zhengzhong Tu[14], Holger Caesar[15]
Alina Roitberg[16], Daniel Watzenig[5]
Mohan M. Trivedi[10], Alois C. Knoll[1]

[1]Technical University of Munich, [2]University of California, Merced, [3]Boston University, [4]Carnegie Mellon University
[5]Graz University of Technology, [6]University of California, Riverside, [7]University of Hong Kong
[8]Tsinghua University, [9]University of Erlangen-Nuremberg (FAU)
[10]University of California, San Diego, [11]University of Sydney, [13]University of California, Los Angeles
[14]Texas A&M University, [15]Technical University of Delft, [16]University of Stuttgart, [17]University of Cambridge

Abstract

With the rapid advancement of autonomous driving, Vehicle-to-Everything (V2X) communication has emerged as a key enabler for extending perception beyond the line of sight and improving safety in complex traffic environments. Cooperative perception leverages information from both vehicle-mounted and infrastructure-mounted sensors to enhance scene understanding, mitigate occlusions, and enable more reliable decision-making. In this work, we study V2X cooperative perception through the DriveX challenge built on the TUMTraf-V2X dataset, a real-world benchmark comprising synchronized multi-modal data from multiple roadside and onboard sensors, including LiDAR and cameras, with large-scale 3D annotations and diverse urban driving scenarios. The dataset captures challenging conditions such as occlusions, traffic violations, near-miss events, and complex maneuvers, making it well-suited for evaluating real-world cooperative perception systems. We analyze representative approaches focusing on cross-modal fusion, multi-agent collaboration, and robustness under heterogeneous sensor configurations. In particular, recent methods leverage attention-based fusion and transformer architectures to effectively integrate complementary information across agents. Furthermore, we identify key research challenges, including communication-efficient feature sharing, synchronization across distributed agents, and generalization under diverse environmental conditions. Our study provides insights into the design of scalable and robust cooperative perception systems and outlines future directions toward real-world deployment of V2X-enabled autonomous driving.

1. Introduction

The rapid progress of intelligent transportation systems (ITS) has made cooperative perception an essential component for improving safety and efficiency in autonomous driving [10]. In this context, Vehicle-to-Everything (V2X) perception extends beyond single-vehicle sensing by enabling vehicles and roadside infrastructure to share information, achieving wider environmental coverage and better awareness under occluded or low-visibility conditions [7]. Despite recent progress in single-agent perception and multimodal fusion, building reliable and scalable cooperative perception models remains challenging due to heterogeneous sensor configurations [3], synchronization difficulties, limited communication bandwidth, and diverse environmental conditions, such as illumination changes and adverse weather.

The 1st DriveX Workshop on V2X Cooperative Perception (DriveX@CVPR 2025) focused on addressing these challenges by providing a standardized benchmark for evaluating cooperative perception systems under realistic traffic conditions. The workshop was organized in conjunction with CVPR 2025 and built upon the TUMTraf-V2X

