UniMM-V2X: MoE-Enhanced Multi-Level Fusion for End-to-End Cooperative Autonomous Driving

Ziyi Song[1], Chen Xia[1], Chenbing Wang[1] Haibao Yu[2], Sheng Zhou[1, 3*], Zhisheng Niu[1]

[1]Department of Electronic Engineering, Tsinghua University
[2]The University of Hong Kong
[3]State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University
{songzy24,xiac23,wcb24}@mails.tsinghua.edu.cn, yuhaibao94@gmail.com,
sheng.zhou@tsinghua.edu.cn, niuzhs@tsinghua.edu.cn

Abstract

Autonomous driving holds transformative potential but remains fundamentally constrained by the limited perception and isolated decision-making with standalone intelligence. While recent multi-agent approaches introduce cooperation, they often focus merely on perception-level tasks, overlooking the alignment with downstream planning and control, or fall short in leveraging the full capacity of the recent emerging end-to-end autonomous driving. In this paper, we present UniMM-V2X, a novel end-to-end multi-agent framework that enables hierarchical cooperation across perception, prediction, and planning. At the core of our framework is a multi-level fusion strategy that unifies perception and prediction cooperation, allowing agents to share queries and reason cooperatively for consistent and safe decision-making. To adapt to diverse downstream tasks and further enhance the quality of multi-level fusion, we incorporate a Mixture-of-Experts (MoE) architecture to dynamically enhance the BEV representations. We further extend MoE into the decoder to better capture diverse motion patterns. Extensive experiments on the DAIR-V2X dataset demonstrate our approach achieves state-of-the-art (SOTA) performance with a 39.7% improvement in perception accuracy, a 7.2% reduction in prediction error, and a 33.2% improvement in planning performance compared with UniV2X, showcasing the strength of our MoE-enhanced multi-level cooperative paradigm.

Code — https://github.com/Souig/UniMM-V2X

Introduction

Traditional autonomous driving pipelines, with their modular structure, suffer from error propagation and limited generalization. As (Li et al. 2024) improved environmental perception through bird's-eye-view (BEV) representations, end-to-end autonomous driving has been widely studied in (Hu et al. 2023; Jiang et al. 2023; Sun et al. 2024). Although end-to-end autonomous driving offers a solution by directly mapping raw sensor data to final control, this standalone-intelligence system is constrained by sensor range and and struggle with rare critical events and predicting other agents' intentions. Vehicle-to-Everything (V2X) communication emerges as a key enabler to overcome these limitations by facilitating real-time information exchange.

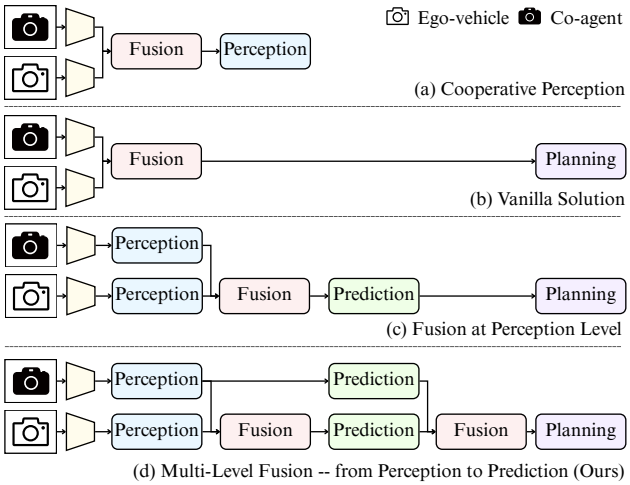


Figure 1: V2X communication modes in the VICAD (Vehicle-to-Infrastructure Cooperation Autonomous Driving) problem (Yu et al. 2022). (a) Cooperative perception methods focus on multi-agent detection and tracking, but may not align with planning objectives. (b) Vanilla solutions fuse features directly to generate planning outputs, with limited interpretability and compromised safety. (c) Module results can be supervised, but only enable perception-level cooperation. (d) Our design employs multi-level, multi-agent cooperation that integrates perception and prediction to enable cooperative decision-making.

As shown in Figure 1(a), V2X communication is widely applied in cooperative perception, improving environmental awareness through multi-agent cooperation (Xu et al. 2022b,a; Chen et al. 2019a; Hu et al. 2022b). CooperNaut (Cui et al. 2022) encodes LiDAR into compact features for transmission but suffers from limited interpretability (Figure 1(b)). UniV2X (Yu et al. 2025) adopts a query-based architecture with sparse-dense hybrid communication but its fusion is restricted to the perception level (Figure 1(c)). With end-to-end autonomous driving becoming the prevail-

*Corresponding author.

