End-to-End Autonomous Driving Through V2X Cooperation

Haibao Yu[1,2*], Wenxian Yang[2*], Jiaru Zhong[2,3*], Zhenwei Yang[2,4], Siqi Fan[2], Ping Luo[1], Zaiqing Nie[2†]

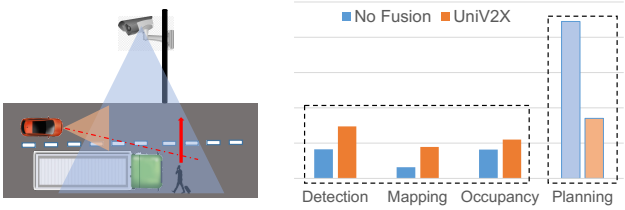
[1]The University of Hong Kong
[2]AIR, Tsinghua University
[3]Beijing Institute of Technology
[4]University of Science and Technology Beijing
yuhaibao94@gmail.com, zaiqing@air.tsinghua.edu.cn

Abstract

Cooperatively utilizing both ego-vehicle and infrastructure sensor data via V2X communication has emerged as a promising approach for advanced autonomous driving. However, current research mainly focuses on improving individual modules, rather than taking end-to-end learning to optimize final planning performance, resulting in underutilized data potential. In this paper, we introduce UniV2X, a pioneering cooperative autonomous driving framework that seamlessly integrates all key driving modules across diverse views into a unified network. We propose a sparse-dense hybrid data transmission and fusion mechanism for effective vehicle-infrastructure cooperation, offering three advantages: 1) Effective for simultaneously enhancing agent perception, online mapping, and occupancy prediction, ultimately improving planning performance. 2) Transmission-friendly for practical and limited communication conditions. 3) Reliable data fusion with interpretability of this hybrid data. We implement UniV2X, as well as reproducing several benchmark methods, on the challenging DAIR-V2X, the real-world cooperative driving dataset. Experimental results demonstrate the effectiveness of UniV2X in significantly enhancing planning performance, as well as all intermediate output performance.

Introduction

Despite significant progress achieved through the integration of deep learning, single-vehicle autonomous driving still faces great safety challenges due to limited perceptual range and inadequate information, especially for vehicles relying on cost-effective cameras. Leveraging external sensors, particularly infrastructure sensors with a broader perception field, has shown promising potential for advancing autonomous driving capacities through Vehicle-to-Everything (V2X) communication (see Figure 1 (a)). Several research studies have investigated the efficacy of external sensor data in diverse tasks such as detection (Wang et al. 2020; Hu et al. 2023a; Yu et al. 2023a; Tianhang et al. 2023; Qiu et al. 2022), tracking (Yu et al. 2023c), segmentation (Xu et al. 2022a), localization (Jiang et al. 2023; Dong et al. 2023),



(a) VICAD (b) Evaluation Results.

Figure 1: (a) VICAD: Infrastructure sensor installed highly has a broad perception field (Yu et al. 2022, 2023c), which can supplement the blind and long-range spots of single vehicle. (b) Performance Enhancement: Compared with No Fusion solution, UniV2X achieves significant gains in various tasks, such as detection (+13%), mapping (+11.4%), occupancy prediction (+5.7%), and collision rate (-0.5%).

and forecasting (Yu et al. 2023c; Ruan et al. 2023; Song et al. 2024). However, existing solutions primarily emphasize individual task optimization, neglecting the overall planning enhancement. This creates challenges in comprehensive data exploitation, driven by a misalignment between individual task goals and final planning objectives. Thus, end-to-end learning exploration, which directly optimizes the final planning output by harnessing both onboard and external sensor data, becomes necessary. In this paper, we focus on vehicle-infrastructure cooperative autonomous driving (VICAD).

The VICAD problem can be formulated as a planning-centric optimization with multiple-view sensor inputs under constrained communication bandwidth. Compared with single-vehicle autonomous driving, VICAD poses additional challenges when addressed through end-to-end learning. Firstly, the transmitted infrastructure data must be *effective*. It should enhance both critical modules and the final planning performance in autonomous driving. These critical modules encompass dynamic obstacle perception, online mapping, and grid occupancy-based general obstacle detection, providing an explicit scene representation crucial for ensuring the safety of autonomous driving. Second, the data must be *friendly*. Driven by real-time requirements and

*Equal contribution. Work done while at AIR, Tsinghua.
†Corresponding author.

Approach	Sensor Data	Transmission				Task				End-to-End
		Type	Effective	Friendly	Reliable	AgentP	Map	Occ	Plan/Control	
V2VNet (Wang et al. 2020)	Point Cloud	BEV Feature	-	Medium	No	✓				No
CoCa3D (Hu et al. 2023a)	Image	BEV Feature	-	Medium	No	✓				No
CoBEVT (Xu et al. 2022a)	Image	BEV Feature	-	Medium	No	✓	✓			No
PP-VIC (Yu et al. 2023c)	Point Cloud	Detected Result	-	High	Yes	✓				No
DeepA (Wang et al. 2023)	Image	BEV Feature	-	Medium	No	✓	✓			No
TransIFF (Chen, Shi, and Jia 2023)	Point Cloud	Instance Feature	-	High	Yes	✓				No
Where2Comm (Hu et al. 2022)	Point Cloud	Instance Feature	-	High	Yes	✓				No
CSA (Valiente et al. 2019)	Image	Raw Image	Yes	Low	Yes				✓	Non-Explicit
CooperNaut (Cui et al. 2022)	Point Cloud	BEV Feature	Yes	Medium	No				✓	Non-Explicit
UniV2X (Ours)	Image	Hybrid Feature	Yes	High	Yes	✓	✓	✓	✓	Explicit

