DH-VLM: Dual-Horizon Cooperative Latent Reasoning for Autonomous Driving

Ziyi Song[1], Chen Xia[1], Hang Yu[1], Sheng Zhou[1,2⋆], and Zhisheng Niu[1]

[1] Department of Electronic Engineering, Tsinghua University
[2] State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University
{songzy24,xiac23,yuhang22}@mails.tsinghua.edu.cn
sheng.zhou@tsinghua.edu.cn, niuzhs@tsinghua.edu.cn

Abstract. Large-scale language models for autonomous driving enable enhanced global understanding and long-horizon planning. However, when deployed in isolated vehicles, limited sensing range and occlusions restrict reliable decision-making, and the substantial computational and latency overhead makes on-board deployment impractical. Cooperative driving provides a potential solution by leveraging external agents for information exchange, but existing methods remain limited in semantic reasoning capability under practical constraints. To address these challenges, we propose DH-VLM, a dual-horizon cooperative latent reasoning framework that enables asymmetric semantic cooperation between the infrastructure and ego vehicle. The infrastructure aggregates multi-layer hidden states to form a global-reasoning horizon latent guidance, which is integrated into the ego model through an Infrastructure-Driven Latent Evolution mechanism for conditional latent refinement. This enables the ego vehicle to leverage long-range contextual understanding while preserving autonomous decision-making within its local planning horizon. Furthermore, we construct a cooperation-oriented question–answer (QA) dataset covering fundamental scene understanding and ego-personalized comprehension to support counterfactual and safety-aware reasoning. Extensive experiments demonstrate that DH-VLM achieves state-of-the-art planning performance, outperforming the previous state of the art by 14.6% in L2 error and 26.9% in collision rate. Compared with query-based end-to-end cooperative driving methods, our approach reduces the communication cost by 57.3% and GPU memory usage by 25.5%, while maintaining strong robustness against infrastructure guidance errors, providing a practical and robust paradigm for cooperative autonomous driving.

Keywords: Autonomous driving · Vision, language, and reasoning · V2X communication

1 Introduction

Traditional autonomous driving has long been relying on modular pipelines, where perception, prediction, and planning are optimized independently. While

⋆ Corresponding author.

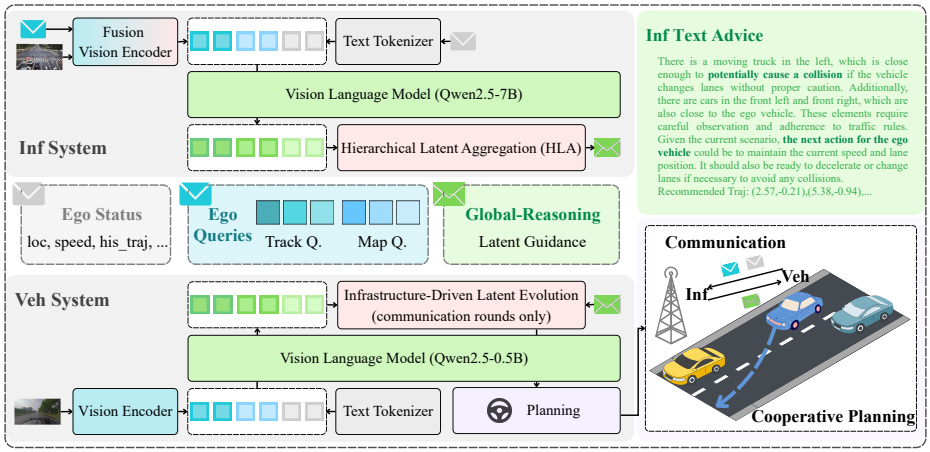


Fig. 1: The overview of DH-VLM. The infrastructure performs global-reasoning horizon understanding by aggregating multi-agent perception queries through a fusion vision encoder, followed by an HLA module to compress multi-layer hidden states into a compact latent cognitive guidance. This information is transmitted via V2X communication to the ego vehicle, where an IDLE module conditionally refines ego hidden states, enabling cooperative latent-level reasoning while preserving local autonomy.

modularity improves interpretability and engineering robustness, it suffers from error accumulation and information bottlenecks across stages. End-to-end autonomous driving addresses these limitations by enabling differentiable learning from raw sensory observations to control commands, demonstrating the benefits of joint optimization across different stages [7, 9, 13, 23]. Nevertheless, these methods remain limited in long-tail scenarios, open-world generalization, and semantic reasoning, as feature maps lack the high-level semantic and causal understanding required for safe decision-making in complex environments.

Vision-Language Models (VLMs) offer strong zero-shot generalization and semantic reasoning capabilities by leveraging massive pretraining corpora, with recent works [17, 21, 25, 27, 31, 39, 42, 43] demonstrating the potential of VLMs for autonomous driving. Beyond explicit action prediction, latent-space reasoning models high-level latent dynamics as an intermediate semantic abstraction between perception and control [6, 24]. Recent works [19, 30] leverage latent-based methods to mitigate compounding errors in autonomous driving, highlighting latent dynamics as a scalable and interpretable foundation for robust autonomous driving systems. However, these approaches typically operate in isolation with limited perception range, increasing safety risks under occlusion or sensor blind spots. In contrast, Vehicle-to-Everything (V2X)-enabled cooperative driving extends perceptual coverage and enhances safety through shared information.

Cooperative autonomous driving has evolved from cooperative perception [2, 3, 8, 14–16, 32, 35] to modular end-to-end frameworks [4, 22, 38] that investigate how cross-agent representation fusion, such as BEV features and queries, influences planning performance. More recent works [5, 36] further incorporate VLMs into cooperative driving. Despite their potential, [36] incurs substantial bandwidth overhead due to raw sensory transmission, while [5] imposes considerable computational demands that may exceed the capabilities of resource-constrained vehicles. These limitations motivate a fundamental question: *how can cooperative autonomous driving systems effectively balance semantic reasoning capacity, communication efficiency under on-board constraints?*

To address these challenges, inspired by dual-system paradigms [18, 25, 30] in single-agent intelligence, we propose DH-VLM, a **Dual-Horizon cooperative latent reasoning VLM architecture** that explicitly decouples global-reasoning horizon understanding from ego local-planning horizon, where a computationally powerful infrastructure provides global semantic guidance to assist a resource-constrained ego vehicle in safety-aware decision-making. Specifically, we introduce a latent-level fusion mechanism, where the infrastructure generates aggregated, semantically structured latent guidance to assist ego decision-making. By performing cooperation at the latent level rather than exchanging raw sensory data or text outputs, DH-VLM enhances semantic expressiveness while reducing communication overhead and preserving decision-making autonomy under on-board computational and temporal constraints.