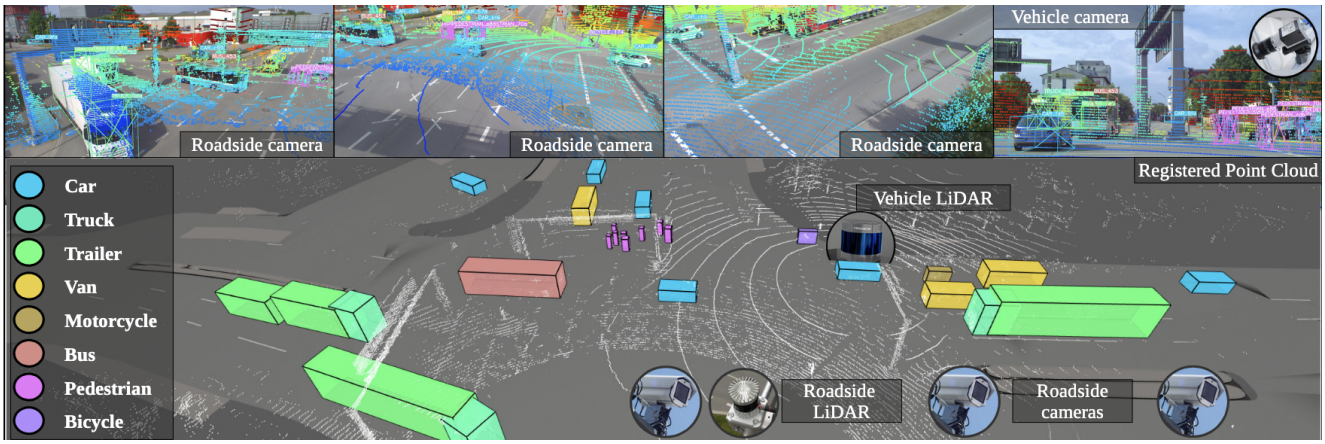

Figure 1. Visualization of the TUMTraf-V2X cooperative perception dataset showing multi-view synchronization between infrastructure and vehicle sensors [26].

Cooperative Perception Dataset [26], which provides synchronized data from vehicle-mounted and infrastructure-mounted sensors, including LiDAR and cameras, collected in diverse urban scenes. The workshop introduced three sub-challenges: (i) LiDAR–Camera 3D Object Detection, (ii) Cross-Modal Cooperative Perception, and (iii) Nighttime and Low-Visibility Generalization. These tasks were designed to evaluate models in different cooperative perception settings, ranging from sensor fusion to cross-domain adaptation.

The challenge attracted broad participation from both academic and industrial teams. The winning entry, HeCoFuse [23], proposed a Cross-Modal Complementary Fusion framework that integrates Hierarchical Attention Fusion (HAF) and Adaptive Spatial Resolution (ASR) modules, achieving the highest mean Average Precision (mAP) of 43.38%. The BUPTMM team [21] introduced a multi-phase training strategy incorporating domain-adaptive nighttime augmentation and a context-aware Non-Maximum Suppression (NMS) approach, achieving 43.35% mAP. The third-place team, KaAI, presented a Transformer based cooperative perception model employing a mixture-query mechanism derived from Uni3DETR, reaching an mAP of 33.04%. These results underscore the growing importance of transformer-based architectures and attention-driven fusion strategies in V2X perception research.

This report summarizes the structure, methodology, and results of the DriveX Challenge. Section 3.1 describes the dataset used for the competition. Section 3 outlines the challenge setup, tracks, and evaluation protocol. Section 4 presents the results and analysis of the submitted methods, and Section 5 concludes with observations and directions for future research in cooperative perception and multimodal fusion.

2. Background

Cooperative perception has emerged as a key paradigm in autonomous driving, enabling multiple agents such as vehicles and roadside infrastructure to collaboratively perceive the environment. Traditional single-agent perception systems are constrained by limited field of view, occlusions, and sensor noise, which degrade performance in complex urban scenarios. By leveraging multi-agent collaboration, cooperative perception significantly enhances spatial coverage and robustness [4, 14].

Vehicle-to-Everything (V2X) communication enables such collaboration by facilitating data exchange across vehicles (V2V), infrastructure (V2I), and other entities. In cooperative settings, perception can be performed at different levels, including raw data fusion, feature-level fusion, and decision-level fusion. However, these approaches introduce challenges such as communication bandwidth constraints, latency, synchronization errors, and heterogeneous sensor configurations across agents [2, 22].

Multi-modal sensor fusion plays a central role in cooperative perception. Early works focused on single-agent fusion of LiDAR and camera data using geometric projection or feature concatenation techniques. More recent approaches extend these ideas to multi-agent systems by sharing intermediate representations, particularly Bird's-Eye View (BEV) features, which provide a unified spatial representation for aggregating information across viewpoints [13, 18]. Transformer-based architectures and attention mechanisms have further advanced cooperative perception by enabling adaptive and scalable fusion strategies. Attention-based fusion allows models to selectively prioritize reliable sensor inputs, while transformer-based methods capture long-range dependencies and contextual relationships across distributed agents. Methods such as V2X-ViT and CoBEVT demonstrate the effectiveness of attention-

driven fusion for improving detection performance in cooperative environments [13, 25].

Robustness under domain shifts remains a critical challenge in real-world deployment. Variations in illumination, weather, and traffic conditions can significantly impact perception accuracy. To address this, recent works incorporate domain adaptation, data augmentation, and curriculum learning strategies to improve generalization across diverse conditions [5, 8].