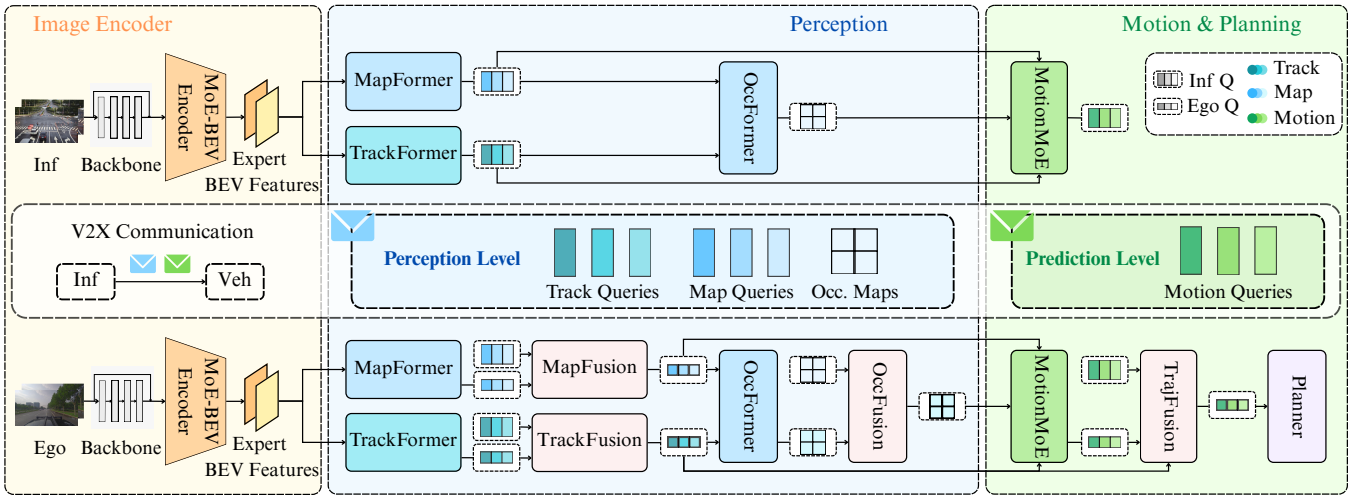


Figure 2: The overview of the UniMM-V2X framework. The system performs explicit multi-level fusion by integrating perception-level and prediction-level information from multiple agents to enhance downstream planning. Both the BEV encoder and motion decoder are equipped with MoE architectures, where the encoder generates task-adaptive BEV features tailored for various downstream tasks, and the decoder employs specialized experts to model diverse motion patterns, enhancing the effectiveness and adaptability of multi-level fusion for more robust planning performance. This unified MoE-enhanced multi-level fusion framework facilitates effective cooperation among agents throughout the entire autonomous driving pipeline.

ing paradigm, a natural question arises: *While multi-agent cooperation has improved perception, is this enough for end-to-end systems where planning is the final goal?*

Due to the complexity of end-to-end autonomous driving, relying solely on perception-level fusion is often insufficient, as accurate multi-agent motion prediction plays a more critical role in ensuring safety and efficiency. To address this, we propose UniMM-V2X, an MoE-enhanced *multi-level* fusion framework that performs cooperative information fusion at both the perception and prediction levels, addressing the VICAD problem identified in (Yu et al. 2025). At the perception level, we exchange track queries, map queries and occupancy probability maps to enable cooperative scene understanding. Building on this foundation, motion queries are further transmitted at the prediction level, allowing agents to reason jointly about future behaviors. Moreover, the interpretability of queries at both the instance and scene levels enhances the reliability of the system.

While multi-level fusion enables coherent information flow from perception to prediction, different downstream tasks have distinct requirements for BEV representations and other features. A shared BEV encoder may struggle to simultaneously meet the needs of perception, prediction, and planning, and the conventional motion decoder may fail to capture diverse agent motions. These concerns raise another natural question: *How can the system adaptively generate specialized representations and predictions to meet these heterogeneous demands?*

To overcome these challenges, we innovatively integrate MoE architecture into both the *BEV encoder* and *motion decoder*. The MoE-enhanced encoder dynamically generates task-specialized BEV representations, allowing each task to leverage features best suited to its objectives. Meanwhile, the MoE-equipped decoder further dynamically generates motion queries via expert branches, each modeling distinct motion patterns such as keeping forward, turning left, or turning right. By complementing multi-level fusion, this design not only improves decision quality but also enhances interpretability and reliability.