For latent guidance extraction and aggregation on the infrastructure side, we introduce a *Hierarchical Latent Aggregation (HLA)* module to distill global semantic representations from intermediate VLM hidden states. Motivated by prior findings [10, 11, 20] that Transformer latent spaces encode structured visual–linguistic concepts, we aggregate multi-layer hidden states via attention and interpret the aggregated result as a continuous semantic guidance that captures global-reasoning horizon dependencies and global context. To incorporate this guidance, the ego vehicle adopts a two-stage forward process. It first performs a standard forward pass on its observations to obtain an initial latent state. When the infrastructure guidance is available, the initial latent state is refined via *Infrastructure-Driven Latent Evolution (IDLE)*, followed by a second forward pass based on the evolved representation. If the guidance is unavailable due to asynchronous updates or communication packet loss, the ego model relies on the initial pass alone. This design preserves decision autonomy while enabling cooperative semantic conditioning and naturally accommodates the asynchronous update rates inherent in the dual-system architecture.

Furthermore, we construct a cooperation-oriented QA dataset built upon existing cooperative driving benchmark [37]. In contrast to generic scene-level QA formulations, our dataset is explicitly tailored for ego-centric reasoning, facilitating guidance that is conditioned on the ego vehicle's state and driving intent. It covers both fundamental semantic understanding (e.g., scene description, traffic regulations, and object presence) and ego-personalized comprehension (e.g., in-view ego masking, counterfactual reasoning, and safety-aware planning). By

integrating structured QA annotations into multi-agent observations, the dataset systematically enhances semantic modeling capacity and promotes safety-aware scene reasoning for cooperative driving.

Our main contributions are summarized as follows:

- We propose DH-VLM, a dual-horizon cooperative latent reasoning framework, where a high-capacity infrastructure assists a resource-limited ego vehicle, enhancing decision quality under on-board efficiency constraints.
- To the best of our knowledge, we are the first to explore latent-level fusion in VLM-based cooperative autonomous driving, enabling cooperative semantic reasoning without compromising ego autonomy.
- We construct a cooperation-oriented QA dataset to facilitate cooperative reasoning. Extensive experiments show that DH-VLM achieves SOTA planning performance with efficient communication and low computational cost.

2 Related Works

2.1 End-to-End Autonomous Driving

Early end-to-end methods [29, 41] map raw sensor data directly to final control via imitation learning. To improve interpretability, [7] jointly models perception, prediction, and planning in a transformer-based architecture; [9] introduces vectorized representations to improve planning consistency; [23] employs sparse query mechanisms to reduce computational cost. However, these methods still struggle with long-tail scenarios and lack semantic reasoning. VLM-based approaches introduce stronger semantic understanding and reasoning processes into planning via large-scale pretraining [27, 31, 42, 43]. Dual-system designs further decouple high-level reasoning from on-board driving [18,25,30,34]. However, these methods remain limited by sensing ranges and occlusion, necessitating the cooperative driving frameworks that integrate multi-agent information exchange to achieve complementary reasoning.

2.2 Cooperative Autonomous Driving

Cooperative autonomous driving leverages V2X communication to overcome the perceptual limits of individual vehicles. Early works focus on cooperative perception [2,8,14–16,28,32,33,35], designing fusion strategies to balance detection accuracy and communication cost. However, perception-oriented objectives are often misaligned with planning [40]. End-to-end cooperative frameworks address this by directly mapping multi-agent observations to ego control via BEV or query fusion [4, 22, 38]. Recent VLM-based methods [5, 36] introduce reasoning into cooperation. However, result-level fusion is sensitive to upstream errors and introduces heavy on-board computational burden [5], while early-level fusion incurs substantial communication overhead [36]. We propose a dual-horizon cooperative framework that assigns global-horizon reasoning to the infrastructure while preserving decision autonomy at the vehicle, enabling semantic guidance through lightweight latent integration.

2.3 Latent Representation and Cooperation

Latent representations have emerged as a compact yet semantically dense medium for reasoning and communication in foundation models. The latent transition paradigm, introduced in [6], enables continuous reasoning directly within the latent space, bypassing the need for discrete token decoding. In autonomous driving, a unified latent space for vision–action alignment has been explored to avoid auto-regressive language generation [19]. Building upon this direction, latent representations are further compressed via knowledge distillation to facilitate on-board deployment [30]. Beyond single-agent systems, recent robotics study on latent-level cooperation [44] explores multi-agent coordination through the final layer of hidden states exchange rather than raw data or text to facilitate efficient cooperation. Motivated by these advances, our dual-system framework transmits compact latent-level information from infrastructure to the ego vehicle, avoiding the error propagation of text-based reasoning and the bandwidth overhead of raw data sharing, while supporting on-board execution.

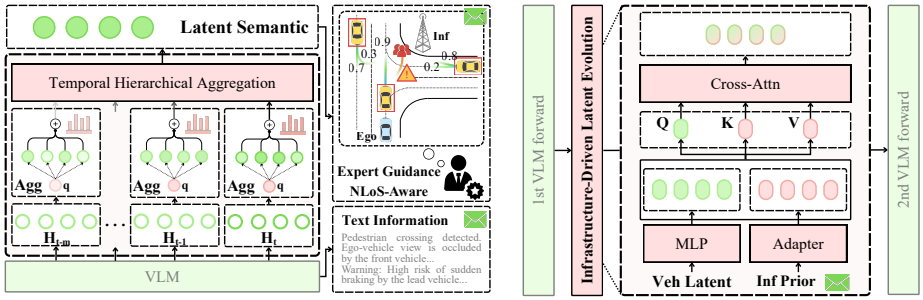
3 Method

We introduce a dual-horizon cooperative latent reasoning system tailored for vehicle-infrastructure autonomous driving. To fully leverage the capabilities of vision–language models and explore efficient information exchange in this emerging paradigm, we study latent-level fusion between the infrastructure and ego systems (Sec. 3.2 and Sec. 3.3). To support our framework, we further construct a cooperation-oriented QA dataset (Sec. 3.4). Additional details, including the Fusion Vision Encoder module and the VLM processing pipeline, are provided in the Appendix. Our approach integrates global semantic guidance while preserving ego autonomy, enabling robust and efficient cooperative driving.