Benchmark datasets have played a crucial role in advancing research in cooperative perception. Early datasets such as OPV2V and V2V4Real introduced realistic multi-agent simulation and real-world scenarios, respectively. More recent datasets, including DAIR-V2X, provide large-scale, synchronized vehicle-infrastructure data for evaluating cooperative perception algorithms under diverse conditions [8, 24]. The TUMTraf-V2X dataset further extends these efforts by offering high-quality multi-modal data with precise synchronization and challenging urban scenarios, making it a comprehensive benchmark for studying cooperative V2X perception.

3. Challenge Setup

The **DriveX@CVPR 2025 V2X Cooperative Perception Challenge** was organized to advance research on multi-agent sensor fusion, cross-modal collaboration, and robustness under real-world driving conditions. The competition was hosted as part of the CVPR 2025 Workshops and built upon the recently released **TUMTraf-V2X Cooperative Perception Dataset** [26]. This dataset enables synchronized and calibrated perception across both vehicle-mounted and infrastructure-mounted sensors, providing a large-scale foundation for evaluating cooperative perception algorithms. The challenge comprised three sub-tracks designed to benchmark models under varying sensor setups and environmental conditions. Each track emphasized robustness, scalability, and generalization across sensor domains.

3.1. Dataset

The **TUMTraf-V2X Cooperative Perception Dataset** serves as the foundation for the DriveX Challenge. It is a large-scale, real-world benchmark designed to enable research in cooperative vehicle–infrastructure perception. The dataset was collected at a busy urban intersection in Munich, Germany, and features both infrastructure-mounted and vehicle-mounted sensing platforms that provide synchronized, high-resolution multimodal data. The infrastructure setup consists of five stationary sensor nodes installed on traffic poles around the intersection, while the mobile setup includes four vehicles equipped with similar sensors to ensure overlapping fields of view and temporal alignment. Each node is outfitted with multiple RGB cameras and LiDAR sensors, enabling comprehensive scene coverage and facilitating multi-view, cross-modal data fusion.

All sensors are calibrated with precise extrinsic and intrinsic parameters to allow for accurate geometric alignment between modalities and viewpoints. Temporal synchronization between LiDAR and camera streams is achieved using GNSS based timestamps and hardware-level triggering to maintain sub frame accuracy. The dataset records a wide range of traffic participants and behaviors, including cars, pedestrians, cyclists, buses, and heavy vehicles under both daytime and nighttime conditions. Scenarios include complex interactions such as overtaking, lane changes, pedestrian crossings, and partial occlusions, making it particularly suitable for evaluating robustness in real-world cooperative perception.

The dataset comprises approximately 2,000 annotated LiDAR frames and 5,000 synchronized image frames, corresponding to over 30,000 labeled 3D bounding boxes across nine object classes (*Car*, *Bus*, *Truck*, *Pedestrian*, *Cyclist*, *Motorcycle*, *Traffic Light*, *Traffic Sign*, and *Other*). All annotations are manually verified and adhere to the OpenLABEL format, providing spatially consistent 3D bounding boxes and metadata for both infrastructure and vehicle viewpoints. The dataset is split into 80% training, 10% validation, and 10% testing subsets, ensuring a balanced representation of object types and environmental conditions.

In addition to raw data, the authors provide a development kit that supports data preprocessing, calibration loading, and performance evaluation, along with standardized protocols for cooperative 3D object detection. The dataset's combination of multimodality, viewpoint diversity, and challenging environmental conditions makes it one of the most comprehensive public benchmarks for studying cooperative V2X perception. Figure 2 compares TUMTraf-V2X with other large-scale cooperative datasets, such as DAIR-V2X-C, V2V4Real, V2X-Seq, and OPV2V, illustrating its advantages in sensor coverage, scene realism, and synchronization quality.

3.2. Data Structure and Format

The TUMTraf-V2X dataset comprises eight multi-sensor scenes, each approximately 10 seconds long, recorded synchronously from both vehicle and infrastructure perspectives using nine sensors, including LiDARs, RGB cameras, GNSS, and IMUs. All data streams are timestamped and spatially aligned, with recordings provided in ROS bag format for reproducibility. The dataset follows the OpenLABEL standard for annotations and includes synchronized LiDAR point clouds, camera frames, and metadata such as weather, time of day, and occlusion level. Each scene directory contains separate subfolders for vehicle-side and infrastructure-side data, along with calibration files specify-