Table 1: Comparison with the existing methods for cooperative autonomous driving. "AgentP" denotes dynamic object perception."Map" denotes online mapping. "Occ" denotes occupancy prediction. "-" denotes that the information is not verified.

limited communication conditions, minimizing transmission costs becomes crucial to mitigate communication bandwidth consumption and reduce latency. Thirdly, the transmitted data must be *reliable*. it should be *reliable*. Vehicles need interpretable information that can be validated and used judiciously to avoid safety issues such as communication attacks or data corruption. Addressing these challenges necessitates a well-designed solution for data transmission and cross-view data fusion.

Here are a few straightforward attempts to address the cooperative driving problem through end-to-end learning. CSA (Valiente et al. 2019) directly shares and feeds raw images received from other vehicles into basic neural networks for control output. CooperNaut (Cui et al. 2022) shares features derived from point clouds among vehicles and inputs them into a basic CNN network for the final output. However, these existing solutions rely on a vanilla approach, utilizing simple networks to optimize planning and control outputs. This paradigm lacks explicit modules, compromising safety assurance and interpretability. Especially within intricate urban settings, this approach falls short in ensuring the reliability of the driving system. More comparisons are in Table 1, and related work is discussed in the appendix.

To this end, we introduce UniV2X, an innovative cooperative autonomous driving framework that seamlessly integrates pivotal modules and cross views into a unified network, as depicted in Figure 2. Beyond the final planning task, we address three common tasks for scene representation in autonomous driving: 1) agent perception, encompassing 3D object detection, tracking, and motion forecasting for dynamic obstacle perception, 2) road element (especially lane) detection for online mapping, and 3) grid-occupancy prediction for general obstacle perception. Inspired by UniAD (Hu et al. 2023b), we adopt a query-based architecture to establish connections across nodes, encompassing internal modules within infrastructure and ego-vehicle systems, as well as cross-view interactions. In transmission and cross-view interaction, we classify agent perception and road element detection as instance-level representation and occupancy prediction as scene-level representation. We transmit agent queries and lane queries for cross-view agent perception interaction and online mapping interaction. We transmit the occupied probability map, recognizing its dense nature at the scene-level occupancy, for cross-view occupancy interaction. This transmission, termed sparse-dense hybrid transmission, balances sparsity and density in spatial and feature dimensions, respectively. Cross-view data fusion, such as agent fusion, mainly involves temporal and spatial synchronization, cross-view data matching and fusion, data adaptation for planning and intermediate outputs. The resulting lightweight approach strengthens dynamic object perception, online mapping, and occupancy modules, thereby enhancing planning performance. Moreover, the interpretability of queries and occupied probability maps at the instance and scene levels, respectively, fortifies the reliability of the VICAD system, bolstering its transmission integrity and fusion safety.

The contributions are summarized as follows:

- We pioneer a first explicitly end-to-end framework that unifies vital modules within a single model, advancing the landscape of cooperative autonomous driving. Notably, UniV2X is the first end-to-end framework for VICAD.
- We design a sparse-dense hybrid transmission and cross-view data interaction approach, aligning with effectiveness, transmission-friendliness, and reliability prerequisites for end-to-end cooperative autonomous driving.
- We reproduce several cooperative methods as benchmarks, as well as instantiating the UniV2X on DAIR-V2X (Yu et al. 2022). Experimental results underscore the efficacy of our end-to-end paradigm (see Figure 1 (b)).

Method

In this section, we introduce UniV2X, an all-in-one solution for vehicle and infrastructure cooperative autonomous driving (VICAD), together with the proposed sparse-dense hybrid data transmission and fusion design, as shown in Figure 2. We start by presenting the VICAD problem and introduce the background. Following that, we describe how to generate sparse-dense hybrid data for transmission and cross-view data fusion. The training process is also outlined.

VICAD Problem Formulation

The VICAD problem is planning-oriented, aiming to improve planning performance by utilizing both infrastruc-