3.1 Overview: Dual-Horizon Cooperative Driving

In DH-VLM, the vehicle performs ego-centric planning and control for on-board decision-making, while the infrastructure provides high-level, global-reasoning horizon semantic guidance under relaxed latency constraints. Instead of directly affecting control outputs, the infrastructure information is encoded as latent guidance and fused into the ego model at the latent level. This forms an asymmetric and loosely coupled cooperation paradigm in which the ego vehicle retains execution authority and the infrastructure provides semantic guidance.

Infrastructure System for Global-Reasoning Horizon. The infrastructure focuses on global reasoning over complex traffic evolution. It aggregates its perception queries with transmitted ego queries via the Fusion Vision Encoder module, producing fused perception tokens Q_A and Q_L. A unified 3D coordinate space is adopted to ensure spatial alignment with the ego vehicle. The fused tokens are processed by a large language model to generate: (i) semantic latent guidance z_t^{inf} encoding global traffic context and driving intent through an HLA module; and (ii) text-based reasoning and advice T_t^{inf} for interpretability.



(a) Hierarchical Latent Aggregation for Guidance Generation (b) Infrastructure-Driven Latent Evolution

Fig. 2: Latent-level interaction between the infrastructure and vehicle systems. (a) illustrates the hierarchical latent guidance generation in the infrastructure system, while (b) depicts how this guidance is utilized in the vehicle system via the proposed IDLE module with an attention-based mechanism.

Vehicle System for Local-Planning Horizon. The ego vehicle conducts continuous self-contained decision making architecture comprising a vision encoder, a lightweight language model, and an IDLE module. The vision encoder follows the same architecture as the infrastructure encoder but omits fusion components, decoupling dual-system interactions. The lightweight language model processes vision and text tokens for understanding and planning. When infrastructure latent guidance is available, the IDLE module performs conditional latent refinement; otherwise, the vehicle system follows its native pipeline, ensuring uninterrupted autonomous operation.

3.2 Hierarchical Latent Aggregation for Infrastructure Guidance

Large language models encode hierarchical semantic abstractions of complex driving scenes through layered hidden representations. Prior studies indicate that intermediate Transformer layers capture rich contextual, structural, and causal semantics, whereas the final layer is primarily optimized for token prediction and often compresses representations into compact decision-oriented embeddings. To provide the ego vehicle with high-level semantic and causal guidance rather than deterministic decisions, we extract latent representations from intermediate reasoning layers of the infrastructure VLM, as shown in Fig. 2(a).

Formally, let the VLM consist of L Transformer layers with hidden states. We select a subset of intermediate layers $\mathcal{S} \subset \{1, \ldots, L-1\}$ and treat their hidden states as a latent reasoning manifold:

$$\mathbf{Z}^{\text{inf}} = \{\mathbf{h}_l \mid l \in \mathcal{S}\}. \tag{1}$$

To obtain a compact and robust infrastructure guidance, we propose an HLA module that consolidates hidden states across hierarchical dimensions. For each

selected layer $l \in \mathcal{S}$, we perform attention-based semantic aggregation to summarize the latent states into a compact representation:

$$\mathrm{Agg}(\mathbf{h}_l^{(t)}) = \sum_{i=1}^{N} \beta_{l,i}^{(t)} \mathbf{h}_{l,i}^{(t)}, \quad \beta_{l,i}^{(t)} = \frac{\exp(\mathbf{q}^\top \mathbf{h}_{l,i}^{(t)})}{\sum_{j=1}^{N} \exp(\mathbf{q}^\top \mathbf{h}_{l,j}^{(t)})}, \tag{2}$$

where $\mathbf{q}$ is a learnable parameter that selects salient tokens for global reasoning. We then aggregate across layers via learnable layer-wise weights:

$$\tilde{\mathbf{z}}_t^{\mathrm{inf}} = \sum_{l \in \mathcal{S}} \alpha_l \, \mathrm{Agg}(\mathbf{h}_l^{(t)}), \quad \alpha_l = \frac{\exp(w_l)}{\sum_{k \in \mathcal{S}} \exp(w_k)}, \tag{3}$$

where w_l is a learnable parameter indicating the relative importance of layer l, allowing the model to adaptively weight semantic representations from different reasoning depths. To reduce temporal fluctuations and enhance stability, we further aggregate the hierarchical representations within a temporal window $\mathcal{T}$:

$$\mathbf{z}_t^{\mathrm{inf}} = \mathcal{G}\left(\{\tilde{\mathbf{z}}_{t-\tau}^{\mathrm{inf}}\}_{\tau \in \mathcal{T}}\right). \tag{4}$$

This produces a temporally consistent latent representation that summarizes the infrastructure's global-reasoning horizon understanding of the scene. The resulting $\mathbf{z}_t^{\mathrm{inf}}$ is transmitted to the ego vehicle as a compact high-level cognitive latent guidance for cooperative reasoning and planning.

3.3 Infrastructure-Driven Latent Evolution on the Vehicle

We introduce the IDLE module that injects infrastructure latent guidance into the ego reasoning, enabling global scene understanding of occluded risks and distant traffic intent, while maintaining local autonomy and robustness.