Dataset	OPV2V [36]	V2XSet [35]	V2X-Sim [20]	V2V4Real [37]	DAIR-V2X-C [41]	V2X-Seq (SPD) [43]	**TUMTraf-V2X (Ours)**
Year	2022	2022	2022	2022	2022	2023	2024
V2X	V2V	V2V&I	V2V&I	V2V	V2I	V2I	V2I
Real data	-	-	-	✓	✓	✓	✓
Annotation range	120 m	120 m	70 m	200 m	280 m	280 m	200 m
Day & night scenes	-	-	-	-	✓	✓	✓
# object classes	1	1	1	5	10	9	8
Track IDs	-	-	✓	✓	-	✓	✓
HD Maps	✓	✓	✓	✓	-	✓	✓
# of sensors (I | V)	- | 6*	- | 6*	5 | 7	- | 8‡	2 | 3	2 | 3	5 | 4
Available worldwide	✓	✓	✓	✓	-	-	✓
Traffic violations	-	-	-	-	-	-	✓
Labeled attributes§	-	-	-	-	-	-	✓
OpenLABEL format	-	-	-	-	-	-	✓
# Point Clouds	11k	11k	10k	20k	39k	15k	2.0k
# Images	44k	44k	60k	40k†	39k	15k	5.0k
# 3D Boxes	233k	233k	26k	240k	464k	10.45k	29.38k
Location	CARLA	CARLA	CARLA	USA	China	China	Germany

† Image dataset has not been released yet.
* Value per vehicle. Multiple Conn. and Autom. Vehicles (CAVs) are used.
‡ Total sensors from 2 CAVs.
§ Weather, time of day, orientation, #3D points, occlusion, color, sub type

Figure 2. Comparison of existing 3D cooperative perception datasets with the proposed TUMTraf-V2X Cooperative Perception dataset.(I = Infrastructure sensors, V = Vehicle sensors.) [26]

ing intrinsic and extrinsic parameters. Annotation files are stored in JSON format and include fields for *track_id*, *object_class*, *3D_bbox*, *visibility*, and *timestamp*. Additionally, GPS and IMU data are provided for precise spatial referencing and motion compensation. This standardized structure facilitates straightforward integration into cooperative perception pipelines and ensures compatibility with the official TUMTraf-V2X development kit.

3.3. Evaluation

We adopt the official TUMTraf-V2X benchmark evaluation protocol to assess model performance on cooperative perception tasks. The primary metric is the mean Average Precision (**mAP**), which quantifies detection accuracy across multiple confidence thresholds. Two complementary measures are employed: the **3D mAP**, based on volumetric intersection-over-union (IoU) between predicted and ground-truth bounding boxes, and the **BEV mAP**, computed using the projected centers in the bird's-eye-view plane. Both follow the nuScenes evaluation protocol and are implemented through the TUMTraf dataset development kit, ensuring consistency across submissions. Additionally, runtime efficiency is measured in frames per second (**FPS**), reflecting real-time feasibility on edge and vehicle-side hardware. The official evaluation kit, publicly available on GitHub[1], provides standardized scripts for result submission, ranking, and analysis.

4. Results and Analysis

The DriveX@CVPR 2025 Challenge received numerous submissions exploring advanced strategies for cooperative 3D perception, sensor fusion, and domain adaptation. This section details the top-performing solutions, **HeCoFuse** (1st place), **BUPTMM** (2nd place), and **KaAI** (3rd place), each of which contributed novel insights toward improving real-world V2X perception on the TUMTraf-V2X benchmark.

4.1. HeCoFuse: Cross-Modal Complementary Fusion with Heterogeneous Sensors

The winning solution, **HeCoFuse** [23], proposed by UC Riverside, addresses the challenge of cooperative perception under heterogeneous sensor configurations, where nodes may have different sensor types such as LiDAR-only, camera-only, or combined LiDAR–camera setups. To handle such variability, *HeCoFuse* introduces a unified *cross-modal cooperative fusion framework* with two principal modules, **Hierarchical Attention Fusion (HAF)** and **Adaptive Spatial Resolution (ASR)**, that dynamically integrate information across diverse sensor configurations.