The integration of multi-level fusion and MoE architecture creates a synergistic effect beyond individual gains. When combined, perception-level fusion benefits from more specialized BEV features generated by task-aware MoE encoders, and prediction-level fusion receives more reliable trajectory candidates guided by expert-decoded motion queries. This close integration enables each stage of the cooperative autonomous driving pipeline to perform more effectively and consistently, ultimately leading to more accurate and robust driving decisions.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to explore multi-level cooperation in multi-agent end-to-end autonomous driving, enabling cooperation across both perception and prediction to significantly improve decision-making reliability under complex scenarios.
- We introduce MoE into both the encoder and decoder of the end-to-end framework, enhancing the flexibility and specialization of the model to adapt to diverse tasks and prediction requirements in autonomous driving.
- Through extensive experiments, we validate that the combination of multi-level fusion and MoE architecture yields a strong complementary effect, facilitating more reliable cooperation and substantially improving decision quality. Compared to single-agent and existing cooperative methods, UniMM-V2X achieves SOTA results in perception, prediction, and planning.

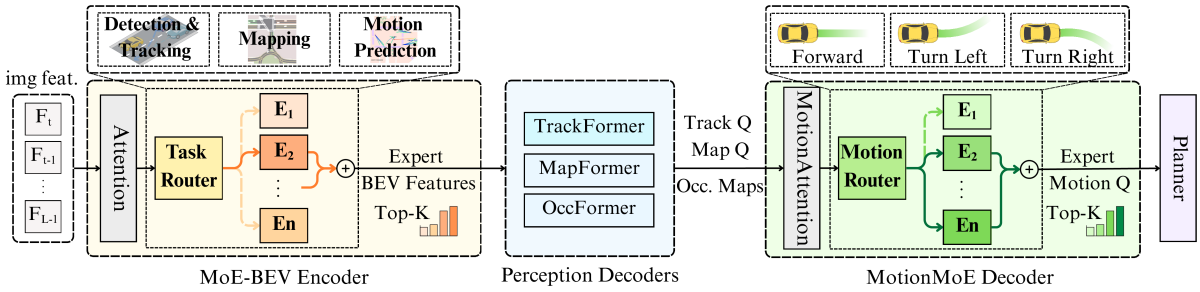


Figure 3: MoE-enhanced encoder and decoder in UniMM-V2X. The encoder enriches BEV feature extraction for diverse downstream tasks (e.g., detection, tracking, mapping, motion prediction), while the decoder generates motion queries through motion-specific experts (e.g., going forward, turning left, turning right) to improve planning quality.

Related Works

End-to-End Autonomous Driving

End-to-end autonomous driving has attracted growing attentions. Early methods (Codevilla et al. 2018, 2019; Zhang et al. 2021) lack interpretability and optimization by skipping intermediate tasks. Furthermore, ST-P3 (Hu et al. 2022a) builds interpretable maps from perception; UniAD (Hu et al. 2023) unifies perception, prediction, and planning via a query-based framework; VAD (Jiang et al. 2023) and SparseDrive (Sun et al. 2024) reduce computational cost through vectorization or sparse design; DiffusionDrive (Liao et al. 2025) employs diffusion models for planning. However, these methods are limited to single-agent inputs. In this work, we extend the paradigm to a multi-agent setting by incorporating cross-agent communication, joint perception-level and prediction-level fusion within a unified end-to-end framework for cooperative planning.

Cooperative Autonomous Driving

V2X communication in autonomous driving has its roots in early frameworks (Chen et al. 2019a,b). With the emergence of Transformer-based architectures, methods like (Liu et al. 2020; Hu et al. 2022b; Xu et al. 2022b) have improved communication strategies by learning when, where and with whom to communicate. CooperNaut (Cui et al. 2022) goes further by linking perception and control into a unified framework. UniV2X (Yu et al. 2025) introduces a sparse-dense hybrid communication protocol, effectively coordinating vehicle-to-infrastructure information. However, these methods either adopt vanilla fusion strategies or perform fusion only at the perception stage, limiting their effectiveness in planning. In contrast, we propose a multi-level fusion framework that operates across both perception and prediction stages, enabling agents to cooperatively reason from spatial observations to motion intents for safer planning.

Mixture of Experts

MoE operates using conditional computation and a learnable gating function. Early work, such as (Shazeer et al. 2018), enabled MoE scaling in Transformers by replacing standard FFNs with sparsely activated experts, an idea later extended to large encoder-decoder models by (Lepikhin et al. 2020; Fedus, Zoph, and Shazeer 2022; Zoph et al. 2022) to address stability and fine-tuning issues. In autonomous driving, DriveMoE (Yang et al. 2025) applies MoE for sensor scheduling and action guidance. However, these methods typically restrict MoE to a single stage, limiting its ability to handle heterogeneous task demands. In contrast, we integrate MoE into both the BEV encoder and the motion decoder, enabling the system to generate task-specialized BEV representations as well as expert-guided motion predictions to improve decision reliability in multi-agent cooperation.