As illustrated in Fig. 2(b), at each time step t, the ego vehicle first extracts its semantic tokens $\tilde{Q}_t^{\mathrm{veh}}$ by fusing on-board visual tokens Q_t^{veh} and textual instructions T_t through the first forward pass of the ego VLM:

$$\tilde{Q}_t^{\mathrm{veh}} = \mathcal{F}_{\mathrm{VLM}}^{\mathrm{veh}}\big(Q_t^{\mathrm{veh}}, T_t\big). \tag{5}$$

Due to heterogeneous capacities and training objectives, the infrastructure and vehicle models produce misaligned latent representations, with discrepancies in feature statistics and semantic embeddings. Direct fusion under such misalignment can introduce noise and degrade cooperative reasoning. To mitigate this issue, we employ a lightweight MLP-based adapter to project the infrastructure latent guidance into the vehicle latent space before fusion:

$$\bar{z}_t^{\mathrm{inf}} = \mathrm{MLP}(z_t^{\mathrm{inf}}). \tag{6}$$

After alignment, an attention-based fusion mechanism then enables knowledge-aware latent integration:

$$\tilde{z}_t^{\mathrm{veh}} = \mathrm{CrossAttn}\big(z_t^{\mathrm{veh}}, \bar{z}_t^{\mathrm{inf}}\big), \quad z_t^{\mathrm{veh}} = \mathrm{MLP}(\tilde{Q}_t^{\mathrm{veh}}), \tag{7}$$

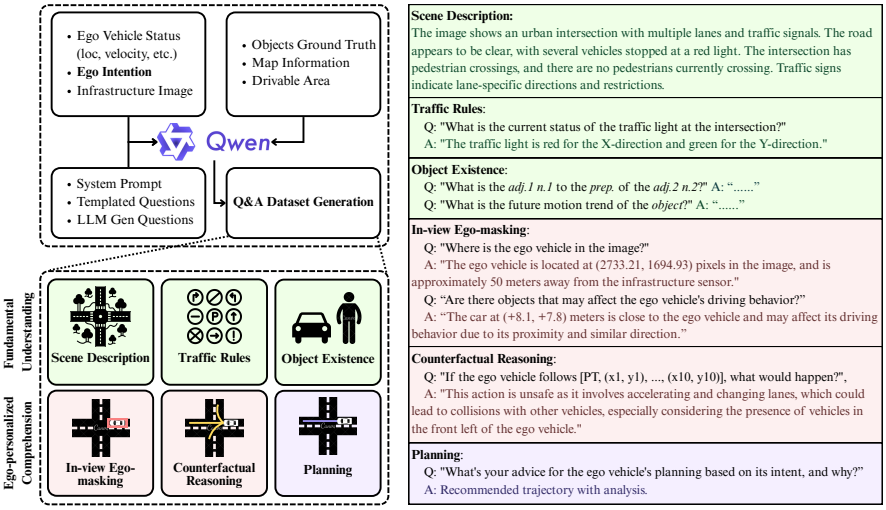


Fig. 3: Construction and contents of the cooperation-oriented dataset.

where z_t^{veh} denotes the ego latent representation derived from vehicle semantic tokens. Through cross-attention, the ego vehicle selectively incorporates complementary global information, producing an enhanced cooperative semantic state $\tilde{z}_t^{\text{veh}}$ while remaining grounded in ego perception.

To further integrate the cooperative information into the reasoning process, we perform a recursive refinement via a second forward pass of the vehicle VLM:

$$\hat{z}_t^{\text{veh}} = \mathcal{F}_{\text{VLM}}^{\text{veh}}\big(\tilde{z}_t^{\text{veh}}, \mathcal{C}_t\big), \tag{8}$$

where $\mathcal{C}_t = \{Q_t^{\text{veh}}, T_t\}$. This refinement step embeds the fused latent into the ego model's internal reasoning dynamics, producing a cooperative semantic representation that reflects both global infrastructure guidance and local contextual understanding. $\hat{z}_t^{\text{veh}}$ is then utilized for downstream planning and trajectory generation, enabling informed and self-contained decision-making.

3.4 Cooperation-Oriented QA Dataset

To support and validate our dual-horizon cooperative framework, we introduce a cooperation-oriented QA benchmark that bridges the gap between sensor annotations and cooperative semantic reasoning. Specifically, this benchmark enhances the infrastructure's semantic understanding capability, facilitating safer and more robust latent guidance for the ego vehicle. As illustrated in Fig. 3, the dataset is generated through a two-stage pipeline with human-in-the-loop verification to ensure annotation quality and reasoning consistency.

Fundamental Scene Understanding. At this stage, we generate objective global scene descriptions that characterize traffic topology, lane constraints, and

the spatial distribution and motion trends of surrounding agents. These annotations provide a consistent global context that facilitates downstream cooperative perception and decision-making.

Ego-Personalized Comprehension. Beyond conventional QA datasets for single-intelligence autonomous driving, we explicitly model the auxiliary and strategic role of infrastructure intelligence in cooperative scenarios. Based on the fundamental stage, we align infrastructure reasoning with the ego vehicle by projecting the ego's global pose into the infrastructure coordinate system, enabling explicit ego grounding in the global scene. Conditioned on this grounding, the infrastructure performs counterfactual reasoning over multiple what-if future trajectories generated under lane constraints and traffic rules to assess driving safety, interaction risks, and global traffic dynamics beyond the ego perception and prediction. Based on this information, the infrastructure outputs future-oriented safety assessments and planning-level strategic suggestions, such as potential collision risks, interaction conflicts, and recommended maneuver intentions. This ego-personalized reasoning further strengthens latent semantic guidance to support vehicle planning.

4 Experiments

4.1 Experimental Setup

The infrastructure system employs Qwen2.5-7B [1] to perform global and long-horizon semantic reasoning, while the vehicle system adopts Qwen2.5-0.5B [1] to meet on-board deployment constraints. The communication ratio is set to 5:1, determined empirically based on computational time. Inference is conducted on NVIDIA GeForce RTX 3090 GPUs, and training is performed in multiple stages on NVIDIA A800 GPUs. Training consists of three stages. First, the infrastructure model is trained for scene understanding and ego-centric comprehension using visual data and our QA dataset with a learning rate of 1×10^{-5}. Second, the ego model is trained using ego visual data with a learning rate of 1×10^{-4}. Finally, the two systems are jointly optimized by training the HLA and IDLE for latent-level cooperation. All models are optimized using AdamW with a weight decay of 0.03. We select $|\mathcal{S}| = 8$ intermediate layers for latent aggregation and set the temporal window size $k = 5$.