[1] https://github.com/tum-traffic-dataset/tum-traffic-dataset-dev-kit

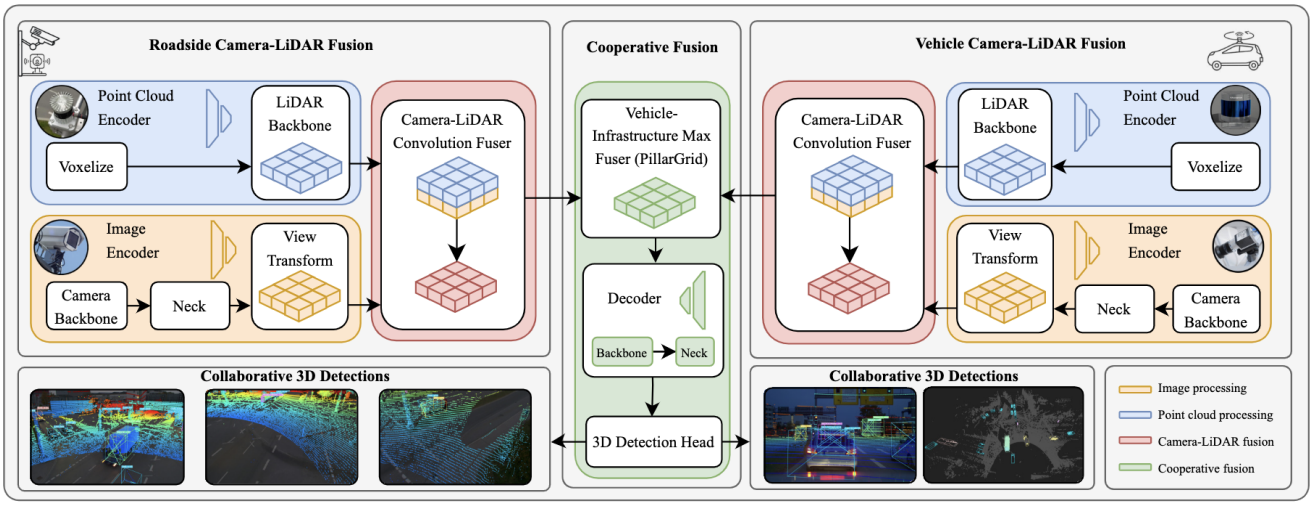

Figure 3. Architecture of the TUMTraf-V2X cooperative perception pipeline showing roadside and vehicle camera–LiDAR fusion, followed by cross-agent cooperative fusion for unified 3D object detection [26].

The model extracts Bird's-Eye View (BEV) features for each node using separate LiDAR and camera encoders. LiDAR data is voxelized and processed using a sparse 3D convolutional backbone, while camera images are projected to BEV space via a depth estimation and view transformation module. In multi-sensor configurations, these features are combined using a convolutional BEV fusion block:

$$\text{BEVFusion}(F_1, F_2) = \phi_{\text{act}}(\phi_{\text{norm}}(\phi_{\text{conv}}([F_1, F_2]))).$$

When a node possesses only a single sensor type, *Pseudo-Fusion* adapts features using modality-specific 1×1 convolutions to maintain dimensional consistency across agents.

For inter-node fusion, the HAF module applies channel-wise and spatial attention weighting to select the most reliable features from each sensor stream. Specifically, the channel attention map $\alpha \in \mathbb{R}^{1\times C\times 1\times 1}$ assigns reliability weights across modalities, while spatial attention maps emphasize modality-specific strengths (e.g., LiDAR for distance, camera for appearance). The ASR module further adjusts feature-map resolution based on configuration—higher resolution for LiDAR, lower for cameras—balancing fidelity and computation cost.

Training is conducted jointly over nine heterogeneous sensor configurations (e.g., L+C, LC+L, L+LC), allowing the model to generalize across missing modalities. Experiments on the TUMTraf-V2X dataset show that HeCoFuse achieves the highest 3D mAP of **43.38%** under the L+LC configuration, outperforming CoopDet3D by +1.17%. The system also attains strong precision (43.73%) and low positional RMSE (0.50). Qualitative results demonstrate consistent detection under occlusions and night-time conditions, confirming the robustness of hierarchical fusion and adaptive resolution mechanisms for real-world V2X applications.

4.2. BUPTMM: Multi-Phase Training and Scene Augmentation

The runner-up, **BUPTMM**, from Beijing University of Posts and Telecommunications, proposed a lightweight yet powerful framework that enhances 3D detection through a staged training pipeline and domain-adaptive data augmentation. The core contribution is a *multi-phase training strategy* that incrementally integrates infrastructure and vehicle sensors, thereby stabilizing learning and improving cross-domain generalization.

