V2V-LLM: Vehicle-to-Vehicle Cooperative Autonomous Driving with Multi-Modal Large Language Models

Hsu-kuang Chiu[1,2] Ryo Hachiuma[1] Chien-Yi Wang[1] Stephen F. Smith[2] Yu-Chiang Frank Wang[1]
Min-Hung Chen[1]

[1]NVIDIA, [2]Carnegie Mellon University

1. Introduction

Cooperative perception algorithms [4, 6, 7, 10–12, 14, 18, 24, 28, 31–33, 36] via vehicle-to-vehicle (V2V) or vehicle-to-everything (V2X) communication have been proposed to improve safety of autonomous driving. Perception information from multiple *Connected Autonomous Vehicles (CAVs)* and roadside units *(RSUs)* are fused to generate better overall detection or tracking results. Relevant datasets have been released to the public [17, 37], including simulated ones [8, 15, 32, 33] and real ones [30, 34, 38, 39]. However, how these algorithms can be used to generate safe cooperative planning results is still under-explored. Other research has attempted to use LLM-based methods to build end-to-end perception and planning algorithms [3, 20, 23, 25–27, 29, 35]. However, these approaches have not yet explored the benefits of cooperative perception and planning.

In this paper, we propose a novel problem setting wherein Multi-Modal LLM-based methods are used to build end-to-end perception and planning algorithms for *Cooperative Autonomous Driving*, as illustrated in Fig. 1. We assume that there are multiple CAVs, RSUs, and a centralized LLM computing node. All CAVs and RSUs share their individual perception information with the LLM. Any CAV can ask the LLM a question in natural language to obtain useful information for driving safety. We create the **Vehicle-to-Vehicle Question-Answering (V2V-QA)** dataset, built upon the V2V4Real [34] and V2X-Real [30] cooperative perception datasets for autonomous driving. Our V2V-QA includes **grounding** (Figs. 2a to 2c), **notable object identification** (Fig. 2d), and **planning** (Fig. 2e) question-answer pairs. The main differences between our V2V-QA and other related datasets are summarized in Tab. 1. To establish a benchmark for the V2V-QA dataset, we propose a strong baseline method: **Vehicle-to-Vehicle Multi-Modal Large Language Model (V2V-LLM)**, as illustrated in Fig. 3. Experimental results show that our proposed V2V-LLM outperforms other baseline methods: *no fusion*, *early fusion*, and *intermediate fusion* [30–34].

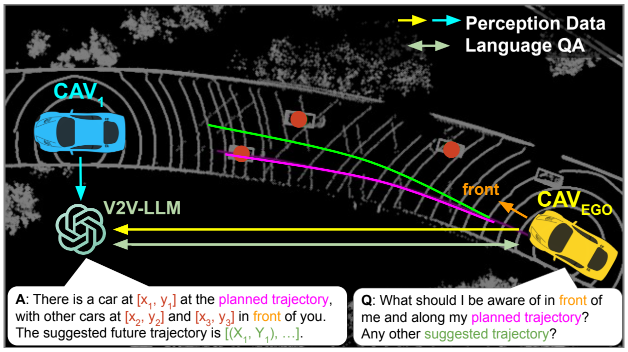

Perception Data
Language QA
CAV$_1$
V2V-LLM
front
CAV$_{EGO}$
A: There is a car at $[x_1, y_1]$ at the planned trajectory, with other cars at $[x_2, y_2]$ and $[x_3, y_3]$ in front of you. The suggested future trajectory is $[(X_1, Y_1), \ldots]$.
Q: What should I be aware of in front of me and along my planned trajectory? Any other suggested trajectory?

Figure 1. Overview of our problem setting of Multi-Modal LLM-based cooperative autonomous driving. All CAVs/RSUs share their perception information with the LLM. Any CAV can ask the LLM a question to obtain useful information for driving safety.

In summary, we create and introduce the V2V-QA dataset to support the development and evaluation of Multi-Modal LLM-based approaches to end-to-end cooperative autonomous driving. Our proposed V2V-LLM outperforms other baseline fusion methods, indicating the potential of being a promising foundation model for cooperative autonomous driving.

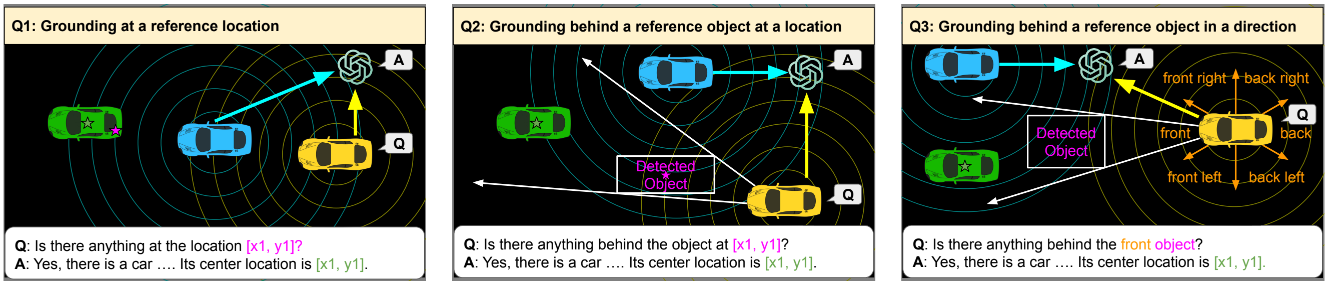
2. V2V-QA Dataset

2.1. Dataset Overview

Our V2V-QA dataset includes **grounding (Q1-3)**, **notable object identification (Q4)**, and **planning (Q5)**, as illustrated in Fig. 2. Our V2V-QA dataset contains two splits: **V2V-split** and **V2X-split**, which are built on top of V2V4Real [34] and V2X-Real [30] datasets, respectively. Tab. 2 summarizes the numbers of QA pairs in our proposed V2V-QA's V2V-split and V2X-split. We have 1.45M QA pairs in total and 30.2 QA pairs per frame on average. More details can be found in the supplementary materials.

Dataset	Publication	# CAVs	RSU	Sim/Real	# Frames	# QA	# QA/frame	Point Cloud	Planning
Cooperative perception in AD									
OPV2V [33]	ICRA 2022	2-7	-	Sim	11K	-	-	✓	
V2X-Sim [15]	RA-L 2022	2-5	✓	Sim	10K	-	-	✓	
V2XSet [32]	ECCV 2022	2-5	✓	Sim	11K	-	-	✓	
DAIR-V2X [38]	CVPR 2022	1	✓	Real	71K	-	-	✓	
V2V4Real [34]	CVPR 2023	2	-	Real	20K	-	-	✓	
TUMTrafV2X [39]	CVPR 2024	1	✓	Real	2K	-	-	✓	
V2X-Real [30]	ECCV 2024	2	✓	Real	33K	-	-	✓	
LLM-based AD									
NuScenes-QA [21]	AAAI 2024	-	-	Real	34K	460K	13.5	✓	
Lingo-QA [19]	ECCV 2024	-	-	Real	28K	420K	15.3		✓
MAPLM-QA [2]	CVPR 2024	-	-	Real	14K	61K	4.4	✓	
DriveLM [23]	ECCV 2024	-	-	Sim+Real	69K	2M	29.1		✓
TOKEN [25]	CoRL 2024	-	-	Real	28K	434K	15.5		✓
OmniDrive [27]	arXiv 2024	-	-	Real	34K	450K	13.2		✓
V2V-QA (Ours)	-	2	✓	Real	48K	**1.45M**	**30.2**	✓	✓

Table 1. Comparison between our V2V-QA and recent related Autonomous Driving (AD) datasets.


Q1: Grounding at a reference location

Q: Is there anything at the location [x1, y1]?
A: Yes, there is a car …. Its center location is [x1, y1].

Q2: Grounding behind a reference object at a location

Q: Is there anything behind the object at [x1, y1]?
A: Yes, there is a car …. Its center location is [x1, y1].

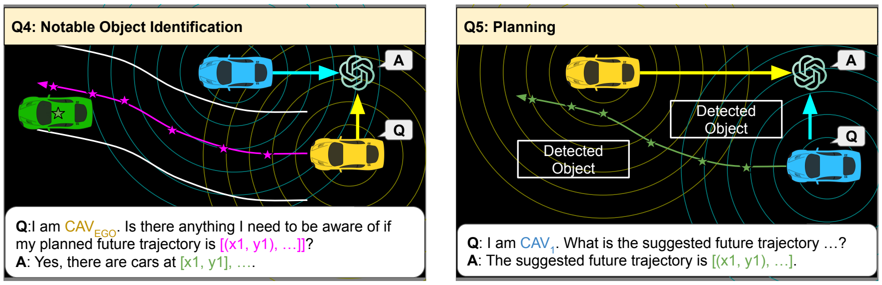
Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front object?
A: Yes, there is a car …. Its center location is [x1, y1].

(a) Q1: Grounding at a reference location.

(b) Q2: Grounding behind a reference object at a location.

(c) Q3: Grounding behind a reference object in a direction.


Q4: Notable Object Identification

Q:I am CAV_{EGO}. Is there anything I need to be aware of if my planned future trajectory is [(x1, y1), …]]?
A: Yes, there are cars at [x1, y1], ….

Q5: Planning

Q: I am CAV_1. What is the suggested future trajectory …?
A: The suggested future trajectory is [(x1, y1), …].

(d) Q4: Notable object identification.

(e) Q5: Planning.

Figure 2. Illustration of V2V-QA's 5 types of QA pairs. The arrows pointing at LLM indicate the perception data from CAVs.

2.2. Question and Answer Pairs Curation

For each frame of V2V4Real [34] and V2X-Real [30] datasets, we create 5 different types of QA pairs, including 3 types of grounding questions, 1 type of notable object identification question, and 1 type of planning question. These QAs are designed for cooperative driving scenarios. To generate instances of these QA pairs, we use V2V4Real [34] and V2X-Real [30]'s ground-truth bounding box annotations, each CAV's ground-truth trajectories, and individual detection results as the source information. Then we use different manually designed rules based on the geometric relationship among the aforementioned entities and text templates to generate our QA pairs. The text template can be seen in Figs. 4 and 5.

Q1. Grounding at a reference location (Fig. 2a): We ask the LLM to identify whether an object that occupies a specific query 2D location exists. We use the center locations of ground-truth boxes and every CAV/RSU's individual detection result boxes as the query locations in the questions. By doing so, we can focus more on evaluating each model's cooperative grounding ability on the potential false positive and false negative detection results.

QA type	V2V-split		V2X-split		Total
	Training	Testing	Training	Testing	
Q1	354820	121383	495290	128711	1100204
Q2	35700	13882	167694	35233	252509
Q3	14339	5097	28740	6465	54641
Q4	12290	3446	6274	1708	23718
Q5	12290	3446	6274	1708	23718
Total	429439	147254	704272	173825	1454790

Table 2. Dataset statistics of our V2V-QA's V2V-split and V2X-split. Q1: Grounding at a reference location. Q2: Grounding behind a reference object at a location. Q3: Grounding behind a reference object in a direction. Q4: Notable object identification. Q5: Planning.

Q2. Grounding behind a reference object at a location (Fig. 2b): When a CAV's field of view is occluded, this CAV may want to ask the centralized LLM to determine whether there exists any object behind that occluding large object. We use the center location of each detection result box as the query locations in these questions. We draw a sector region based on the relative pose of the asking CAV and the reference object, and select the closest ground-truth object in the region as the answer.

Q3. Grounding behind a reference object in a direction (Fig. 2c): We further challenge the LLM on language and spatial understanding ability by replacing Q2's reference 2D location with a reference directional keyword. We first get the closest detection result box in each of the 6 directions of a CAV as the reference object. Then we follow the same approach in Q2 to get the closest ground-truth box in the corresponding sector region as the answer.

Q4. Notable object identification (Fig. 2d): More critical abilities of autonomous vehicles involve identifying notable objects near planned future trajectories. We extract 6 waypoints from the ground-truth trajectory in the next 3 seconds as the reference future waypoints in the questions. Then we get, at most, the 3 closest ground-truth objects within 10 meters of the reference future trajectory as the answer.

Q5. Planning (Fig. 2e): Planning is critical because the ultimate goal of autonomous vehicles is to navigate through complex environments safely and avoid potential collisions. To generate the planning QAs, we extract 6 future waypoints, evenly distributed in the next 3 seconds, from each CAV's ground-truth future trajectory as the answer.

3. V2V-LLM

We propose a competitive baseline model, **V2V-LLM**, for this LLM-based collaborative driving problem, as shown in Fig. 3. Our model is a Multi-Modal LLM (MLLM) that takes the individual perception features of every CAV/RSU as the vision input, a question as the language input, and generates an answer as the language output.

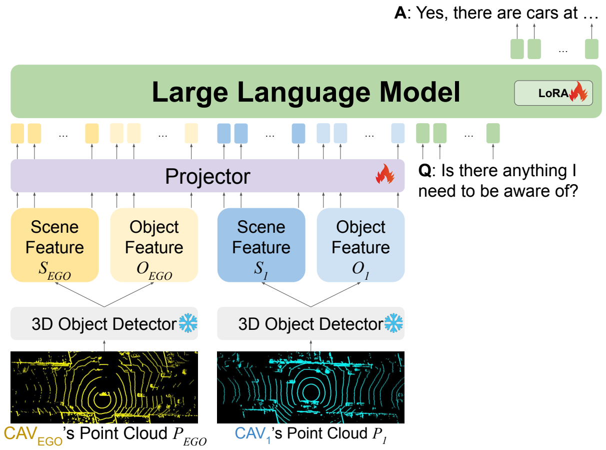


Figure 3. Model diagram of our proposed V2V-LLM for cooperative autonomous driving.

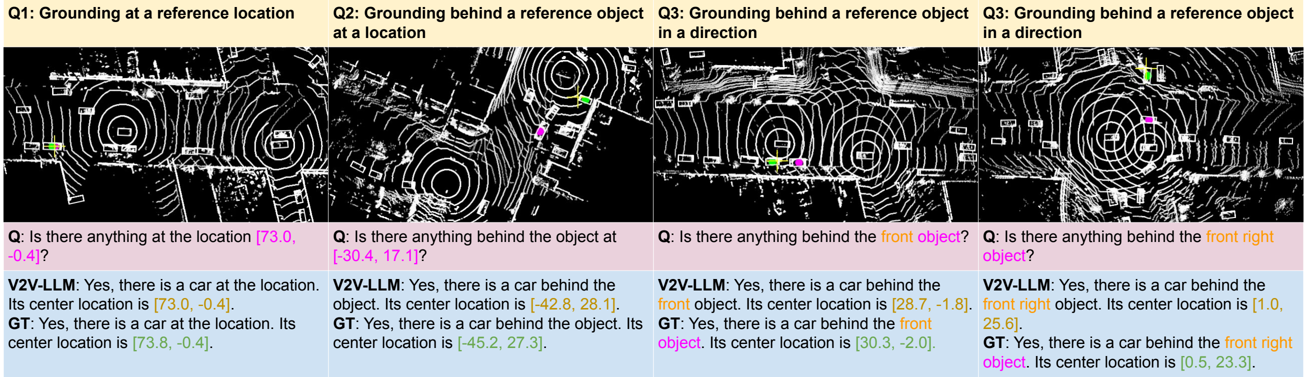
For extracting the perception input features, each CAV/RSU applies a 3D object detection model to its individual LiDAR point cloud: P_{EGO} and P_1. We extract the scene-level feature map S_{EGO} and S_1 from the 3D object detection model and transform the 3D object detection results as the object-level feature vectors O_{EGO} and O_1. We utilize LLaVA [16] to develop our MLLM. Instead of using LLaVA [16]'s CLIP [22] image feature encoder, we use PointPillars [13] LiDAR-based 3D object detector as the point cloud feature encoder. We then feed the resulting features to a multi-layer perceptron-based projector network for feature alignment from the point cloud embedding space to the language embedding space. The aligned perception features are the input perception tokens digested by the LLM together with the input language tokens from the question. Finally, the LLM aggregates the perception information from all CAVs and RSUs to answer the question.

4. Experiment

We establish a benchmark for our proposed V2V-QA dataset with experiments on baseline methods: **no fusion**, **early fusion**, **intermediate fusion**, such as CoBEVT [31], V2X-ViT [32], and AttFuse [33], and our proposed **V2V-LLM** (Fig. 3). The baseline methods also adopt the same projector and LLM architecture as in our V2V-LLM but with different point cloud feature encoders. We follow [25, 27] to use F1 scores, L2-errors and collision rates as metrics. Tab. 3 shows V2V-LLM's testing performance in V2V-QA's V2V-split and V2X-split in comparison with baseline methods. Overall, V2V-LLM achieves the best results in the notable object identification and planning tasks, which are critical in autonomous driving applications. V2V-LLM also achieves competitive results in the grounding tasks. Figs. 4 and 5 show our V2V-LLM's qualitative results on V2V-QA's testing set. Overall, the outputs of our model closely align with the ground-truth answers across all question types, indicating its robustness in various cooperative autonomous driving tasks.