The cooperation-oriented QA dataset is constructed on DAIR-V2X [37], a large-scale multi-modal benchmark for vehicle–infrastructure cooperative driving, containing approximately 100 scenes across 28 complex urban intersections. The question–answer pairs are generated using Qwen2.5-VL-32B [26], enabling semantically grounded supervision tailored to cooperative reasoning scenarios. The final dataset consists of over 80k QA pairs covering diverse cooperative driving scenarios. We further validate our framework on V2X-Sim [12], a simulated benchmark for cooperative autonomous driving. Additional implementation details and results are provided in the Appendix.

Table 1: Planning performance and safety evaluation.

Method	L2 Error (m)↓						Collision Rate (%)↓					
	$1s$	$2s$	$3s$	$4s$	$5s$	Avg.	$1s$	$2s$	$3s$	$4s$	$5s$	Avg.
Single-Intelligence Methods												
VAD [9]	1.65	2.72	3.80	4.65	5.78	3.72	0.86	1.21	1.28	1.46	1.95	1.35
UniAD [7]	1.26	2.22	3.06	3.97	5.22	3.15	0.88	1.18	1.32	1.04	1.45	1.17
SparseDrive [23]	1.02	1.69	2.37	2.99	3.87	2.39	0.46	1.23	1.28	1.34	1.58	1.18
OpenDriveVLA [42]	0.64	1.51	2.69	4.05	5.57	2.89	0.15	0.59	0.59	0.89	**0.44**	0.53
Multi-Agent Cooperative Methods												
V2VNet [28]	1.96	2.37	3.41	3.91	4.73	3.28	0.74	0.88	1.03	1.24	1.58	1.09
CooperNaut [4]	2.69	4.07	5.50	6.02	7.94	5.24	1.18	1.32	1.76	1.97	1.69	1.58
UniV2X [38]	1.45	2.19	3.04	3.96	5.24	3.18	0.15	0.15	0.44	0.59	0.74	0.41
UniMM-V2X [22]	0.78	1.63	2.05	3.54	5.00	2.60	**0.05**	0.15	0.15	0.59	1.18	0.42
V2X-VLM [36]	0.49	1.57	2.89	4.08	5.34	2.87	**0.05**	**0.05**	0.15	0.35	0.68	0.26
LangCoop [5]	0.30	0.93	1.96	3.17	4.58	2.19	**0.05**	0.15	0.44	0.87	1.02	0.51
DH-VLM (Ours)	**0.26**	**0.85**	**1.76**	**2.74**	**3.72**	**1.87**	**0.05**	**0.05**	**0.11**	**0.28**	**0.44**	**0.19**

Table 2: On-board deployment, inference, and communication cost analysis.

Method	#Params (M)↓	Memory (GB)↓	Trans. Cost (BPS)↓
V2VNet [28]	**262.1**	6.09	8.19×10^{7}
CooperNaut [4]	**262.1**	6.09	8.19×10^{7}
UniV2X [38]	264.2	6.11	8.09×10^{5}
UniMM-V2X [22]	267.8	6.14	9.32×10^{5}
V2X-VLM [36]	776.5	10.92	1.24×10^{7}
LangCoop [5]	7384.5	17.74	$\sim$ **4×10^{3}**
DH-VLM (Ours)	640.6	**4.54**	3.45×10^{5}

4.2 Main Results

The planning results are reported in Tab. 1. We compare DH-VLM with single-agent baselines [7,9,23,42] and cooperative methods [4,5,22,28,36,38]. DH-VLM achieves the best performance in both L2 error and collision rate. It reduces L2 error by **14.6%** over [5] and lowers collision rate by **26.9%** compared with [36], highlighting the benefit of latent-level cooperation. The improvement is more significant in long-horizon planning: relative to [38], DH-VLM reduces L2 error by **29.0%** and collision rate by **49.5%**, highlighting the critical role of infrastructure-guided global reasoning in long-term decision making.

To assess deployment cost and communication efficiency, we compare model size, inference cost, and transmission cost across cooperative methods. Although DH-VLM does not reduce total parameter count compared with modular end-to-end frameworks [4, 28, 38], it lowers GPU memory usage by **25.5%**, as runtime memory is dominated by intermediate activations rather than parameters. By replacing dense feature interaction with compact latent guidance, DH-VLM reduces activation and attention overhead, resulting in a reduced on-board inference footprint that better aligns with efficiency-constrained vehicles. Communication is required only at cooperation frames, where compact latent guid-

Table 3: Ablation study on the infrastructure system for ego-centric reasoning.

Fusion in Vision Encoder	QA Level F	QA Level E	L2 Error (m)↓ 3s	4s	5s	Avg.	Collision Rate (%)↓ 3s	4s	5s	Avg.
-	-	-	3.84	4.97	6.77	5.19	2.07	2.07	1.04	1.73
✓	-	-	2.37	3.86	5.52	3.92	0.59	1.19	1.30	1.03
✓	✓	-	1.58	2.46	3.53	2.52	**0.30**	0.74	0.89	0.64
✓	✓	✓	**0.99**	**1.67**	**2.51**	**1.72**	0.44	**0.44**	**0.44**	**0.44**

Note: "F" and "E" denote fundamental scene understanding and ego-personalized comprehension component of our cooperation-oriented QA dataset, respectively. Results are directly evaluated from infrastructure-generated text outputs.