The training proceeds in three phases: (1) training on infrastructure LiDAR data to establish a strong geometric feature backbone; (2) joint training on infrastructure and vehicle LiDARs to refine spatial consistency; and (3) fine-tuning the camera module while freezing LiDAR weights to balance spatial and semantic learning. This staged adaptation ensures gradual feature alignment and mitigates overfitting during multi-modal integration.

To enhance robustness under low-visibility conditions, the team introduced a *domain-adaptive nighttime augmentation* pipeline. Synthetic dark-scene data were generated using OpenCV's `convertScaleAbs()` and Pillow-based photometric transformations, simulating multiple levels of illumination reduction and sensor noise. This augmented dataset increased generalization to nighttime and adverse conditions without additional real data collection.

Finally, a **context-aware Non-Maximum Suppression (NMS)** was designed to dynamically adjust IoU thresholds based on scene density—using a higher threshold (0.7) in sparse scenes and a lower one (0.5) in dense re-

gions—effectively balancing precision and recall. Ablation studies demonstrate that multi-phase training yields a +1.2% mAP improvement, while adaptive NMS adds a further +0.1%. The final model achieves a 3D mAP of **43.35%**, surpassing the CoopDet3D baseline by +1.3%. Visual analysis confirms improved detection of small and low-reflectance objects under dark conditions.

4.3. KaAI: Transformer-Based Cooperative Detection with Mixture Queries

The **KaAI** team from Naver Labs and Kookmin University developed a simplified *Transformer-based cooperative perception model* extending Uni3DETR's mixture-query mechanism into the CoopDet3D framework. The model leverages a combination of learnable, non-learnable, and random queries in the Transformer decoder, promoting feature diversity and improving generalization. Using synchronized LiDAR–camera inputs, the network learns complementary contextual relationships across vehicle and infrastructure nodes while avoiding overfitting.

The approach achieved a 3D mAP of **33.04%**, precision of **33.38%**, and recall of **41.89%**, with positional RMSE of **0.5208** and rotational RMSE of **0.0478**. Despite its compact design, KaAI's results highlight the promise of query-based Transformers for efficient cooperative 3D detection, offering a direction toward scalable, real-time V2X perception.

5. Conclusion

The **DriveX@CVPR 2025 V2X Cooperative Perception Challenge** established a unified benchmark for advancing multi-modal, multi-agent perception within real-world traffic environments. By leveraging the **TUMTraf-V2X Cooperative Perception Dataset**, the competition successfully evaluated models under heterogeneous sensor configurations, varied illumination, and complex urban conditions. The results demonstrated clear trends in the evolution of cooperative perception research: attention-based fusion frameworks and staged adaptation strategies continue to outperform traditional convolutional or early-fusion baselines.

Among the top submissions, **HeCoFuse** introduced an efficient cross-modal complementary fusion framework that effectively handled sensor heterogeneity through hierarchical attention and adaptive spatial resolution mechanisms. The **BUPTMM** system validated the benefits of curriculum-style multi-phase training and domain-adaptive augmentation for robustness in low-visibility environments. Meanwhile, **KaAI** highlighted the growing potential of transformer-based query architectures in cooperative detection tasks. Together, these solutions provide complementary perspectives on improving generalization, efficiency, and reliability in V2X perception.

Moving forward, the DriveX workshop aims to expand cooperative perception benchmarks by incorporating temporal reasoning, communication latency modeling, and large vision-language model (VLM) integration. These directions are essential for bridging the gap between academic research and full-scale deployment of cooperative autonomy in next-generation intelligent transportation systems (ITS). The dataset, codebase, and leaderboard remain publicly available to encourage continued collaboration and reproducibility within the community.

6. Future Directions

Despite significant progress, cooperative perception for V2X systems remains an open research area with several critical challenges. One key direction is the integration of temporal reasoning into cooperative perception pipelines. Most existing approaches focus on frame-level fusion, ignoring temporal consistency across sequences. Incorporating temporal modeling through recurrent architectures or spatio-temporal transformers can improve robustness under occlusions and dynamic traffic scenarios [16, 20].

Another important direction is communication-efficient perception. Current methods often assume high-bandwidth communication between agents, which is impractical in real-world deployments. Future systems must optimize what information to share, when to share it, and how to compress it effectively. Techniques such as feature sparsification, adaptive communication policies, and bandwidth-aware learning are promising avenues to reduce communication overhead while maintaining performance [9, 15].