Method

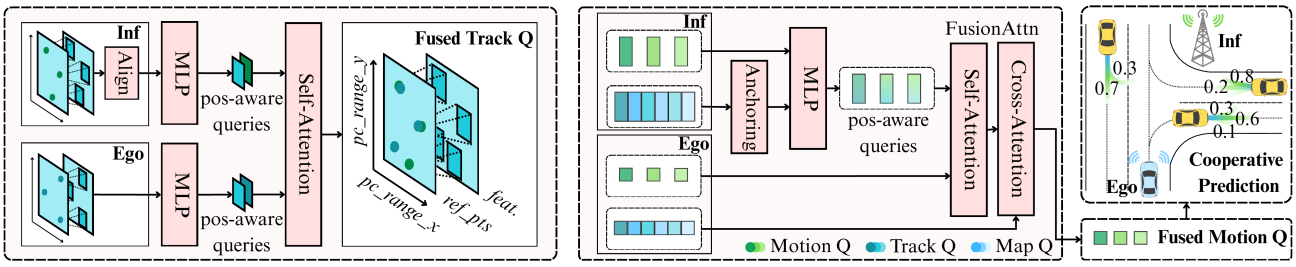
Overview

The overall framework of UniMM-V2X is illustrated in Figure 2. It performs explicit *multi-level* fusion across agents by integrating information at both the perception level and the prediction level, thereby enhancing the safety and robustness of downstream planning decisions. The MoE architecture is integrated into both the BEV encoder and the motion decoder, strengthening the effectiveness of multi-level fusion across perception and prediction. The encoder generates feature representations that are better adapted to the distinct needs of various downstream tasks, while the decoder exploits expert specialization to more precisely capture diverse motion patterns, ultimately delivering more robust and planning-aware trajectory outputs.

The framework consists of three main components: image encoders, a cooperative perception module, and a cooperative motion and planning module. The image encoder incorporates the MoE architecture to extract task-adaptive BEV features. The perception module performs cooperative detection, tracking, mapping, and occupancy map generation. The motion and planning module generates motion predictions via MoE-based decoder and fuses multi-agent predictions for planning decisions. Together, the perception-level and prediction-level fusion form a unified multi-level fusion framework that enables effective cooperation across agents throughout the decision-making process.

MoE for Adaptive Feature and Motion Modeling

To effectively address the complex joint demands of perception, prediction and planning, we place the MoE architecture



(a) Perception-Level Fusion (b) Prediction-Level Fusion

Figure 4: Multi-level fusion in UniMM-V2X. (a) Perception-level fusion introduces positional priors via reference point embeddings and uses attention-based dynamic fusion across agents. (b) Prediction-level fusion employs anchor-based embedding and dynamic fusion to support motion reasoning in complex multi-agent settings.

in both the BEV encoder (MoE-BEV Encoder) and motion decoder (MotionMoE), as illustrated in Figure 3. We adopt a standard sparse MoE design (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022), in which traditional FFNs are replaced by a set of expert networks:

$$\text{MoE}(x) = \sum_{i \in \mathcal{I}_k(x)} \tilde{G}_i(x) \cdot f_i(x), \tag{1}$$

where f_i is the i-th expert, $\tilde{G}_i(x)$ the normalized routing weight, and $\mathcal{I}_k(x)$ the selected experts. To avoid expert collapse and ensure balanced usage, we add a load balancing loss (Fedus, Zoph, and Shazeer 2022):

$$\mathcal{L}_{\text{moe}} = \lambda \left(\text{Var}(p) + \text{Var}(l) \right), \tag{2}$$

where p are the routing probabilities, l the expert loads, and λ a weighting factor, encouraging uniform expert activation.

Within the MoE-BEV encoder, we replace the FFNs with the MoE block to enable adaptive and specialized generation of BEV feature representations:

$$\mathbf{z}^{(l+1)} = \text{MoE}(\text{CrossAttn}(\text{SelfAttn}(\mathbf{z}^{(l)}))), \tag{3}$$

where $\mathbf{z}^{(l)}$ is the BEV feature representations at layer l, and the MoE module selectively activates top-k experts (e.g., $k = 2$) to process the attended BEV features. Similarly, in the decoder, we replace the FFNs with the MoE architecture in the MotionMoE module to generate motion queries Q_M^{veh} and Q_M^{other} that adapt to diverse motion patterns.

Multi-Agent Perception-Level Fusion

In perception-level fusion, we incorporate track fusion, map fusion, and occupancy fusion. Among them, track fusion is particularly critical, because the resulting track queries function as crucial contributing inputs for downstream tasks. To enhance their quality, we introduce *TrackFusion* module, which dynamically builds associations between agents, as shown in Figure 4(a). The map fusion and occupancy fusion modules are provided in the Appendix, where the former uses an MLP and the latter adopts a max operation.