Table 4: Robustness of ego-vehicle planning under noisy infrastructure guidance.

w./o.	Level T	Level L	Inf Per	L2 Error (m)↓ 1s	2s	3s	4s	5s	Avg.	Collision Rate (%)↓ 1s	2s	3s	4s	5s	Avg.
-	-	-	-	0.64	1.51	2.69	4.05	5.57	2.89	0.15	0.59	0.59	0.89	**0.44**	0.53
✓	✓	-	●	**0.19**	**0.73**	**1.59**	**2.66**	3.98	**1.83**	**0.05**	**0.05**	**0.05**	**0.25**	0.45	**0.17**
✓	✓	-	○	0.44	1.26	2.31	3.42	4.98	2.48	0.15	0.25	0.34	0.69	1.27	0.54
✓	✓	-	△	2.26	3.84	5.63	7.40	9.15	5.66	0.43	0.75	1.84	3.87	4.99	2.38
✓	-	✓	●	0.26	0.85	1.76	2.74	**3.72**	1.87	**0.05**	**0.05**	0.11	0.28	**0.44**	0.19
✓	-	✓	○	0.24	0.94	1.82	2.91	3.99	1.98	0.05	0.15	0.24	0.36	0.58	0.28
✓	-	✓	△	0.26	0.95	2.07	3.17	4.50	2.19	0.15	0.25	0.57	0.79	1.02	0.56

Note: "T" and "L" denote text-level and latent-level guidance, respectively.
●, ○, and △ indicate average 5s L2 error of infrastructure-generated trajectories (< 2m, $2 \sim 4$m, > 4m), indicating the infrastructure's latent reasoning capability for the ego vehicle.

ance is transmitted, leading to a **57.3%** reduction compared with the query-based method [38]. A detailed communication cost analysis is provided in the Appendix. While [5] achieves lower bandwidth by sending natural language summaries, it is more sensitive to inaccurate cross-agent information. In contrast, latent-level communication preserves structured semantic guidance with improved robustness, as validated in Sec. 4.3.

4.3 Ablation Study

Performance of the Infrastructure in Assisting the Vehicle. As shown in Tab. 3, the Fusion Vision Encoder is essential for holistic scene understanding at the infrastructure side. Since the ego vehicle may lie partially or entirely outside the infrastructure's field of view, ego-centric reasoning is infeasible without multi-agent feature alignment and fusion. Without this fusion, the infrastructure often produces incomplete or inaccurate guidance, which can mislead ego planning. We further examine the effect of different components in the cooperation-oriented QA dataset. The Fundamental Scene Understanding data provides global traffic context and spatial structure, while the Ego-Personalized Comprehension data explicitly connects global perception to ego-centric decision-making and long-horizon safety reasoning. These findings indicate that both multi-agent visual fusion and structured cooperative QA supervision are critical for reliable global reasoning and high-quality infrastructure guidance.

Table 5: Impact of different communication ratios and the number of selected intermediate layers in HLA for latent aggregation on driving performance.

Comm. Ratio	# Fused Layers†	L2 Error (m)↓						Collision Rate (%)↓					
		1s	2s	3s	4s	5s	Avg.	1s	2s	3s	4s	5s	Avg.
5:1	8	0.26	0.85	1.76	2.74	**3.72**	1.87	0.05	**0.05**	0.11	0.28	0.44	0.19
5:1	16	0.28	0.84	1.79	2.75	3.74	1.88	0.05	**0.05**	0.15	0.35	0.55	0.23
5:1	4	0.33	0.91	1.86	2.83	3.73	1.93	0.05	0.15	0.15	0.35	0.60	0.26
5:1	1	0.45	0.97	2.01	2.97	3.91	2.06	0.15	0.15	0.25	0.45	0.58	0.32
10:1	8	0.36	1.02	2.18	2.95	4.56	2.21	0.05	0.15	0.35	0.35	0.53	0.29
10:1	1	0.53	1.25	2.35	3.87	4.92	2.58	0.15	0.35	0.48	0.55	0.75	0.46
1:1*	8	**0.23**	**0.79**	**1.68**	**2.57**	3.73	**1.80**	**0.03**	**0.05**	**0.05**	**0.19**	**0.33**	**0.13**
1:1*	1	0.35	0.98	1.97	2.95	3.87	2.02	0.10	0.10	0.25	0.28	0.67	0.28

† Fused layer number = 1 denotes using the final hidden layer directly.
* Theoretical setting beyond practical deployment constraints.

Table 6: Ablation study on module effectiveness. 'w/o' denotes replacing the corresponding module with an MLP-based variant.

Method	L2 Avg. (m)↓	CR Avg. (%)↓
DH-VLM (Full)	**1.87**	**0.19**
w/o HLA	2.31	0.39
w/o IDLE	2.48	0.34
w/o both	2.59	0.47

Robustness of Infrastructure Guidance. As shown in Tab. 4, although the infrastructure possesses stronger global reasoning capability, its text-based guidance T_t^{inf} inevitably contains noise due to long-horizon uncertainty and asynchronous sensing. Direct injection propagates these errors to the ego planning process, sometimes degrading performance beyond single-agent baselines (+95.8% L2 error, +349.1% collision rate), indicating that naive cooperation can be harmful. In contrast, our latent-level fusion treats infrastructure guidance as a soft prior, allowing the ego model to selectively absorb useful semantic cues while suppressing unreliable signals. Notably, it achieves performance comparable to the best text-level fusion, with improved robustness and safety.

Communication Frequency and Layer Selection. As shown in Tab. 5, a 1:1 communication ratio yields the best performance, but it is impractical for deployment. Aligning with the inference speeds of the systems, a 5:1 ratio achieves a better balance between communication cost and performance, while lower frequencies under-utilize infrastructure assistance. For latent aggregation, increasing the number of intermediate layers brings marginal gains due to redundancy, whereas too few layers provide insufficient semantic richness. Notably, transmitting only the final layer causes significant semantic degradation, as it is already close to text decoding and diverges from the visual semantic space, resulting in substantial information loss.

Component Analysis. We analyze HLA and IDLE by replacing each with a simple MLP while keeping the remaining architecture unchanged. As shown in

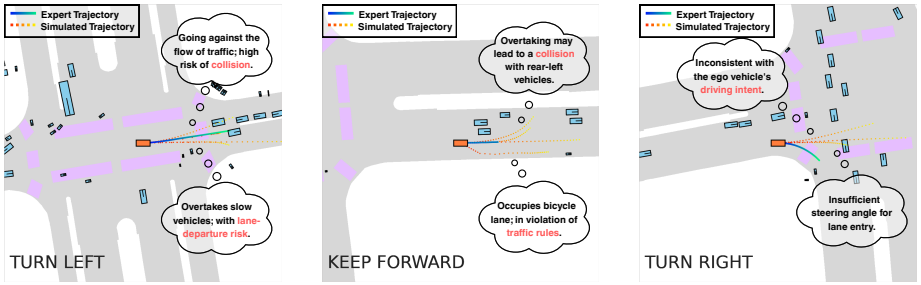


Fig. 4: Counterfactual reasoning conducted by the infrastructure.