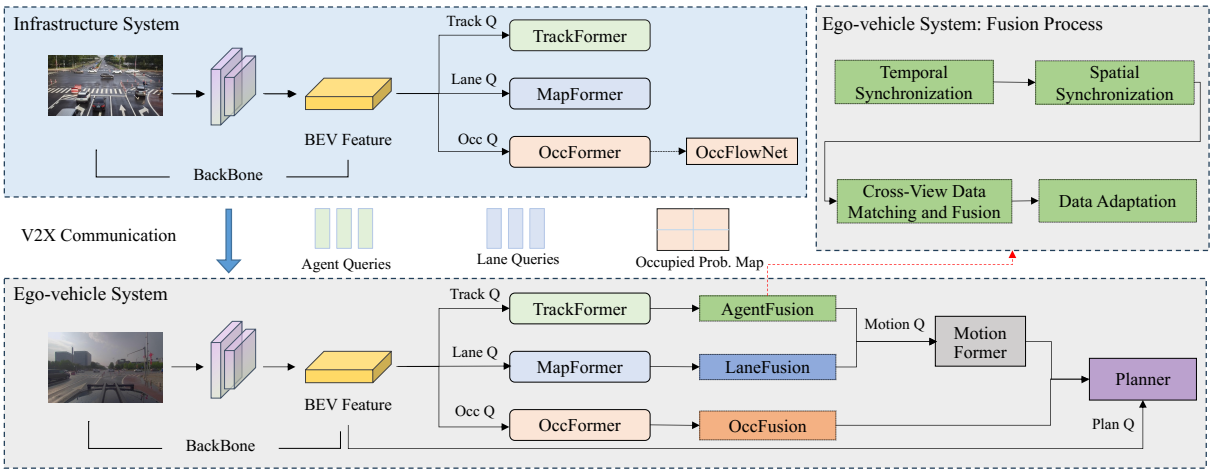


Figure 2: Pipeline of Unified Autonomous Driving through V2X Cooperation (UniV2X). UniV2X aims to connect and jointly optimize all essential modules across diverse views for enhanced planning performance. Cross-view data interaction bolsters pivotal components in autonomous driving like agent perception, online mapping, and occupancy prediction. Additional flow prediction enables minimizing transmission costs for transmitting occupied probability map. Cross-view data fusion involves temporal and spatial synchronization, cross-view data matching and fusion, and data adaptation.

ture sensor data and ego-vehicle sensor data through V2X communication. This paper focuses on the images as inputs. The input of VICAD consists of two parts: (a) Ego-vehicle images $\{I_v(t)|t \leq t_v\}$ and the relative pose $M_v(t_v)$ at the current vehicle timestamp t_v. (b) Infrastructure images $\{I_i(t)|t \leq t_i\}$ and the relative pose $M_i(t_i)$ at the current infrastructure timestamp t_i. Note that in a practical scenario, the timestamp t_i should be earlier than the timestamp t_v due to the communication latency. The output of VICAD is to predict future coordinates of ego vehicle for time steps $t = t_v + 1, \cdots, t_{pred}$.

Evaluation Metrics. We evaluate the planning performance with L2 Error, Collision Rate and Off-Road Rate, and measure transmission cost with Bytes Per Second (BPS), as suggested in (Yu et al. 2022, 2023c). We provide detailed explanations for these metrics in the Appendix.

Challenges. Compared with single-vehicle autonomous driving, VICAD presents additional challenges: (1) Limited by practical communication conditions, fewer infrastructure data should be transmitted to vehicles to minimize bandwidth usage and reduce latency. (2) Wireless communication causes latency, potentially leading to temporal misalignment in data fusion. (3) Potential communication attacks and data corruption can render transmitted data untrustworthy. This highlights the need for interpretable transmitted data.

Data for Transmission. It involves three primary types in V2X cooperation: raw data like raw images, perception outputs like detection results, and intermediate-level data such as Bird's Eye View (BEV) features and queries (Fan et al. 2024; Chen, Shi, and Jia 2023; Zhong et al. 2024). Compared to raw data and detection results, intermediate-level data achieves a balance between preserving valuable information and reducing redundant transmission. To ensure effective, transmission-friendly, and reliable transmitted data, we propose a sparse-dense hybrid transmission mechanism. Queries, as lightweight instance-level features, enhance agent perception and online mapping, as dynamic obstacles and lanes can be treated as instance-level representations. Occupied probability maps, channel-sparse scene-level features, improve occupancy prediction. Compared to less interpretable and high-cost BEV features, occupied probability maps offer pixel-level interpretability and lower transmission costs.

Sparse-Dense Hybrid Data Generation

This part illustrates how to generate sparse-dense hybrid data for transmission in infrastructure system.

BEVFormer(Li et al. 2022b) is adopted as the backbone to extract image features and transform them into bird's-eye-view (BEV) features B_{inf} with size of (200, 200, 256) by incorporating spatial cross-attention and temporal self-attention. TrackFormer is based on DETR (Carion et al. 2020), which optimizes detection and multi-object tracking together, eliminating the need for non-differentiable post-processing like NMS (Bodla et al. 2017). The ultimate filtered output from TrackFormer contains N_a^{inf} valid agent queries $\{Q_A^{inf}\}$ with a feature dimension of 256 and their corresponding assigned tracking IDs and reference points. MapFormer is based on Panoptic SegFormer (Li et al. 2022c). We mainly focus on the lane line and crosswalk elements. During transmission, we filter out low-scoring queries using boxes generated from the classification decoder, and exclusively transmit N_l^{inf} valid lane queries $\{Q_L^{inf}\}$ with a feature dimension of 256, along with their corresponding reference points. Original OccFormer in UniAD (Hu et al. 2023b) solely considers instance-level occupancy associated with agent queries, predicting multiple steps. However, occupancy serves as a complementary

factor to object perception for general obstacle detection, and transmitting multiple probability maps incurs significant transmission costs. To address these challenges, we retain the dense feature obtained through pixel-level attention with a size of (200, 200, 256). Initially, a Multi-layer Perception (MLP) is employed to transform the dense feature into BEV occupied probability map denoted as P^{inf} with a size of (200, 200). Subsequently, adopting the feature flow prediction approach (Yu et al. 2023b,c), an additional probability flow module is utilized to represent T-step maps via a linear operation as

$$P_{future}(t) = P_0 + t * P_1, \tag{1}$$

where P_0 signifies the present BEV probability map, and P_1 indicates the corresponding BEV probability flow. Transmitting T-step occupied probability maps requires T*200*200 floats, while UniV2X only requires 2*200*200 floats.