In the TrackFusion block, an attention mechanism is employed to model complex inter-agent query relationships and perform weighted feature fusion based on learned relevance scores, overcoming the limitations of hard matching methods that rely on fixed distance thresholds in previous works. Initially, queries from other agents Q_A^{other} are transformed into the ego-vehicle's coordinate system using an MLP:

$$Q_A^{\text{other}} = \text{MLP}([Q_A^{\text{other}}, \mathcal{R}]), \tag{4}$$

where $\mathcal{R}$ is the rotation matrix. Subsequently, the reference point information P_A^{other} and P_A^{veh} are integrated as spatial contextual priors into the dynamic feature correlation learning process, as formulated below:

$$Q_A = \text{MHSA}(X_A + \text{MLP}(P_A)), \tag{5}$$

$$X_A = \text{Concat}(Q_A^{\text{veh}}, Q_A^{\text{other}}), \tag{6}$$

$$P_A = \text{Concat}(P_A^{\text{veh}}, P_A^{\text{other}}). \tag{7}$$

We employ an MLP to embed the spatial coordinates of each agent into a learnable representation. These spatial embeddings are concatenated with agent-specific queries and jointly fed into a multi-head self-attention (MHSA) mechanism. This design allows the model to capture semantic dependencies across agents while incorporating their relative spatial positions, enabling context-aware and spatially sensitive feature fusion that enhances cooperative understanding.

Cross-View Prediction-Level Fusion

In prediction-level fusion, as shown in Figure 4(b), we fuse motion queries from multiple agents through *TrajFusion* module to enable cooperative motion prediction, which finally improves the performance of planning decisions.

The fusion process begins with other agents transmitting their motion queries Q_M^{other} to the ego agent via inter-agent communication. To spatially align the heterogeneous trajectory data, we first transform the agent-level anchors P_{anchor}, derived from Q_A^{other}, into the coordinate frame of the ego-vehicle using the rotation matrix $\mathcal{R}$:

$$P_M^{\text{other}} = \text{MLP}([P_{\text{anchor}}, \mathcal{R}]). \tag{8}$$

The transformed positional information is then projected through an MLP for position embedding:

$$\tilde{Q}_M^{\text{other}} = \text{MLP}([Q_M^{\text{other}}, P_M^{\text{other}}]). \tag{9}$$

Method	L2 Error(m)↓				Collision Rate(%)↓				Trans. Cost
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	(BPS)↓
VAD* (Jiang et al. 2023)	1.65	2.72	3.80	2.72	0.86	1.21	1.28	1.12	-
UniAD* (Hu et al. 2023)	1.26	2.22	3.06	2.18	0.88	1.18	1.32	1.13	-
SparseDrive* (Sun et al. 2024)	1.02	1.69	2.37	1.69	0.46	1.23	1.28	0.99	-
Vanilla	1.36	2.29	3.32	2.32	1.03	0.88	1.32	1.08	8.19×10^{7}
V2VNet (Wang et al. 2020)	1.96	2.37	3.41	2.58	0.74	0.88	1.03	0.88	8.19×10^{7}
CooperNaut (Cui et al. 2022)	2.69	4.07	5.50	4.09	1.18	1.32	1.76	1.42	8.19×10^{7}
UniV2X (Yu et al. 2025)	1.45	2.19	3.04	2.23	0.15	**0.15**	0.44	0.25	$8.09\times\mathbf{10^{5}}$
UniMM-V2X	**0.78**	**1.63**	**2.05**	**1.49**	**0.05**	**0.15**	**0.15**	**0.12**	$9.32\times\mathbf{10^{5}}$

Table 1: **Planning** performance. *: Single-agent no fusion method. We achieve improvements in reducing L2 error and collision rate, enhancing overall system safety.

Method	Detection	Tracking	Mapping		Trans. Cost
	mAP↑	AMOTA↑	Lane(%)↑	Crossing(%)↑	(BPS)↓
UniAD* (Hu et al. 2023)	0.181	0.197	13.3	8.7	-
SparseDrive* (Sun et al. 2024)	0.324	0.130	-	5.2	-
Early Fusion	0.243	0.209	16.7	17.8	8.19×10^{7}
Late Fusion	0.236	0.263	13.4	9.1	$\mathbf{6.60}\times10^{2}$
CoAlign† (Lu et al. 2023)	0.261	0.234	-	-	8.19×10^{7}
Where2comm† (Hu et al. 2022b)	0.221	0.106	-	-	5.40×10^{5}
CoBEVT† (Xu et al. 2022a)	0.264	0.243	15.6	16.4	2.56×10^{6}
V2X-ViT† (Xu et al. 2022b)	0.261	0.287	-	-	2.56×10^{6}
UniV2X (Yu et al. 2025)	0.302	0.241	17.7	19.7	2.17×10^{5}
UniMM-V2X	**0.422**	**0.427**	**17.9**	**20.3**	2.17×10^{5}

Table 2: **Perception** performance. *: Single-agent no fusion method. †: Cooperative perception methods. We significantly improve all performance metrics without increasing transmission cost.