Method	V2V-split							V2X-split							Comm(MB) ↓
	Q1	Q2	Q3	Q_{Gr}	Q4	Q5		Q1	Q2	Q3	Q_{Gr}	Q4	Q5		
	F1 ↑	F1 ↑	F1 ↑	F1 ↑	F1 ↑	L2 (m) ↓	CR (%) ↓	F1 ↑	F1 ↑	F1 ↑	F1 ↑	F1 ↑	L2 (m) ↓	CR (%) ↓	
No Fusion	66.6	22.6	17.2	35.5	47.3	6.55	4.57	55.7	21.4	25.2	34.1	64.4	2.31	9.21	**0**
Early Fusion	**73.5**	23.3	20.8	39.2	53.9	6.20	3.55	59.7	23.3	26.1	36.4	67.6	2.12	8.61	1.9208
Intermediate Fusion															
AttFuse [33]	70.7	26.4	18.4	38.5	56.9	6.83	4.12	58.9	23.9	26.3	36.4	65.9	2.19	8.39	0.4008
V2X-ViT [32]	70.8	28.0	**22.6**	40.5	57.6	7.08	4.33	59.6	24.2	26.1	36.6	65.0	2.29	8.86	0.4008
CoBEVT [31]	72.2	29.3	21.3	**40.9**	57.6	6.72	3.88	-	-	-	-	-	-	-	0.4008
LLM Fusion															
V2V-LLM (Ours)	70.0	**30.8**	21.2	40.7	**59.7**	**4.99**	**3.00**	**60.5**	**25.3**	**26.7**	**37.5**	**69.3**	**1.71**	**6.89**	0.4068

Table 3. V2V-LLM's testing performance in V2V-QA's V2V-split and V2X-split in comparison with baseline methods. Q1: Grounding at a reference location. Q2: Grounding behind a reference object at a location. Q3: Grounding behind a reference object in a direction. Q_{Gr}: Average of grounding (Q1, Q2, and Q3). Q4: Notable object identification. Q5: Planning. L2: L2 distance error. CR: Collision rate. Comm: Communication cost. In each column, the **best** results are in boldface, and the second-best results are in underline. More detailed performance evaluation can be seen in the supplementary material.



Q1: Grounding at a reference location

Q: Is there anything at the location [73.0, -0.4]?

V2V-LLM: Yes, there is a car at the location. Its center location is [73.0, -0.4].
GT: Yes, there is a car at the location. Its center location is [73.8, -0.4].

Q2: Grounding behind a reference object at a location

Q: Is there anything behind the object at [-30.4, 17.1]?

V2V-LLM: Yes, there is a car behind the object. Its center location is [-42.8, 28.1].
GT: Yes, there is a car behind the object. Its center location is [-45.2, 27.3].

Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front object?

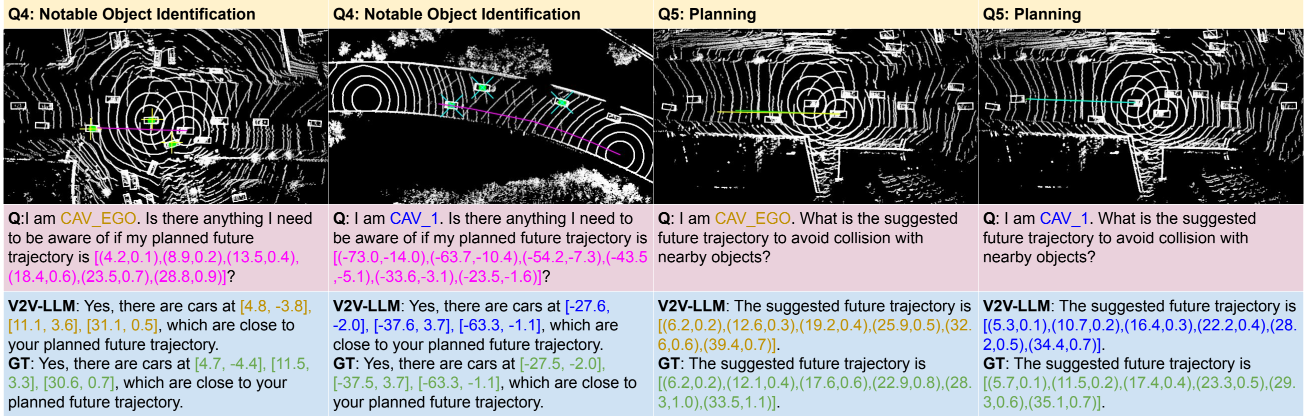
V2V-LLM: Yes, there is a car behind the front object. Its center location is [28.7, -1.8].
GT: Yes, there is a car behind the front object. Its center location is [30.3, -2.0].

Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front right object?

V2V-LLM: Yes, there is a car behind the front right object. Its center location is [1.0, 25.6].
GT: Yes, there is a car behind the front right object. Its center location is [0.5, 23.3].

Figure 4. V2V-LLM's *grounding* results on V2V-QA's testing set. Magenta ○: reference locations in questions. Yellow +: model output locations. Green ○: ground-truth answers.



Q4: Notable Object Identification

Q:I am CAV_EGO. Is there anything I need to be aware of if my planned future trajectory is [(4.2,0.1),(8.9,0.2),(13.5,0.4),(18.4,0.6),(23.5,0.7),(28.8,0.9)]?

V2V-LLM: Yes, there are cars at [4.8, -3.8], [11.1, 3.6], [31.1, 0.5], which are close to your planned future trajectory.
GT: Yes, there are cars at [4.7, -4.4], [11.5, 3.3], [30.6, 0.7], which are close to your planned future trajectory.

Q4: Notable Object Identification

Q: I am CAV_1. Is there anything I need to be aware of if my planned future trajectory is [(-73.0,-14.0),(-63.7,-10.4),(-54.2,-7.3),(-43.5,-5.1),(-33.6,-3.1),(-23.5,-1.6)]?

V2V-LLM: Yes, there are cars at [-27.6, -2.0], [-37.6, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.
GT: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.

Q5: Planning

Q: I am CAV_EGO. What is the suggested future trajectory to avoid collision with nearby objects?

V2V-LLM: The suggested future trajectory is [(6.2,0.2),(12.6,0.3),(19.2,0.4),(25.9,0.5),(32.6,0.6),(39.4,0.7)].
GT: The suggested future trajectory is [(6.2,0.2),(12.1,0.4),(17.6,0.6),(22.9,0.8),(28.3,1.0),(33.5,1.1)].

Q5: Planning

Q: I am CAV_1. What is the suggested future trajectory to avoid collision with nearby objects?

V2V-LLM: The suggested future trajectory is [(5.3,0.1),(10.7,0.2),(16.4,0.3),(22.2,0.4),(28.2,0.5),(34.4,0.7)].
GT: The suggested future trajectory is [(5.7,0.1),(11.5,0.2),(17.4,0.4),(23.3,0.5),(29.3,0.6),(35.1,0.7)].

Figure 5. V2V-LLM's *notable object identification* and *planning* results on V2V-QA's testing set. For notable object identification, Magenta curve: planned future trajectories in questions. Green ○: ground-truth notable object locations. Yellow + and Cyan ×: model identification outputs corresponding to CAV_EGO and CAV_1, respectively. For planning, Green line: future trajectories in ground-truth answers. Yellow curve and Cyan curve: model planning outputs corresponding to CAV_EGO and CAV_1, respectively.

5. Conclusion

We introduce a novel problem setting of Multi-Modal LLM-based cooperative autonomous driving and create V2V-QA dataset and benchmark. Our proposed V2V-LLM outperforms baselines and can be a promising unified multimodal architecture for cooperative autonomous driving.

6. Acknowledgement

The authors thank Boyi Li, Boris Ivanovic, and Marco Pavone for valuable discussions and comments.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[2] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James Rehg, and Chao Zheng. Maplm: A real-world large-scale vision-language dataset for map and traffic scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[3] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1

[4] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE Symposium on Edge Computing (SEC)*, 2019. 1

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[6] Hsu-kuang Chiu and Stephen F. Smith. Selective communication for cooperative perception in end-to-end autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA) Workshop*, 2023. 1

[7] Hsu-kuang Chiu, Chien-Yi Wang, Min-Hung Chen, and Stephen F. Smith. Probabilistic 3d multi-object cooperative tracking for autonomous driving via differentiable multi-sensor kalman filter. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1

[8] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[10] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[11] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Tao Huang, Jianan Liu, Xi Zhou, Dinh C Nguyen, Mostafa Rahimi Azghadi, Yuxuan Xia, Qing-Long Han, and Sumei Sun. V2x cooperative perception for autonomous driving: Recent advances and challenges. *arXiv preprint arXiv:2310.03525*, 2023. 1

[13] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[14] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[15] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 1, 2

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3, 1

[17] Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *IEEE Transactions on Intelligent Vehicles (T-IV)*, 2024. 1

[18] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[19] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[20] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[21] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3

[23] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger,

and Hongyang Li. Drivelm: Driving with graph visual question answering. In *Europian Conference on Computer Vision (ECCV)*, 2024. 1, 2

[24] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *2024 IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2024. 1

[25] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *Conference on Robot Learning (CoRL)*, 2024. 1, 2, 3

[26] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[27] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. OmniDrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv:2405.01533*, 2024. 1, 2, 3

[28] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, James Tu, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[29] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 1

[30] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In *Europian Conference on Computer Vision (ECCV)*, 2024. 1, 2

[31] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning (CoRL)*, 2022. 1, 3, 4, 2

[32] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4

[33] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 1, 2, 3, 4

[34] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4

[35] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters (RA-L)*, 2024. 1

[36] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[37] Melih Yazgan, Mythra Varun Akkanapragada, and J Marius Zöllner. Collaborative perception datasets in autonomous driving: A survey. In *IEEE Intelligent Vehicles Symposium (IV)*, 2024. 1

[38] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[39] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

V2V-LLM: Vehicle-to-Vehicle Cooperative Autonomous Driving with Multi-Modal Large Language Models

Supplementary Material

7. Model Training Details

We use 8 NVIDIA A100-80GB GPUs to train our model. Our V2V-LLM uses LLaVA-v1.5-7b [16]'s Vicuna [5] as the LLM backbone. To train our model, we initialize it by loading the pre-trained LLaVA-v1.5-7b [16]'s checkpoint. We freeze the LLM and the point cloud feature encoder, and finetune the projector and the LoRA [9] parts of the model. During training, we use batch size 32. Adam optimizer is adopted for training with a starting learning rate $2e-5$ and a cosine learning rate scheduler with a 3% warm-up ratio. For all other training settings and hyperparameters, we use the same ones from LLaVA-v1.5-7b [16].

8. Detailed Evaluation Results

Tabs. 4 and 5 summarize the detailed evaluation results of our V2V-LLM and other baseline methods in V2V-QA's V2V-split and V2X-split. In addition, Tabs. 6 and 7 show the detailed planning performance. For the grounding task, our V2V-LLM achieves competitive results in V2V-split and outperforms all other baseline methods in V2X-split. More importantly, for the notable object identification task and the planning task, our V2V-LLM outperforms all other baseline methods in both V2V-split and V2X-split.

9. Detailed Communication Cost and Scaling Analysis

In our centralized setting, assume that there is one centralized LLM computing node, N_v CAVs, and each CAV asks N_q questions at each timestep. Each CAV sends one scene-level feature map ($\leq$ 0.2MB), one set of individual object detection result parameters ($\leq$ 0.003MB), N_q questions (each $\leq$ 0.0002MB) to the LLM and receives N_q answers (each $\leq$ 0.0002MB) at each timestep. Note that each CAV only needs to send the same features to the LLM once at each timestep because the LLM node can save and reuse them to answer multiple questions from the same or different CAVs at the same timestep. The communication cost of each CAV is: $0.2 + 0.003 + (0.0002 + 0.0002)N_q = (0.203 + 0.0004N_q)$ MB. The LLM receives N_v scene-level feature maps, N_v set of individual object detection result parameters, N_qN_v questions and returns N_qN_v answers. The communication cost of the centralized LLM is $(0.2 + 0.003 + (0.0002 + 0.0002)N_q)N_v = (0.203N_v + 0.0004N_qN_v)$ MB.

Alternatively, one can also consider a decentralized setting that deploys one LLM in each CAV. In this setting, each CAV receives the features from all other CAVs and does not need to send or receive any questions or answers. The communication cost of each CAV is $(0.2 + 0.003)(N_v - 1) = 0.203(N_v - 1)$ MB. Tab. 8 summarizes the communication cost and scaling analysis in the aforementioned settings. There could be more different decentralized settings. Which setting works best in terms of communication costs is beyond the current focus of our work.

10. Planning Results with Temporal Inputs

In the main paper, all experiments use point clouds at a single frame from each CAV as the visual input to the models. In this section, we experiment with feeding visual features from 3 consecutive frames, the current one and the previous two, as the visual input to the models. Tab. 9 shows the planning results of the new setting together with the original setting from the main paper. In general, using visual inputs from multiple frames improves planning performance.

11. Detailed Ablation Results

Tab. 10 shows the detailed ablation results when using only the scene-level features or only the object-level features as input to our V2V-LLM. Both types of input features contribute to the final performance, and object-level features are easier for LLM to digest. Training from scratch achieves worse performance, meaning that pre-training with LLaVA's VQA tasks improves our V2V-LLM's performance in V2V-QA.

12. Additional Dataset Statistics

Our V2V-QA dataset contains two splits: **V2V-split** and **V2X-split**, which are built on top of V2V4Real [34] and V2X-Real [30] datasets, respectively. In V2V4Real [34], the training set has 32 driving sequences and a total of 7105 frames of data per CAV, and the testing set has 9 driving sequences and a total of 1993 frames of data per CAV. In V2X-Real [30], the training set has 43 driving sequences and a total of 5772 frames of data per CAV, and the testing set has 9 driving sequences and a total of 1253 frames of data per CAV. The frame rate is 10Hz.

For the grounding questions (Q1, Q2, Q3) and the notable object identification question (Q4), a QA pair can be categorized into either a positive case or a negative case. If at least one object exists that satisfies the condition specified in the questions, the corresponding QA pair is a positive case. Otherwise, it is a negative case. Tabs. 11 and 12

Method	Q1			Q2			Q3			Q_{Gr}	Q4			Q5		Comm(MB) ↓
	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	F1 ↑	P ↑	R ↑	$L2_{avg}$ (m) ↓	CR_{avg} (%) ↓	
No Fusion	66.6	77.9	58.2	22.6	29.4	18.4	17.2	17.4	16.9	35.5	47.3	49.2	45.6	6.55	4.57	**0**
Early Fusion	**73.5**	**82.2**	66.5	23.3	29.1	19.5	20.8	22.7	19.3	39.2	53.9	55.4	52.6	6.20	3.55	1.9208
Intermediate Fusion																
AttFuse [33]	70.7	79.6	63.6	26.4	31.6	22.7	18.4	19.6	17.4	38.5	56.9	57.2	56.6	6.83	4.12	0.4008
V2X-ViT [32]	70.8	81.1	62.8	28.0	33.9	23.9	**22.6**	**25.2**	20.5	40.5	57.6	57.0	**58.2**	7.08	4.33	0.4008
CoBEVT [31]	72.2	76.8	**68.1**	29.3	34.7	25.3	21.3	22.1	20.6	**40.9**	57.6	57.2	58.1	6.72	3.88	0.4008
LLM Fusion																
V2V-LLM (Ours)	70.0	80.1	62.2	**30.8**	**36.3**	**26.7**	21.2	21.5	**20.8**	40.7	**59.7**	**61.9**	57.6	**4.99**	**3.00**	0.4068

Table 4. V2V-LLM's testing performance in V2V-QA's V2V-split in comparison with baseline methods. Q1: Grounding at a reference location. Q2: Grounding behind a reference object at a location. Q3: Grounding behind a reference object in a direction. Q_{Gr}: Average of grounding (Q1, Q2, and Q3). Q4: Notable object identification. Q5: Planning. P: Precision. R: Recall. L2: L2 distance error. CR: Collision rate. Comm: Communication cost. In each column, the **best** results are in boldface, and the second-best results are in underline.