Cross-View Data Fusion (Agent Fusion)

In the ego-vehicle system, the BEV features B_{veh} are first extracted from the images captured by onboard sensors. We also adopt TrackFormer, MapFormer, and OccFormer to generate the corresponding agent queries $\{Q_A^{veh}\}$, lane queries $\{Q_L^{veh}\}$, and the occupied probability map P^{veh}. The network for these modules aligns with that of the infrastructure system. In this section, we describe how to implement cross-view agent fusion. Cross-view agent fusion is mainly composed of temporal synchronization for latency compensation, spatial synchronization to unify the cross-view coordinates, data matching and fusing, and data adaptation for planning and intermediate outputs.

Temporal Synchronization with Flow Prediction. The transmission delay in wireless communication, as t_i is earlier than t_v, is significant in complex traffic systems, especially for the busy intersection scenario. Due to the movement of dynamic objects, there is a temporal misalignment when fusing data from different sources. To address that, we incorporate feature prediction into infrastructure agent queries to mitigate latency, following feature flow prediction as (Yu et al. 2023b,c). Specifically, we input both agent query Q_A^{inf} and query associated in the previous frame into QueryFlowNet, a three-layer Multi-Layer Perceptron (MLP), to generate the agent query flow Q_{AFlow}^{inf}. The dimensions of the agent query flow match those of the agent query. Subsequently, a linear operation forecasts future features can be used to mitigate latency $t_v - t_i$, depicted as

$$Q_A^{inf}(t_v) = Q_A^{inf}(t_i) + (t_v - t_i) * Q_{AFlow}^{inf}. \tag{2}$$

Notably, QueryFlowNet of the Flow Prediction module is not trained in an end-to-end manner in UniV2X. We adopt self-supervised learning following (Yu et al. 2023b,c).

Spatial Synchronization with Rotation-Aware Query Transformation. We initially transform the reference points of infrastructure agent queries Q_A^{inf} from the infrastructure to the ego-vehicle using the relative pose $[R, T]$ between the infrastructure system and ego-vehicle system. Here, the relative pose is generated from the global relative poses of the two systems, with R representing a rotation matrix and T denoting translation. However, each object inherently possesses 3D information about its location, size, and rotation. In the context of a query representing a 3D object, the location is explicitly denoted by reference points, while the rotation is implicitly encoded within the query's feature, as illustrated in Figure 3. To address this issue, we propose a solution termed rotation-aware query transformation to achieve spatial synchronization. This involves inputting the infrastructure query, along with its rotation R in the relative pose, into a three-layer MLP to update the feature with rotation awareness, achieving explicit spatial synchronization as

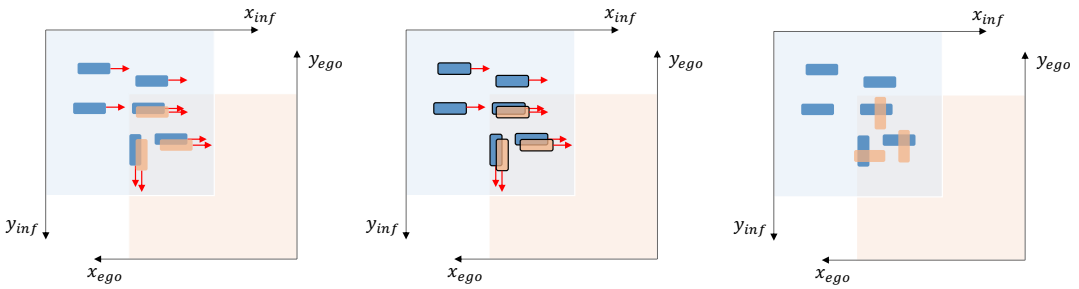
$$\text{spatial\_update}(Q_A^{inf}) = \text{MLP}([Q_A^{inf}, R]), \tag{3}$$

where the rotation matrix R is reshaped into 9 dimensions. Finally, we transform the infrastructure agent query data into the ego-vehicle coordinate system.

Cross-View Query Matching and Fusion. At this stage, cross-view agent queries are temporally and spatially synchronized. To match corresponding queries from different sides, we calculate the Euclidean distance of their reference points and employ the Hungarian method (Kuhn 1955). For the matched query pairs Q_A^{inf} and Q_A^{veh}, they are fed into a three-layer MLP to generate the cooperated query Q_A, which is used to update the ego-vehicle agent query Q_A^{veh}. For the unmatched queries from infrastructure, they are utilized to be added to the ego-vehicle queries. Finally, we assign tracking IDs and filter out cross-view fused queries with low detection confidence, obtaining the final agent queries.