The ego-agent motion queries and the positionally enhanced queries from other agents are then concatenated and processed via an attention-based mechanism:

$$F_M = \text{Concat}(Q_M^{\text{veh}}, \tilde{Q}_M^{\text{other}}), \tag{10}$$

$$Q_M = \text{MHCA}(\text{MHSA}(F_M), Q_A). \tag{11}$$

Here, MHSA captures inter-agent dependencies within combined motion queries F_M, and MHCA integrates perception-aware context by attending to the fused perception queries Q_A, which are historically enriched and semantically aligned, thereby providing strong priors for motion reasoning in complex multi-agent scenarios.

Learning

The overall training objective is to jointly optimize multiple sub-tasks involved in end-to-end cooperative autonomous driving. Specifically, the loss function consists of six components: detection and tracking, online mapping, occupancy prediction, motion prediction, planning, and the auxiliary load balancing term introduced by the MoE module.

$$\mathcal{L} = \mathcal{L}_{\text{track}} + \mathcal{L}_{\text{map}} + \mathcal{L}_{\text{occ}} + \mathcal{L}_{\text{mot}} + \mathcal{L}_{\text{plan}} + \mathcal{L}_{\text{moe}}. \tag{12}$$

All components are jointly optimized in an end-to-end manner to achieve unified perception, prediction, and planning.

Experiments

Experimental Settings

The overall framework is trained with the DAIR-V2X dataset (Yu et al. 2022), which comprises approximately 100 scenes captured at 28 complex traffic intersections in the real world. We use the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01. We train the perception stage for 40 epochs on 8 NVIDIA A800 GPUs, and subsequently perform end-to-end motion and planning training for 20 epochs using the same GPU setup. During training, the MoE layers select the top-2 experts for each token to balance specialization and computational efficiency. Evaluation metrics of each task are described in the Appendix. We also implement UniMM-V2X on the V2X-Sim dataset (Li et al. 2022), a large-scale simulation benchmark with diverse traffic scenarios for cooperative autonomous driving, and results are provided in the Appendix.

Main Results

We compare UniMM-V2X with several single-agent end-to-end autonomous driving models (Jiang et al. 2023; Hu et al. 2023; Sun et al. 2024) as well as multi-agent cooperative driving frameworks. For the cooperative baselines, we evaluate both cooperative perception methods (Lu et al. 2023;

Method	IoU-n(%)↑	IoU-f(%)↑
UniAD* (Hu et al. 2023)	16.3	13.1
UniV2X (Yu et al. 2025)	22.2	**26.0**
UniMM-V2X	**23.0**	23.7

(a) **Occupancy prediction** performance. "n" and "f" denote near (30×30m) and far (50×50m) ranges.

Method	minADE(m)↓	minFDE(m)↓	MR↓
UniAD* (Hu et al. 2023)	0.78	0.82	0.21
SparseDrive* (Sun et al. 2024)	1.02	1.87	0.34
UniV2X (Yu et al. 2025)	0.69	0.74	0.17
UniMM-V2X	**0.64**	**0.69**	**0.13**

(b) **Moion prediction** performance.

Table 3: **Prediction** performance. *: Single-agent no fusion method.

Multi-Level Fusion		MoE		Perception		Motion Prediction	Planning L2 Error(m)				Coll.(%)
P-Level	M-Level	Enc.	Dec.	mAP↑	AMOTA↑	minADE(m)↓	1s	2s	3s	Avg.↓	Avg.↓
-	-	-	-	0.181	0.197	0.78	1.26	2.22	3.06	2.18	1.13
✓	-	-	-	0.352	0.328	0.69	1.09	2.25	2.75	2.03	0.68
-	✓	-	-	0.191	0.193	0.67	1.13	1.71	2.71	1.85	0.50
✓	✓	-	-	0.351	0.328	0.66	1.14	1.76	2.71	1.87	0.47
-	-	✓	-	0.238	0.269	0.81	1.28	1.97	2.92	2.06	0.39
-	-	-	✓	0.179	0.198	0.78	1.24	1.84	2.98	2.02	0.54
-	-	✓	✓	0.240	0.267	0.75	1.02	1.73	2.82	1.85	0.24
✓	-	✓	✓	**0.427**	**0.427**	0.74	0.91	1.78	2.47	1.72	0.40
-	✓	✓	✓	0.238	0.271	0.65	0.96	**1.53**	2.08	1.52	0.15
✓	✓	✓	✓	0.422	**0.427**	**0.64**	**0.78**	1.63	**2.05**	**1.49**	**0.12**

Table 4: **Ablation study results.** We conduct experiments to evaluate the effectiveness of multi-level fusion and the MoE mechanism. P-Level and M-Level refer to the perception level fusion and motion prediction level fusion, while Enc. and Dec. indicate applying MoE to the BEV encoder and motion decoder, respectively.

Hu et al. 2022b; Xu et al. 2022a,b) and end-to-end cooperative driving approaches (Cui et al. 2022; Yu et al. 2025).