Method	Q1			Q2			Q3			Q_{Gr}	Q4			Q5		Comm(MB) ↓
	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	F1 ↑	P ↑	R ↑	$L2_{avg}$ (m) ↓	CR_{avg} (%) ↓	
No Fusion	55.7	**71.6**	45.5	21.4	33.2	15.8	25.2	26.2	24.2	34.1	64.4	66.1	62.7	2.31	9.21	**0**
Early Fusion	59.7	70.6	51.8	23.3	34.0	17.7	26.1	28.0	24.5	36.4	67.6	69.3	66.0	2.12	8.61	1.9208
Intermediate Fusion																
AttFuse [33]	58.9	71.1	50.3	23.9	34.3	18.4	26.3	**28.3**	24.6	36.4	65.9	67.0	64.9	2.19	8.39	0.4008
V2X-ViT [32]	59.6	69.6	52.1	24.2	33.2	19.1	26.1	28.2	24.3	36.6	65.0	64.8	65.3	2.29	8.86	0.4008
LLM Fusion																
V2V-LLM (Ours)	**60.5**	69.5	**53.6**	**25.3**	**34.9**	**19.8**	**26.7**	27.0	**26.4**	**37.5**	**69.3**	**71.9**	**66.8**	**1.71**	**6.89**	0.4068

Table 5. V2V-LLM's testing performance in V2V-QA's V2X-split in comparison with baseline methods. Q1: Grounding at a reference location. Q2: Grounding behind a reference object at a location. Q3: Grounding behind a reference object in a direction. Q_{Gr}: Average of grounding (Q1, Q2, and Q3). Q4: Notable object identification. Q5: Planning. P: Precision. R: Recall. L2: L2 distance error. CR: Collision rate. Comm: Communication cost. In each column, the **best** results are in boldface, and the second-best results are in underline.

Method	L2 (m)				CR (%)			
	1s ↓	2s ↓	3s ↓	average ↓	1s ↓	2s ↓	3s ↓	average ↓
No Fusion	3.84	6.52	9.30	6.55	1.31	4.76	7.63	4.57
Early Fusion	3.68	6.19	8.74	6.20	0.96	3.86	5.83	3.55
Intermediate Fusion								
AttFuse [33]	4.06	6.78	9.64	6.83	1.42	4.41	6.53	4.12
V2X-ViT [32]	4.21	7.05	9.99	7.08	1.33	4.82	6.85	4.33
CoBEVT [31]	3.97	6.71	9.47	6.72	0.93	3.74	6.96	3.88
LLM Fusion								
V2V-LLM (ours)	**2.96**	**4.97**	**7.05**	**4.99**	**0.55**	**3.19**	**5.25**	**3.00**

Table 6. V2V-LLM's planning performance in V2V-QA's V2V-split in comparison with baseline methods. L2: L2 distance error. CR: Collision rate. In each column, the **best** results are in boldface, and the second-best results are in underline.

summarizes the numbers of QA pairs in each category, for V2V-QA's V2V-split and V2X-split respectively. This table shows that V2V-QA has sufficient positive and negative data samples in both training and testing sets for each of these QA pairs. The planning task (Q5) is excluded from Tabs. 11 and 12, as each planning QA pair inherently includes a ground-truth future trajectory as the answer.

We also visualize our V2V-split distribution of ground truth answer locations relative to the asking CAV for the grounding questions (Q1, Q2, Q3) and the notable object identification question (Q4), as shown in Figs. 6 to 9. In our coordinate system, x is the CAV's front direction, and y is the CAV's right direction. For the planning question (Q5), we show the distribution of the ending waypoints in the ground truth answer future trajectories, as shown in Fig. 10. We visualize the location distribution of V2V-QA's V2X-splitin Figs. 11 to 15. These figures indicate that our V2V-QA has diverse spatial distributions in the driving scenes. Compared to NuScenes [1], our V2V-QA has larger ranges and standard deviations of the ground-truth ending waypoints, as shown in Tab. 13. Therefore, the planning task in our V2V-QA could be challenging.

Method	L2 (m)				CR (%)			
	1s ↓	2s ↓	3s ↓	average ↓	1s ↓	2s ↓	3s ↓	average ↓
No Fusion	1.33	2.28	3.31	2.31	2.52	9.54	15.57	9.21
Early Fusion	<u>1.24</u>	<u>2.10</u>	<u>3.00</u>	<u>2.12</u>	3.51	<u>8.37</u>	13.93	8.61
Intermediate Fusion								
AttFuse [33]	1.27	2.17	3.11	2.19	2.40	9.07	<u>13.70</u>	<u>8.39</u>
V2X-ViT [32]	1.34	2.27	3.25	2.29	**1.41**	9.89	15.28	8.86
LLM Fusion								
V2V-LLM (ours)	**0.99**	**1.70**	**2.45**	**1.71**	<u>2.17</u>	**6.79**	**11.71**	**6.89**

Table 7. V2V-LLM's planning performance in V2V-QA's V2X-split in comparison with baseline methods. L2: L2 distance error. CR: Collision rate. In each column, the **best** results are in boldface, and the <u>second-best</u> results are in underline.

Setting	Each CAV	Centralized LLM
Centralized	$0.203 + 0.0004N_q$	$0.203N_v + 0.0004N_qN_v$
Decentralized	$0.203(N_v - 1)$	-

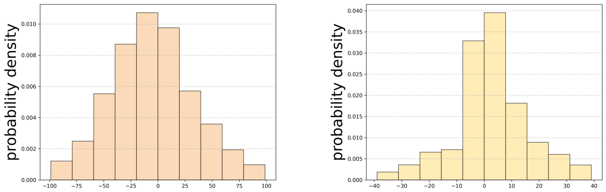
Table 8. Communication cost (MB) and scaling analysis. N_v: number of CAVs. N_q: number of questions asked by each CAV at each timestep.

Method	1 input frame		3 input frames	
	L2 (m) ↓	CR (%) ↓	L2 (m) ↓	CR (%) ↓
No Fusion	6.55	4.57	5.94	3.77
Early Fusion	<u>6.20</u>	<u>3.55</u>	<u>5.13</u>	<u>3.04</u>
Intermediate Fusion				
AttFuse [33]	6.83	4.12	6.46	3.50
V2X-ViT [32]	7.08	4.33	5.52	3.84
CoBEVT [31]	6.72	3.88	6.02	3.40
LLM Fusion				
V2V-LLM (ours)	**4.99**	**3.00**	**4.82**	**2.93**

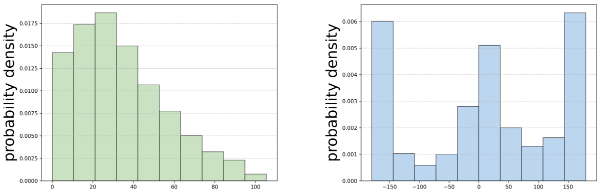
Table 9. V2V-LLM's planning performance in V2V-QA's V2V-split in comparison with baseline methods. L2: L2 distance error. CR: Collision rate. In each column, the **best** results are in boldface. and the <u>second-best</u> results are in underline.

13. Additional Qualitative Results

We show more qualitative results of our proposed V2V-LLM and other baseline methods in the testing set of V2V-QA's grounding task in Figs. 16 to 19, notable object identification task in Figs. 20 to 21, and planning task in Figs 22 to 23. The baseline methods include no-fusion, early-fusion, and intermediate-fusion: AttFuse [33], V2X-ViT [32], and CoBEVT [31]. Results of V2X-split can be seen in Figs. 24 to 30. In general, our proposed V2V-LLM's outputs are closer to the ground-truth answers, in comparison to other baseline methods' results.

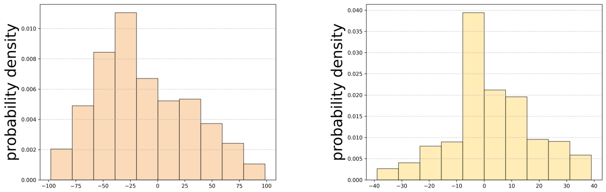


(a) x (meters) (b) y (meters)

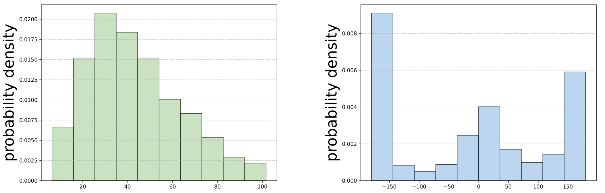


(c) distance (meters) (d) angle (degrees)

Figure 6. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2V-split Q1: Grounding at a reference location.



(a) x (meters) (b) y (meters)



(c) distance (meters) (d) angle (degrees)

Figure 7. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2V-split Q2: Grounding behind a reference object at a location.

14. Limitation

Fig. 31 shows failure cases of V2V-LLM's *planning* results on V2V-QA's testing set. In a few frames, the model generates future trajectories in the lane of the opposite traffic direction. A potential solution and future work is to include HD map information as additional input to the model. Currently, this approach is not feasible because the base dataset

Method	Q1			Q2			Q3			Q_{Gr}	Q4			Q5		Comm (MB) ↓
	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	F1 ↑	P ↑	R ↑	$L2_{avg}$ (m) ↓	CR_{avg} (%) ↓	
Scene-level only	69.9	74.9	**65.5**	15.4	19.9	12.6	17.9	**26.9**	13.5	34.4	43.2	40.2	46.7	7.21	15.55	0.4008
Object-level only	69.0	**80.9**	60.1	26.9	34.7	21.9	17.6	18.3	16.9	37.8	52.6	57.3	48.6	5.24	7.78	**0.0068**
Scratch	67.6	77.6	60.0	26.5	26.4	26.5	17.2	16.4	18.2	37.1	49.3	52.7	46.3	6.30	5.01	0.4068
V2V-LLM (ours)	**70.0**	80.1	62.2	**30.8**	**36.3**	**26.7**	**21.2**	21.5	**20.8**	**40.7**	**59.7**	**61.9**	**57.6**	**4.99**	**3.00**	0.4068

Table 10. Ablation study in V2V-QA's V2V-split. Q1: Grounding at a reference location. Q2: Grounding behind a reference object at a location. Q3: Grounding behind a reference object in a direction. Q_{Gr}: Average of grounding (Q1, Q2, and Q3). Q4: Notable object identification. Q5: Planning. P: Precision. R: Recall. L2: L2 distance error. CR: Collision rate. Comm: Communication cost.

QA type	Train-Pos	Train-Neg	Test-Pos	Test-Neg	Total
Q1	217403	137417	76522	44861	476203
Q2	17859	17841	8391	5491	49582
Q3	7197	7142	3082	2015	19436
Q4	9911	2379	2517	929	15736
Total	252370	164779	90512	53296	560957

Table 11. Dataset statistics of our V2V-QA's V2V-split on positive and negative samples.

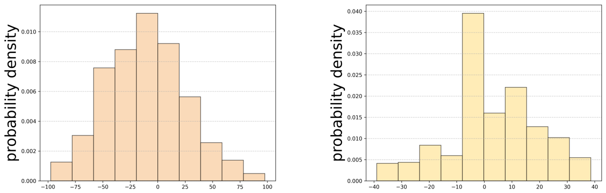
QA type	Train-Pos	Train-Neg	Test-Pos	Test-Neg	Total
Q1	247447	247843	62332	66379	624001
Q2	84005	83689	18297	16936	202927
Q3	14346	14394	3421	3044	35205
Q4	4624	1650	1172	536	7982
Total	350422	347576	85222	86895	870115

Table 12. Dataset statistics of our V2V-QA's V2X-split on positive and negative samples.

Dataset	x: forward			y: right		
	min	max	std	min	max	std
NuScenes	-0.9	39.7	10.4	-11.0	11.1	1.9
V2V-QA (ours)	-2.1	177.0	28.1	-24.3	12.0	2.4

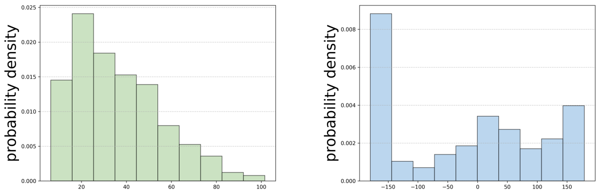
Table 13. Ranges and standard deviations of ground-truth ending waypoints.

V2V4Real [34] has not released its HD map to the public.



(a) x (meters)

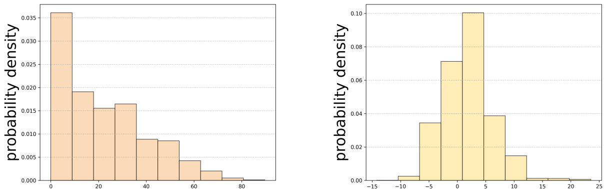
(b) y (meters)



(c) distance (meters)

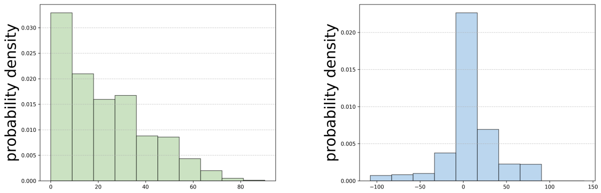
(d) angle (degrees)

Figure 8. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2V-split Q3: Grounding behind a reference object in a direction.



(a) x (meters)

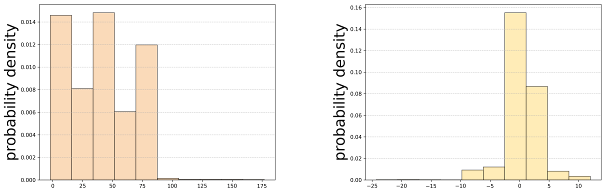
(b) y (meters)



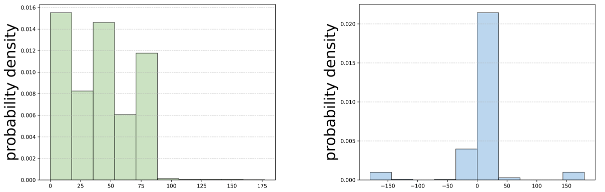
(c) distance (meters)

(d) angle (degrees)

Figure 9. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2V-split Q4: Notable object identification.

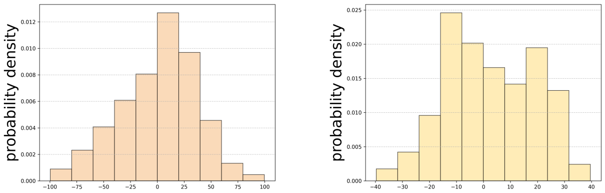


(a) x (meters) (b) y (meters)

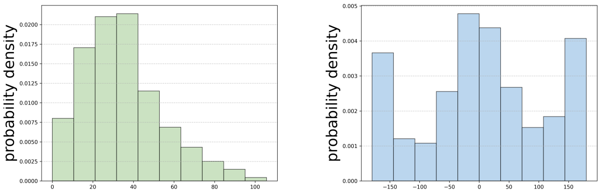


(c) distance (meters) (d) angle (degrees)

Figure 10. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2V-split Q5: Planning.

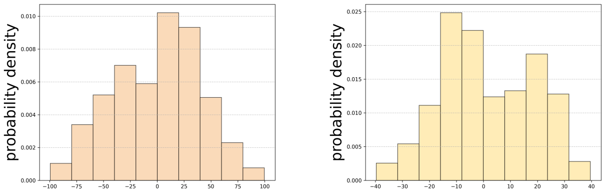


(a) x (meters) (b) y (meters)

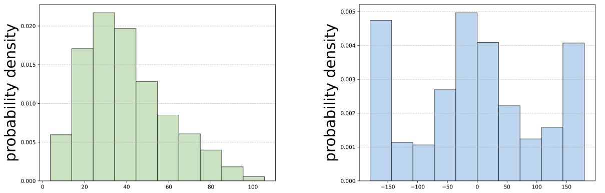


(c) distance (meters) (d) angle (degrees)

Figure 11. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q1: Grounding at a reference location.

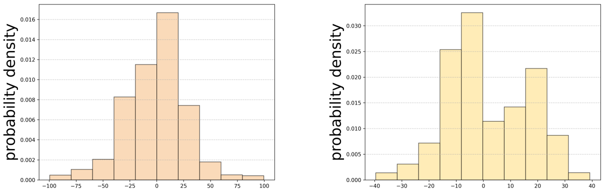


(a) x (meters) (b) y (meters)

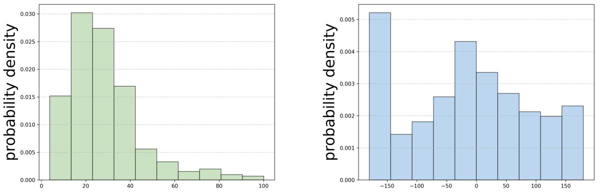


(c) distance (meters) (d) angle (degrees)

Figure 12. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q2: Grounding behind a reference object at a location.

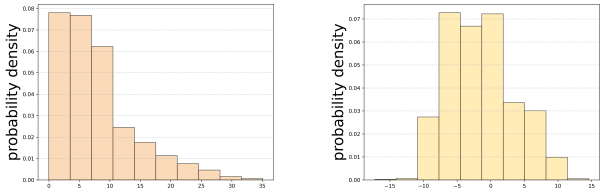


(a) x (meters) (b) y (meters)

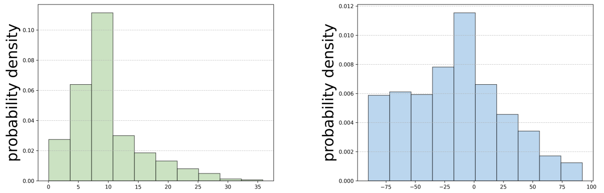


(c) distance (meters) (d) angle (degrees)

Figure 13. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q3: Grounding behind a reference object in a direction.