Modeling communication latency and asynchrony is also critical for real-world V2X systems. Most benchmarks assume perfect synchronization between agents; however, in practice, delays and packet losses are inevitable. Designing robust perception models that can operate under asynchronous inputs and partial observations is an emerging challenge [17].

The integration of large-scale foundation models, including vision-language models (VLMs) and multimodal large language models (MLLMs), presents another promising research direction. These models can provide high-level semantic reasoning, enabling tasks such as scene understanding, intent prediction, and interactive decision-making in cooperative environments. Leveraging pretrained multimodal representations can improve generalization across domains and reduce the need for large labeled datasets [12, 19].

Scalability to large numbers of agents is another open problem. Most current approaches are evaluated with a small number of vehicles and infrastructure nodes. Future work should explore decentralized and graph-based architectures that can scale to dense traffic scenarios while maintaining computational efficiency and real-time perfor-

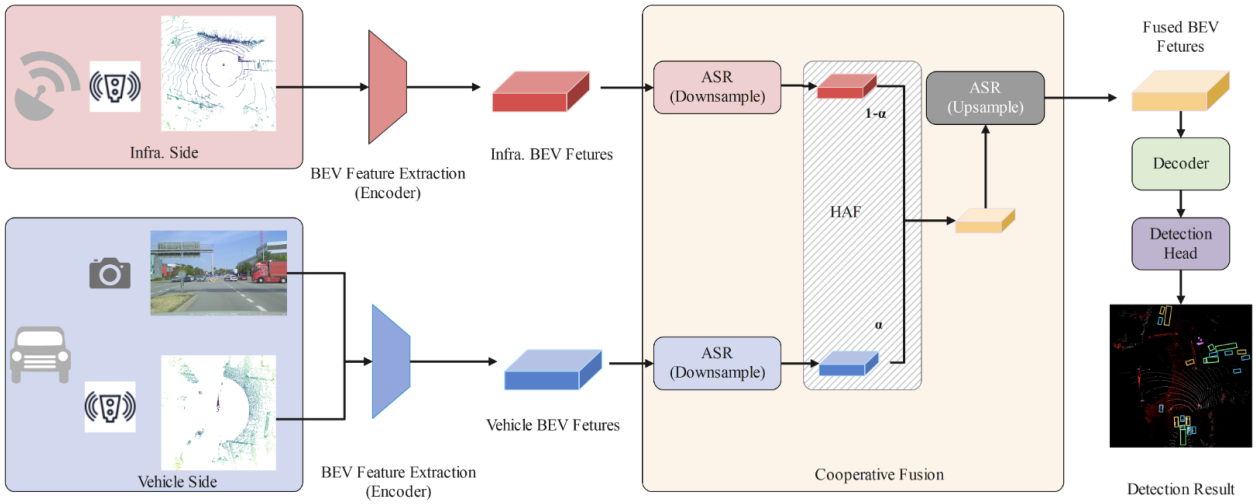

Figure 4. Overview of the HeCoFuse framework. The LC+L configuration (vehicle: LiDAR + Camera; infrastructure: LiDAR) is shown as an example. For other vehicle–infrastructure sensor pairs, the node configurations vary accordingly. The framework performs feature integration through Hierarchical Attention Fusion (HAF) and adapts spatial fidelity via Adaptive Spatial Resolution (ASR) across diverse scenarios.

Team	3D mAP (%)	Precision (%)	Recall (%)	Pos. RMSE	Rot. RMSE
HetoFuse	**43.38**	43.73	42.91	0.50	0.04
BUPTMM	**43.35**	43.63	44.55	0.45	0.03
KaAI_8 (eunl@)	**33.70**	34.27	42.10	0.53	0.05
cv123	43.25	43.53	44.32	0.45	0.03
team_12138	43.23	43.52	44.35	0.45	0.03
KaAI	43.22	43.51	44.35	0.45	0.03
Wang-Zhe	43.20	43.48	44.37	0.38	0.03
sssuuuun (test555)	40.35	40.71	41.13	0.47	0.05
Bentley (vifl)	39.17	39.65	38.53	0.52	0.05
tlab	28.71	29.13	37.11	0.55	0.05

Table 1. Quantitative results from the DriveX@CVPR 2025 Challenge on the TUMTraf-V2X dataset. The top three teams are highlighted.

mance [11]. Finally, safety, reliability, and interpretability remain essential for deployment in real-world systems. Cooperative perception models must be robust to adversarial inputs, sensor failures, and environmental uncertainties. Developing uncertainty-aware models, explainable fusion mechanisms, and standardized evaluation protocols will be crucial for ensuring trustworthiness in safety-critical applications [1, 6]. Advancing these directions will be critical for bridging the gap between research prototypes and real-world deployment of cooperative perception systems in next-generation intelligent transportation systems.

References

[1] Dario Amodei et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 7

[2] Eduardo Arnold, Omar Al-Jarrah, and Mehrdad Dianati. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2

[3] Sayed Pedram Haeri Boroujeni, Niloufar Mehrabi, Hazim Alzorgan, Ahmad Sarlak, Mahlagha Fazeli, and Abolfazl Razi. All you need for object detection: From pixels, points, and prompts to next-gen fusion and multimodal llms/vlms in autonomous vehicles, 2025. 1

[4] Qi Chen, Xinpeng Ma, Siheng Tang, et al. Cooper: Cooperative perception for connected autonomous vehicles. In *IEEE Intelligent Vehicles Symposium*, 2019. 2

[5] Yuhua Chen et al. Domain adaptive object detection for autonomous driving. In *CVPR*, 2021. 3

[6] Di Feng et al. Deep multi-modal object detection and semantic segmentation for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 7

[7] Ruiyang Hao, Haibao Yu, Jiaru Zhong, Chuanye Wang, Jiahao Wang, Yiming Kan, Wenxian Yang, Siqi Fan, Huilin Yin, Jianing Qiu, Yao Mu, Jiankai Sun, Li Chen, Walter Zimmer, Dandan Zhang, Shanghang Zhang, Mac Schwager, Ping Luo, and Zaiqing Nie. Research challenges and progress in the end-to-end v2x cooperative autonomous driving competition, 2025. 1

[8] Yue Hu, Rui Fang, et al. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. *IEEE Robotics and Automation Letters*, 2023. 3

[9] Q. Huang et al. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *NeurIPS*, 2022. 6

[10] Tao Huang, Jianan Liu, Xi Zhou, Dinh C. Nguyen, Mostafa Rahimi Azghadi, Yuxuan Xia, Qing-Long Han, and Sumei Sun. Vehicle-to-everything cooperative perception for autonomous driving. *Proceedings of the IEEE*, 113(5): 443–477, 2025. 1

[11] H. Jiang et al. Graph neural networks for scalable cooperative perception in v2x. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 7

[12] Junnan Li et al. Blip-2: Bootstrapping language-image pre-training. In *ICML*, 2023. 6

[13] Ming Liang, Bin Yang, et al. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*, 2022. 2, 3

[14] Yue Liu, Jun Wang, and Xiaozhi Chen. A survey on cooperative perception for connected autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2

[15] Yue Liu et al. Who2com: Collaborative perception via learnable communication. In *NeurIPS*, 2020. 6

[16] Y. Liu et al. Temporal bev fusion for multi-agent perception. *arXiv preprint arXiv:2309.XXXXX*, 2023. 6

[17] X. Lu et al. Asynchronous cooperative perception for v2x systems. *arXiv preprint arXiv:2401.XXXXX*, 2024. 6

[18] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 2

[19] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[20] Pei Sun et al. Streaming object detection for 3d perception. In *CVPR*, 2023. 6

[21] BUPTMM Team. Multi-phase training and scene augmentation for v2x object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, 2024. To appear / manuscript. 2

[22] Ting Wang, Xiaoxiao Zhu, et al. V2vnet: Vehicle-to-vehicle communication for joint perception. In *ECCV*, 2020. 2

[23] Chuheng Wei, Ziye Qin, Walter Zimmer, Guoyuan Wu, and Matthew J. Barth. Hecofuse: Cross-modal complementary v2x cooperative perception with heterogeneous sensors, 2025. 2, 4

[24] Hao Xu et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *CVPR*, 2021. 3

[25] Runsheng Xu, Zhengzhong Tu, et al. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *CoRL*, 2022. 3

[26] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois Knoll. Tumtraf v2x cooperative perception dataset. *arXiv preprint arXiv:2403.01316*, 2024. 2, 3, 4, 5