Planning. The planning results are summarized in Table 1. UniMM-V2X achieves the lowest average L2 error of **1.49m**, reducing by **33.2%** compared with UniV2X (Yu et al. 2025), outperforming all the baselines including advanced single-agent and existing cooperative methods. More importantly, UniMM-V2X demonstrates superior safety, attaining the lowest average collision rate of **0.12%**, which represents a **52.0%** reduction compared to UniV2X (Yu et al. 2025). Although our approach introduces slightly higher communication overhead due to the transmission of motion queries, the improvements in the planning performance clearly justify the additional cost.

Perception. Table 2 presents the performance of UniMM-V2X on perception tasks. Compared to the SOTA no-fusion baseline (Sun et al. 2024), our method achieves a **+0.098** improvement in mAP and a **+0.297** improvement in AMOTA, demonstrating the effectiveness of cooperation. Compared to the SOTA end-to-end cooperative driving framework (Yu et al. 2025), our method achieves an improvement of **39.7%** in mAP and **77.2%** in AMOTA, without introducing additional communication cost at the perception level.

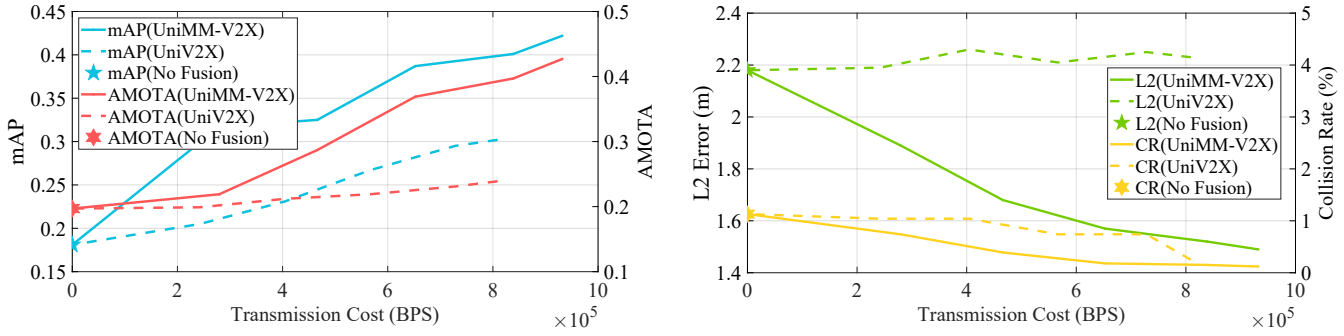
Prediction. The motion prediction results are shown in Table 3b. UniMM-V2X achieves the best performance with **0.64m** minADE, **0.69m** minFDE and **13.2%** MissRate, reducing errors by **7.2%** and **6.8%** on minADE and minFDE respectively compared with UniV2X (Yu et al. 2025). These improvements contribute significantly to the improvement of the final planning performance mentioned above.

Ablation Study

Effect of Multi-Level Fusion. As shown in Table 4, perception-level fusion improves detection and tracking performance but has limited effect on motion prediction and planning, probably due to the misalignment between perception accuracy and planning requirements. In contrast, prediction-level fusion enhances planning safety by providing supplementary motion cues for occluded objects and refining uncertain trajectories, but perception performance remains similar to the single-agent baseline due to the lack of early-stage cooperation. These observations indicate that single-level fusion alone is insufficient to optimize all driving tasks. Multi-level fusion ensures the propagation of high-quality intermediate features throughout the pipeline, resulting in consistent improvements across all modules.

Effect of MoE. As shown in Table 4, integrating MoE into the BEV encoder enhances environmental understanding, improving both perception and planning performance for the single vehicle. Using MoE only in the motion decoder yields limited gains, likely due to insufficient task-specific BEV features for accurate motion prediction. The best results are achieved when MoE is applied to both the encoder and the decoder, where the encoder produces task-aware BEV features and the decoder leverages expert specialization to capture complex motion behaviors.

Interaction between Multi-Level Fusion and MoE. Applying multi-level fusion individually does not achieve opti-



(a) Perception performance.

(b) Planning performance.

Figure 5: Performance under different communication constraints.

mal performance, suggesting that traditional fusion can saturate when used independently at each level. In contrast, integrating MoE in both the encoder and decoder significantly enhances multi-level fusion, resulting in gains (AMOTA +0.230, L2 error -0.69m, Table 4). This indicates that multi-level fusion is crucial for capturing hierarchical interactions between perception and prediction, while MoE further enhances this capability by enabling task-specific specialization. Combined, they produce substantial improvements across perception, prediction, and planning.