Ego Identification and Removing. This module is used to eliminate the problem of false detection in the ego-vehicle area. From the infrastructure view, the ego vehicle can be perceived either as a distinct obstacle in agent perception or as part of the occupied area in occupancy prediction. Following cross-view data fusion, there is a possibility of generating an obstacle query within the region where the ego vehicle is located, thereby marking the ego-vehicle area as occupied. Such an occurrence can significantly disrupt decision-making processes and ultimately impact decision-making performance. To mitigate this issue, we define the ego-vehicle area as a rectangle, filter queries within this area, and designate this region as unoccupied. However, this straightforward solution may not consistently perform optimally due to relative position errors between infrastructure and ego vehicle, stemming from positioning and calibration inaccuracies (Yang et al. 2023; Gu et al. 2023). Further exploration and refinement are essential to contribute to cooperative autonomous driving.

Decoder Input Augmentation for Intermediate Output. Through cross-attention between ultimately fused agent queries and the output from the encoder in the ego-vehicle TrackFormer, we can obtain intermediate outputs for agents, such as 3D detection outputs, to enhance the interpretability of UniV2X. However, the output of the encoder is all generated from ego-vehicle sensor data information, rendering queries from the infrastructure unable to produce corresponding agent outputs. To address this issue, we use



(a) BEV Features Transformation (b) 3D Boxes Transformation (c) Queries Transformation

Figure 3: Object orientation is explicitly encoded in BEV feature maps (a) and bounding box (b), while the orientation is implicitly embedded in the feature of queries (c), resulting in the challenge of cross-view rotation alignment in spatial synchronization.

synchronized infrastructure queries to enhance ego-vehicle BEV features, the output of the encoder, as:

$$\text{update}(B^{veh}) = B^{veh} + \text{MLP}(\text{synchronized}(Q_A^{inf})). \quad (4)$$

Cross-View Data Fusion (Lane Fusion)

The LaneFusion module is utilized to fuse lane queries across different sides. In this context, we omit temporal synchronization in lane fusion, as the road lane elements remain unaffected by latency and maintain stability. Similar to AgentFusion, LaneFusion incorporates spatial synchronization through the rotation-aware query transformation. This process converts infrastructure lane queries, comprising reference points and query features, into the ego-vehicle coordinate system. We then match and fuse the synchronized infrastructure lane queries with ego-vehicle lane queries, as done in AgentFusion. To accelerate training, we also choose to directly concatenate synchronized queries with ego-vehicle lane queries. The synchronized queries are also used for decoder input augmentation.

Cross-View Data Fusion (Occupancy Fusion)

We first generate multiple-step infrastructure occupied probability maps through linear operations, aligning them with ego-vehicle multiple-step occupancy predictions. Leveraging the explicit representation of rotation in the dense probability map, we directly transform the infrastructure occupied probability maps to the ego-vehicle system using the relative pose. Subsequently, we fuse the synchronized occupied probability maps with ego-vehicle occupied probability maps using simple max operations, generating the fused probability map $\hat{P}$. Grids with a probability exceeding a certain threshold are marked as occupied.

Planning Output

With the fused agent queries, lane queries, and occupancy features, we first generate rough future waypoints by reusing the implementations in UniAD (Hu et al. 2023b). MotionFormer is used to generate a set of N_a motion queries with a prediction horizon of t_{pred}. These queries are created by capturing the interactions among agents, lanes, and goals. Notably, these agent queries encompass the ego-vehicle query, thereby enabling MotionFormer to generate ego-vehicle queries with multimodal intentions. The BEV occupied probability map $\hat{P}$ is utilized to create a binary occupancy map $\hat{O}$. In the planning phase, the ego-vehicle query obtained from MotionFormer is combined with command embeddings to shape a "plan query". These commands comprise turning left, turning right, and moving forward. This plan query, along with the BEV feature, is input into the decoder to produce future waypoints.

The final planning trajectory is determined by: 1) adjusting the future waypoints on the roads to ensure adherence to traffic rules and staying within driving areas with the generated lanes and other road elements, and 2) minimizing a cost function to avoid collisions with occupied grids $\hat{O}$.

Experiments

In this section, we implement UniV2X, alongside reproducing various perception, online mapping, and end-to-end methods on DAIR-V2X (Yu et al. 2022, 2023c). More implementation details, ablation studies, visualizations and analysis are provided in the Appendix. We also conduct UniV2X on more V2X datasets such as V2X-Sim (Li et al. 2022a), and present the experiment results in the Appendix.

Experiment Settings

DAIR-V2X Dataset comprises approximately 100 scenes captured at 28 complex traffic intersections, recorded using both infrastructure and vehicle sensors. Each scene has a duration ranging from 10 to 25 seconds, capturing data at a rate of 10 Hz, and is equipped with a high-definition (HD) map. This dataset provides a diverse range of driving behaviors, including actions such as moving forward, turning left, and turning right. To align with nuScenes (Caesar et al. 2020), we categorize object classes into four categories (car, bicycle, pedestrian, traffic_cone).