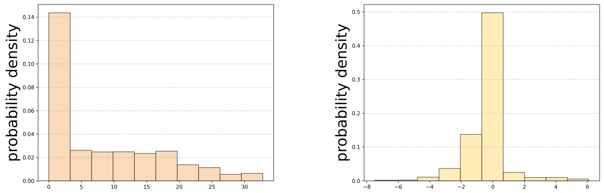


(a) x (meters) (b) y (meters)

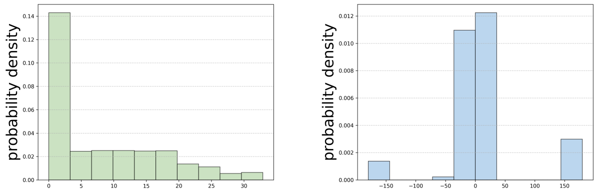


(c) distance (meters) (d) angle (degrees)

Figure 14. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q4: Notable object identification.

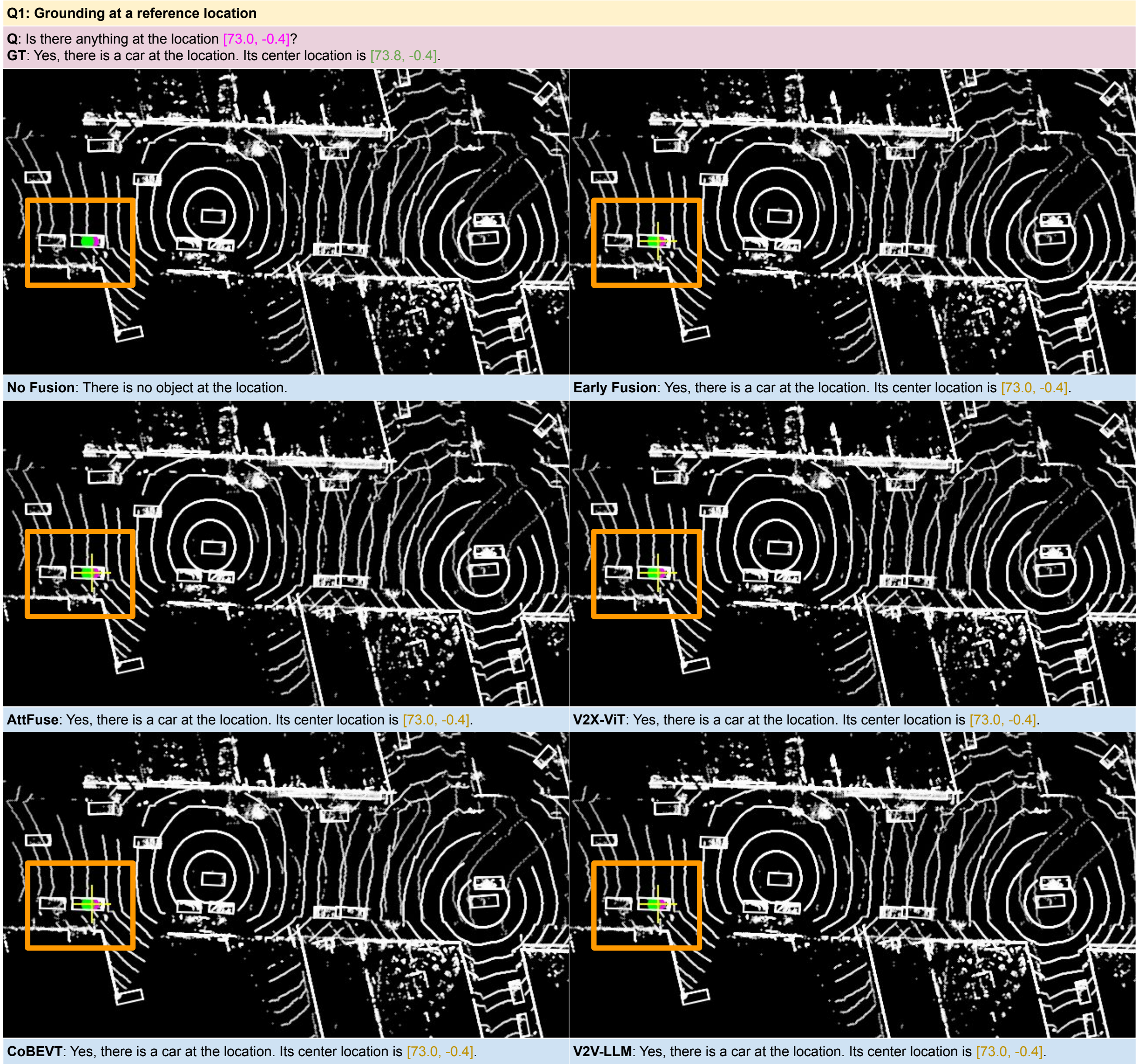


(a) x (meters) (b) y (meters)



(c) distance (meters) (d) angle (degrees)

Figure 15. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q5: Planning.



Q1: Grounding at a reference location

Q: Is there anything at the location [73.0, -0.4]?

GT: Yes, there is a car at the location. Its center location is [73.8, -0.4].

No Fusion: There is no object at the location.

Early Fusion: Yes, there is a car at the location. Its center location is [73.0, -0.4].

AttFuse: Yes, there is a car at the location. Its center location is [73.0, -0.4].

V2X-ViT: Yes, there is a car at the location. Its center location is [73.0, -0.4].

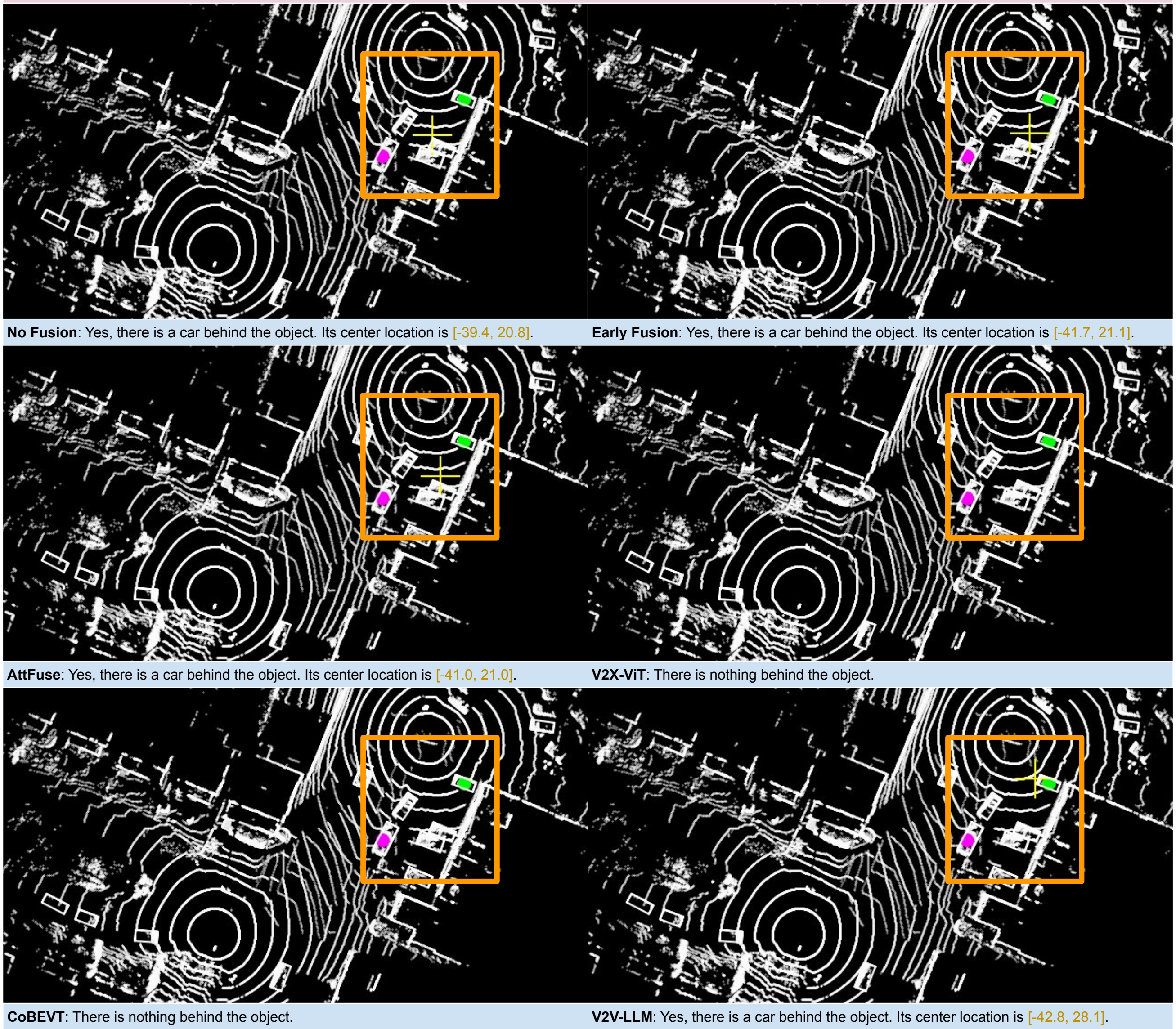
CoBEVT: Yes, there is a car at the location. Its center location is [73.0, -0.4].

V2V-LLM: Yes, there is a car at the location. Its center location is [73.0, -0.4].

Figure 16. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2V-split testing set. Magenta ○: reference locations in questions. Yellow +: model output locations. Green ○: ground-truth answers.

Q2: Grounding behind a reference object at a location

Q: Is there anything behind the object at [-30.4, 17.1]?
GT: Yes, there is a car behind the object. Its center location is [-45.2, 27.3].



No Fusion: Yes, there is a car behind the object. Its center location is [-39.4, 20.8].

Early Fusion: Yes, there is a car behind the object. Its center location is [-41.7, 21.1].

AttFuse: Yes, there is a car behind the object. Its center location is [-41.0, 21.0].

V2X-ViT: There is nothing behind the object.

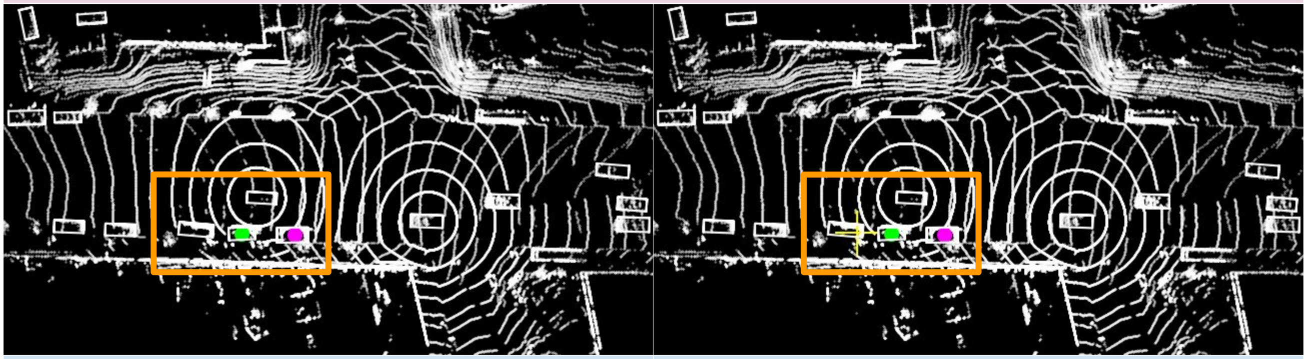
CoBEVT: There is nothing behind the object.

V2V-LLM: Yes, there is a car behind the object. Its center location is [-42.8, 28.1].

Figure 17. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2V-split testing set. Magenta ○: reference locations in questions. Yellow +: model output locations. Green ○: ground-truth answers.

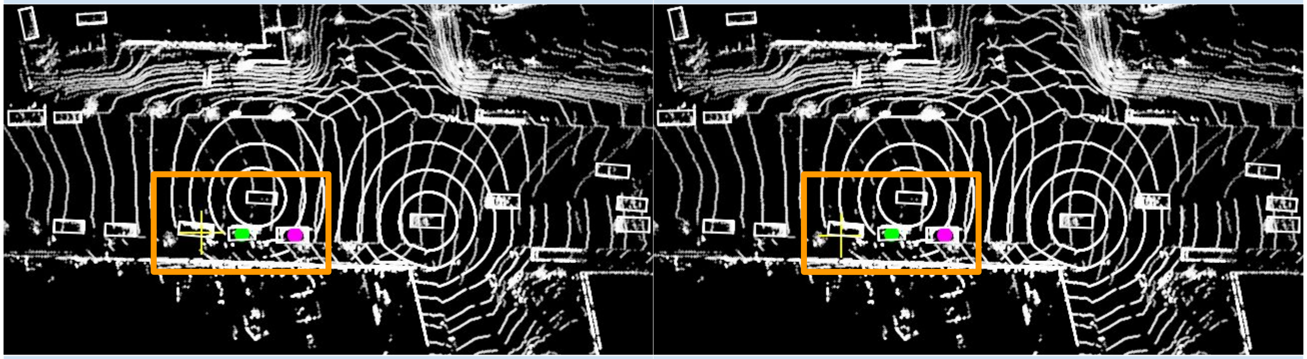
Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front object?
GT: Yes, there is a car behind the front object. Its center location is [30.3, -2.0].



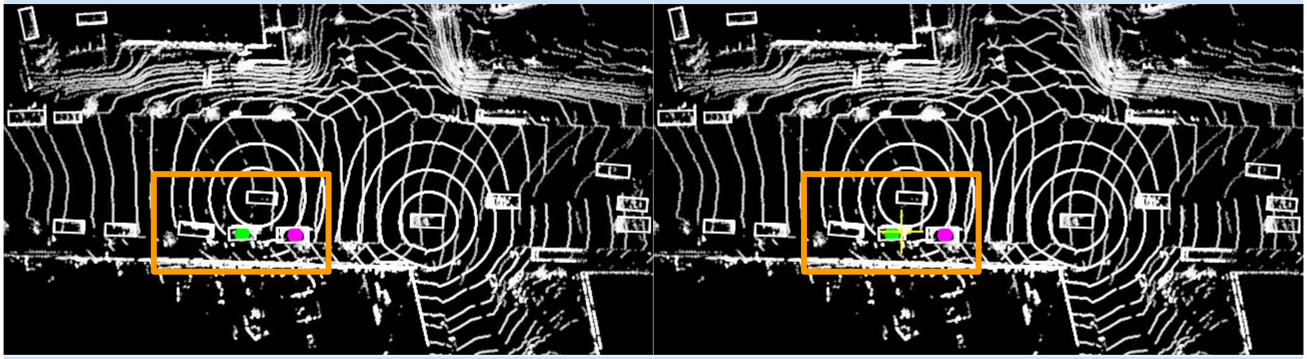
No Fusion: There is nothing behind the front object.

Early Fusion: Yes, there is a car behind the front object. Its center location is [35.9, -1.9].



AttFuse: Yes, there is a car behind the front object. Its center location is [36.9, -1.9].

V2X-ViT: Yes, there is a car behind the front object. Its center location is [38.5, -2.4]



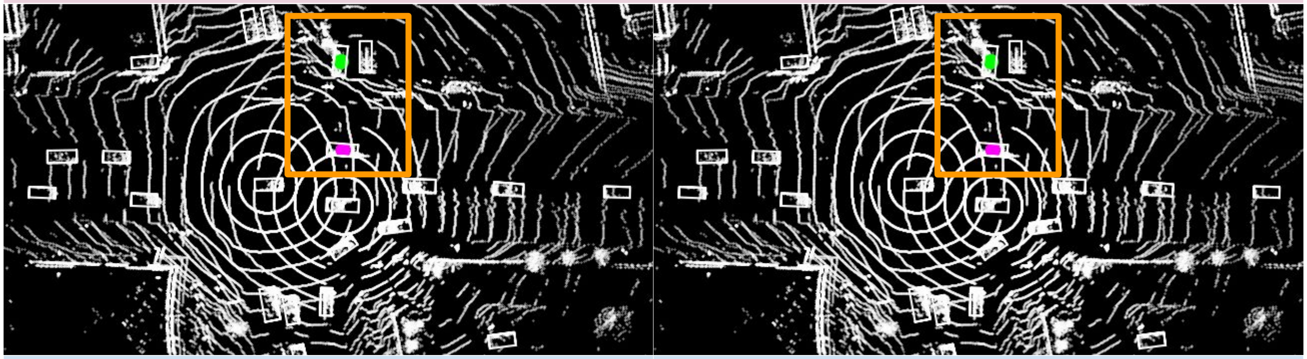
CoBEVT: There is nothing behind the front object.

V2V-LLM: Yes, there is a car behind the front object. Its center location is [28.7, -1.8].

Figure 18. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2V-split testing set. Magenta ○: reference locations in questions. Yellow +: model output locations. Green ○: ground-truth answers.