Num.	P	M	mAP↑	AMOTA↑	L2(m)↓	CR(%)↓
4	-	-	0.240	0.229	1.77	0.39
4	✓	-	0.235	0.230	1.85	0.34
4	-	✓	0.237	0.231	1.60	0.20
4	✓	✓	0.230	0.229	1.78	0.37
8	-	-	0.421	0.425	1.66	0.20
8	✓	-	0.366	0.347	1.75	0.43
8	-	✓	**0.422**	**0.427**	**1.49**	**0.12**
8	✓	✓	0.368	0.351	1.68	0.25
16	-	-	0.403	0.374	1.71	0.36
16	✓	-	0.359	0.341	1.76	0.41
16	-	✓	0.401	0.374	1.64	0.15
16	✓	✓	0.361	0.339	1.72	0.20

Table 5: **Ablation on MoE expert number and decoder placement** conducted within the multi-level fusion framework. All the experiments utilize the MoE-based encoder. "P" indicates MoE applied to the perception decoder and "M" denotes its placement in the motion decoder.

MoE Configuration. Since all decoder modules are Transformer-based, we explore replacing their FFNs with MoE. All variants employ the MoE-based encoder, where spatial experts consistently improve performance. Results in Table 5 show that using 8 experts in the motion decoder achieves the best trade-off. Too few experts limit the specialization, while too many experts cause data sparsity and under-utilization. Furthermore, limited token diversity in the perception decoder reduces MoE benefits and may introduce gating noise. In contrast, placing MoE in the motion decoder, which is more tightly coupled with the planner, enables better adaptation to diverse behaviors, leading to more flexible and accurate planning.

Practicality and Reliability of the System

We assess the practicality and efficiency of our method by comparing communication cost and inference latency (BPS and FPS). Unlike bandwidth-heavy BEV methods, our query-based design drastically reduces communication cost by 87.9× without sacrificing planning quality. UniMM-V2X achieves an FPS of 5.4, which is a slight decrease compared to UniV2X's 5.8 FPS due to the integration of MoE and multi-level fusion, along with a modest increase in communication cost from enriched motion queries. However, these minor costs are strongly justified by significant improvements in planning safety and reliability, reflecting an excellent cost–benefit profile. Further evaluation under variable bandwidth conditions in Figure 5 shows that UniMM-V2X consistently outperforms UniV2X across all settings. While UniV2X's fusion offers negligible benefits to planning, performing close to the No Fusion baseline, our multi-level fusion with MoE approach can effectively leverage available communication for cooperative planning, ensuring reliability and scalability in real-world autonomous driving.

Conclusion

In this work, we propose UniMM-V2X, an end-to-end framework for robust multi-agent cooperative driving. By explicitly fusing information at both perception and prediction levels, and integrating MoE modules in the BEV encoder and motion decoder, the system adaptively handles diverse driving tasks and motion patterns. Extensive evaluations on the DAIR-V2X benchmark show that UniMM-V2X achieves state-of-the-art performance, with +39.7% in detection, +77.2% in tracking, -7.2% motion prediction error, -33.2% L2 error, and -52.0% collision rate compared to previous SOTA method, while maintaining comparable communication cost. The framework demonstrates reliability under different bandwidth constraints, highlighting its practical deployability for real-world cooperative driving. Future work will extend UniMM-V2X to closed-loop evaluation, explore more communication-efficient fusion strategies, and further improve the robustness of multi-agent cooperation.

Acknowledgments

This work is sponsored in part by the project of Tsinghua University-Toyota Joint Research Center for AI Technology of Automated Vehicle.

References

Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; and Fu, S. 2019a. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 88–100.

Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2019b. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 514–524. IEEE.

Codevilla, F.; Müller, M.; López, A.; Koltun, V.; and Dosovitskiy, A. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, 4693–4700. IEEE.

Codevilla, F.; Santana, E.; López, A. M.; and Gaidon, A. 2019. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9329–9338.

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.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022a. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

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

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

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

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2024. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.

Liu, Y.-C.; Tian, J.; Ma, C.-Y.; Glaser, N.; Kuo, C.-W.; and Kira, Z. 2020. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 6876–6883. IEEE.

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.

Shazeer, N.; Cheng, Y.; Parmar, N.; Tran, D.; Vaswani, A.; Koanantakool, P.; Hawkins, P.; Lee, H.; Hong, M.; Young, C.; et al. 2018. Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems*, 31.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2024. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*.

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, Z.; Chai, Y.; Jia, X.; Li, Q.; Shao, Y.; Zhu, X.; Su, H.; and Yan, J. 2025. DriveMoE: Mixture-of-Experts for Vision-Language-Action Model in End-to-End Autonomous Driving. arXiv:2505.16278.

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.; Yang, W.; Zhong, J.; Yang, Z.; Fan, S.; Luo, P.; and Nie, Z. 2025. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9598–9606.

Zhang, Z.; Liniger, A.; Dai, D.; Yu, F.; and Van Gool, L. 2021. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15222–15232.

Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; and Fedus, W. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.