Implementation. We establish the interest range of the ego vehicle as [-50, 50, -50, 50] meters. The ego-vehicle BEV range shares the same area spanning [-50, 50, -50, 50] meters, with each grid measuring 0.25m by 0.25m. The infrastructure BEV range is set as [0, 100, -50, 50] meters, accounting for the camera's forward sensing range and facilitating more effective utilization of infrastructure data. The

Method	L2 Error (m)↓				Col. Rate (%)↓				Off-Road Rate (%)↓				Transm. Cos (BPS)↓
	2.5s	3.5s	4.5s	Avg.	2.5s	3.5s	4.5s	Avg.	2.5s	3.5s	4.5s	Avg.	
No Fusion	2.58	3.37	4.36	3.44	0.15	1.04	1.48	0.89	0.44	0.59	2.22	1.08	0
Vanilla	2.33	3.69	5.12	3.71	0.59	2.07	3.70	2.12	0.15	1.33	4.74	2.07	8.19×10^7
BEV Feature Fusion	2.31	3.29	4.31	3.30	0.0	1.04	1.48	0.83	0.44	0.44	1.91	0.93	8.19×10^7
CooperNaut (Cui et al. 2022)	3.84	5.33	6.87	5.35	0.44	1.33	1.93	1.23	0.15	0.15	1.33	0.54	8.19×10^7
UniV2X (Ours)	2.59	3.35	4.49	3.48	0.0	0.44	0.59	**0.34**	0.74	0.74	1.19	**0.89**	**8.09**$\times10^5$

Table 2: Planning Evaluation Results. We do not report the results at 0.5s and 1.5s because most of the collision rate is zero. **Although CooperNaut achieves a lower off-road rate, it has a much larger L2 error compared to other methods.** This is because its planning length is relatively conservative, ensuring it is easier to remain within the drivable area over a given period.

experiments are conducted utilizing 8 NVIDIA A100 GPUs. More implementation details are provided in the appendix.

Baseline Settings. No Fusion only utilizes ego-vehicle images as sensor data input, without any infrastructure data input. In Vanilla approach, we employ a simple CNN to fuse infrastructure and ego-vehicle BEV features. The fused BEV feature is reshaped into one dimension and subsequently fed into a Multi-Layer Perceptron (MLP) to generate the planning path. In BEV Feature Fusion, we use a CNN to fuse two-side BEV features into a new ego-vehicle BEV feature, and send this new feature into UniAD (Hu et al. 2023b). CooperNaut (Cui et al. 2022) originally employs Point Transformer to aggregate cross-view feature by using the sparse characteristics of point clouds. However, the image is a dense representation, and conducting similar sparse operations, as seen with Where2comm (Hu et al. 2022) in Table 3, results in poor performance. Therefore, we directly transmit dense BEV features and use CNN to fuse features from both sides, with only one frame input each time to achieve better comparison. Given the significant role of ego status, such as ego-vehicle velocity, in open-loop end-to-end autonomous driving, as illustrated in (Li et al. 2024), we remove the ego-vehicle velocity embedding in all baseline settings for a fair comparison. Additionally, we explore the role of ego-vehicle velocity for UniV2X in the appendix.

Experiment Results on DAIR-V2X

Planning Results. We report the planning results in Table 2. Compared to No Fusion, UniV2X achieves a 61% reduction in the average collision rate and a 9.3% reduction in the average off-road rate, as shown in Table 2. Notably, as the planning time increases, the performance improvement becomes more pronounced. This result effectively demonstrates that utilizing infrastructure information can enhance autonomous driving performance, particularly for low-cost monocular solutions. When compared to Vanilla and BEV Feature Fusion methods, UniV2X significantly outperforms them in terms of average collision rate (0.34 *vs* 2.12, 0.34 *vs* 0.83) and average off-road rate (0.89 *vs* 2.07, 0.89 *vs* 0.93). Even when compared to CooperNaut, UniV2X still achieves a better average collision rate (0.34 *vs* 1.23). However, there is an abnormal phenomenon where CooperNaut exhibits a much larger L2 error than all other methods up to 5.35m. This is because its planning length is relatively conservative, resulting in a shorter planning path that more easily stays within the drivable area, thereby achieving a lower off-road rate. Furthermore, UniV2X requires significantly less transmission cost compared to the baseline solutions (8.09×10^5 vs 8.19×10^7), making it far more transmission-efficient and transmission-friendly.

Agent Perception Results. We employ various fusion strategies on DAIR-V2X, including No Fusion, Early Fusion (fusing raw infrastructure BEV feature), and Late Fusion (fusing infrastructure detection results with Hungarian method(Kuhn 2010)). Additionally, we reproduce current SOTA cooperative perception methods on DAIR-V2X, namely V2X-ViT (Xu et al. 2022b), Where2comm (Hu et al. 2022), DiscoNet (Li et al. 2021), and CoAlign (Lu et al. 2023). For a fair comparison, we standardize inputs (image-only) and evaluation settings. All methods, except for CoCa3D (Hu et al. 2023a) based on depth estimation, are re-implemented using BEVFormer (Li et al. 2022b).

Method	mAP ↑	AMOTA ↑	Trans. Cost ↓
No Fusion	0.165	0.163	0
Early Fusion	0.243	0.209	8.19×10^7
Late Fusion	0.196	**0.263**	**6.60**$\times10^2$
CoAlign	0.240	0.234	8.19×10^7
CoCa3D	0.226	-	4.63×10^6
V2X-ViT	0.268	**0.287**	2.56×10^6
Where2comm	0.162	0.106	5.40×10^5
DiscoNet	0.216	0.203	1.60×10^5
V2X-ViT+Where2comm	0.178	0.071	7.22×10^4
UniV2X (Ours)	**0.295** (+0.13)	**0.239** (+0.076)	**6.96**$\times10^4$

Table 3: Detection and Multi-Object Tracking Evaluation Results.