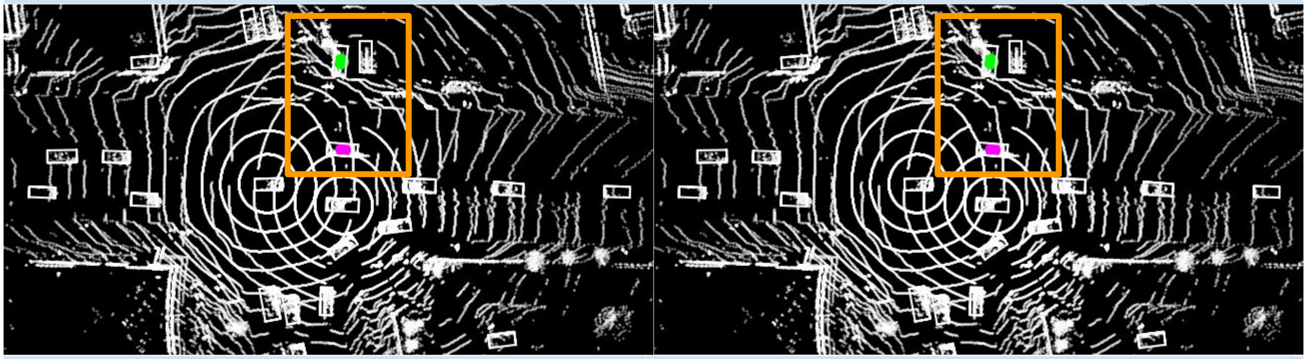
Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front right object?
GT: Yes, there is a car behind the front right object. Its center location is [0.5, 23.3].



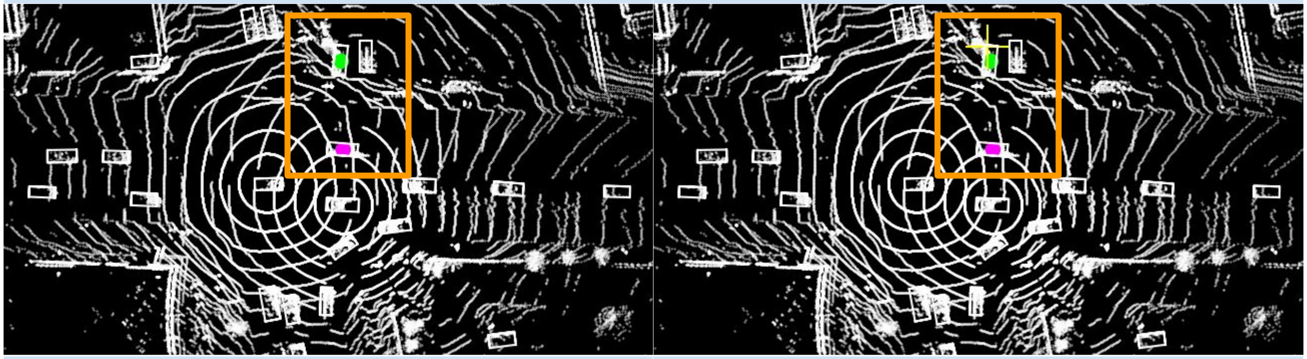
No Fusion: There is nothing behind the front right object.

Early Fusion: There is nothing behind the front right object.



AttFuse: There is nothing behind the front right object.

V2X-ViT: There is nothing behind the front right object.



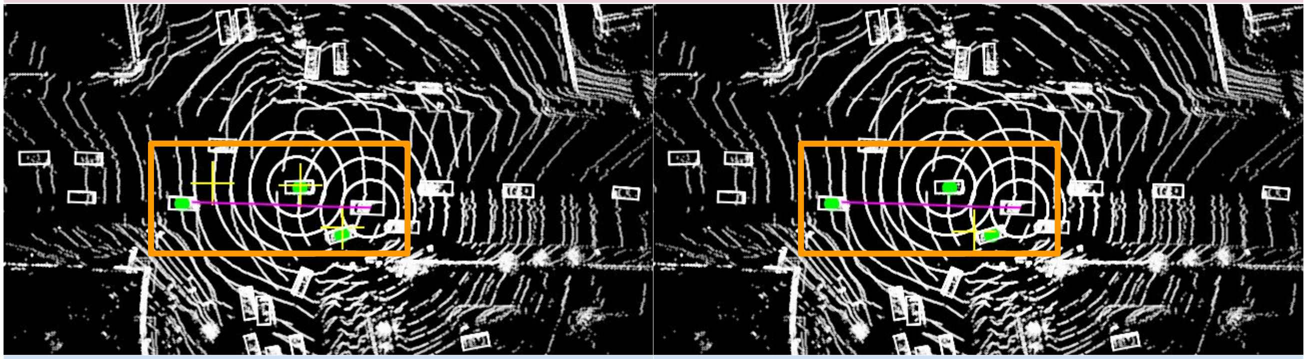
CoBEVT: There is nothing behind the front right object.

V2V-LLM: Yes, there is a car behind the front right object. Its center location is [1.0, 25.6].

Figure 19. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2V-split testing set. Magenta ○: reference locations in questions. Yellow +: model output locations. Green ○: ground-truth answers.

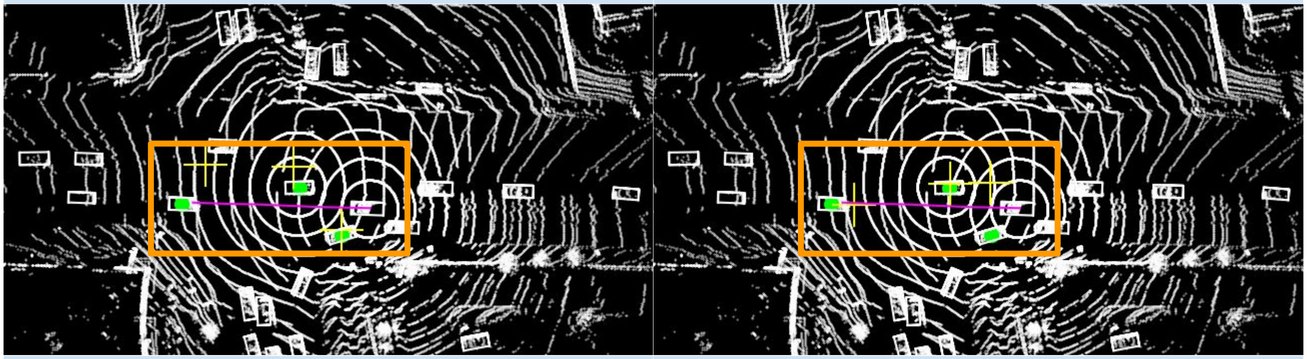
Q4: Notable Object Identification

Q: I am CAV_EGO. Is there anything I need to be aware of if my planned future trajectory is [(4.2,0.1),(8.9,0.2),(13.5,0.4),(18.4,0.6),(23.5,0.7),(28.8,0.9)]?
GT: Yes, there are cars at [4.7, -4.4], [11.5, 3.3], [30.6, 0.7], which are close to your planned future trajectory.



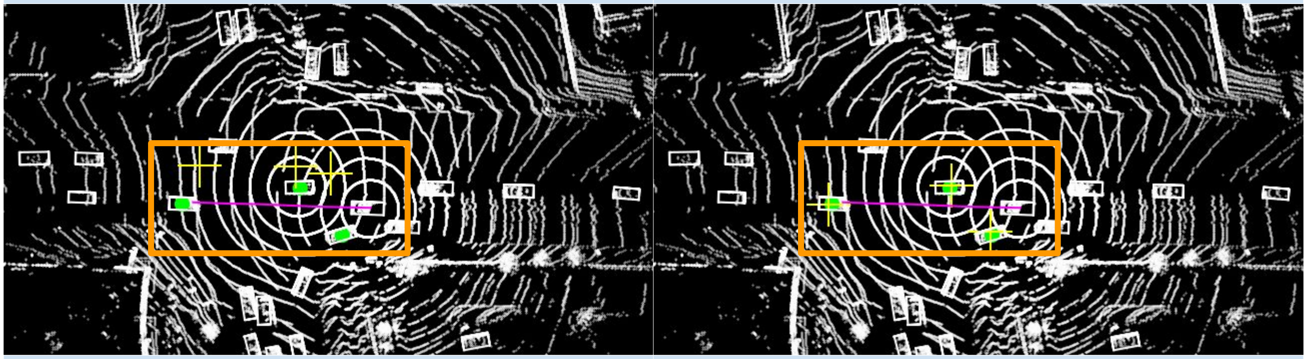
No Fusion: Yes, there are cars at [4.6, -3.2], [11.4, 3.6], [25.7, 4.0], which are close to your planned future trajectory.

Early Fusion: Yes, there is a car at [7.4, -3.8], which is close to your planned future trajectory.



AttFuse: Yes, there are cars at [4.7, -3.6], [12.5, 6.7], [26.7, 7.1], which are close to your planned future trajectory.

V2X-ViT: Yes, there are cars at [4.8, 4.1], [11.4, 3.9], [26.9, 0.5], which are close to your planned future trajectory.



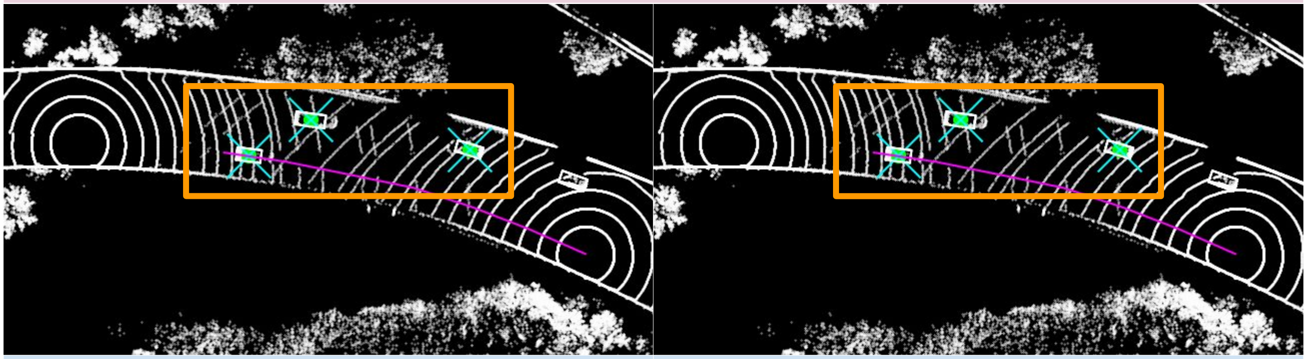
CoBEVT: Yes, there are cars at [6.4, 5.6], [12.2, 6.7], [27.7, 6.9], which are close to your planned future trajectory.

V2V-LLM: Yes, there are cars at [4.8, -3.8], [11.1, 3.6], [31.1, 0.5], which are close to your planned future trajectory.

Figure 20. V2V-LLM and baseline methods' *notable object identification* results on V2V-QA's V2V-split testing set. Magenta curve: planned future trajectories in questions. Green ○: ground-truth notable object locations. Yellow +: model identification outputs.

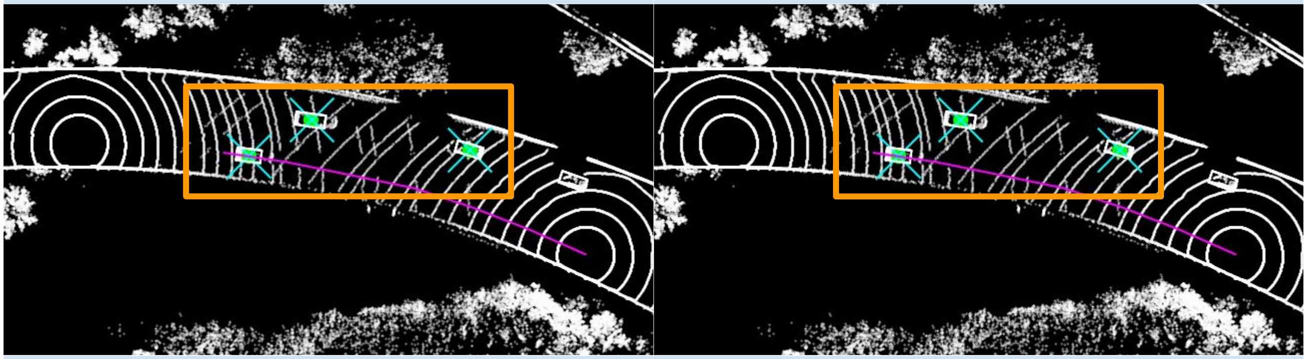
Q4: Notable Object Identification

Q: I am CAV_1. Is there anything I need to be aware of if my planned future trajectory is [(-73.0,-14.0),(-63.7,-10.4),(-54.2,-7.3),(-43.5,-5.1),(-33.6,-3.1),(-23.5,-1.6)]?
GT: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.



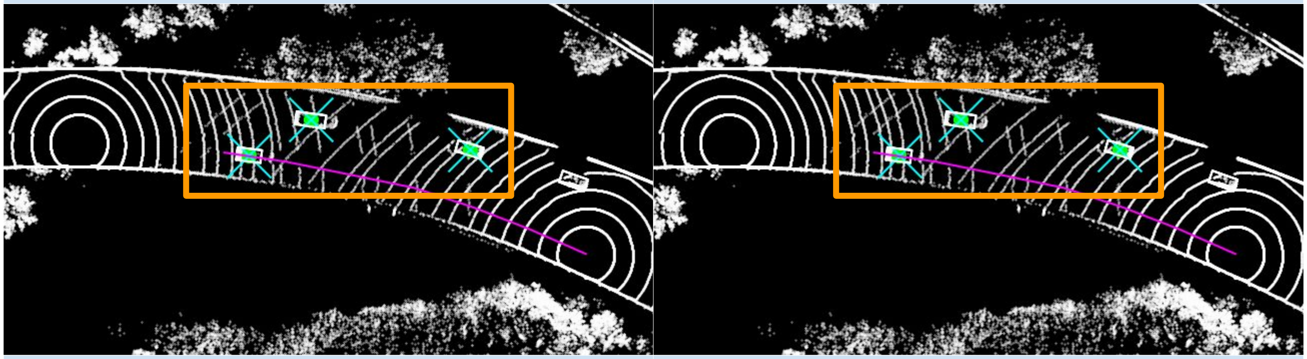
No Fusion: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.

Early Fusion: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.



AttFuse: Yes, there are cars at [-27.5, -2.0], [-37.7, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.

V2X-ViT: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.



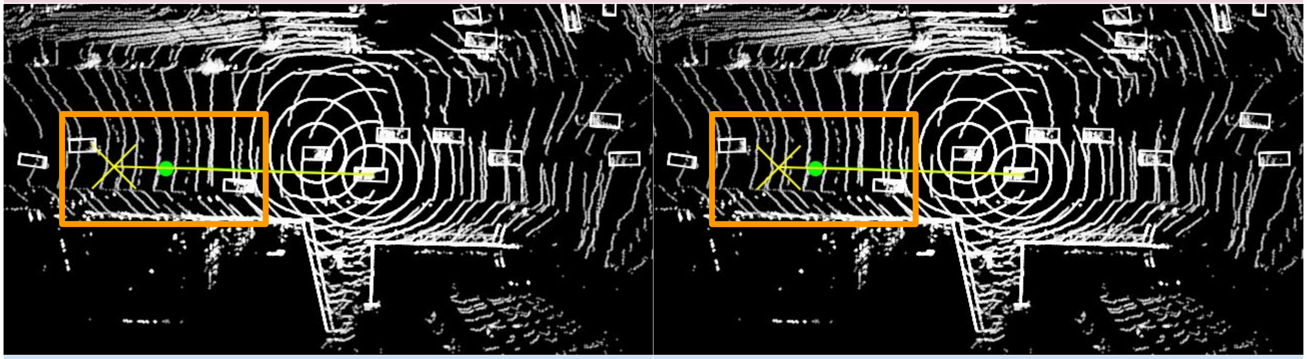
CoBEVT: Yes, there are cars at [-27.5, -2.0], [-37.5, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.

V2V-LLM: Yes, there are cars at [-27.6, -2.0], [-37.6, 3.7], [-63.3, -1.1], which are close to your planned future trajectory.

Figure 21. V2V-LLM and baseline methods' *notable object identification* results on V2V-QA's V2V-split testing set. Magenta curve: planned future trajectories in questions. Green ○: ground-truth notable object locations. Cyan ×: model identification outputs.