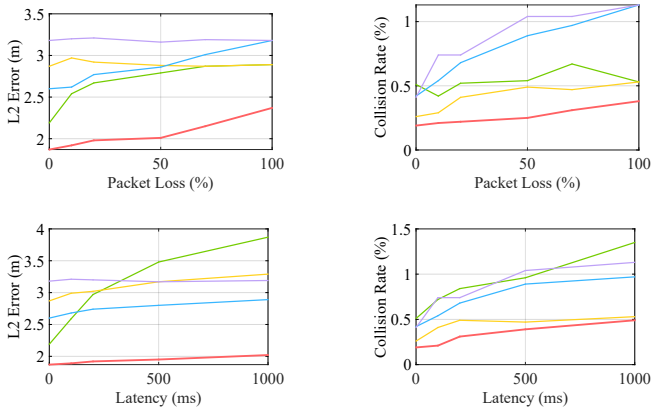


Fig. 5: Communication robustness experiments. Methods: DH-VLM (ours), LangCoop, V2X-VLM, UniMM-V2X, UniV2X.

Tab. 6, replacing HLA increases the L2 error by 23.5% and doubles the collision rate, highlighting the role of historical latent aggregation in robust cooperative planning. Replacing IDLE increases the L2 error by 32.6% and the collision rate by 78.9%, demonstrating the importance of infrastructure-guided latent refinement for accurate trajectory generation. These observations highlight the complementary roles of HLA and IDLE and attribute the performance gains to the proposed latent guidance instead of increased model complexity.

4.4 System Stability and Degraded Communication Analysis

To evaluate robustness under realistic communication conditions, we analyze DH-VLM under varying packet loss rates and communication latencies. As shown in Fig. 5, DH-VLM maintains stable planning performance across a wide range of communication degradations. This robustness is enabled by asynchronous communication and temporal latent compensation. Specifically, DH-VLM adopts a

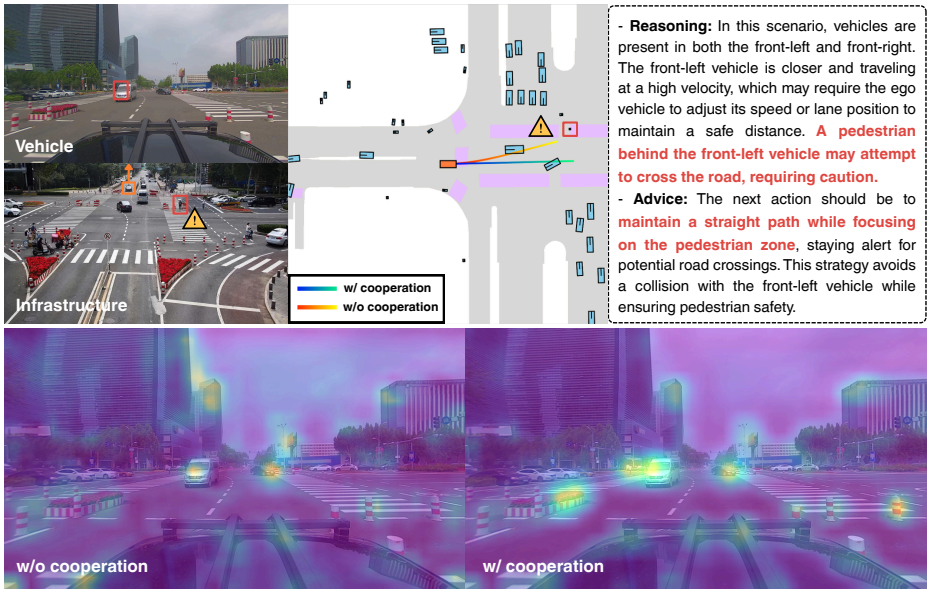


Fig. 6: Cooperative driving results and latent-derived spatial activations of DH-VLM, with infrastructure textual reasoning shown for interpretability.

sparse 5:1 communication ratio, reducing bandwidth requirements and relaxing synchronization constraints. Instead of transmitting raw sensor data [36], intermediate features [22, 38], or text-based messages [5], latent-level communication exchanges compact semantic representations, improving resilience to communication impairments while mitigating performance degradation and safety risks, further validating the infrastructure guidance robustness analysis in Sec. 4.3. When infrastructure guidance is delayed or unavailable, e.g., under 100% packet loss or latencies exceeding 500ms, the ego vehicle operates independently using its onboard model, enabling stable asynchronous cooperation.

4.5 Qualitative Visualization

Ego-Personalized Comprehension. To qualitatively evaluate the proposed infrastructure-assisted reasoning mechanism, we visualize the ego-centric counterfactual reasoning process in Fig. 4. In left-turn, forward-driving, and right-turn scenarios, the infrastructure generates multiple candidate trajectories and selects the safest one conditioned on the intended maneuver of the ego vehicle. The chosen trajectories consistently follow traffic rules and avoid potential collision regions. These observations suggest that the counterfactual reasoning module enables intention-aware safety assessment beyond reactive collision avoidance. By reasoning over alternative futures conditioned on ego intent, the

infrastructure gains a clearer understanding of complex traffic interactions and potential risks, thereby supporting more reliable cooperative planning decisions.

Latent Evolution and Planning under Cooperation. Fig. 6 presents qualitative planning results of DH-VLM, along with infrastructure-generated textual outputs for interpretability. To analyze how infrastructure guidance affects ego latent evolution, we also visualize spatial activation maps by projecting ego latent representations through a lightweight perception head. In heavily occluded intersections, the ego-only model may fail to detect pedestrians hidden behind foreground vehicles. With infrastructure support, the latent representations encode a more complete and globally consistent scene understanding, enabling the ego vehicle to anticipate pedestrian motion and adjust its trajectory. The spatial activation maps show more concentrated and semantically structured responses around critical regions, particularly near occluded pedestrians and interaction areas, suggesting that the latent space captures safety-critical contextual cues that are difficult to infer from ego-only perception.