We present the evaluation results (car class) for detection and tracking in Table 3. (1) UniV2X demonstrates a notable enhancement of **+7.6** and **+3.0** in AMOTA(%) compared to No Fusion and Early Fusion. (2) UniV2X outperforms CoCa3D, Where2comm, and DiscoNet at similar or less transmission cost. (3) UniV2X achieves inferior tracking performance compared to tracking-by-detection methods with complex association, such as Late Fusion+AB3DMOT (0.239 *vs* 0.263 at AMOTA), but it significantly outperforms this tracking-by-detection solution in detection (0.295 *vs* 0.196). It is important to note that this tracking-by-detection solution is not suitable for end-to-end autonomous driving. (4) V2X-ViT exhibits better performance than UniV2X at

Method	IoU-Lane (%)↑	IoU-Crosswalk (%)↑	Trans. Cost (BPS)↓
No Fusion	6.4	2.7	0
Ealry Fusion	16.7	17.8	8.19×10^7
CoBEVT	15.6	16.4	2.56×10^6
UniV2X (Ours)	**17.8** (+11.4)	**19.8** (+17.1)	**1.47**$\times10^5$

Table 4: Online Mapping Evaluation Results.

AMOTA (0.287 *vs* 0.239), but it requires much more transmission cost (2.56×10^6 *vs* 6.94×10^4). When we further compress the V2X-ViT transmission to a level similar to UniV2X with Where2comm, there is a significant performance drop (from 0.287 to 0.071 at AMOTA). These outcomes underscore the capability of our infrastructure agent queries and agent fusion module in enhancing agent perception ability with light transmission cost.

Online Mapping Results. We implement No Fusion, Early Fusion, and CoBEVT(Xu et al. 2022a) for online mapping on DAIR-V2X. All methods are re-implemented using BEVFormer (Li et al. 2022b). The mapping performance is reported with Segmentation Intersection over Union (IoU) (%) as the evaluation metric in Table 4. UniV2X demonstrates notable improvements in lane perception and crossing perception compared No Fusion, Early Fusion and CoBEVT, respectively. Moreover, compared with Early Fusion and CoBEVT, UniV2X requires less than 1/10th of the transmission cost. These results indicate that infrastructure lane queries and cross-view lane fusion are effective in enhancing online mapping ability.

Occupancy Prediction Results. Concerning the evaluation of occupancy prediction, as depicted in Table 5, UniV2X exhibits notably superior performance compared to No Fusion in both near and far regions. Particularly, UniV2X achieves **+5.7** and **+13.4** improvement in IoU-n (%) and IoU-f (%) respectively. Here, "IoU-n" and "IoU-f" denote evaluation ranges of 30×30m and 50×50m, respectively. These results underscore the effectiveness of our sparse-dense hybrid transmitted data in significantly enhancing occupancy prediction.

Method	IoU-n (%)↑	IoU-f (%)↑
No Fusion	16.3	13.1
UniV2X	**22.0** (+5.7)	**26.5** (+13.4)

Table 5: Occupancy Prediction Evaluation Results.

Ablation Study on Reliability

We evaluate UniV2X under various communication conditions. Here we assess the impact of data transmission corruption. Additionally, we assess the robustness of UniV2X across different communication bandwidths and latencies, as detailed in the Appendix. We specifically utilize agent queries and fusion to illustrate this reliability.

When assessing UniV2X, our initial step involves randomly discarding 10%, 30%, 50%, 70%, and 100% of infrastructure agent queries during transmission to simulate

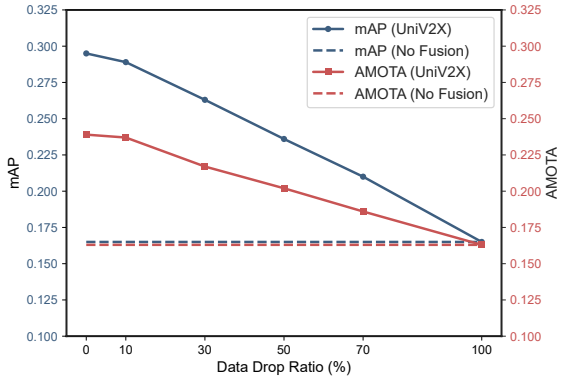


Figure 4: Reliability on Data Corruption.

data corruption. Following this, the retained queries are utilized for cross-view query transmission and interaction, and the performance of agent perception, encompassing object detection and tracking, is evaluated accordingly. The evaluation results depicted in Figure 4 unveil a gradual degradation in agent perception as the data corruption ratio increases. When the data corruption ratio reaches 100%, meaning only ego-vehicle sensor data can be used, the performance of UniV2X becomes comparable to that of the No Fusion model. This decline in performance is anticipated, as data corruption diminishes the complementary information crucial for ego-vehicle autonomous driving. Moreover, even in the absence of certain data due to corruption, UniV2X can maintain a basic level of performance comparable to the No Fusion model. This underscores the reliability of our transmission and cross-view data fusion mechanism.