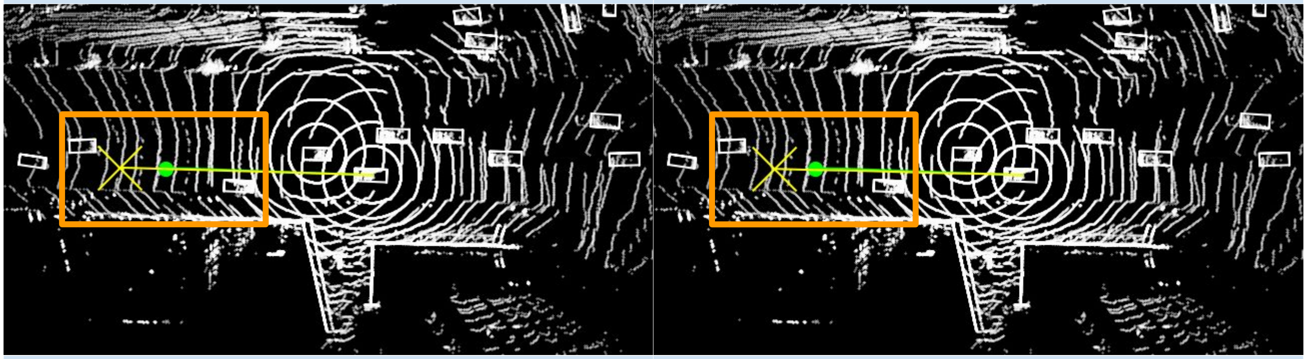
Q5: Planning

Q: I am CAV_EGO. What is the suggested future trajectory to avoid collision with nearby objects?
GT: The suggested future trajectory is [(6.2,0.2),(12.1,0.4),(17.6,0.6),(22.9,0.8),(28.3,1.0),(33.5,1.1)].



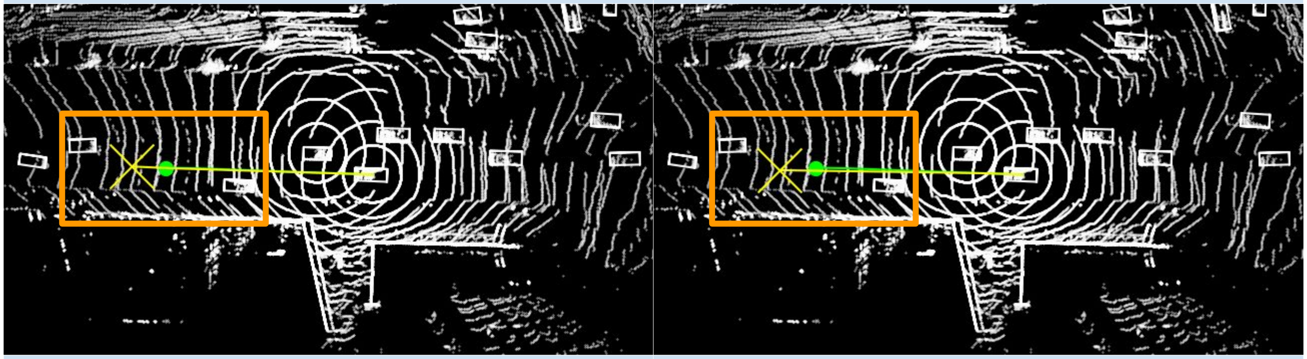
No Fusion: The suggested future trajectory is [(7.2,0.2),(14.6,0.4),(21.4,0.6),(28.4,0.9),(35.2,1.1),(42.0,1.3)].

Early Fusion: The suggested future trajectory is [(6.2,0.2),(12.7,0.4),(19.3,0.6),(25.9,0.8),(32.7,1.0),(39.6,1.2)].



AttFuse: The suggested future trajectory is [(6.4,0.2),(13.1,0.4),(20.1,0.6),(27.4,0.8),(34.2,1.0),(41.1,1.2)].

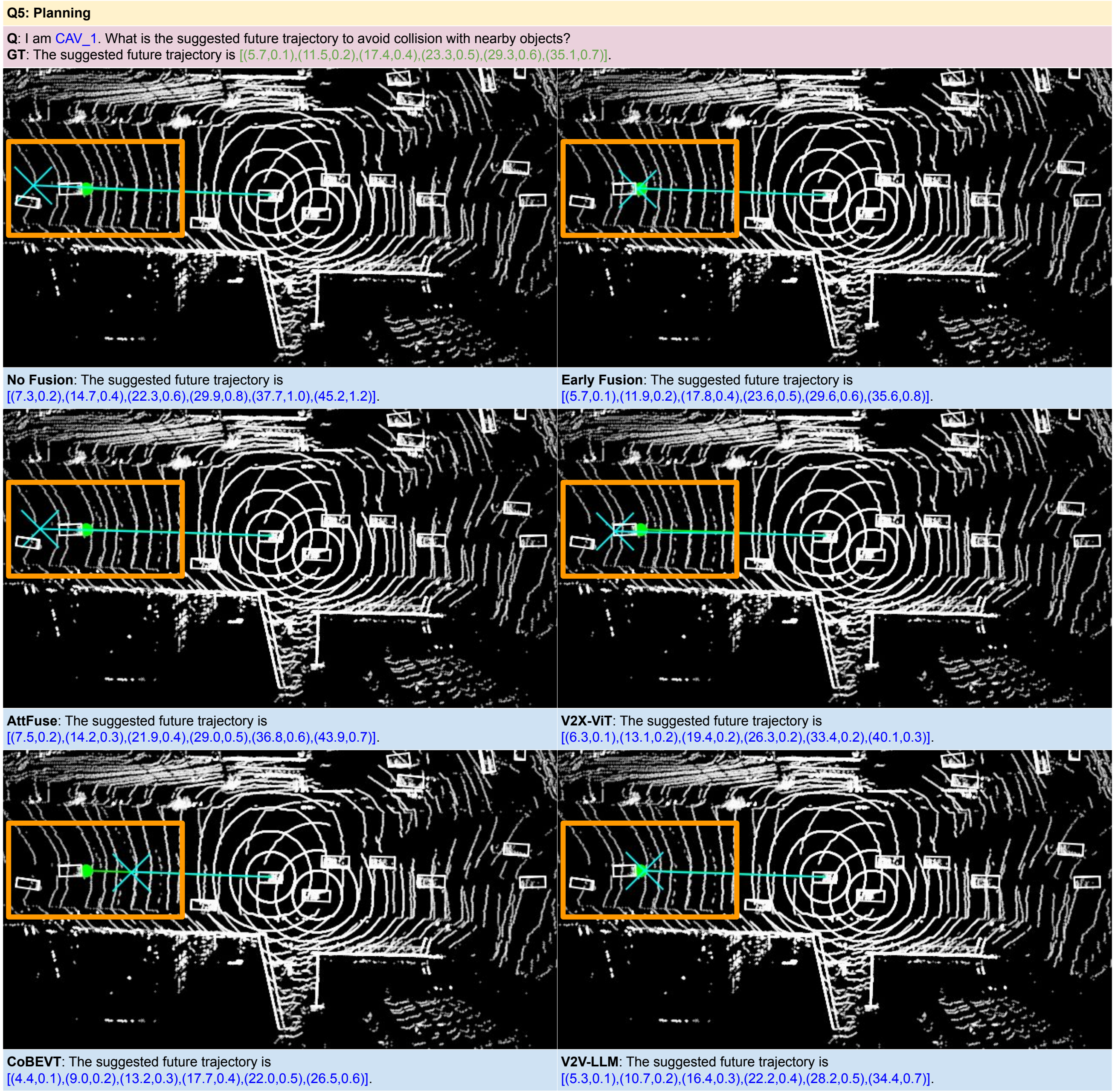
V2X-ViT: The suggested future trajectory is [(6.4,0.2),(13.0,0.3),(20.1,0.5),(26.7,0.7),(33.5,0.8),(40.3,1.0)].



CoBEVT: The suggested future trajectory is [(6.3,0.2),(12.9,0.4),(19.4,0.6),(26.0,0.8),(32.5,1.0),(39.1,1.3)].

V2V-LLM: The suggested future trajectory is [(6.2,0.2),(12.6,0.3),(19.2,0.4),(25.9,0.5),(32.6,0.6),(39.4,0.7)].

Figure 22. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2V-split testing set. Green curve: future trajectories in ground-truth answers. Green ○: ending waypoints in ground-truth answers. Yellow curve: model planning outputs. Yellow ×: ending waypoints in model outputs.



Q5: Planning

Q: I am CAV_1. What is the suggested future trajectory to avoid collision with nearby objects?
GT: The suggested future trajectory is [(5.7,0.1),(11.5,0.2),(17.4,0.4),(23.3,0.5),(29.3,0.6),(35.1,0.7)].

No Fusion: The suggested future trajectory is [(7.3,0.2),(14.7,0.4),(22.3,0.6),(29.9,0.8),(37.7,1.0),(45.2,1.2)].

Early Fusion: The suggested future trajectory is [(5.7,0.1),(11.9,0.2),(17.8,0.4),(23.6,0.5),(29.6,0.6),(35.6,0.8)].

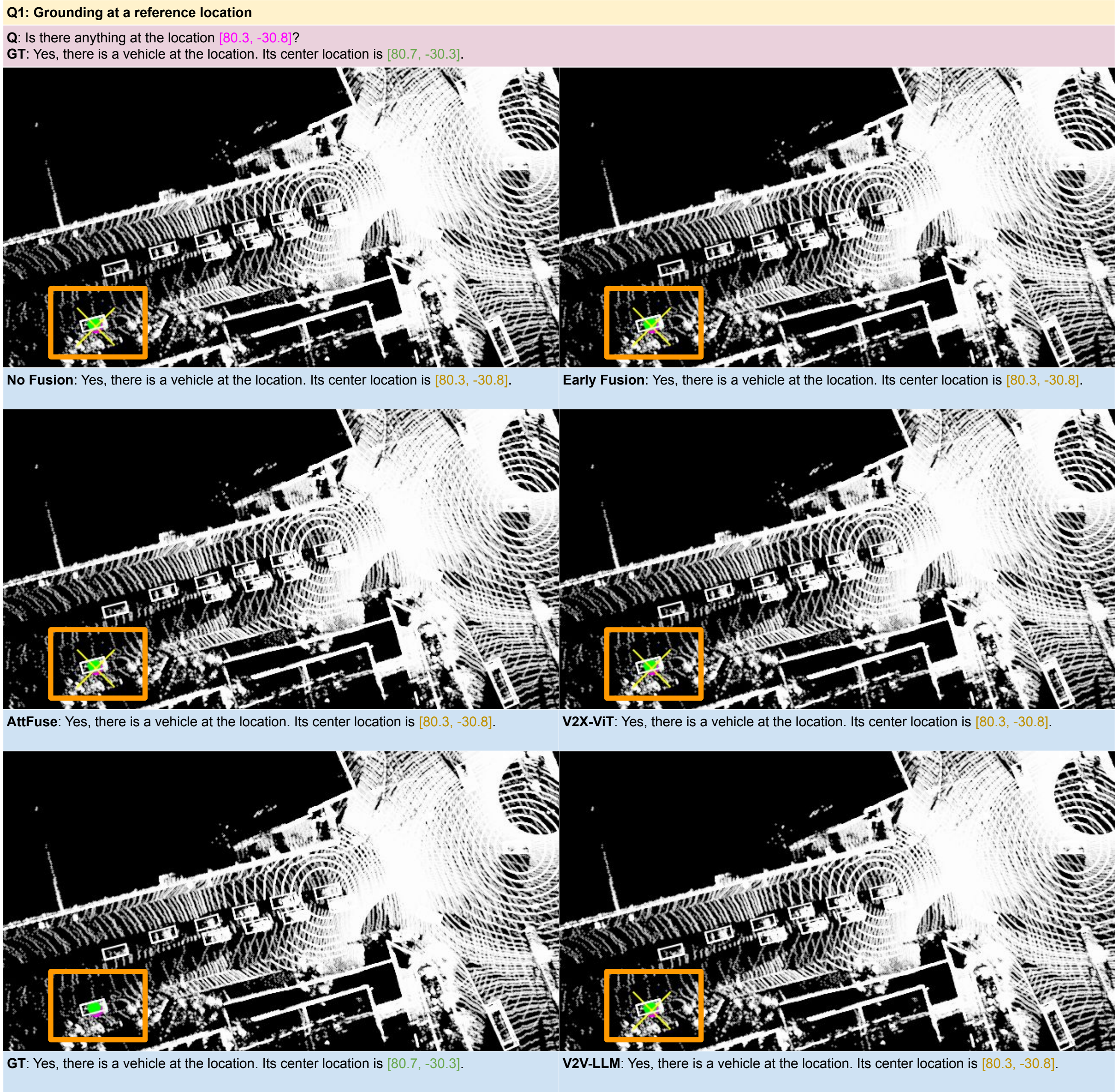
AttFuse: The suggested future trajectory is [(7.5,0.2),(14.2,0.3),(21.9,0.4),(29.0,0.5),(36.8,0.6),(43.9,0.7)].

V2X-ViT: The suggested future trajectory is [(6.3,0.1),(13.1,0.2),(19.4,0.2),(26.3,0.2),(33.4,0.2),(40.1,0.3)].

CoBEVT: The suggested future trajectory is [(4.4,0.1),(9.0,0.2),(13.2,0.3),(17.7,0.4),(22.0,0.5),(26.5,0.6)].

V2V-LLM: The suggested future trajectory is [(5.3,0.1),(10.7,0.2),(16.4,0.3),(22.2,0.4),(28.2,0.5),(34.4,0.7)].

Figure 23. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2V-split testing set. Green curve: future trajectories in ground-truth answers. Green ○: ending waypoints in ground-truth answers. Cyan curve: model planning outputs. Cyan ×: ending waypoints in model outputs.



Q1: Grounding at a reference location

Q: Is there anything at the location [80.3, -30.8]?
GT: Yes, there is a vehicle at the location. Its center location is [80.7, -30.3].

No Fusion: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

Early Fusion: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

AttFuse: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

V2X-ViT: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

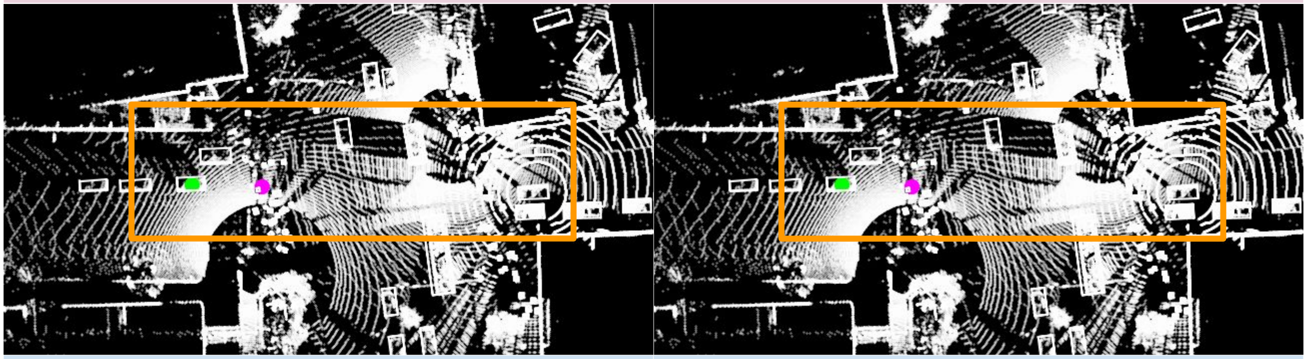
GT: Yes, there is a vehicle at the location. Its center location is [80.7, -30.3].

V2V-LLM: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

Figure 24. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2X-split testing set. Magenta ○: reference locations in questions. Yellow ×: model output locations. Green ○: ground-truth answers.

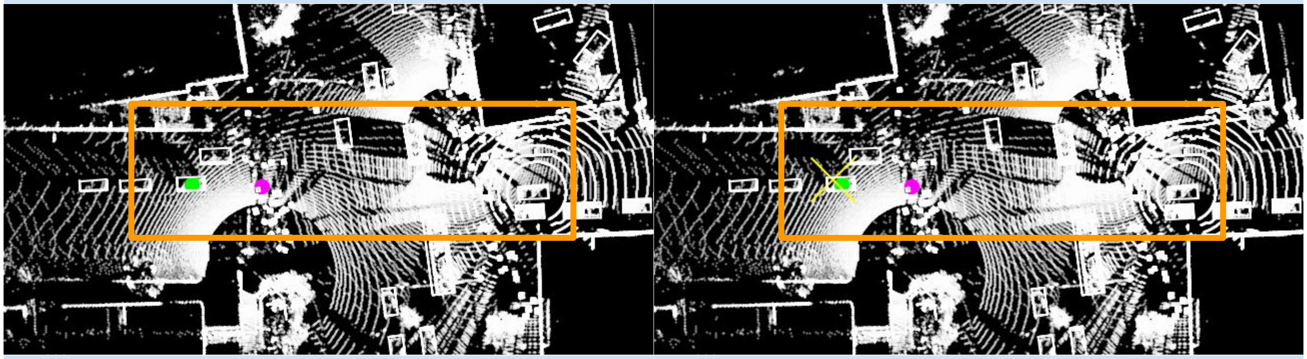
Q2: Grounding behind a reference object at a location

Q: Is there anything behind the object at [44.5, 0.9]?
GT: Yes, there is a vehicle behind the object. Its center location is [55.9, 1.2].



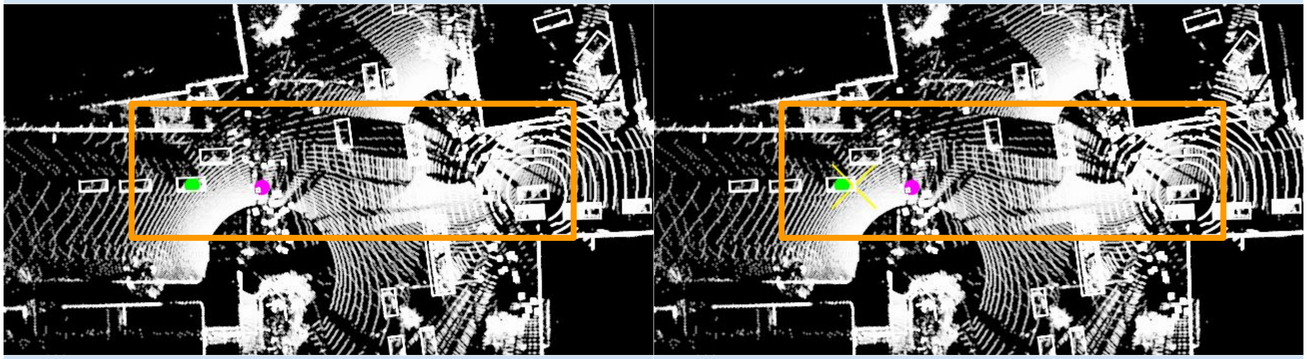
No Fusion: There is nothing behind the object.

Early Fusion: There is nothing behind the object.



AttFuse: There is nothing behind the object.

V2X-ViT: Yes, there is a vehicle behind the object. Its center location is [57.2, 2.0].



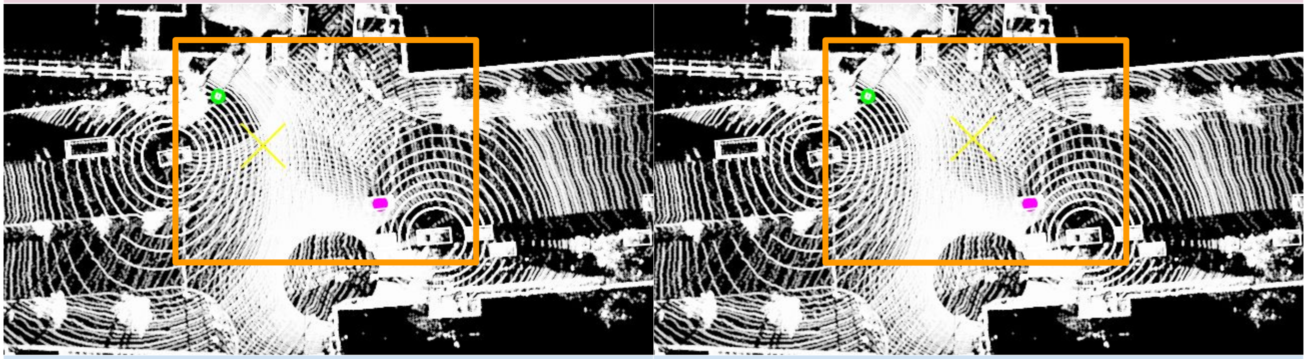
GT: Yes, there is a vehicle behind the object. Its center location is [55.9, 1.2].

V2V-LLM: Yes, there is a vehicle behind the object. Its center location is [53.8, 1.0].

Figure 25. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2X-split testing set. Magenta ○: reference locations in questions. Yellow ×: model output locations. Green ○: ground-truth answers.

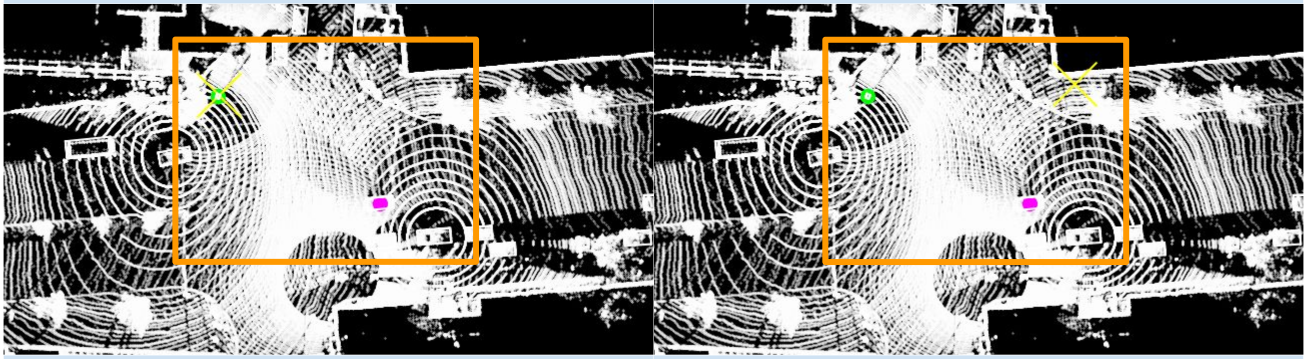
Q3: Grounding behind a reference object in a direction

Q: Is there anything behind the front right object?
GT: Yes, there is a pedestrian behind the front right object. Its center location is [35.3, 22.7].



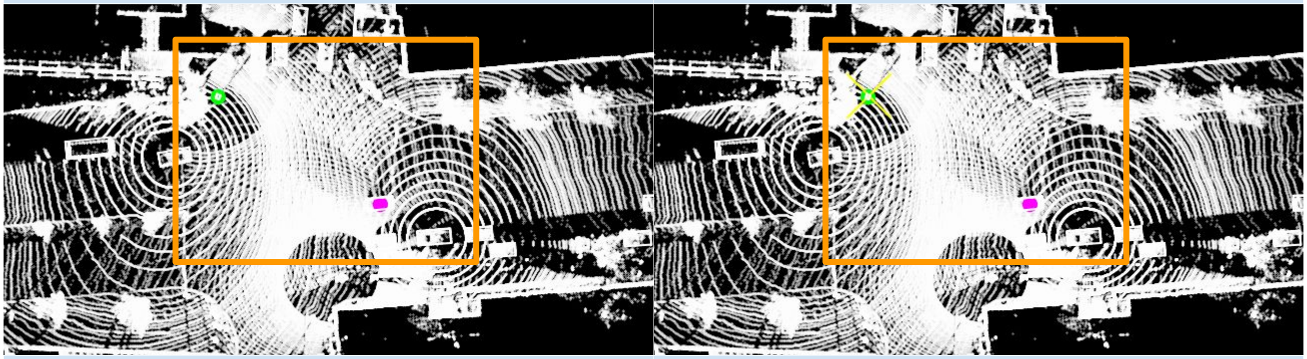
No Fusion: Yes, there is a pedestrian behind the front right object. Its center location is [28.1, 14.8].

Early Fusion: Yes, there is a pedestrian behind the front right object. Its center location is [18.3, 16.0].



AttFuse: Yes, there is a pedestrian behind the front right object. Its center location is [35.2, 23.0].

V2X-ViT: Yes, there is a pedestrian behind the front right object. Its center location is [1.7, 24.7].



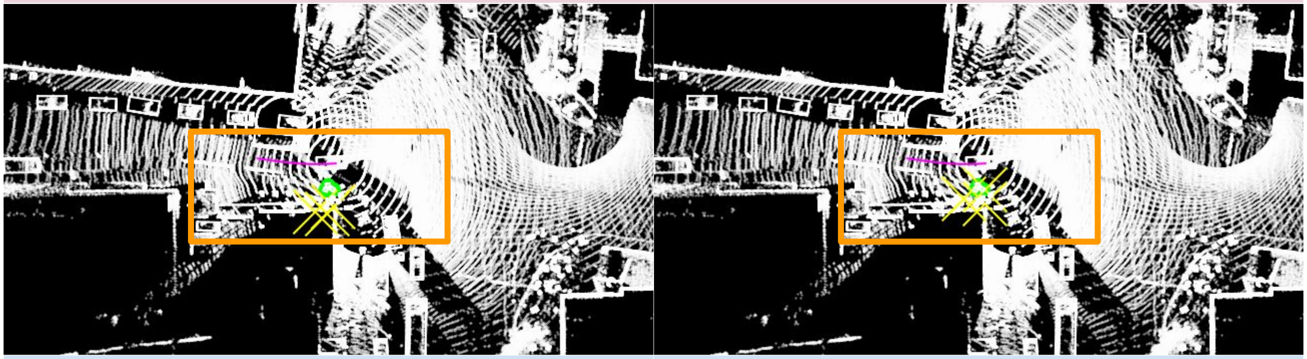
GT: Yes, there is a pedestrian behind the front right object. Its center location is [35.3, 22.7].

V2V-LLM: Yes, there is a pedestrian behind the front right object. Its center location is [35.2, 23.0].

Figure 26. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2X-split testing set. Magenta ○: reference locations in questions. Yellow ×: model output locations. Green ○: ground-truth answers.

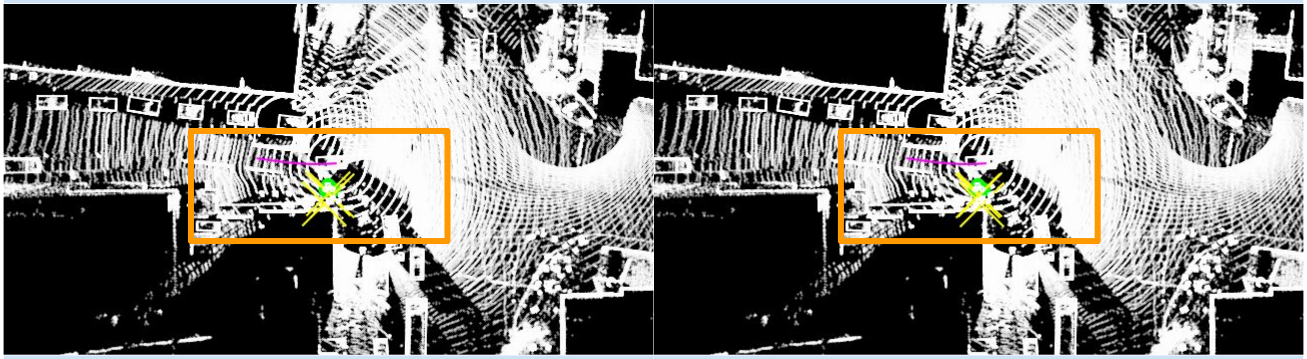
Q4: Notable Object Identification

Q: I am CAV_EGO. Is there anything I need to be aware of if my planned future trajectory is [(2.0,-0.1),(4.1,-0.1),(6.1,0.0),(8.2,0.2),(10.4,0.5),(12.6,0.8)]?
GT: Yes, there are pedestrians at [1.3, -3.5], [0.5, -4.2], [1.6, -4.2] close to your planned future trajectory.



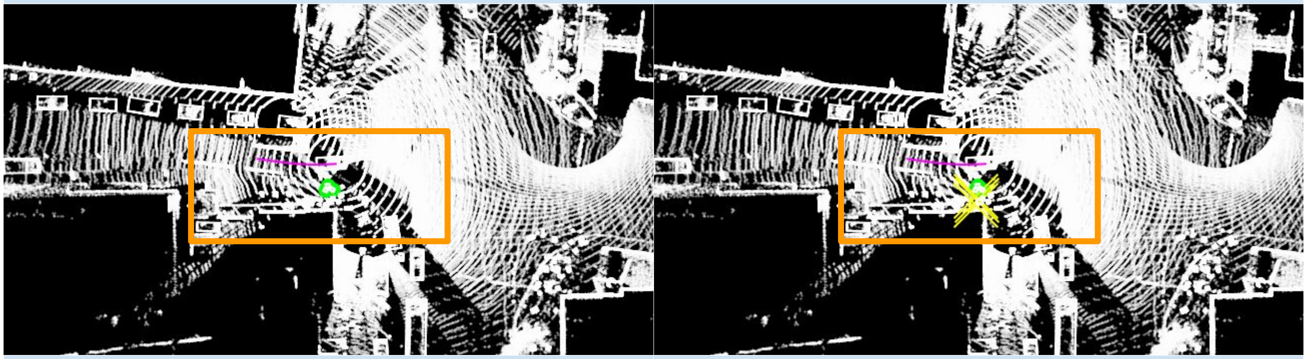
No Fusion: Yes, there are pedestrians at [0.5, -7.0], [3.4, -7.7], [1.4, -8.1] close to your planned future trajectory.

Early Fusion: Yes, there are pedestrians at [0.1, -4.1], [0.2, -5.8], [2.9, -5.7] close to your planned future trajectory.



AttFuse: Yes, there are pedestrians at [1.3, -4.9], [0.6, -5.1], [2.0, -6.5] close to your planned future trajectory.

V2X-ViT: Yes, there are cars at [0.9, -5.0], [1.2, -5.2], [0.7, -6.6] close to your planned future trajectory.



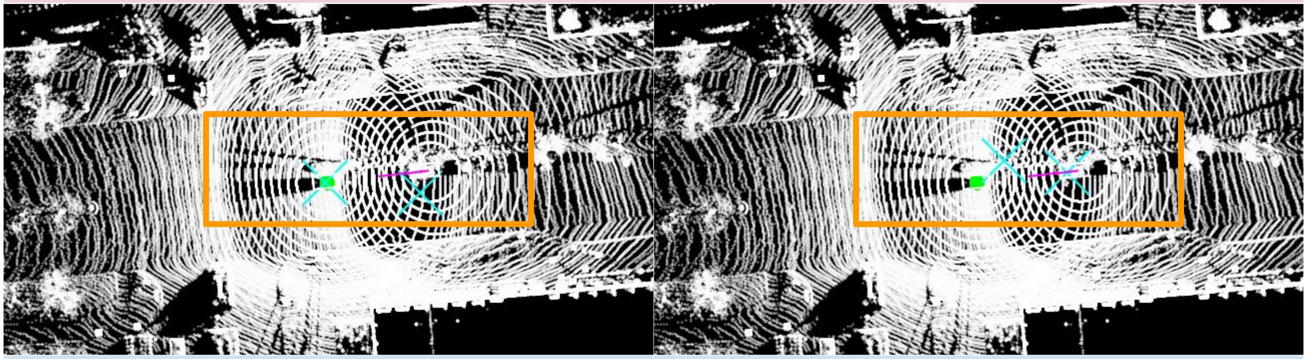
GT: Yes, there are pedestrians at [1.3, -3.5], [0.5, -4.2], [1.6, -4.2] close to your planned future trajectory.

V2V-LLM: Yes, there are pedestrians at [1.6, -5.2], [1.4, -5.8], [1.6, -6.6] close to your planned future trajectory.

Figure 27. V2V-LLM and baseline methods' *notable object identification* results on V2V-QA's V2X-split testing set. Magenta curve: planned future trajectories in questions. Green ○: ground-truth notable object locations. Yellow ×: model identification outputs.

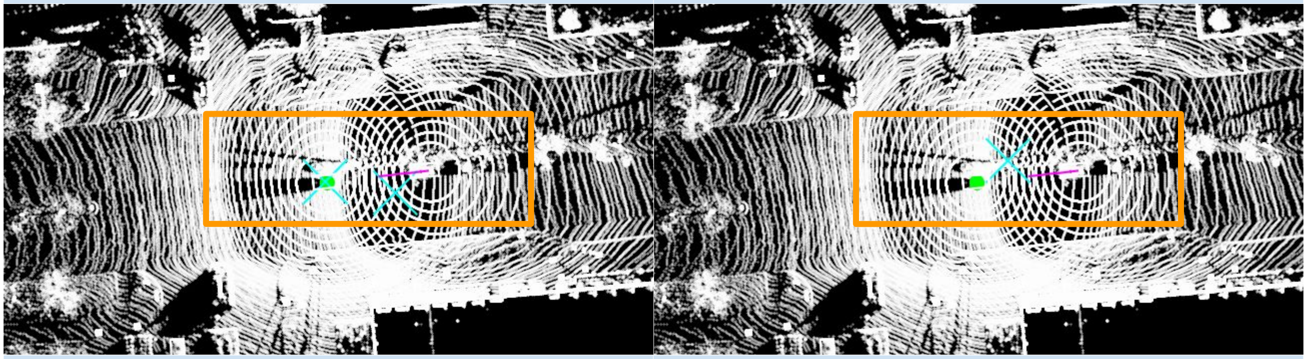
Q4: Notable Object Identification

Q: I am CAV_1. Is there anything I need to be aware of if my planned future trajectory is [(-15.1,1.7),(-13.5,1.5),(-12.1,1.4),(-10.8,1.2),(-9.7,1.1),(-8.9,1.0)]?
GT: Yes, there is a vehicle at [-0.3, 0.1] close to your planned future trajectory.



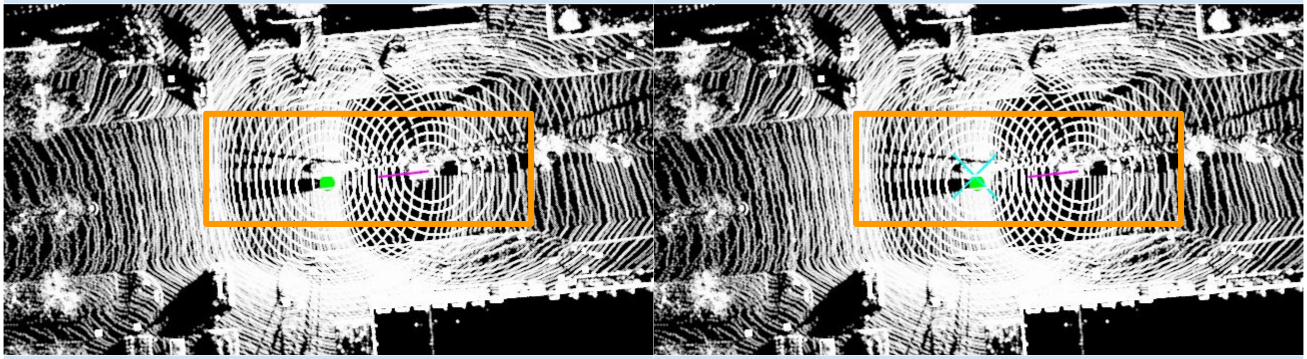
No Fusion: Yes, there are vehicles at [-0.1, 0.1], [-15.4, -1.6] close to your planned future trajectory.

Early Fusion: Yes, there are vehicles at [-9.1, 1.0], [-15.7, -1.6] close to your planned future trajectory.



AttFuse: Yes, there are vehicles at [-0.1, 0.2], [-11.4, -1.6] close to your planned future trajectory.

V2X-ViT: Yes, there is a vehicle at [-5.4, 3.6] close to your planned future trajectory.



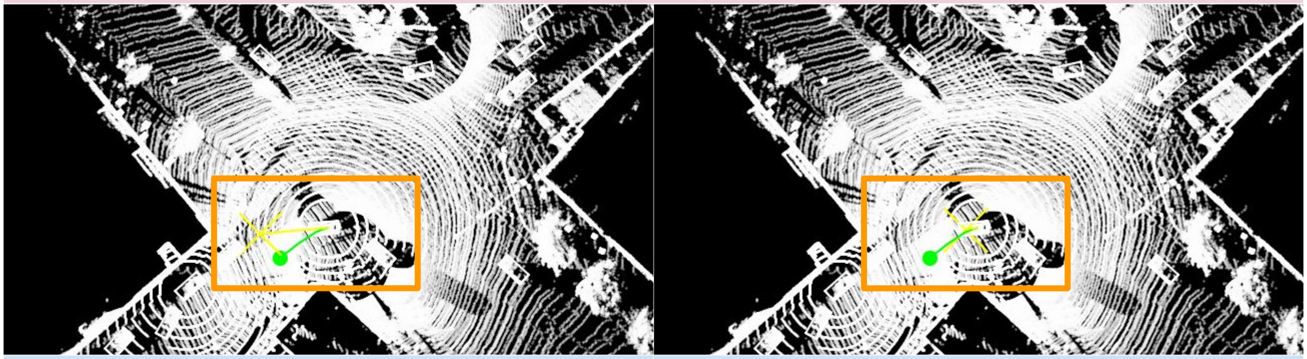
GT: Yes, there is a vehicle at [-0.3, 0.1] close to your planned future trajectory.

V2V-LLM: Yes, there is a vehicle at [-0.1, 1.2] close to your planned future trajectory.

Figure 28. V2V-LLM and baseline methods' *notable object identification* results on V2V-QA's V2X-split testing set. Magenta curve: planned future trajectories in questions. Green ○: ground-truth notable object locations. Cyan ×: model identification outputs.

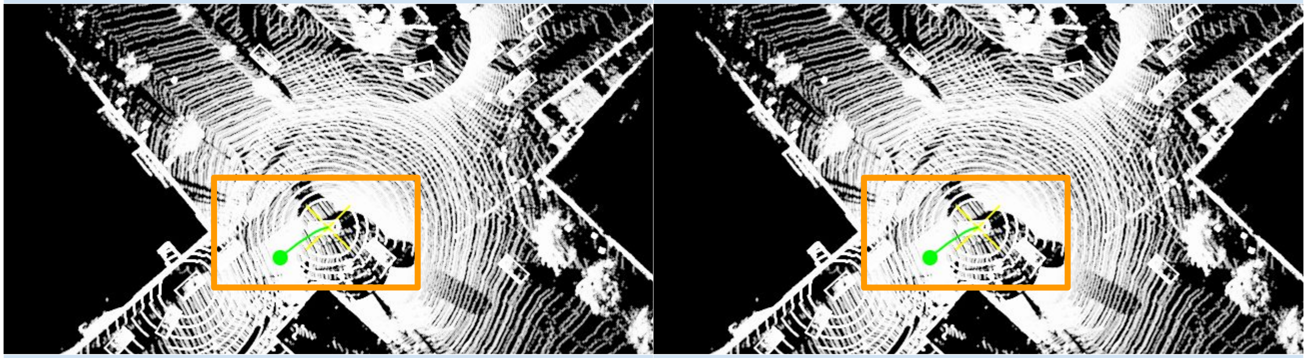
Q5: Planning

Q: I am CAV_EGO. What is the suggested future trajectory to avoid collision with nearby objects?
GT: The suggested future trajectory is [(1.0,-0.3),(2.1,-0.8),(3.3,-1.6),(4.7,-2.5),(6.1,-3.6),(7.7,-4.8)].



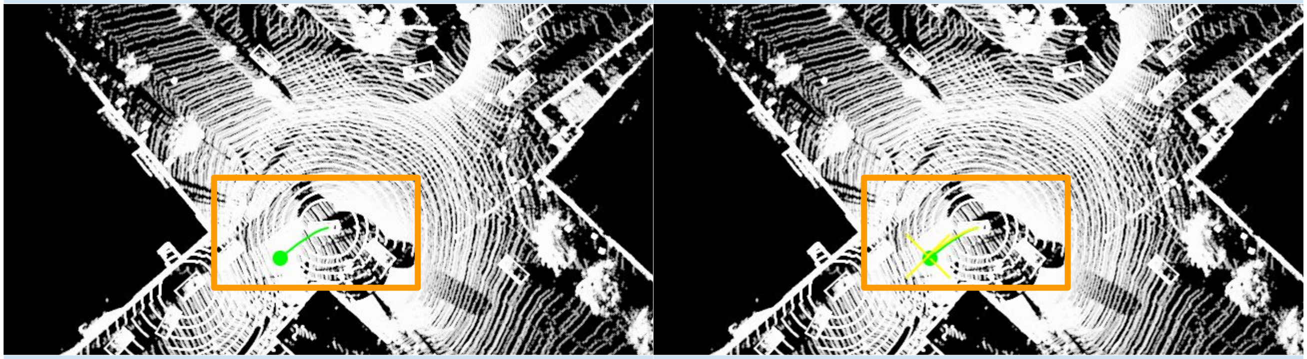
No Fusion: The suggested future trajectory is
[(1.5,-0.1),(3.2,-0.3),(5.0,-0.4),(6.9,-0.6),(8.9,-0.8),(11.0,-1.0)].

Early Fusion: The suggested future trajectory is
[(0.3,-0.0),(0.6,-0.1),(0.9,-0.1),(1.2,-0.1),(1.6,-0.2),(2.0,-0.2)].



AttFuse: The suggested future trajectory is
[(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0)].

V2X-ViT: The suggested future trajectory is
[(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0)].



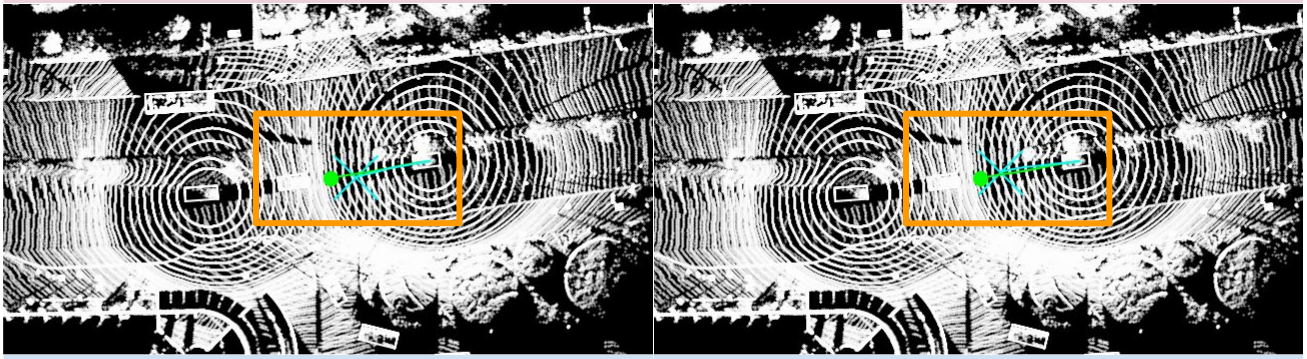
GT: The suggested future trajectory is
[(1.0,-0.3),(2.1,-0.8),(3.3,-1.6),(4.7,-2.5),(6.1,-3.6),(7.7,-4.8)].

V2V-LLM: The suggested future trajectory is
[(1.1,-0.3),(2.3,-0.8),(3.6,-1.5),(5.0,-2.4),(6.5,-3.4),(8.1,-4.5)].

Figure 29. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2X-split testing set. Green curve: future trajectories in ground-truth answers. Green ○: ending waypoints in ground-truth answers. Yellow curve: model planning outputs. Yellow ×: ending waypoints in model outputs.

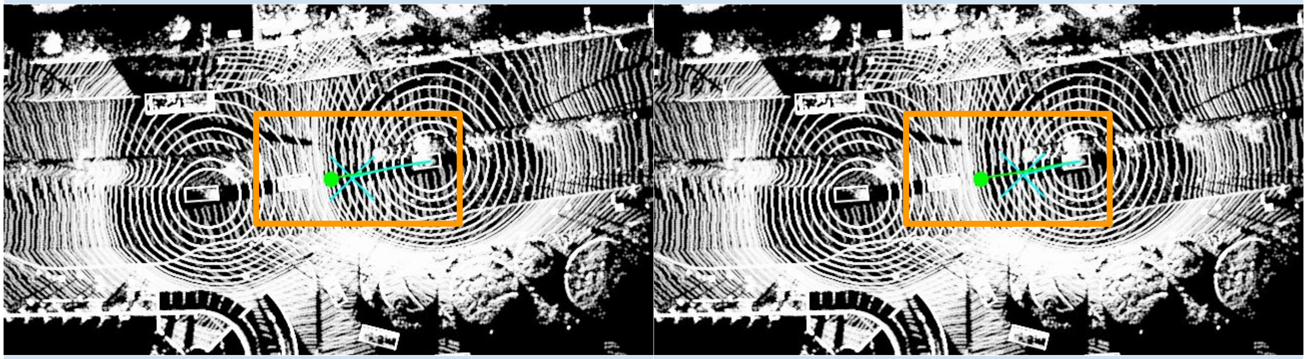
Q5: Planning

Q: I am CAV_1. What is the suggested future trajectory to avoid collision with nearby objects?
GT: The suggested future trajectory is [(2.3,0.1),(4.8,0.1),(7.4,0.0),(10.3,0.0),(13.3,0.1),(16.3,0.1)].



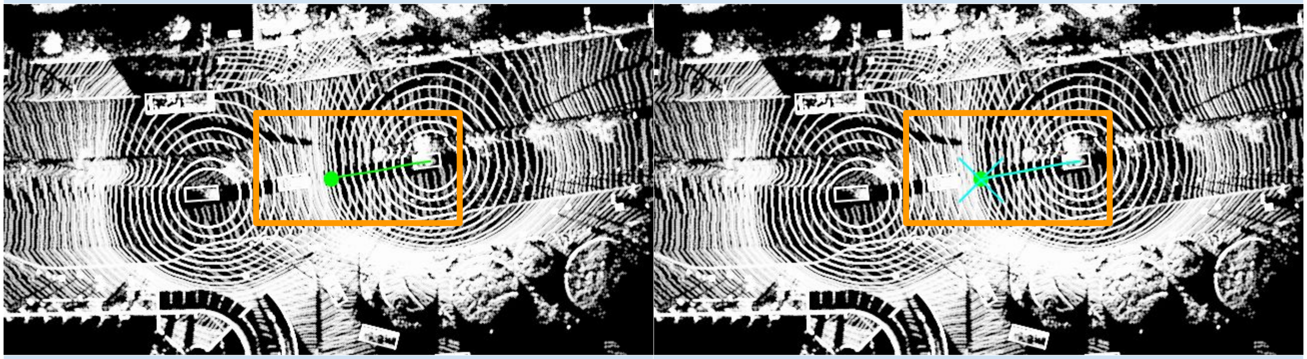
No Fusion: The suggested future trajectory is
[(2.2,0.0),(4.4,0.0),(6.5,-0.0),(8.5,-0.2),(10.5,-0.3),(12.3,-0.5)].

Early Fusion: The suggested future trajectory is
[(2.2,0.1),(4.4,0.3),(6.6,0.4),(8.8,0.6),(10.9,0.6),(12.9,0.5)].



AttFuse: The suggested future trajectory is
[(2.2,0.0),(4.4,0.1),(6.6,0.1),(8.8,-0.0),(10.9,-0.2),(13.0,-0.3)].

V2X-ViT: The suggested future trajectory is
[(1.9,-0.0),(3.7,-0.1),(5.3,-0.2),(6.8,-0.3),(8.2,-0.4),(9.5,-0.5).



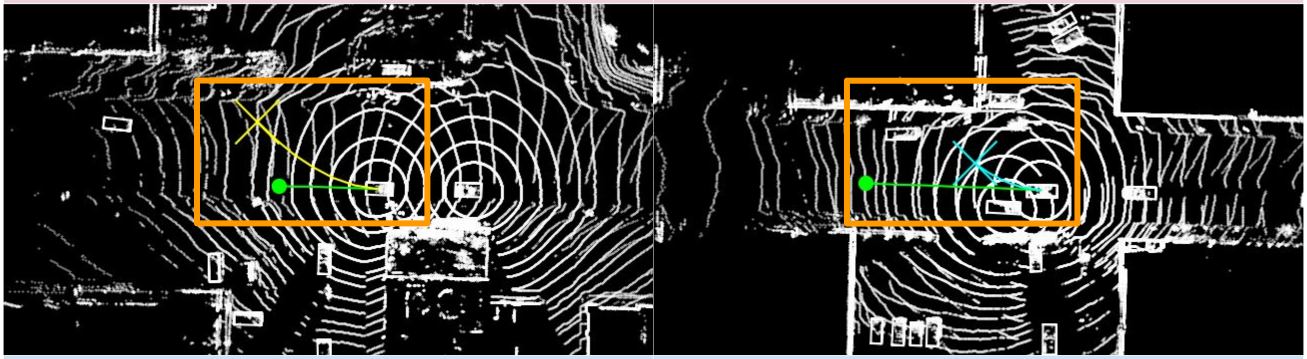
GT: The suggested future trajectory is
[(2.3,0.1),(4.8,0.1),(7.4,0.0),(10.3,0.0),(13.3,0.1),(16.3,0.1)].

V2V-LLM: The suggested future trajectory is
[(2.7,0.1),(5.3,0.1),(8.0,0.1),(10.8,0.0),(13.6,-0.0),(16.4,-0.1)].

Figure 30. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2X-split testing set. Green curve: future trajectories in ground-truth answers. Green ○: ending waypoints in ground-truth answers. Cyan curve: model planning outputs. Cyan ×: ending waypoints in model outputs.

Q5: Planning

Q: I am CAV_EGO. What is the suggested future trajectory to avoid collision with nearby objects?	**Q**: I am CAV_1. What is the suggested future trajectory to avoid collision with nearby objects?
GT: The suggested future trajectory is [(2.5,0.1),(5.2,0.2),(7.9,0.3),(10.6,0.4),(13.4,0.4),(16.1,0.5)].	**GT**: The suggested future trajectory is [(4.7,0.1),(9.2,0.2),(14.0,0.2),(18.6,0.3),(23.6,0.3),(28.5,0.4)].



V2V-LLM: The suggested future trajectory is [(3.7,0.6),(7.2,1.8),(10.4,3.5),(13.6,5.6),(16.6,8.1),(19.6,11.0)].	**V2V-LLM**: The suggested future trajectory is [(1.5,0.2),(3.1,0.4),(4.8,0.9),(6.6,1.6),(8.7,2.7),(10.9,4.1)].

Figure 31. Failure cases of V2V-LLM's *planning* results on V2V-QA's testing set. Green curve: future trajectories in ground-truth answers. Green ○: ending waypoints in ground-truth answers. Yellow curve and Cyan curve: model planning outputs corresponding to CAV_EGO and CAV_1, respectively. Yellow × and Cyan ×: ending waypoints in model outputs corresponding to CAV_EGO and CAV_1, respectively.