5 Conclusion

This paper presents DH-VLM, a dual-horizon cooperative driving architecture that assigns global-reasoning horizon understanding to the infrastructure while preserving local-planning horizon at the vehicle. To bridge heterogeneous computational capacities, DH-VLM enables latent-level cooperation, allowing the infrastructure to provide semantic guidance with high communication efficiency and robustness while minimizing the impact of communication noise and imperfect infrastructure guidance on ego vehicle's driving decisions. In addition, we construct a cooperation-oriented QA dataset, comprising fundamental understanding and ego-personalized comprehension tasks to explicitly enhance the system's ability to reason for the ego vehicle. Extensive experiments demonstrate that DH-VLM achieves state-of-the-art performance. With a 57.3% reduction in communication bandwidth, our method reduces the 5s L2 planning error by 14.6% and the collision rate by 26.9%. Notably, the ego vehicle only requires a lightweight onboard model, significantly improving the practical deployability of cooperative autonomous driving systems. However, due to the lack of standardized closed-loop evaluation methods for cooperative driving, our current experiments are limited to open-loop evaluation. Moreover, the present framework remains within the VLM paradigm. Future work will extend the approach toward a VLA-based cooperative architecture.

Acknowledgements

This work is supported in part by the State Key Laboratory of Internet of Things for Smart City (University of Macau) Open Research Project under Grant SKL-IoTSC(UM)/ORP04/2026, and in part by the Project of Tsinghua University-Toyota Joint Research Center for AI Technology of Automated Vehicle under Grant TTAD-2025-08.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: 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. pp. 88–100 (2019)
3. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524. IEEE (2019)
4. Cui, J., Qiu, H., Chen, D., Stone, P., Zhu, Y.: Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17252–17262 (2022)
5. Gao, X., Wu, Y., Wang, R., Liu, C., Zhou, Y., Tu, Z.: Langcoop: Collaborative driving with language. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4226–4237 (2025)
6. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)
7. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
8. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
9. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
10. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
11. Li, K., Hopkins, A.K., Bau, D., Viégas, F., Pfister, H., Wattenberg, M.: Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382 (2022)
12. Li, Y., Ma, D., An, Z., Wang, Z., Zhong, Y., Chen, S., Feng, C.: V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. IEEE Robotics and Automation Letters **7**(4), 10914–10921 (2022)
13. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
14. Liu, Y.C., Tian, J., Glaser, N., Kira, Z.: When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4106–4115 (2020)
15. Liu, Y.C., Tian, J., Ma, C.Y., Glaser, N., Kuo, C.W., Kira, Z.: Who2com: Collaborative perception via learnable handshake communication. In: 2020 IEEE Inter-

national Conference on Robotics and Automation (ICRA). pp. 6876–6883. IEEE (2020)
16. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023)
17. Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415 (2023)
18. Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., Ren, L.: Vlp: Vision language planning for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14760–14769 (2024)
19. Peng, Q., Chen, X., Yang, C., Shi, S., Li, H.: Colavla: Leveraging cognitive latent reasoning for hierarchical parallel trajectory planning in autonomous driving. arXiv preprint arXiv:2512.22939 (2025)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024)
22. Song, Z., Xia, C., Wang, C., Yu, H., Zhou, S., Niu, Z.: Unimm-v2x: Moe-enhanced multi-level fusion for end-to-end cooperative autonomous driving. arXiv preprint arXiv:2511.09013 (2025)
23. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. arXiv preprint arXiv:2405.19620 (2024)
24. Teoh, J., Tomar, M., Ahn, K., Hu, E.S., Sharma, P., Islam, R., Lamb, A., Langford, J.: Next-latent prediction transformers learn compact world models. arXiv preprint arXiv:2511.05963 (2025)
25. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
26. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
27. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. CoRR (2024)
28. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: 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. pp. 605–621. Springer (2020)
29. Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. Advances in Neural Information Processing Systems **35**, 6119–6132 (2022)
30. Xie, C., Sun, B., Li, T., Wu, J., Hao, Z., Lang, X., Li, H.: Latentvla: Efficient vision-language models for autonomous driving via latent action prediction. arXiv preprint arXiv:2601.05611 (2026)

31. Xing, S., Qian, C., Wang, Y., Hua, H., Tian, K., Zhou, Y., Tu, Z.: Openemma: Open-source multimodal model for end-to-end autonomous driving. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1001–1009 (2025)
32. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022)
33. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
34. Xu, Y., Hu, Y., Zhang, Z., Meyer, G.P., Mustikovela, S.K., Srinivasa, S., Wolff, E.M., Huang, X.: Vlm-ad: End-to-end autonomous driving through vision-language model supervision. arXiv preprint arXiv:2412.14446 (2024)
35. Yang, D., Yang, K., Wang, Y., Liu, J., Xu, Z., Yin, R., Zhai, P., Zhang, L.: How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. Advances in Neural Information Processing Systems **36**, 25151–25164 (2023)
36. You, J., Shi, H., Jiang, Z., Huang, Z., Gan, R., Wu, K., Cheng, X., Li, X., Ran, B.: V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. arXiv preprint arXiv:2408.09251 (2024)
37. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21361–21370 (2022)
38. Yu, H., Yang, W., Zhong, J., Yang, Z., Fan, S., Luo, P., Nie, Z.: End-to-end autonomous driving through v2x cooperation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9598–9606 (2025)
39. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X., Guo, N.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
40. Zeng, W., Luo, W., Suo, S., Sadat, A., Yang, B., Casas, S., Urtasun, R.: End-to-end interpretable neural motion planner. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8660–8669 (2019)
41. Zhang, Z., Liniger, A., Dai, D., Yu, F., Van Gool, L.: End-to-end urban driving by imitating a reinforcement learning coach. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15222–15232 (2021)
42. Zhou, X., Han, X., Yang, F., Ma, Y., Tresp, V., Knoll, A.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. arXiv preprint arXiv:2503.23463 (2025)
43. Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J.: Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. arXiv preprint arXiv:2506.13757 (2025)
44. Zou, J., Yang, X., Qiu, R., Li, G., Tieu, K., Lu, P., Shen, K., Tong, H., Choi, Y., He, J., et al.: Latent collaboration in multi-agent systems. arXiv preprint arXiv:2511.20639 (2025)