Conclusion

This paper presents UniV2X, a novel end-to-end framework that integrates crucial tasks from various perspectives into a single network. With a planning-oriented approach, it leverages raw sensor data while ensuring network interpretability for cooperative autonomous driving. Additionally, a sparse-dense hybrid data transmission strategy is devised to harness cross-view data and enhance overall planning performance. This transmission approach is both communication-friendly and reliable, aligning with V2X communication requirements. Empirical results on the DAIR-V2X dataset validate the efficacy of our proposed approach.

Limitations and Future Work. The framework involves multiple modules and different agent perspectives, resulting in a high degree of complexity. As a result, several interaction fusion modules within the framework remain in preliminary stages. Further refinement is essential for optimizing the internal design of the subsequent framework. In this work, we only consider open-loop evaluation for end-to-end autonomous driving. We will conduct more closed-loop experiments to evaluate our UniV2X.

Acknowledgments

This work is supported by the Wuxi Research Institute of Applied Technologies at Tsinghua University under Grant No. 20242001120, as well as the General Research Fund of Hong Kong under Grants No. 17200622 and 17209324.

Code — https://github.com/AIR-THU/UniV2X

References

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

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

Chen, Z.; Shi, Y.; and Jia, J. 2023. TransIFF: An Instance-Level Feature Fusion Framework for Vehicle-Infrastructure Cooperative 3D Detection with Transformers.

Cui, J.; Qiu, H.; Chen, D.; Stone, P.; and Zhu, Y. 2022. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17252–17262.

Dong, J.; Chen, Q.; Qu, D.; Lu, H.; Ganlath, A.; Yang, Q.; Chen, S.; and Labi, S. 2023. LiDAR-based Cooperative Relative Localization. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, 1–8. IEEE.

Fan, S.; Yu, H.; Yang, W.; Yuan, J.; and Nie, Z. 2024. Quest: Query stream for vehicle-infrastructure cooperative perception. *IEEE international conference on robotics and automation (ICRA)*.

Gu, J.; Zhang, J.; Zhang, M.; Meng, W.; Xu, S.; Zhang, J.; and Zhang, X. 2023. FeaCo: Reaching Robust Feature-Level Consensus in Noisy Pose Conditions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3628–3636.

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

Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; and Wang, Y. 2023a. Collaboration Helps Camera Overtake LiDAR in 3D Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9243–9252.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023b. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Jiang, Y.; Javanmard, E.; Nakazato, J.; Tsukada, M.; and Esaki, H. 2023. Roadside LiDAR Assisted Cooperative Localization for Connected Autonomous Vehicles. *ACM Intelligent Computing and its Emerging Applications (ICEA)*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kuhn, H. W. 2010. The Hungarian Method for the Assignment Problem. In *50 Years of Integer Programming*.

Li, Y.; Ma, D.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; and Feng, C. 2022a. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4): 10914–10921.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P.; and Lu, T. 2022c. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1280–1289.

Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Qiu, C.; Yadav, S.; Squicciarini, A.; Yang, Q.; Fu, S.; Zhao, J.; and Xu, C. 2022. Distributed data-sharing consensus in cooperative perception of autonomous vehicles. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, 1212–1222. IEEE.

Ruan, H.; Yu, H.; Yang, W.; Fan, S.; Tang, Y.; and Nie, Z. 2023. Learning Cooperative Trajectory Representations for Motion Forecasting. *arXiv preprint arXiv:2311.00371*.

Song, R.; Liang, C.; Cao, H.; Yan, Z.; Zimmer, W.; Gross, M.; Festag, A.; and Knoll, A. 2024. Collaborative Semantic Occupancy Prediction with Hybrid Feature Fusion in Connected Automated Vehicles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tianhang, W.; Guang, C.; Kai, C.; Zhengfa, L.; Bo, Z.; Alois, K.; and Jiang, C. 2023. UMC: A Unified Bandwidth-efficient and Multi-resolution based Collaborative Perception Framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23383–23392.

Valiente, R.; Zaman, M.; Ozer, S.; and Fallah, Y. P. 2019. Controlling steering angle for cooperative self-driving vehicles utilizing CNN and LSTM-based deep networks. In *2019 IEEE intelligent vehicles symposium (IV)*, 2423–2428. IEEE.

Wang, T.; Kim, S.; Ji, W.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Luo, P. 2023. DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving. *arXiv preprint arXiv:2304.01168*.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 605–621. Springer.

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

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

Yang, K.; Yang, D.; Zhang, J.; Li, M.; Liu, Y.; Liu, J.; Wang, H.; Sun, P.; and Song, L. 2023. Spatio-temporal domain awareness for multi-agent collaborative perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23383–23392.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. 2022. 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*, 21361–21370.

Yu, H.; Tang, Y.; Xie, E.; Mao, J.; Luo, P.; and Nie, Z. 2023a. Flow-based Feature Fusion for Vehicle-Infrastructure Cooperative 3D Object Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yu, H.; Tang, Y.; Xie, E.; Mao, J.; Yuan, J.; Luo, P.; and Nie, Z. 2023b. Vehicle-infrastructure cooperative 3d object detection via feature flow prediction. *arXiv preprint arXiv:2303.10552*.

Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N.; et al. 2023c. 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*, 5486–5495.

Zhong, J.; Yu, H.; Zhu, T.; Xu, J.; Yang, W.; Nie, Z.; and Sun, C. 2024. Leveraging temporal contexts to enhance vehicle-infrastructure cooperative perception. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE.