

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

Hsu-kuang Chiu^{1,2} Ryo Hachiuma¹ Chien-Yi Wang¹ Stephen F. Smith² Yu-Chiang Frank Wang¹
Min-Hung Chen¹

¹NVIDIA, ²Carnegie Mellon University

Abstract

Current autonomous driving vehicles rely mainly on their individual sensors to understand surrounding scenes and plan for future trajectories, which can be unreliable when the sensors are malfunctioning or occluded. To address this problem, cooperative perception methods via vehicle-to-vehicle (V2V) communication have been proposed, but they have tended to focus on perception tasks like detection or tracking. How those approaches contribute to overall cooperative planning performance is still under-explored. Inspired by recent progress using Large Language Models (LLMs) to build autonomous driving systems, we propose a novel problem setting that integrates a Multi-Modal LLM into cooperative autonomous driving, with the proposed Vehicle-to-Vehicle Question-Answering (V2V-QA) dataset and benchmark. We also propose our baseline method Vehicle-to-Vehicle Multi-Modal Large Language Model (V2V-LLM), which uses an LLM to fuse perception information from multiple connected autonomous vehicles (CAVs) and answer various types of driving-related questions: grounding, notable object identification, and planning. Experimental results show that our proposed V2V-LLM can be a promising unified model architecture for performing various tasks in cooperative autonomous driving, and outperforms other baseline methods that use different fusion approaches. Our work also creates a new research direction that can improve the safety of future autonomous driving systems. The code and data will be released to the public to facilitate open-source research in this field.

1. Introduction

Autonomous driving technology has advanced significantly due to the evolution of deep learning algorithms, computing infrastructures, and the release of large-scale real-world driving datasets and benchmarks [3, 13, 38]. However, the perception and planning systems of autonomous vehicles

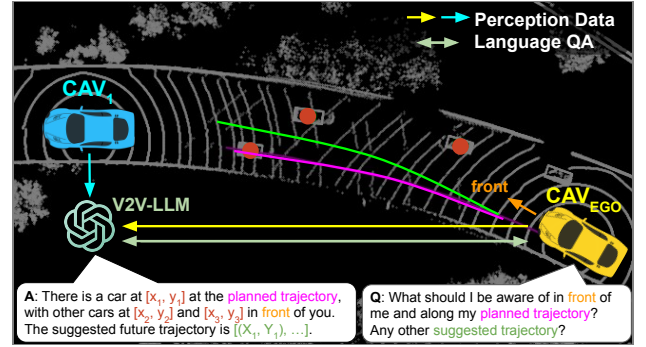


Figure 1. Overview of our problem setting of LLM-based cooperative autonomous driving. All CAVs share their perception information with the LLM. Any CAV can ask the LLM a question to obtain useful information for driving safety.

in daily operation rely mainly on their local LiDAR sensors and cameras to detect notable nearby objects and plan for future trajectories. This approach may encounter safety-critical problems when the sensors are occluded by nearby large objects. In such situations, autonomous driving vehicles are unable to accurately detect all nearby notable objects, making subsequent planning results unreliable.

To address this safety-critical issue, recent research proposes cooperative perception algorithms [6, 9, 44, 50–52] via vehicle-to-vehicle (V2V) communication. In cooperative driving scenarios, multiple *Connected Autonomous Vehicles* (CAVs) driving nearby to each other share their perception information via V2V communication. The received perception data from multiple CAVs is then fused to generate better overall detection or tracking results. A number of cooperative autonomous driving datasets have been released to the public, including simulated ones [10, 24, 51, 52] and real ones [48, 53, 59, 60]. These datasets also establish benchmarks to evaluate the performance of cooperative perception algorithms. However, to date, cooperative driving research and datasets have mostly focused on perception tasks. How these state-of-the-art cooperative perception models can be connected with the downstream plan-

ning models to generate good cooperative planning results is still under-explored.

Other recent research has attempted to use LLM-based methods to build end-to-end perception and planning algorithms for an individual autonomous vehicle [5, 34, 37, 39, 40, 43, 46, 54] due to their common-sense reasoning and generalization ability from large-scale pre-trained data. These LLM-based models encode the raw sensor inputs into visual features, and then perform visual understanding and answer varieties of driving-related perception and planning questions. These approaches have shown some promise but have not yet explored the benefits of cooperative perception and planning. LLM-based driving algorithms without cooperative perception could also face safety-critical issues when the individual vehicle’s sensory capability is limited.

In this paper, we propose and explore a novel problem setting wherein LLM-based methods are used to build end-to-end perception and planning algorithms for *Cooperative Autonomous Driving*, as illustrated in Fig. 1. In this problem setting, we assume that there are multiple CAVs and a centralized LLM computing node. All CAVs 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. To enable the study of this problem setting, we first create the **Vehicle-to-Vehicle Question-Answering (V2V-QA)** dataset, built upon the V2V4Real [53] and V2X-Real [48] 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. There are several differences between our novel problem setting and other existing LLM-based driving research [4, 33, 35, 37, 39, 43]. First, our LLM can fuse multiple perception information from different CAVs and provide answers to different questions from any CAV, rather than just serving a single self-driving car. Second, our grounding questions are specially designed to focus on the potential occluded regions of each individual CAV. More differences between our V2V-QA and other related datasets are summarized in Tab. 1.

To establish a benchmark for the V2V-QA dataset, we first propose a strong baseline method: **Vehicle-to-Vehicle Multi-Modal Large Language Model (V2V-LLM)** for cooperative autonomous driving, as illustrated in Fig. 3. Each CAV extracts its own perception features and shares them with V2V-LLM. The V2V-LLM fuses the scene-level feature maps and object-level feature vectors, and then performs vision and language understanding to provide the answer to the input driving-related questions in V2V-QA. We also compare V2V-LLM with other baseline methods corresponding to different feature fusion methods: *no fusion*, *early fusion*, and *intermediate fusion* [48, 50–53]. The results show that V2V-LLM achieves the best performance in

the more important notable object identification and planning tasks and competitive performance in the grounding tasks, achieving strong performance for the overall autonomous driving system.

Our contribution can be summarized as follows:

- We create and introduce the V2V-QA dataset to support the development and evaluation of LLM-based approaches to end-to-end cooperative autonomous driving. V2V-QA includes grounding, notable object identification, and planning question-answering tasks.
- We propose a baseline method V2V-LLM for cooperative autonomous driving to provide an initial benchmark for V2V-QA. This method fuses scene-level feature maps and object-level feature vectors provided by multiple CAVs, and answers different CAV’s driving-related questions.
- We create a benchmark for V2V-QA and show that V2V-LLM outperforms other baseline fusion methods on the notable object identification and planning tasks and achieves competitive results on the grounding tasks, indicating the potential of V2V-LLM to be a foundation model for cooperative autonomous driving.

2. Related Work

2.1. Cooperative Perception in Autonomous Driving

Cooperative perception [16] algorithms were proposed to address the potential occlusion problem in individual autonomous vehicles. Pioneering work F-Cooper [6] proposes the first intermediate fusion approach that merges feature maps to achieve good cooperative detection performance. V2VNet [44] builds graph neural networks for cooperative perception. DiscoNet [23] adopts a knowledge distillation approach. More recent work, AttFuse [52], V2X-ViT [51], and CoBEVT [50] integrate attention-based models to aggregate features. Another group of works [8, 15, 29, 55] focuses on developing efficient communication approaches.

From a dataset perspective [28, 57], simulation datasets, OPV2V [52], V2X-Sim [24], and V2XSet [51] were first generated for cooperative perception research. More recently, real datasets have been collected. V2V4Real [53] is the first worldwide available real vehicle-to-vehicle cooperative perception dataset with perception benchmarks. V2X-Real [48], DAIR-V2X [59], and TUMTraf-V2X [60] further include sensor data from roadside infrastructures.

Different from this group of research, our problem setting and proposed V2V-QA dataset include both perception and planning question-answering tasks for multiple CAVs. Our proposed V2V-LLM also adopts a novel LLM-based fusion approach.

2.2. LLM-based Autonomous Driving

Language-based planning models [21, 31, 32] first transforms the driving scene, object detection results, and ego-

| Dataset | Publication | # CAVs | Sim/Real | # Frames | # QA | # QA/frame | Point Cloud | Planning |
|-------------------------------------|-------------|--------|----------|----------|--------------|-------------|-------------|----------|
| <i>AD</i> | | | | | | | | |
| NuScenes [3] | CVPR 2020 | - | Real | 400K | - | - | ✓ | |
| Waymo [38] | CVPR 2020 | - | Real | 200K | - | - | ✓ | |
| <i>Cooperative perception in AD</i> | | | | | | | | |
| OPV2V [52] | ICRA 2022 | 2-7 | Sim | 11K | - | - | ✓ | |
| V2XSet [51] | ECCV 2022 | 2-5 | Sim | 11K | - | - | ✓ | |
| V2V4Real [53] | CVPR 2023 | 2 | Real | 20K | - | - | ✓ | |
| V2X-Real [48] | ECCV 2024 | 2 | Real | 33K | - | - | ✓ | |
| <i>LLM-based AD</i> | | | | | | | | |
| NuScenes-QA [35] | AAAI 2024 | - | Real | 34K | 460K | 13.5 | ✓ | |
| Lingo-QA [33] | ECCV 2024 | - | Real | 28K | 420K | 15.3 | | ✓ |
| MAPLM-QA [4] | CVPR 2024 | - | Real | 14K | 61K | 4.4 | ✓ | |
| DriveLM [37] | ECCV 2024 | - | Sim+Real | 69K | 2M | 29.1 | | ✓ |
| TOKEN [39] | CoRL 2024 | - | Real | 28K | 434K | 15.5 | | ✓ |
| OmniDrive-nuScenes [43] | 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.

vehicle’s state into text input to the LLM. Then the LLM generates text output including the suggested driving action or the planned future trajectory. More recent approaches [37, 39, 40, 43, 45, 54, 56] use Multi-Modal Large Language Models (MLLMs) [1, 2, 17, 22, 25–27, 36, 41, 42] to encode point clouds or images into visual features. Then, the visual features are projected to the language embedding space for LLM to perform visual understanding and question-answering tasks.

From a dataset perspective, several LLM-based autonomous driving datasets have been built on top of existing autonomous driving datasets. For example, Talk2Car [11], NuPrompt [47], NuScenes-QA [35], NuInstruct [49], and Reason2Drive [34] create captioning, perception, prediction, and planning QA pairs based on the NuScenes [3] dataset. BDD-X [18] is extended from BDD100K [58]. DriveLM [37] adopts real data from NuScenes [3] and simulated data from CARLA [12] to have larger-scale and more diverse driving QAs. Other datasets curated independently focus on different types of QA tasks. HAD [19] contains human-to-vehicle advice data. DRAMA [30] introduces joint risk localization and captioning. Lingo-QA [33] proposed counterfactual question-answering tasks. MAPLM-QA [4] emphasizes map and traffic scene understanding.

Different from those LLM-based driving research that only supports individual autonomous vehicles, our problem setting and proposed V2V-QA dataset are designed for cooperative driving scenarios with multiple CAVs. In our problem setting, the LLM can aggregate perception features from multiple CAVs and provide answers to questions from different CAVs. Our V2V-QA is also designed to focus on the potential occluded regions. In addition, our V2V-QA contains both highway and urban cooperative driving

scenarios, making our planning task more challenging than prior works based on the NuScenes [3] dataset.

3. V2V-QA Dataset

To enable the research in our proposed novel problem setting: LLM-based cooperative autonomous driving, we create the **Vehicle-to-Vehicle Question-Answering (V2V-QA)** dataset to benchmark different models’ performance on fusing perception information and answering safety-critical driving-related questions.

3.1. Problem Setting

Our proposed V2V cooperative autonomous driving with LLM problem is illustrated in Fig. 1. In this setting, we assume there are multiple Connected Autonomous Vehicles (CAVs) and a centralized LLM computing node. All CAVs share their individual perception information, such as scene-level or object-level features, with the centralized LLM. Any CAV can ask the LLM a question in natural language to obtain information for driving safety. The LLM aggregates the received perception information from multiple CAVs and provides a natural language answer to the CAV’s question. In this research, the questions and answers include **grounding (Q1-3)**, **notable object identification (Q4)**, and **planning (Q5)**, as illustrated in Fig. 2.

3.2. Dataset Details

Our V2V-QA dataset contains two splits: **V2V-split** and **V2X-split**, which are built on top of V2V4Real [53] and V2X-Real [48] datasets, respectively. These base datasets are collected by driving two vehicles with LiDAR sensors simultaneously near to each other. These datasets also includes 3D bounding box annotations for other objects in the

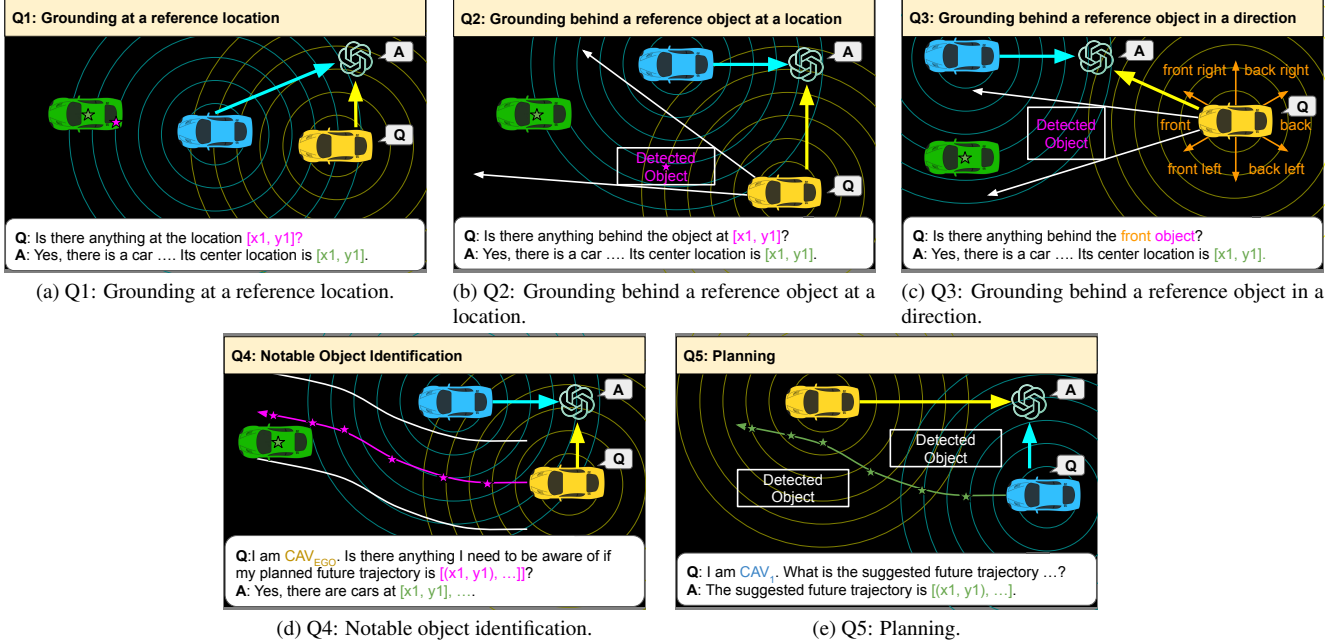


Figure 2. Illustration of V2V-QA’s 5 types of QA pairs. The arrows pointing at LLM indicate the perception data from CAVs.

| 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. In V2V4Real [53], 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 [48], 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. In V2X-Real [48], some driving scenes also provide lidar point clouds from roadside infrastructures. In our V2X-split, we also include them as perception inputs to the LLM with the same approach as using CAVs’ lidar point clouds to answer CAVs’ questions. We follow the same training and testing settings from V2V4Real [53] and V2X-Real [48] when building our V2V-split and V2X-split.

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.

3.3. Question and Answer Pairs Curation

For each frame of V2V4Real [53] and V2X-Real [48] 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 [53] and V2X-Real [48]’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. 5 and 6. The generation rule of each QA type is described as follows.

Q1. Grounding at a reference location (Fig. 2a): In this type of question, we ask the LLM to identify whether an object that occupies a specific query 2D location exists. If so, the LLM is expected to provide the center location of the object. Otherwise, the LLM should indicate that there is nothing at the reference location. We use the center locations of ground-truth boxes and every CAV’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.

Q2. Grounding behind a reference object at a location (Fig. 2b): When a CAV’s field of view is occluded by a nearby large detected object, this CAV may want to ask the centralized LLM to determine whether there exists any ob-

ject behind that occluding large object given the fused perception information from all CAVs. If so, the LLM is expected to return the object’s location and the asking CAV may need to drive more defensively or adjust its planning. Otherwise, the LLM should indicate that there is nothing behind the reference 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): The aforementioned grounding tasks are intermediate tasks in the autonomous driving pipeline. More critical abilities of autonomous vehicles involve both identifying notable objects near planned future trajectories and adjusting future planning to avoid potential collisions. 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 important because the ultimate goal of autonomous vehicles is to navigate through complex environments safely and avoid any potential collision in the future. 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. Our V2V-QA’s planning task is more challenging than other NuScenes [3]-based LLM-driving related works for a couple reasons. First, we support multiple CAVs in cooperative driving scenarios. The LLM model needs to provide different answers depending on which CAV is asking, while prior works only need to generate planning results for a single autonomous vehicle. Second, our V2V-QA’s ground-truth planning trajectories are more diverse. V2V-QA contains both urban and highway driving scenarios, while NuScenes [3] only includes urban driving scenarios. Detailed dataset statistics and comparison can be seen in the supplementary material.

3.4. Evaluation Metrics

We follow prior works [39, 43]’s approach to evaluate model performance. For the grounding questions (Q1, Q2, Q3) and the notable object identification question (Q4), the evaluation metrics are F1 score, precision, and recall. The ground-truth answers and model outputs contain objects’

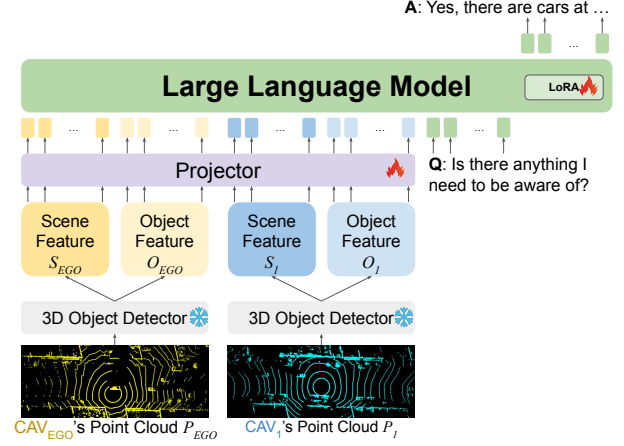


Figure 3. Model diagram of our proposed V2V-LLM for cooperative autonomous driving.

center locations. If the center distance between the ground-truth answer and the model output is less than a threshold value, this output is considered as a true positive. We set the threshold value to be 4 meters, a typical length of a vehicle.

For the planning question (Q5), the evaluation metrics are L2 errors and collision rates. The ground-truth answers and model outputs contain 6 future waypoints, so we calculate the L2 errors on those waypoints. When calculating collision rates, we assume each of the CAVs’ bounding box sizes is 4 meters in length, 2 meters in width, and 1.5 meters in height. We place each CAV’s bounding box at the model’s output future waypoints and calculate the Intersection-over-Union (IOU) between the CAV’s bounding box and every ground-truth object’s annotation bounding box in those future frames. If the IOU is larger than 0, it is considered as a collision.

4. V2V-LLM

We also 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 as the vision input, a question as the language input, and generates an answer as the language output.

4.1. LiDAR-based Input Features

For extracting the perception input features, each CAV applies a 3D object detection model to its individual LiDAR point cloud: P_{EGO} and P_I . We extract the scene-level feature map S_{EGO} and S_I from the 3D object detection model and transform the 3D object detection results as the object-level feature vectors O_{EGO} and O_I . Following prior works V2V4Real [53] and V2X-Real [48], we use PointPillars [20] as the 3D object detector for fair comparisons.

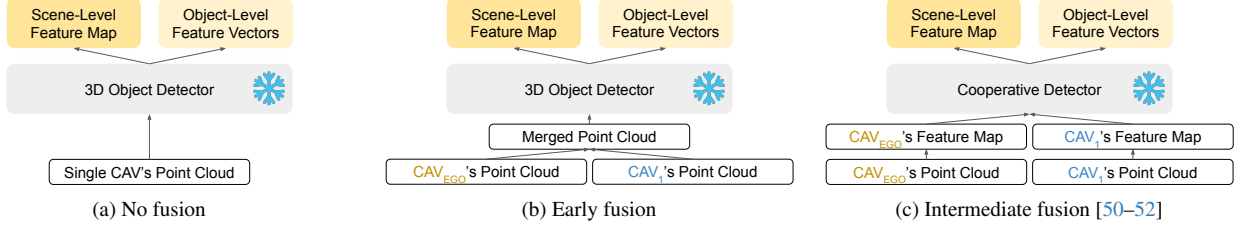


Figure 4. Feature encoder diagrams of the baseline methods from different fusion approaches.

4.2. LiDAR-based Multi-Modal LLM

Model architecture: We utilize LLaVA [25] to develop our MLLM, given its superior performance on visual question-answering tasks. However, since the perception features of our cooperative driving tasks are LiDAR-based instead of RGB images used by LLaVA [25], we use a LiDAR-based 3D object detector as the point cloud feature encoder, as described in the previous section, instead of LLaVA [25]’s CLIP [36] image feature encoders. 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 returns an answer based on the question.

Training: We use 8 NVIDIA A100-80GB GPUs to train our model. Our V2V-LLM uses LLaVA-v1.5-7b [25]’s Vicuna [7] as the LLM backbone. To train our model, we initialize it by loading the pre-trained LLaVA-v1.5-7b [25]’s checkpoint. We freeze the LLM and the point cloud feature encoder, and finetune the projector and the LoRA [14] 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 [25].

5. Experiment

5.1. Baseline Methods

We follow V2V4Real [53] and V2X-Real [48] to establish a benchmark for our proposed V2V-QA dataset with experiments on baseline methods using different fusion approaches: **no fusion**, **early fusion**, **intermediate fusion**, and our proposed baseline, **LLM fusion** (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, as shown in Fig. 4. In some driving sequences of V2X-split that have point clouds from roadside infrastructures, we also include them as perception input in the same way as using CAVs’ point clouds.

No fusion: Only a single CAV’s LiDAR point cloud is fed

to a single 3D object detector to extract the scene-level and object-level features as the LLM’s visual input. The performance is expected to be worse than all other cooperative perception approaches.

Early fusion: The LiDAR point cloud from two CAVs is merged first. Then the merged point cloud is used as input to a 3D object detector to extract the visual features as the visual input to the LLM. This approach requires much higher communication cost and is less practical for deployment on real-world autonomous vehicles.

Intermediate fusion: Prior work CoBEVT [50], V2X-ViT [51], and AttFuse [52] propose cooperative detection models that merge feature maps from multiple CAVs via attention mechanisms. Such approaches require less communication cost and can still achieve good performance. In our benchmark, we extract the features from those cooperative detection models as the input tokens to the LLM.

LLM fusion: We categorize our proposed V2V-LLM as a new type of fusion method, *LLM fusion*, which lets each CAV perform its individual 3D object detection to extract the scene-level feature maps and object-level feature vectors, and uses the Multi-Modal LLM to fuse the features from multiple CAVs. This approach is related to the traditional *late fusion* method that performs individual 3D object detection and aggregates the results by non-maximum suppression (NMS). Instead of applying NMS, our method adopts LLM to perform more tasks than just detection.

5.2. Experimental Results

5.2.1. Grounding

Our V2V-LLM and baseline methods’ performance on V2V-QA’s 3 types of grounding questions can be seen in Tab. 3 for V2V-split and V2X-split, respectively. CoBEVT [50] is not included in V2X-split’s result because V2X-Real [48] does not release its CoBEVT [50] baseline model. In average, V2V-LLM achieves similar performance in V2V-split and outperforms all other baseline methods in V2X-split. Such results indicate that our V2V-LLM has a promising capability of fusing perception features from multiple CAVs to answer grounding questions.

5.2.2. Notable Object Identification

Tab. 3 show the performance on the notable object identification task (Q4). Our proposed V2V-LLM outperforms

| 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 (%) ↓ | |
| <i>No Fusion</i> | 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 |
| <i>Early Fusion</i> | 73.5 | 23.3 | 20.8 | 39.2 | 53.9 | <u>6.20</u> | <u>3.55</u> | <u>59.7</u> | 23.3 | 26.1 | 36.4 | <u>67.6</u> | <u>2.12</u> | 8.61 | 1.9208 |
| <i>Intermediate Fusion</i> | | | | | | | | | | | | | | | |
| AttFuse [52] | 70.7 | 26.4 | 18.4 | 38.5 | 56.9 | 6.83 | 4.12 | 58.9 | 23.9 | <u>26.3</u> | 36.4 | 65.9 | 2.19 | <u>8.39</u> | <u>0.4008</u> |
| V2X-ViT [51] | 70.8 | 28.0 | 22.6 | 40.5 | <u>57.6</u> | 7.08 | 4.33 | 59.6 | <u>24.2</u> | 26.1 | <u>36.6</u> | 65.0 | 2.29 | 8.86 | <u>0.4008</u> |
| CoBEVT [50] | <u>72.2</u> | <u>29.3</u> | <u>21.3</u> | 40.9 | <u>57.6</u> | 6.72 | 3.88 | - | - | - | - | - | - | - | <u>0.4008</u> |
| <i>LLM Fusion</i> | | | | | | | | | | | | | | | |
| V2V-LLM (Ours) | 70.0 | 30.8 | 21.2 | <u>40.7</u> | 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.

all other methods in both V2V-split and V2X-split. Compared with the aforementioned grounding tasks, this notable object identification task requires more spatial understanding ability to identify the objects close to the planned future waypoints. For such a task, our V2V-LLM, which lets the Multi-Modal LLM perform both perception feature fusion and question answering, achieves the best results.

5.2.3. Planning

Tab. 3 show the performance of the planning task (Q5) for V2V-split and V2X-split, respectively. Our proposed V2V-LLM outperforms other methods in this safety-critical task to generate a future trajectory that aims to avoid potential collisions. More planning performance evaluation can be seen in the supplementary material.

5.2.4. Communication Cost and Scaling Analysis

In our centralized setting, each CAV sends one scene-level feature map ($\leq 0.2\text{MB}$), one set of individual object detection result parameters ($\leq 0.003\text{MB}$), one question ($\leq 0.0002\text{MB}$) to the LLM computing node and receives one answer ($\leq 0.0002\text{MB}$) at each timestep. If there are N_v CAVs and each asks N_q questions, the communication cost of each CAV is $(0.2 + 0.003 + (0.0002 + 0.0002)N_q) = (0.203 + 0.0004N_q)$ MB, and the cost of the LLM is $(0.2 + 0.003 + (0.0002 + 0.0002)N_q)N_v = (0.203N_v + 0.0004N_qN_v)$ MB, as shown in Tab. 4. Note that each CAV only needs to send the same features to the LLM computing node 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. Detailed scaling analysis can be seen in the supplementary material.

5.2.5. Summary

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

| Setting | Each CAV | Centralized LLM |
|-------------|---------------------|---------------------------|
| Centralized | $0.203 + 0.0004N_q$ | $0.203N_v + 0.0004N_qN_v$ |

Table 4. Communication cost (MB) and scaling analysis. N_v : number of CAVs. N_q : number of questions asked by each CAV at each timestep.

| Method | Q1 | Q2 | Q3 | Q _{Gr} | Q4 | Q5 | |
|-------------|-------------|-------------|-------------|-----------------|-------------|-------------|-------------|
| | F1 ↑ | F1 ↑ | F1 ↑ | F1 ↑ | F1 ↑ | L2 (m) ↓ | CR (%) ↓ |
| Scene only | 69.9 | 15.4 | 17.9 | 34.4 | 43.2 | 7.21 | 15.55 |
| Object only | 69.0 | 26.9 | 17.6 | 37.8 | 52.6 | 5.24 | 7.78 |
| Scratch | 67.6 | 26.5 | 17.2 | 37.1 | 49.3 | 6.30 | 5.01 |
| V2V-LLM | 70.0 | 30.8 | 21.2 | 40.7 | 59.7 | 4.99 | 3.00 |

Table 5. Ablation study in V2V-QA’s V2V-split testing set.

competitive results in the grounding tasks. In terms of communication costs, V2V-LLM only increases communication costs by 1.5% in comparison to other intermediate fusion baseline methods. More detailed evaluation and analysis can be seen in the supplementary material.

5.3. Ablation Study

Input Features: We experiment with variants of our V2V-LLM model that use either only the scene-level feature maps or only the object-level feature vectors as the visual input. The ablation results can be seen in Tab. 5. Both types of features contribute to final performance in all QA tasks. In general, the object-level-only model outperforms the scene-level-only model. This implies that the object-level features are easier for LLM to digest, which is consistent with the results observed in the previous work with the TOKEN model [39].

Training from Scratch: Tab. 5 shows that training from scratch achieves worse performance, meaning that pre-training with LLaVA’s VQA tasks improves our V2V-LLM’s performance in V2V-QA. More detailed ablation results can be seen in the supplementary material.

| Q1: Grounding at a reference location | Q2: Grounding behind a reference object at a location | Q3: Grounding behind a reference object in a direction | Q3: Grounding behind a reference object in a direction |
|---|---|--|--|
| | | | |
| 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]. | 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]. | 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]. | 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 5. V2V-LLM’s *grounding* results on V2V-QA’s testing set. Magenta \circ : reference locations in questions. Yellow $+$: model output locations. Green \circ : ground-truth answers.

| Q4: Notable Object Identification | Q4: Notable Object Identification | Q5: Planning | Q5: Planning |
|---|---|---|---|
| | | | |
| 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. | Q: I am CAV_I. 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. | 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)]. | Q: I am CAV_I. 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 6. 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 \circ : ground-truth notable object locations. Yellow $+$ and Cyan \times : model identification outputs corresponding to CAV_EGO and CAV_I, 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_I, respectively.

5.4. Qualitative Results

Fig. 5 shows our V2V-LLM’s *grounding* results and the ground truth with visualization on V2V-QA’s testing set. We can observe that our V2V-LLM is able to locate the objects given the provided reference information for each of the 3 types of grounding questions. Fig. 6’s left part shows our V2V-LLM’s *notable object identification* results. V2V-LLM demonstrate its capability of identifying multiple objects near the planned future trajectories specified in the questions for each CAV. Fig. 6’s right part shows V2V-LLM’s *planning* results. Our model is able to suggest future trajectories that avoid potential collisions with nearby objects. Overall, the outputs of our model closely align with the ground-truth answers across all question types, indicating its robustness in cooperative autonomous driving tasks.

6. Conclusion

In this work, we expand the research scope of cooperative autonomous driving by integrating the use of Multi-Modal

LLM-based methods, aimed at improving the safety of future autonomous driving systems. We propose a new problem setting and create a novel V2V-QA dataset and benchmark that includes grounding, notable object identification, and planning question-answering tasks designed for varieties of cooperative driving scenarios. We propose a baseline model V2V-LLM that fuses each CAV’s individual perception information and performs visual and language understanding to answer driving-related questions from any CAV. Our proposed V2V-LLM outperforms all other baselines adopted from state-of-the-art cooperative perception algorithms in the notable object identification and planning tasks, and achieves competitive performance in the grounding tasks. These experimental results indicate that V2V-LLM is promising as a unified multi-modal foundation model that can effectively perform perception and planning tasks for cooperative autonomous driving. We will publicly release our V2V-QA dataset and code to facilitate the open-source research, and believe it will bring the cooperative driving research to the next stage.

7. Acknowledgement

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

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3
- [3] 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 multi-modal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 5
- [4] 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
- [5] Long Chen, Oleg Sinavski, Jan Hünemann, 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. 2
- [6] 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, 2
- [7] 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. 6
- [8] 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. 2
- [9] 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
- [10] 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
- [11] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 3
- [12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, 2017. 3
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1
- [14] 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. 6
- [15] 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. 2
- [16] 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. 2
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3
- [18] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *European Conference on Computer Vision (ECCV)*, 2018. 3
- [19] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [20] 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. 5
- [21] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023. 3

- [23] 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. 2
- [24] Yiming Li, Dekun Ma, Ziyun 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
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3, 6
- [26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3
- [28] 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. 2
- [29] 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. 2
- [30] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3
- [31] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop (Foundation Models for Decision Making)*, 2023. 2
- [32] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *Conference On Language Modeling (COLM)*, 2024. 2
- [33] Ana-Maria Marcu, Long Chen, Jan Hünemann, 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, 3
- [34] 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. 2, 3
- [35] 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, 3
- [36] 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, 6
- [37] 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 *European Conference on Computer Vision (ECCV)*, 2024. 2, 3
- [38] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3
- [39] 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. 2, 3, 5, 7
- [40] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivelm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2, 3
- [41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. 3
- [42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shriti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [43] 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. 2, 3, 5
- [44] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyan 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, 2
- [45] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 3
- [46] Wenhao Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivelm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 2

- [47] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint*, 2023. 3
- [48] Hoa 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 large-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 4, 5, 6
- [49] Ding Xinpeng, Han Jinahua, Xu Hang, Laing Xiaodan, Hang Xu, Zhang Wei, and Li Xiaomeng. Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3
- [50] 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, 2, 6, 7, 3
- [51] 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, 6, 7
- [52] 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, 6, 7
- [53] 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, 3, 4, 5, 6
- [54] 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. 2, 3
- [55] Dingkan 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. 2
- [56] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint*, 2023. 3
- [57] Melih Yazgan, Mythra Varun Akkanapragada, and J Marius Zöllner. Collaborative perception datasets in autonomous driving: A survey. In *IEEE Intelligent Vehicles Symposium (IV)*, 2024. 2
- [58] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [59] 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
- [60] 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

8. Detailed Evaluation Results

Tabs. 6 and 7 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. 8 and 9 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 ($\leq 0.2\text{MB}$), one set of individual object detection result parameters ($\leq 0.003\text{MB}$), N_q questions (each $\leq 0.0002\text{MB}$) to the LLM and receives N_q answers (each $\leq 0.0002\text{MB}$) 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. 10 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. 11 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. 12 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

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. 13 and 14 summarize 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. 13 and 14, as each planning QA pair inherently includes a ground-truth future trajectory in its corresponding 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. 7 to 10. 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. 11. We visualize the location distribution of V2V-QA’s V2X-split in Figs. 12 to 16. These figures indicate that our V2V-QA has diverse spatial distributions in the driving scenes. Compared to NuScenes [3], our V2V-QA has larger ranges and standard deviations of the ground-truth ending waypoints, as shown in Tab. 15. Therefore, the planning task in our V2V-QA is more challenging.

| 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} (%) ↓ | |
| <i>No Fusion</i> | 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 |
| <i>Early Fusion</i> | 73.5 | 82.2 | <u>66.5</u> | 23.3 | 29.1 | 19.5 | 20.8 | <u>22.7</u> | 19.3 | 39.2 | 53.9 | 55.4 | 52.6 | <u>6.20</u> | <u>3.55</u> | 1.9208 |
| <i>Intermediate Fusion</i> | | | | | | | | | | | | | | | | |
| AttFuse [52] | 70.7 | 79.6 | 63.6 | 26.4 | 31.6 | 22.7 | 18.4 | 19.6 | 17.4 | 38.5 | 56.9 | <u>57.2</u> | 56.6 | 6.83 | 4.12 | <u>0.4008</u> |
| V2X-ViT [51] | 70.8 | <u>81.1</u> | 62.8 | 28.0 | 33.9 | 23.9 | 22.6 | 25.2 | 20.5 | 40.5 | <u>57.6</u> | <u>57.0</u> | 58.2 | 7.08 | 4.33 | <u>0.4008</u> |
| CoBEVT [50] | <u>72.2</u> | 76.8 | 68.1 | <u>29.3</u> | <u>34.7</u> | <u>25.3</u> | <u>21.3</u> | 22.1 | <u>20.6</u> | 40.9 | <u>57.6</u> | <u>57.2</u> | <u>58.1</u> | 6.72 | 3.88 | <u>0.4008</u> |
| <i>LLM Fusion</i> | | | | | | | | | | | | | | | | |
| V2V-LLM (Ours) | 70.0 | 80.1 | 62.2 | 30.8 | 36.3 | 26.7 | 21.2 | 21.5 | 20.8 | <u>40.7</u> | 59.7 | 61.9 | 57.6 | 4.99 | 3.00 | 0.4068 |

Table 6. 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} (%) ↓ | |
| <i>No Fusion</i> | 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 |
| <i>Early Fusion</i> | <u>59.7</u> | 70.6 | 51.8 | 23.3 | 34.0 | 17.7 | 26.1 | 28.0 | 24.5 | 36.4 | <u>67.6</u> | <u>69.3</u> | <u>66.0</u> | <u>2.12</u> | 8.61 | 1.9208 |
| <i>Intermediate Fusion</i> | | | | | | | | | | | | | | | | |
| AttFuse [52] | 58.9 | <u>71.1</u> | 50.3 | 23.9 | <u>34.3</u> | 18.4 | <u>26.3</u> | 28.3 | <u>24.6</u> | 36.4 | 65.9 | 67.0 | 64.9 | 2.19 | <u>8.39</u> | <u>0.4008</u> |
| V2X-ViT [51] | 59.6 | 69.6 | <u>52.1</u> | <u>24.2</u> | 33.2 | <u>19.1</u> | 26.1 | <u>28.2</u> | <u>24.3</u> | <u>36.6</u> | 65.0 | 64.8 | 65.3 | 2.29 | 8.86 | <u>0.4008</u> |
| <i>LLM Fusion</i> | | | | | | | | | | | | | | | | |
| 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 7. 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 ↓ |
| <i>No Fusion</i> | 3.84 | 6.52 | 9.30 | 6.55 | 1.31 | 4.76 | 7.63 | 4.57 |
| <i>Early Fusion</i> | <u>3.68</u> | <u>6.19</u> | <u>8.74</u> | <u>6.20</u> | 0.96 | 3.86 | <u>5.83</u> | <u>3.55</u> |
| <i>Intermediate Fusion</i> | | | | | | | | |
| AttFuse [52] | 4.06 | 6.78 | 9.64 | 6.83 | 1.42 | 4.41 | 6.53 | 4.12 |
| V2X-ViT [51] | 4.21 | 7.05 | 9.99 | 7.08 | 1.33 | 4.82 | 6.85 | 4.33 |
| CoBEVT [50] | 3.97 | 6.71 | 9.47 | 6.72 | <u>0.93</u> | <u>3.74</u> | 6.96 | 3.88 |
| <i>LLM Fusion</i> | | | | | | | | |
| V2V-LLM (ours) | 2.96 | 4.97 | 7.05 | 4.99 | 0.55 | 3.19 | 5.25 | 3.00 |

Table 8. 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.

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. 17 to 20, notable object identification task in Figs. 21 to 22, and planning task in Figs 23 to 24. The baseline methods include no-fusion, early-fusion, and intermediate-fusion: AttFuse [52], V2X-ViT [51], and CoBEVT [50]. Results of V2X-split can be seen in Figs. 25 to 31. In general, our proposed V2V-LLM’s outputs are closer to the ground-truth answers, in

comparison to other baseline methods’ results.

14. Limitation

Fig. 32 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 V2V4Real [53] has not released its HD map to the public.

| Method | L2 (m) | | | | CR (%) | | | |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|
| | 1s ↓ | 2s ↓ | 3s ↓ | average ↓ | 1s ↓ | 2s ↓ | 3s ↓ | average ↓ |
| <i>No Fusion</i> | 1.33 | 2.28 | 3.31 | 2.31 | 2.52 | 9.54 | 15.57 | 9.21 |
| <i>Early Fusion</i> | <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 |
| <i>Intermediate Fusion</i> | | | | | | | | |
| AttFuse [52] | 1.27 | 2.17 | 3.11 | 2.19 | 2.40 | 9.07 | <u>13.70</u> | <u>8.39</u> |
| V2X-ViT [51] | 1.34 | 2.27 | 3.25 | 2.29 | 1.41 | 9.89 | 15.28 | 8.86 |
| <i>LLM Fusion</i> | | | | | | | | |
| V2V-LLM (ours) | 0.99 | 1.70 | 2.45 | 1.71 | <u>2.17</u> | 6.79 | 11.71 | 6.89 |

Table 9. 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 second-best 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 10. 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 (%) ↓ |
| <i>No Fusion</i> | 6.55 | 4.57 | 5.94 | 3.77 |
| <i>Early Fusion</i> | <u>6.20</u> | <u>3.55</u> | <u>5.13</u> | <u>3.04</u> |
| <i>Intermediate Fusion</i> | | | | |
| AttFuse [52] | 6.83 | 4.12 | 6.46 | 3.50 |
| V2X-ViT [51] | 7.08 | 4.33 | 5.52 | 3.84 |
| CoBEVT [50] | 6.72 | 3.88 | 6.02 | 3.40 |
| <i>LLM Fusion</i> | | | | |
| V2V-LLM (ours) | 4.99 | 3.00 | 4.82 | 2.93 |

Table 11. 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.

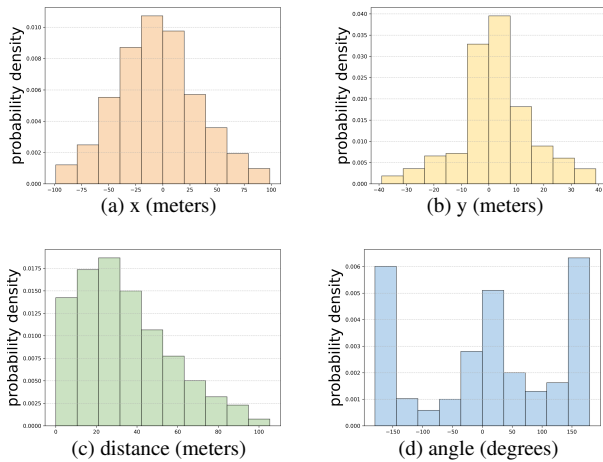


Figure 7. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2V-split Q1: Grounding at a reference location.

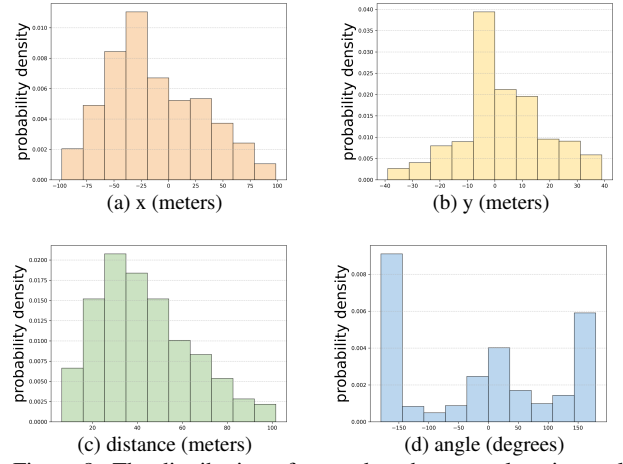


Figure 8. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2V-split Q2: Grounding behind a reference object at a location.

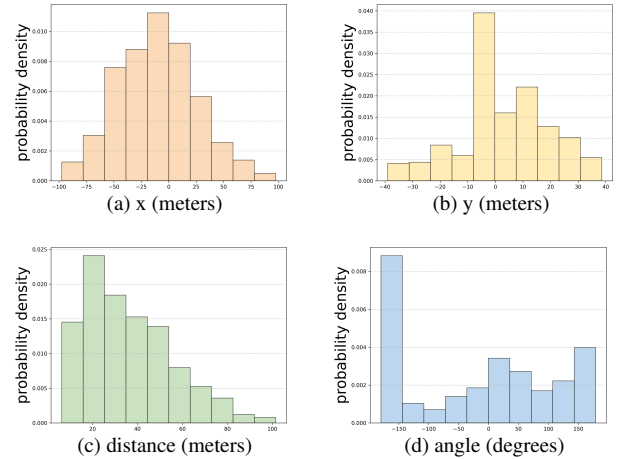


Figure 9. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2V-split Q3: Grounding behind a reference object in a direction.

| 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 12. 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 13. 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 14. 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 15. Ranges and standard deviations of ground-truth ending waypoints.

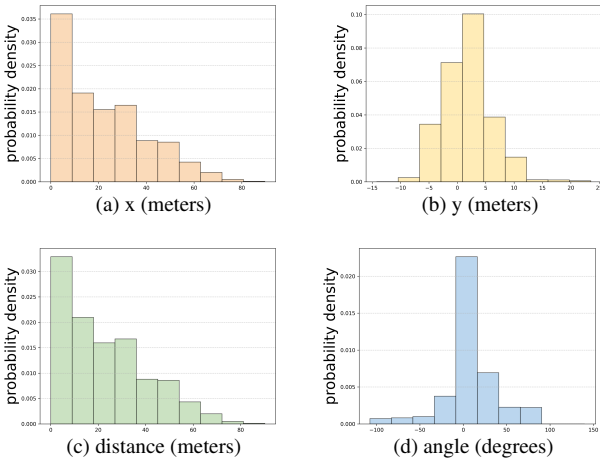


Figure 10. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2V-split Q4: Notable object identification.

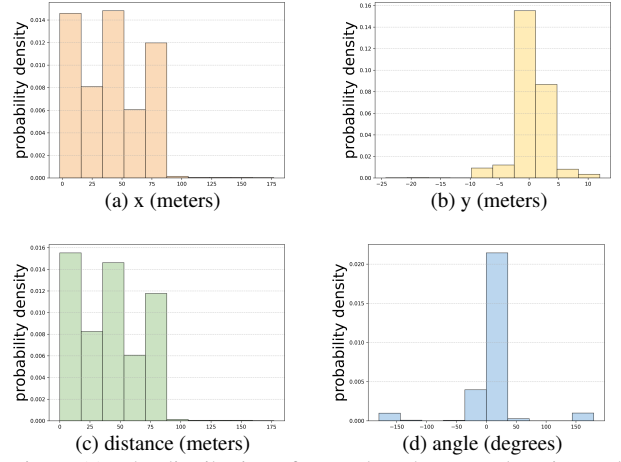


Figure 11. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2V-split Q5: Planning.

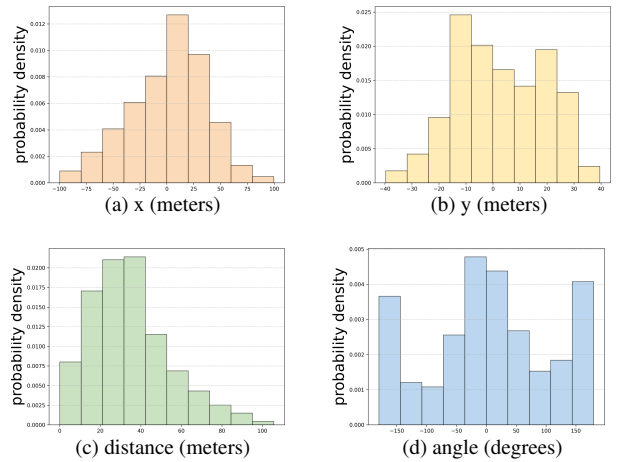


Figure 12. The distribution of ground-truth answer locations relative to CAV in V2V-QA’s V2X-split Q1: Grounding at a reference location.

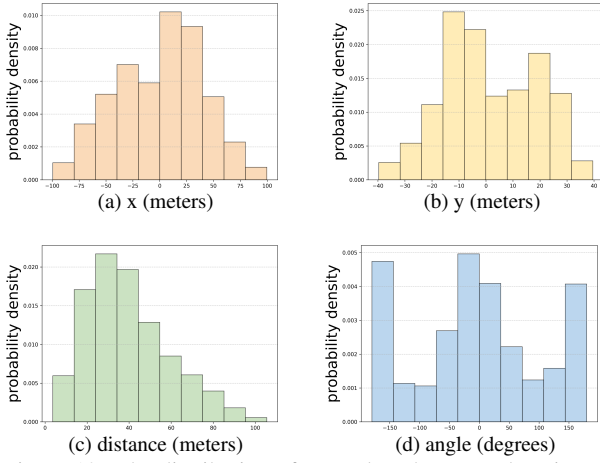


Figure 13. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q2: Grounding behind a reference object at a location.

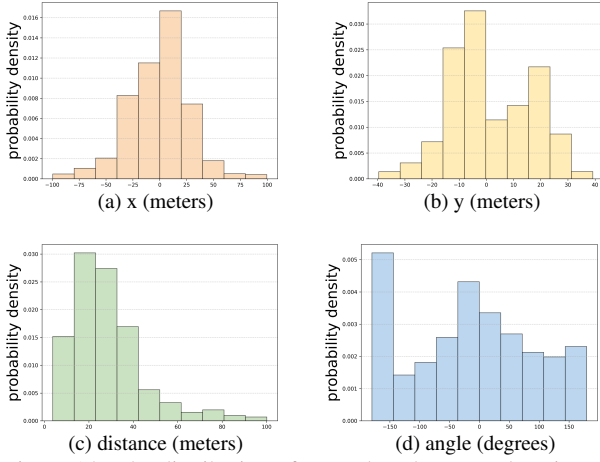


Figure 14. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q3: Grounding behind a reference object in a direction.

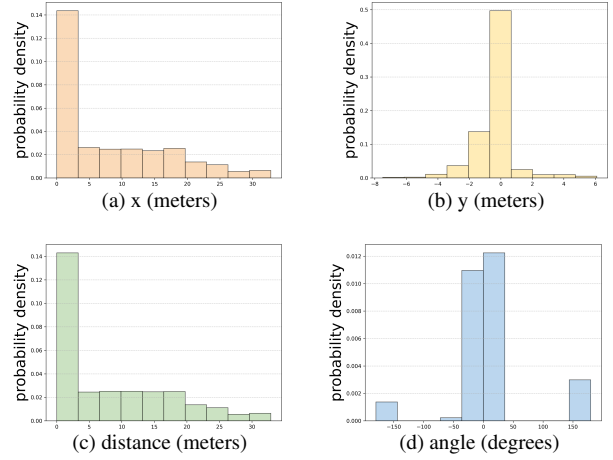


Figure 16. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q5: Planning.

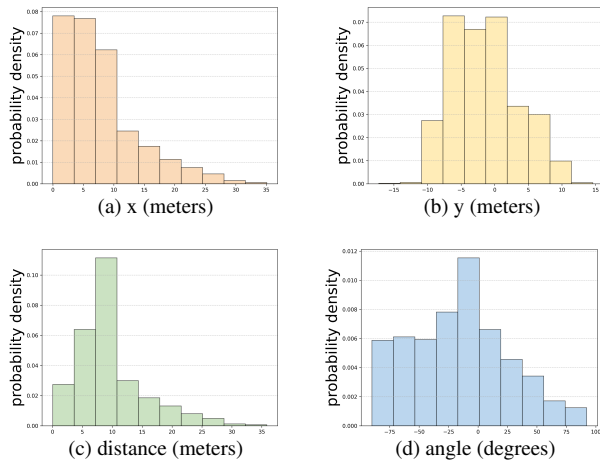
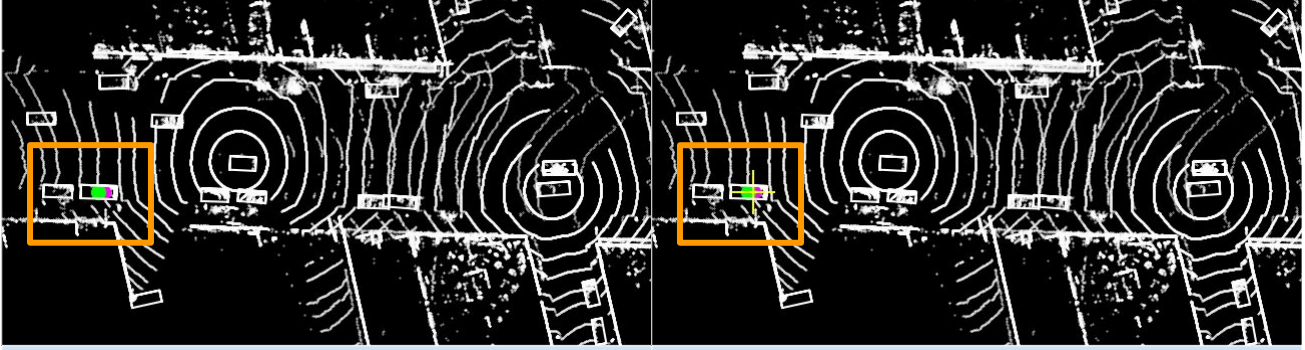


Figure 15. The distribution of ground-truth answer locations relative to CAV in V2V-QA's V2X-split Q4: Notable object identification.

Q1: Grounding at a reference location

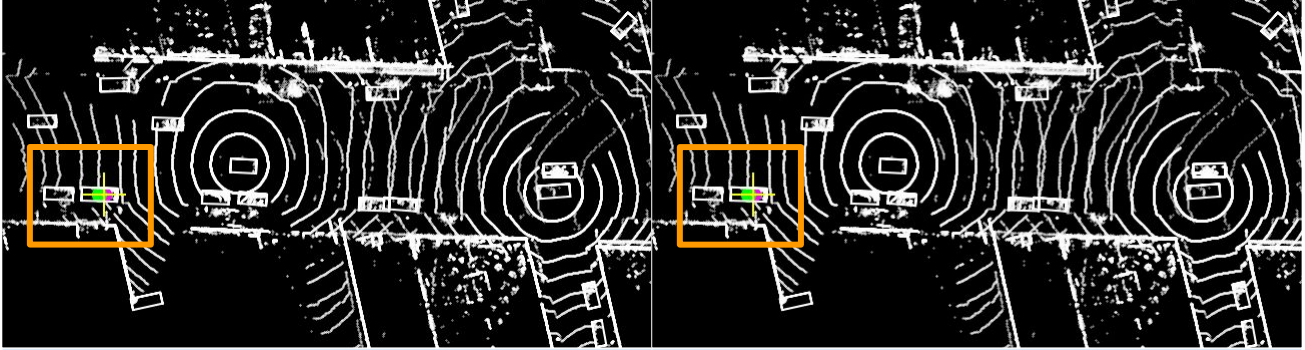
Q: Is there anything at the location [73.0, -0.4]?

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



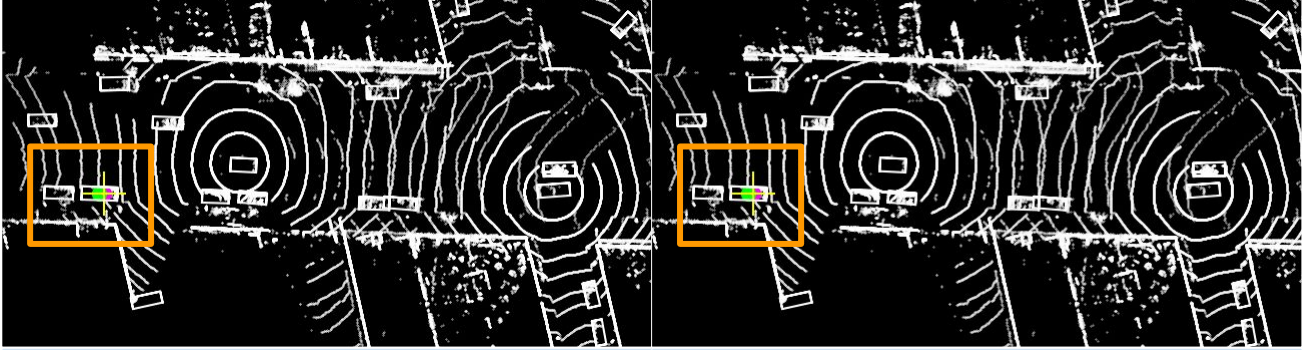
No Fusion: There is no object at the location.

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].

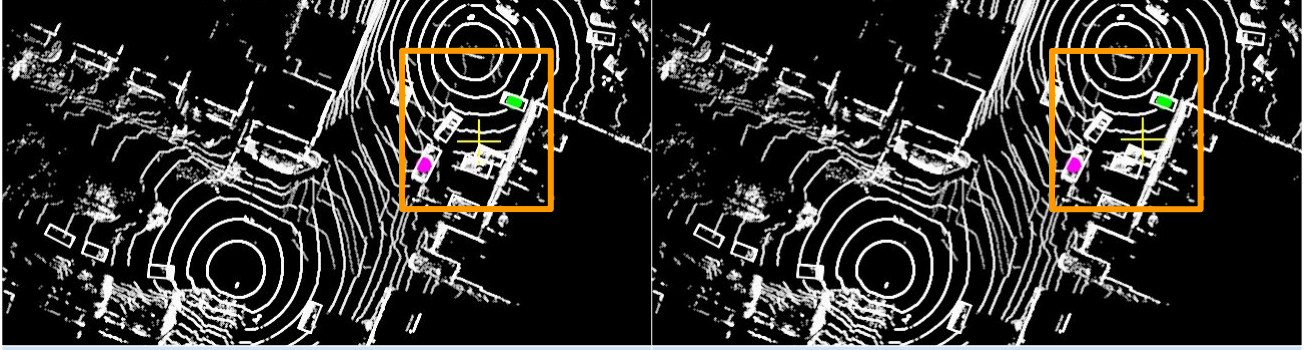
V2V-LLM: Yes, there is a car at the location. Its center location is [73.0, -0.4].

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

Q2: Grounding behind a reference object at a location

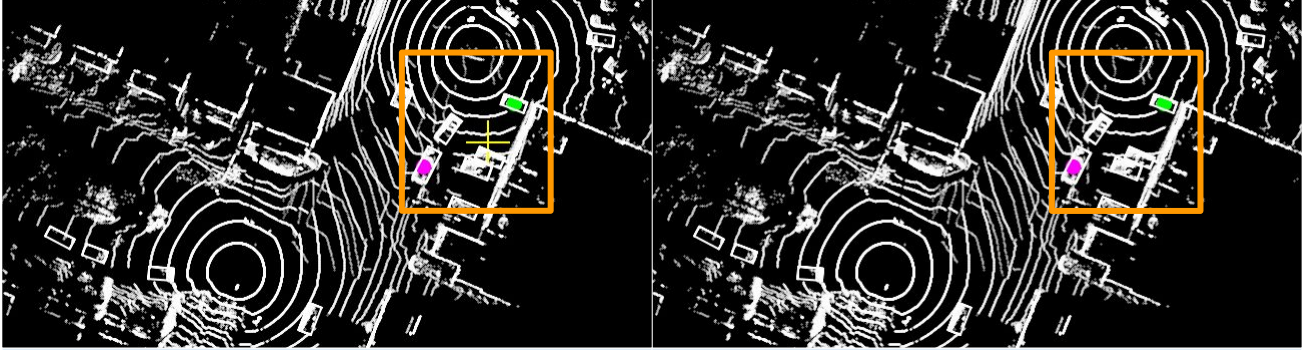
Q: Is there anything behind the object at [-30.4, 17.1]?

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



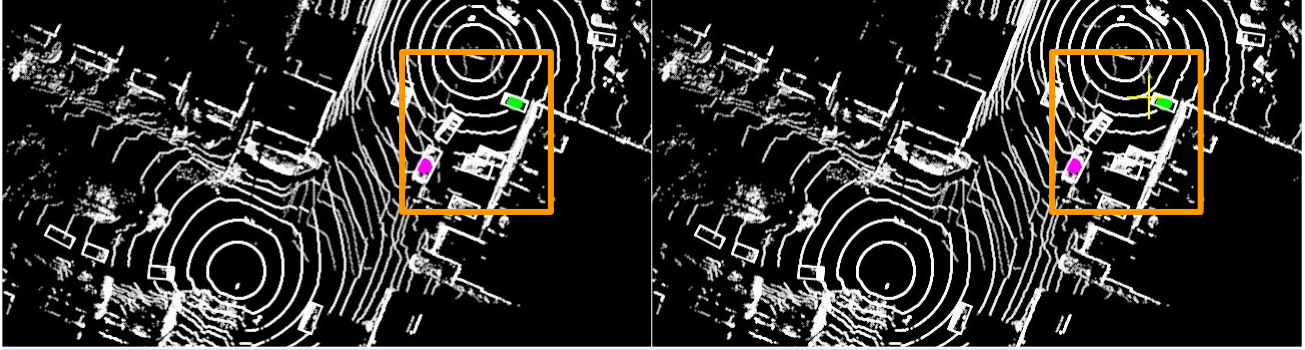
No Fusion: Yes, there is a car behind the object. Its center location is [-39.4, 20.8].

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.

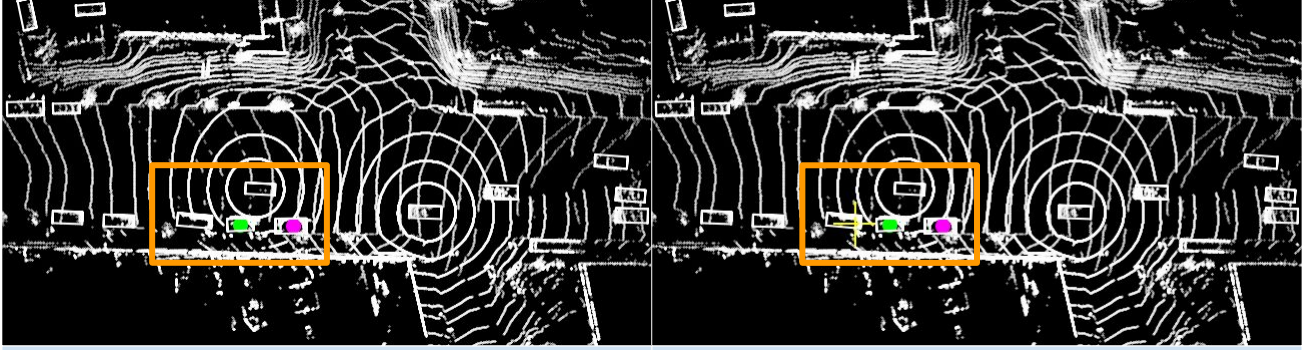
V2V-LLM: Yes, there is a car behind the object. Its center location is [-42.8, 28.1].

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

Q3: Grounding behind a reference object in a direction

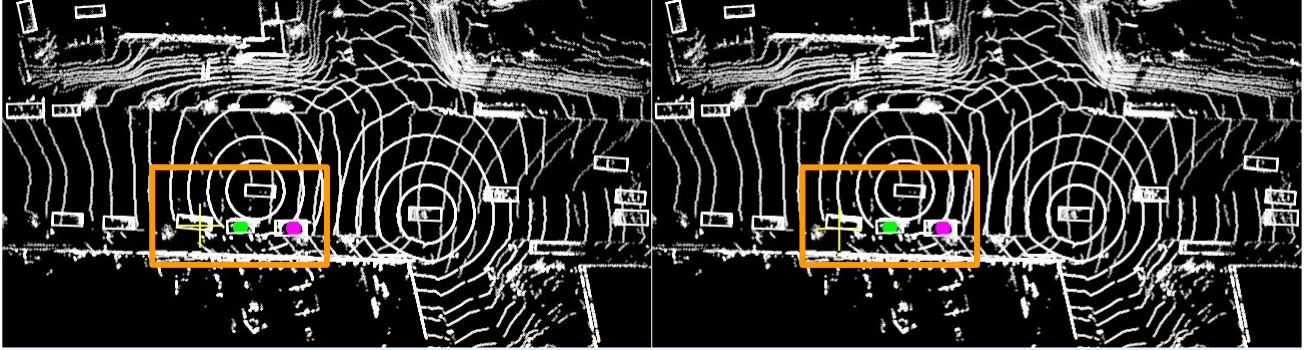
Q: Is there anything behind the **front object**?

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



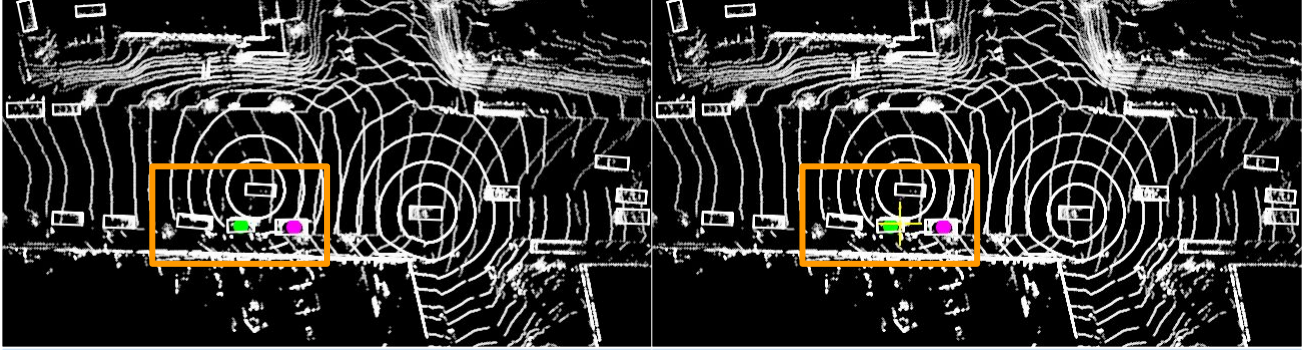
No Fusion: There is nothing behind the **front object**.

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**.

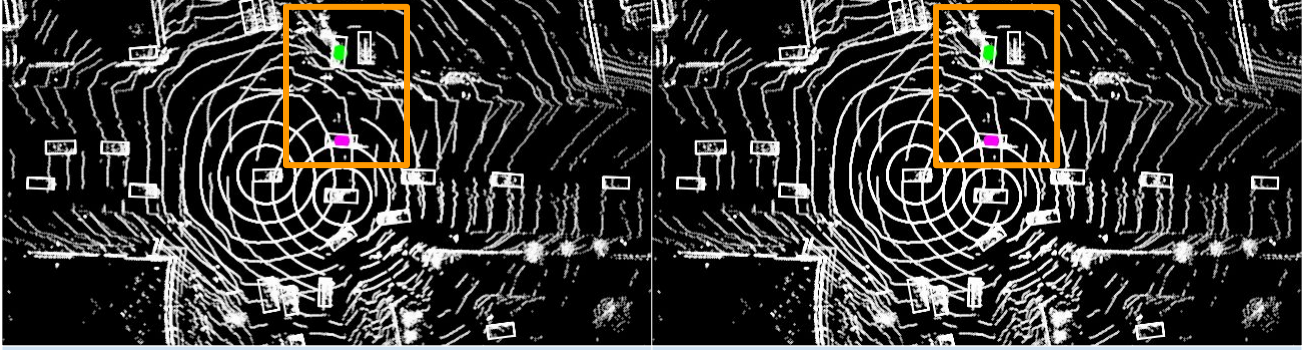
V2V-LLM: Yes, there is a car behind the **front object**. Its center location is [28.7, -1.8].

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

Q3: Grounding behind a reference object in a direction

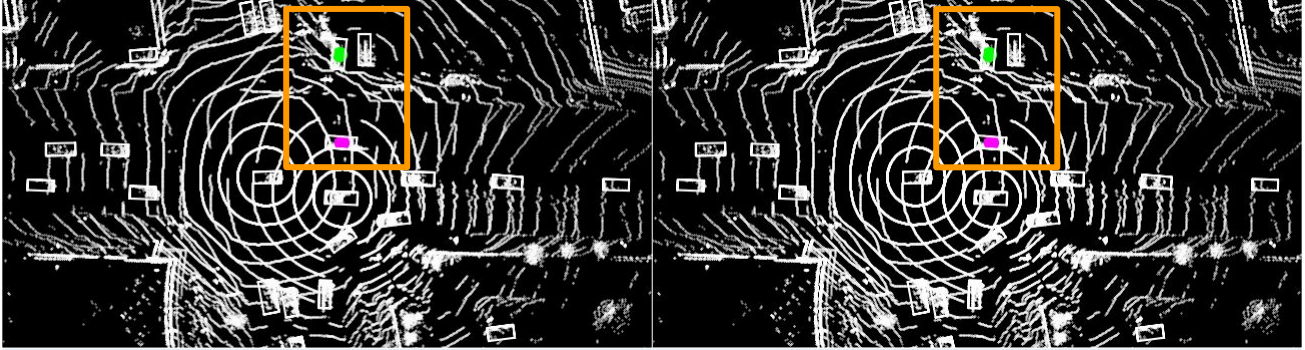
Q: Is there anything behind the **front right object**?

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



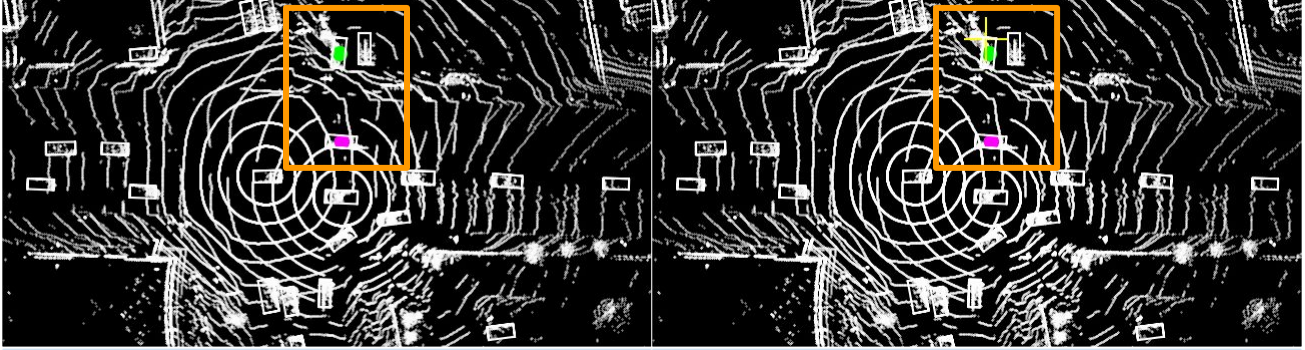
No Fusion: There is nothing behind the **front right object**.

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**.

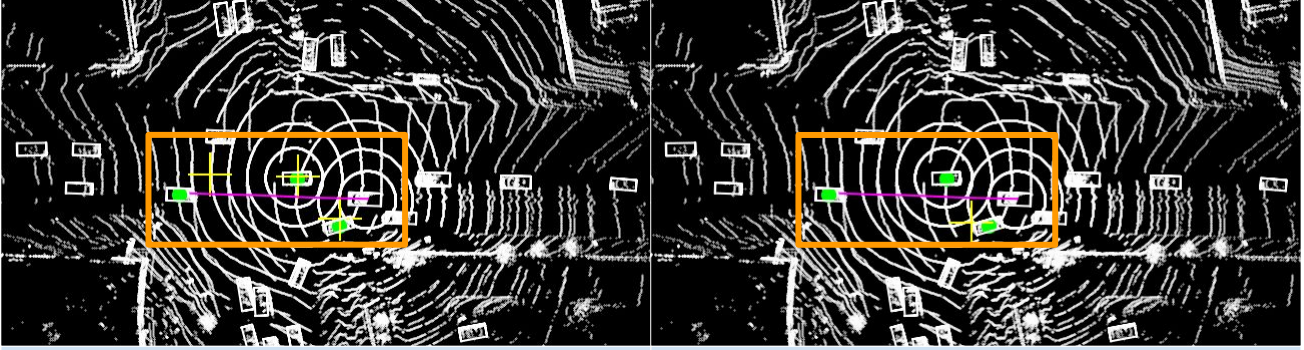
V2V-LLM: Yes, there is a car behind the **front right object**. Its center location is [1.0, 25.6].

Figure 20. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2V-split testing set. **Magenta** \circ : reference locations in questions. **Yellow** $+$: model output locations. **Green** \circ : ground-truth answers.

Q4: Notable Object Identification

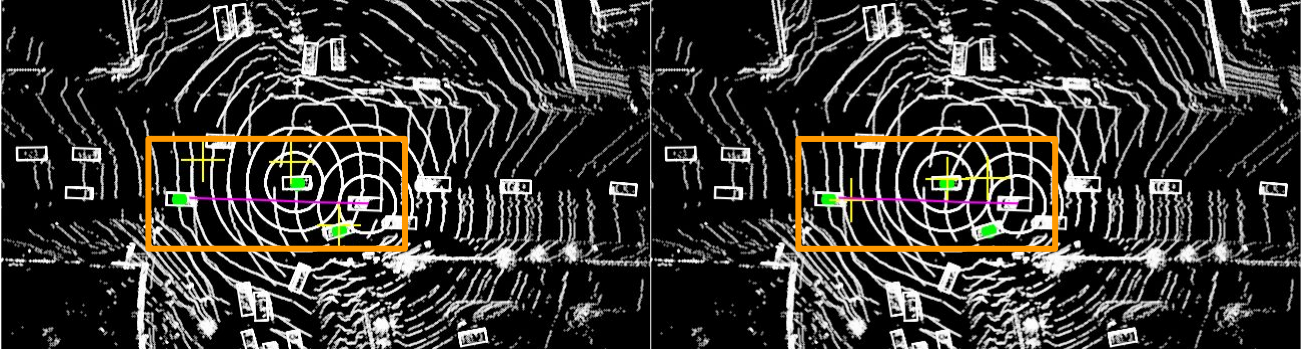
Q: I am **CAV_EGO**. Is there anything I need to be aware of if my planned future trajectory is $[(4.2, 0.1), (8.9, 0.2), (13.5, 0.4), (18.4, 0.6), (23.5, 0.7), (28.8, 0.9)]$?

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.



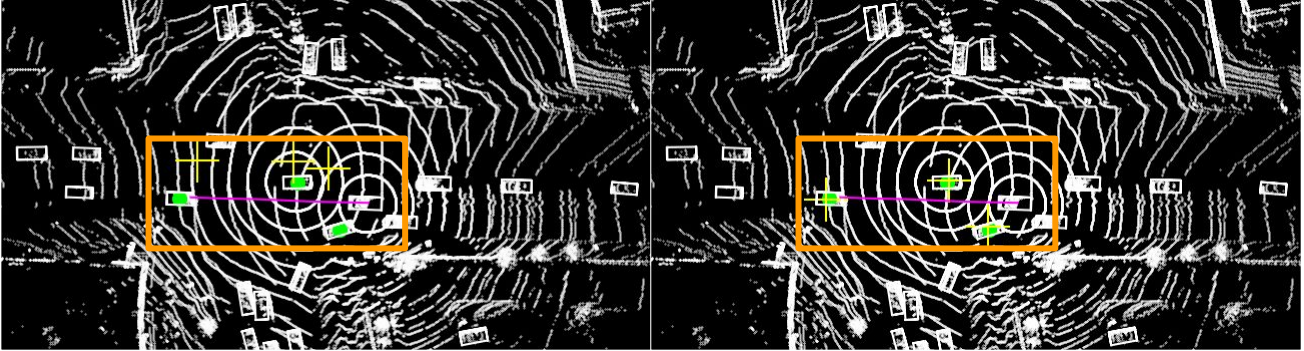
No Fusion: Yes, there are cars at $[4.6, -3.2]$, $[11.4, 3.6]$, $[25.7, 4.0]$, which are close to your planned future trajectory.

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.

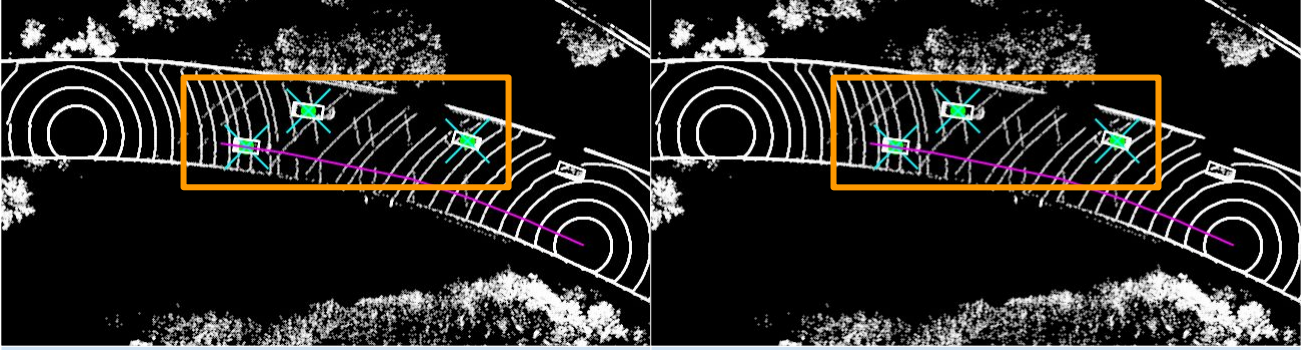
V2V-LLM: Yes, there are cars at $[4.8, -3.8]$, $[11.1, 3.6]$, $[31.1, 0.5]$, which are close to your planned future trajectory.

Figure 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 o:** ground-truth notable object locations. **Yellow +:** model identification outputs.

Q4: Notable Object Identification

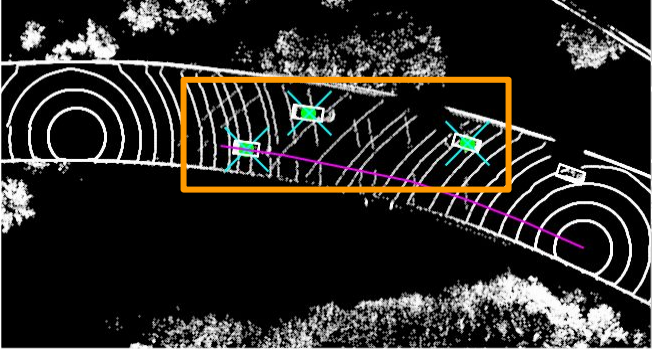
Q: I am CAV_1. Is there anything I need to be aware of if my planned future trajectory is $[(-73.0, -14.0), (-63.7, -10.4), (-54.2, -7.3), (-43.5, -5.1), (-33.6, -3.1), (-23.5, -1.6)]$?

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.

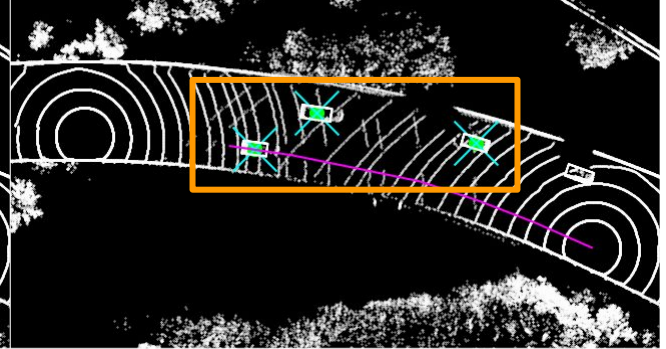


No Fusion: Yes, there are cars at $[-27.5, -2.0]$, $[-37.5, 3.7]$, $[-63.3, -1.1]$, which are close to your planned future trajectory.

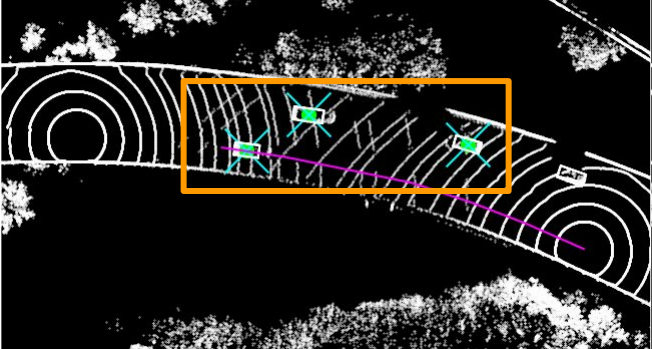
Early Fusion: Yes, there are cars at $[-27.5, -2.0]$, $[-37.5, 3.7]$, $[-63.3, -1.1]$, which are close to your planned future trajectory.



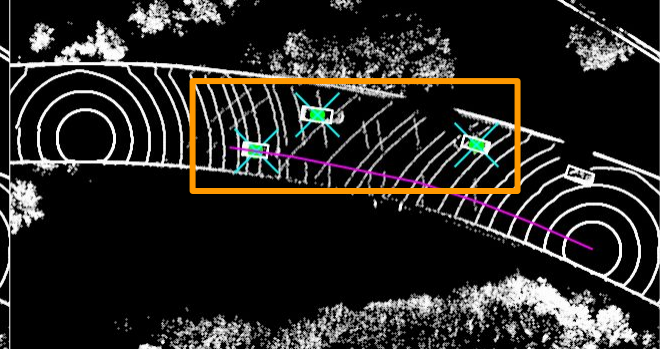
AttFuse: Yes, there are cars at $[-27.5, -2.0]$, $[-37.7, 3.7]$, $[-63.3, -1.1]$, which are close to your planned future trajectory.



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.



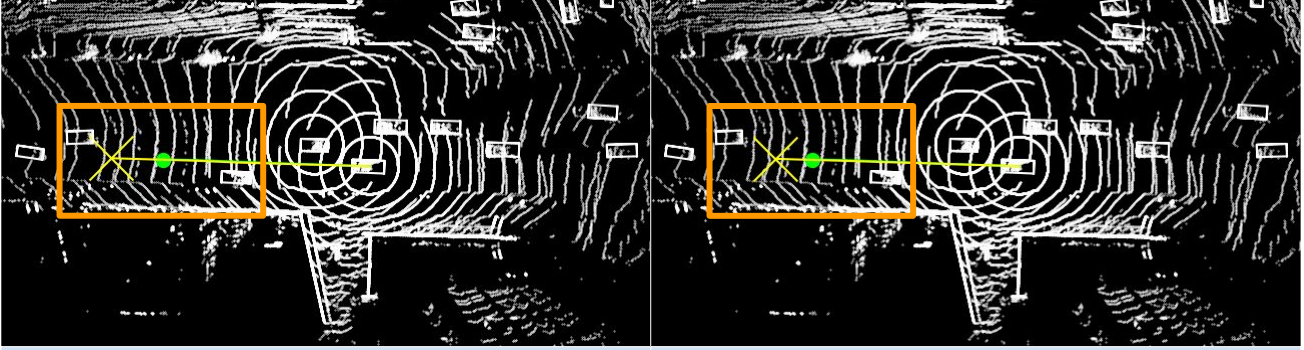
V2V-LLM: Yes, there are cars at $[-27.6, -2.0]$, $[-37.6, 3.7]$, $[-63.3, -1.1]$, which are close to your planned future trajectory.

Figure 22. 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 o:** ground-truth notable object locations. **Cyan x:** model identification outputs.

Q5: Planning

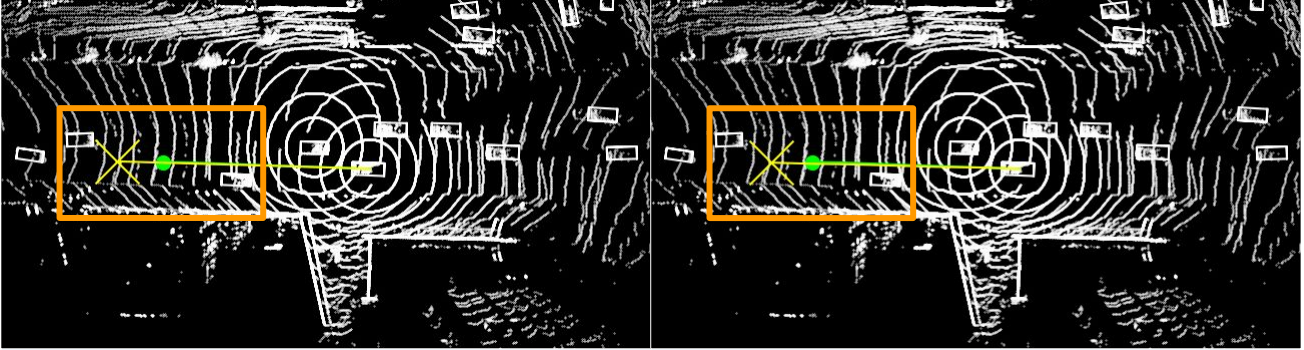
Q: I am **CAV_EGO**. What is the suggested future trajectory to avoid collision with nearby objects?

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)].



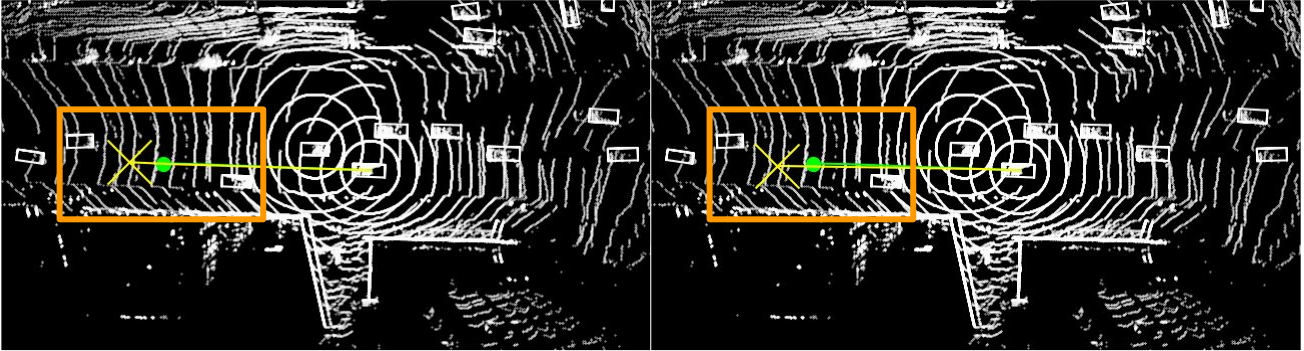
No Fusion: The suggested future trajectory is [(7.2,0.2),(14.6,0.4),(21.4,0.6),(28.4,0.9),(35.2,1.1),(42.0,1.3)].

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)].

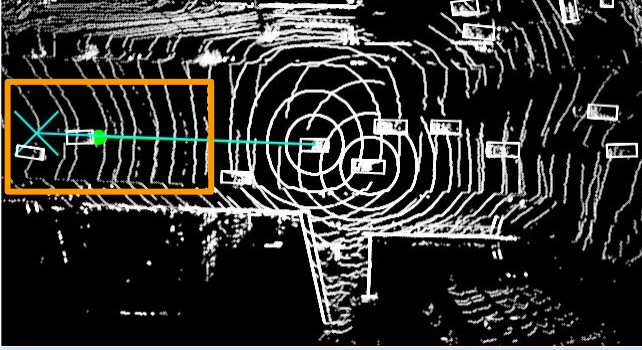
V2V-LLM: The suggested future trajectory is [(6.2,0.2),(12.6,0.3),(19.2,0.4),(25.9,0.5),(32.6,0.6),(39.4,0.7)].

Figure 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 o:** ending waypoints in ground-truth answers. **Yellow curve:** model planning outputs. **Yellow x:** ending waypoints in model outputs.

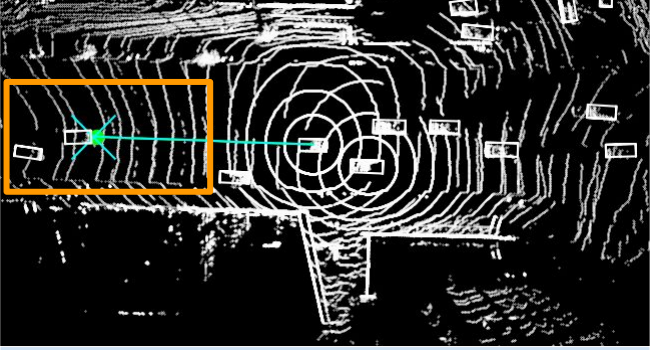
Q5: Planning

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

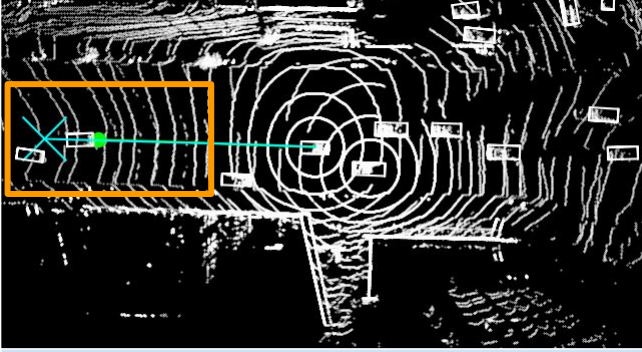
GT: The suggested future trajectory is [(5.7,0.1),(11.5,0.2),(17.4,0.4),(23.3,0.5),(29.3,0.6),(35.1,0.7)].



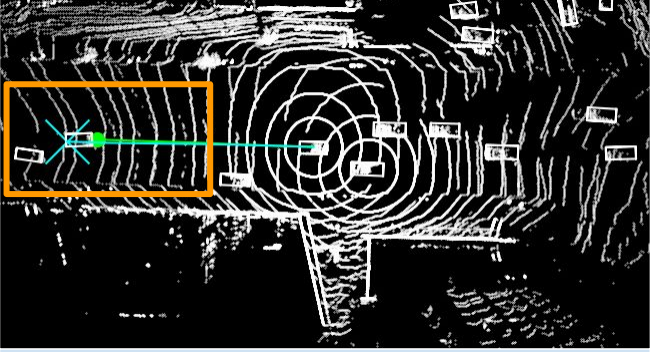
No Fusion: The suggested future trajectory is [(7.3,0.2),(14.7,0.4),(22.3,0.6),(29.9,0.8),(37.7,1.0),(45.2,1.2)].



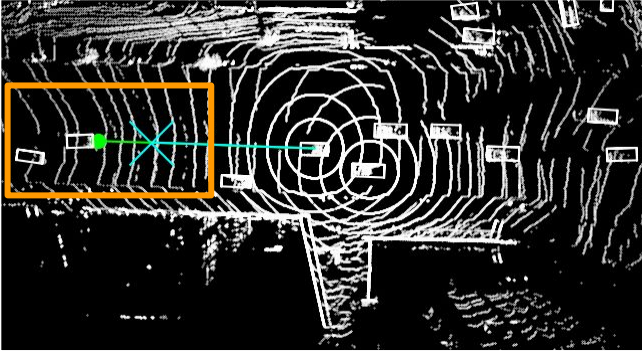
Early Fusion: The suggested future trajectory is [(5.7,0.1),(11.9,0.2),(17.8,0.4),(23.6,0.5),(29.6,0.6),(35.6,0.8)].



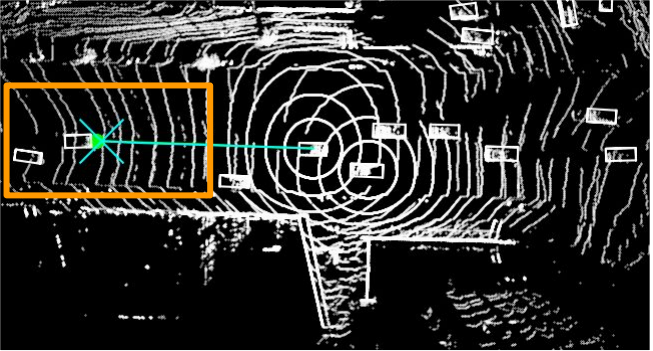
AttFuse: The suggested future trajectory is [(7.5,0.2),(14.2,0.3),(21.9,0.4),(29.0,0.5),(36.8,0.6),(43.9,0.7)].



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)].



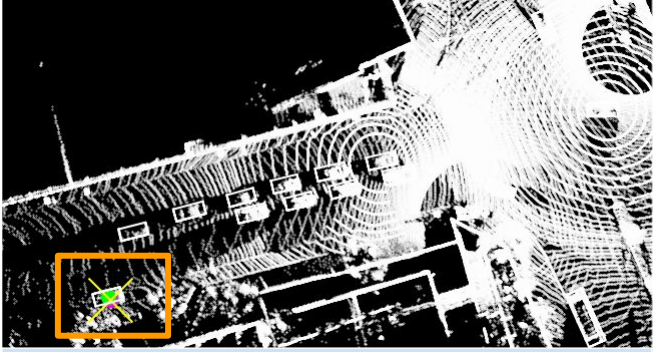
V2V-LLM: The suggested future trajectory is [(5.3,0.1),(10.7,0.2),(16.4,0.3),(22.2,0.4),(28.2,0.5),(34.4,0.7)].

Figure 24. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2V-split testing set. **Green curve:** future trajectories in ground-truth answers. **Green o:** ending waypoints in ground-truth answers. **Cyan curve:** model planning outputs. **Cyan x:** ending waypoints in model outputs.

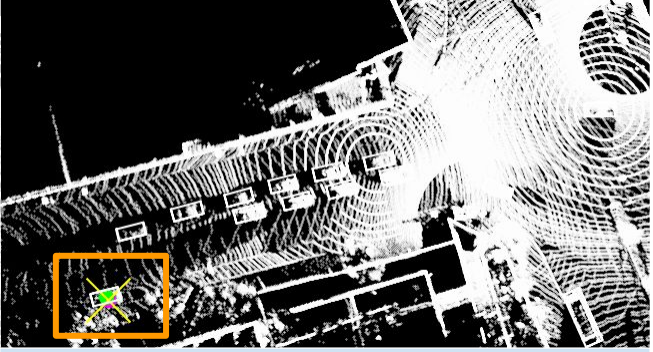
Q1: Grounding at a reference location

Q: Is there anything at the location [80.3, -30.8]?

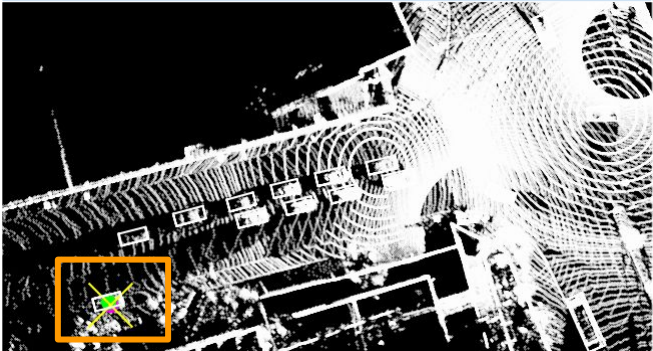
GT: Yes, there is a vehicle at the location. Its center location is [80.7, -30.3].



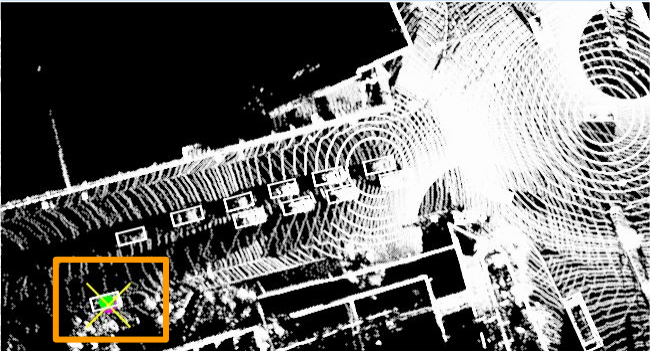
No Fusion: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].



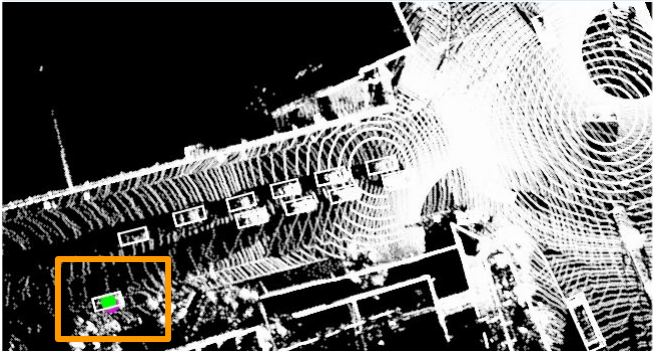
Early Fusion: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].



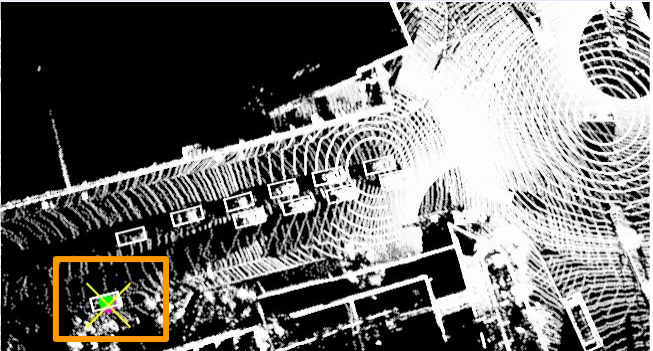
AttFuse: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].



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].



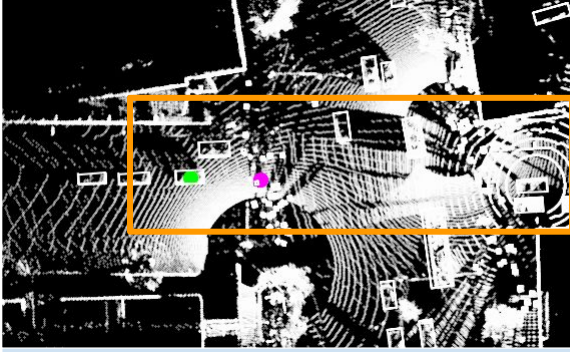
V2V-LLM: Yes, there is a vehicle at the location. Its center location is [80.3, -30.8].

Figure 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.

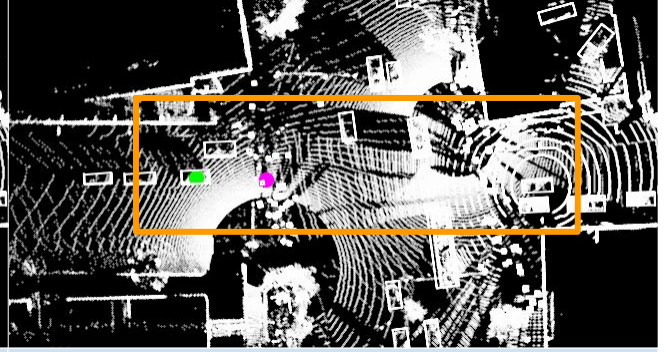
Q2: Grounding behind a reference object at a location

Q: Is there anything behind the object at [44.5, 0.9]?

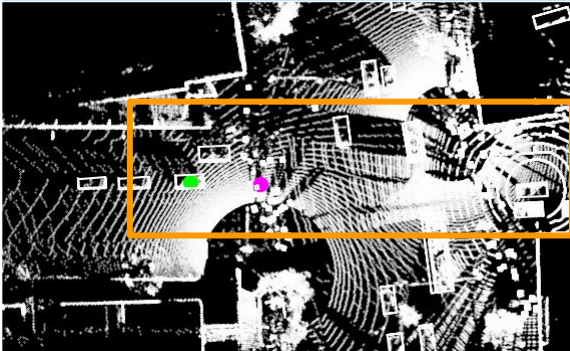
GT: Yes, there is a vehicle behind the object. Its center location is [55.9, 1.2].



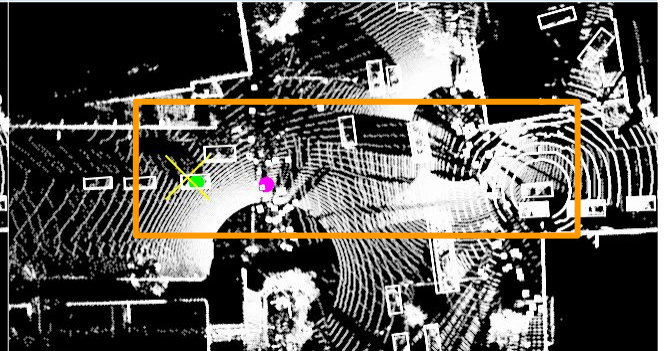
No Fusion: There is nothing behind the object.



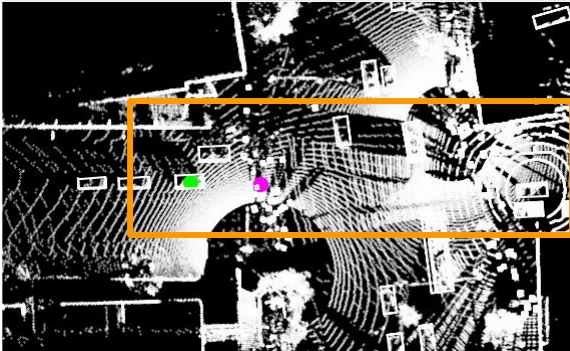
Early Fusion: There is nothing behind the object.



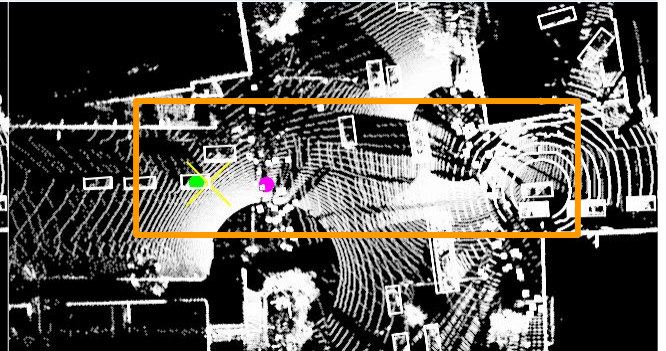
AttFuse: There is nothing behind the object.



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 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.

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].

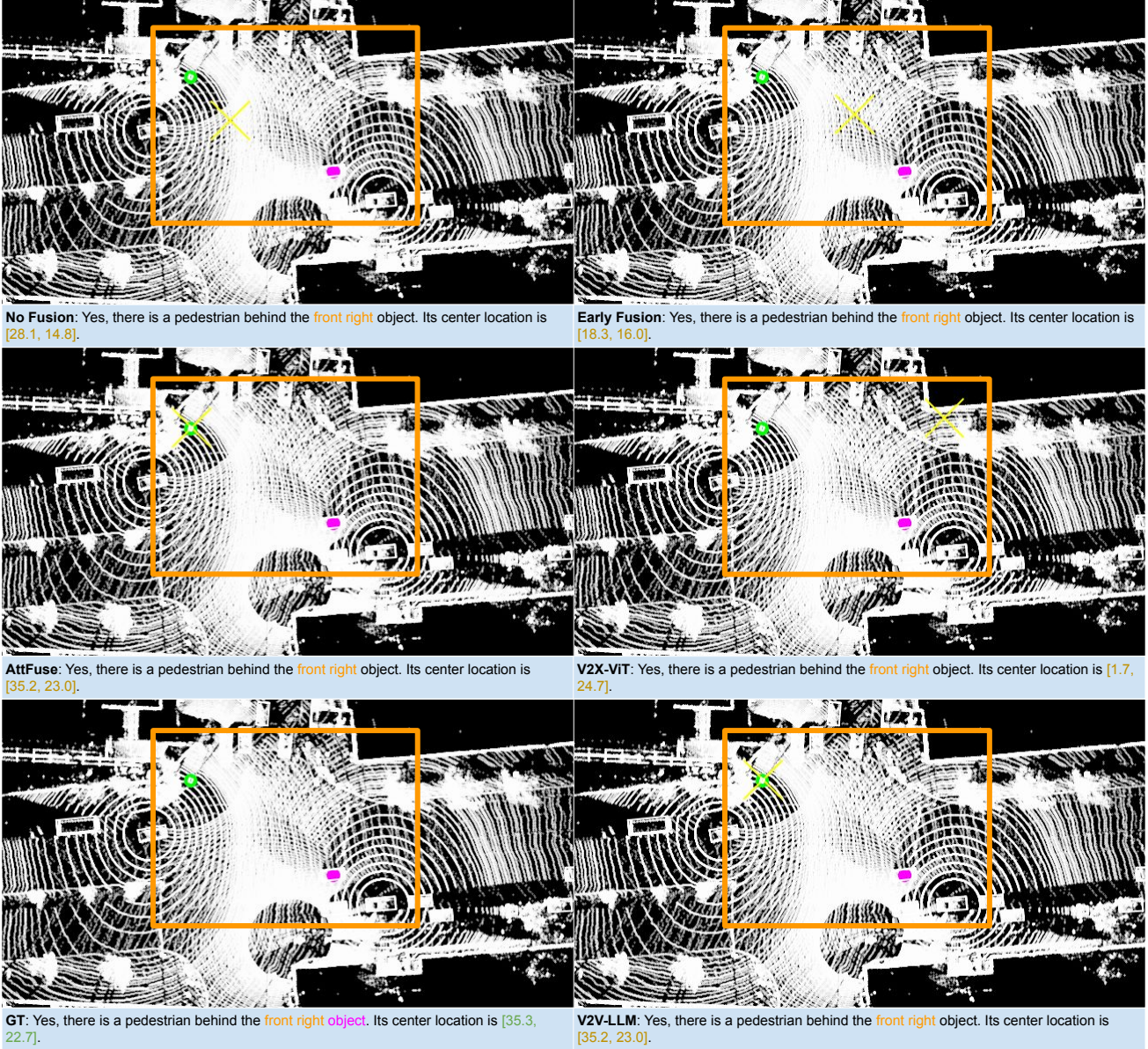
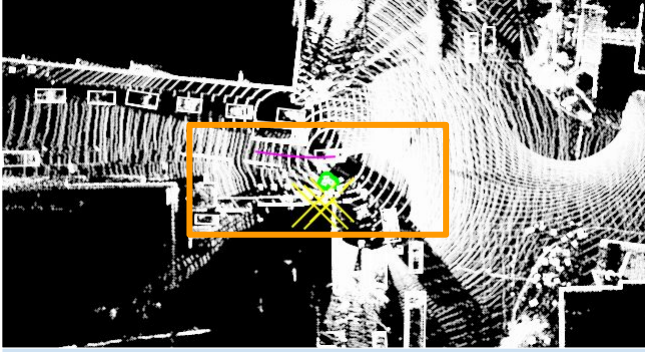


Figure 27. V2V-LLM and baseline methods' *grounding* results on V2V-QA's V2X-split testing set. **Magenta** \circ : reference locations in questions. **Yellow** \times : model output locations. **Green** \circ : ground-truth answers.

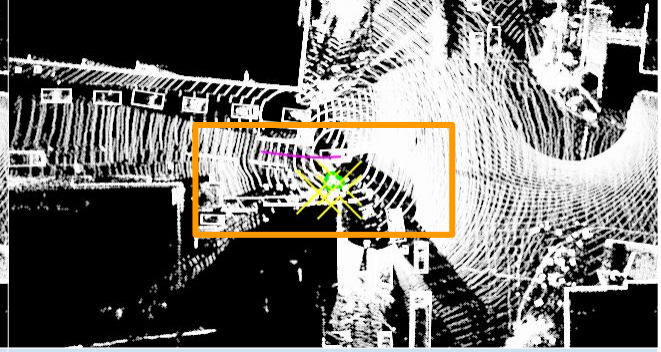
Q4: Notable Object Identification

Q: I am **CAV_EGO**. Is there anything I need to be aware of if my planned future trajectory is $[(2.0, -0.1), (4.1, -0.1), (6.1, 0.0), (8.2, 0.2), (10.4, 0.5), (12.6, 0.8)]$?

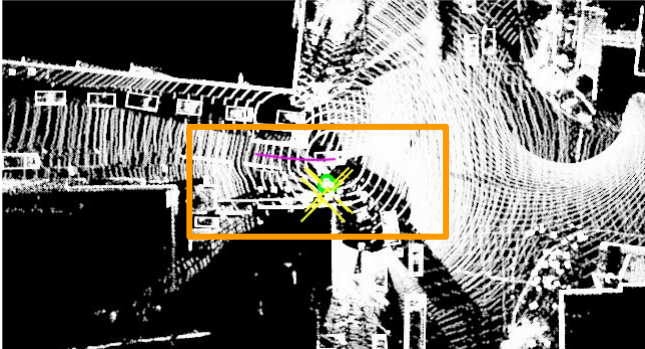
GT: Yes, there are pedestrians at $[1.3, -3.5]$, $[0.5, -4.2]$, $[1.6, -4.2]$ close to your planned future trajectory.



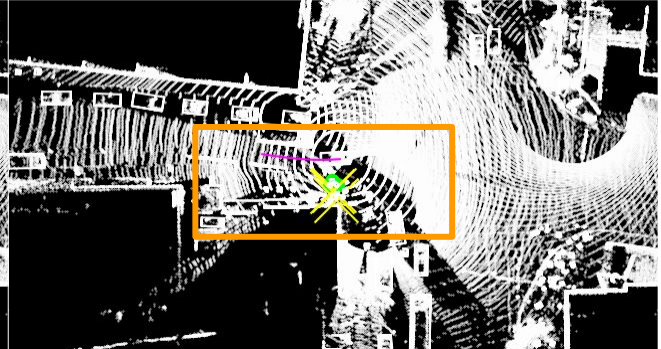
No Fusion: Yes, there are pedestrians at $[0.5, -7.0]$, $[3.4, -7.7]$, $[1.4, -8.1]$ close to your planned future trajectory.



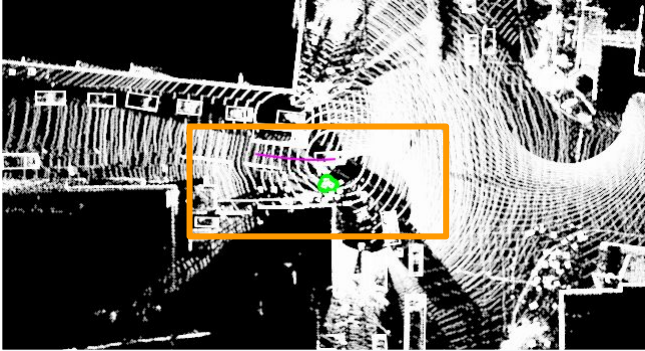
Early Fusion: Yes, there are pedestrians at $[0.1, -4.1]$, $[0.2, -5.8]$, $[2.9, -5.7]$ close to your planned future trajectory.



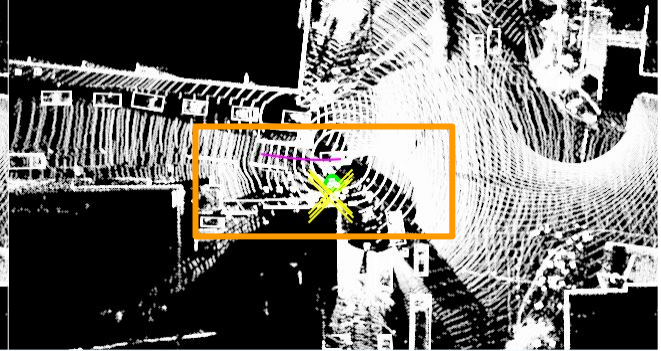
AttFuse: Yes, there are pedestrians at $[1.3, -4.9]$, $[0.6, -5.1]$, $[2.0, -6.5]$ close to your planned future trajectory.



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.



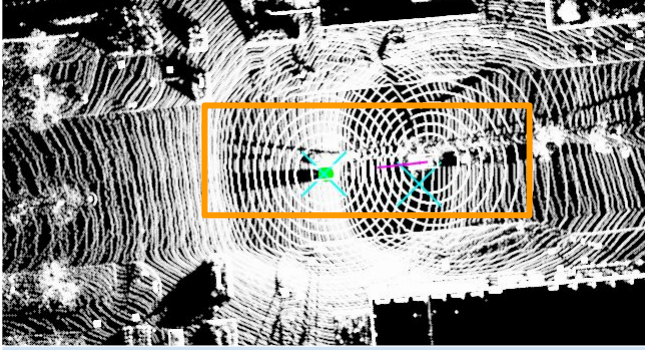
V2V-LLM: Yes, there are pedestrians at $[1.6, -5.2]$, $[1.4, -5.8]$, $[1.6, -6.6]$ close to your planned future trajectory.

Figure 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 o:** ground-truth notable object locations. **Yellow x:** model identification outputs.

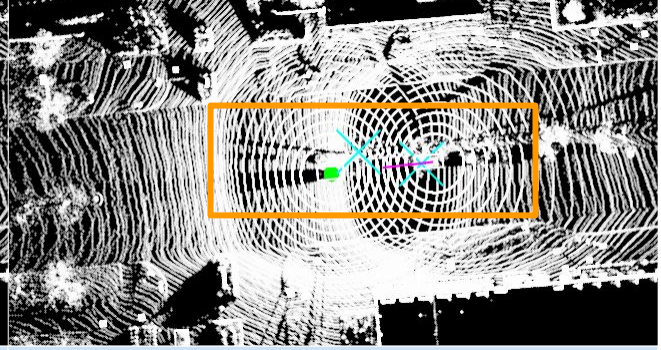
Q4: Notable Object Identification

Q: I am CAV_1. Is there anything I need to be aware of if my planned future trajectory is $[(-15.1, 1.7), (-13.5, 1.5), (-12.1, 1.4), (-10.8, 1.2), (-9.7, 1.1), (-8.9, 1.0)]$?

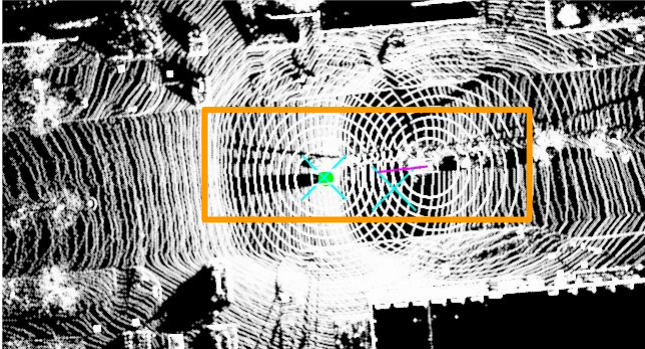
GT: Yes, there is a vehicle at $[-0.3, 0.1]$ close to your planned future trajectory.



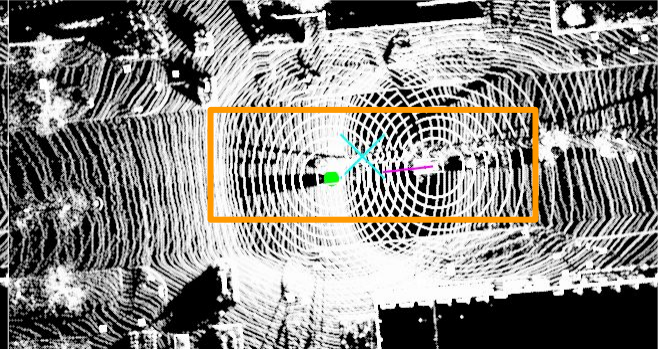
No Fusion: Yes, there are vehicles at $[-0.1, 0.1]$, $[-15.4, -1.6]$ close to your planned future trajectory.



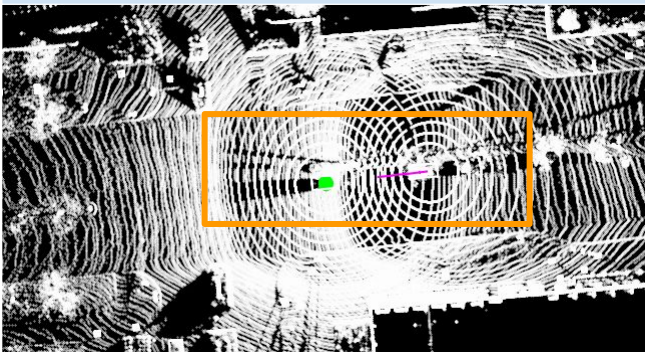
Early Fusion: Yes, there are vehicles at $[-9.1, 1.0]$, $[-15.7, -1.6]$ close to your planned future trajectory.



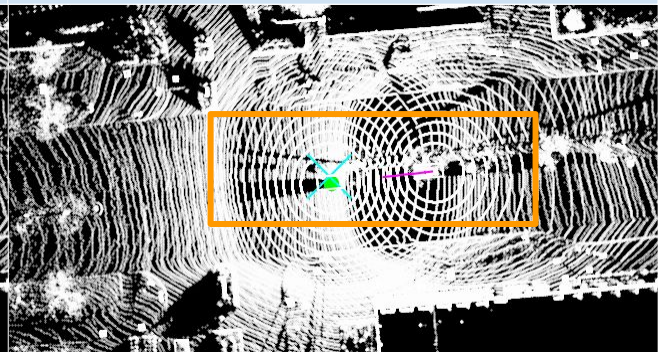
AttFuse: Yes, there are vehicles at $[-0.1, 0.2]$, $[-11.4, -1.6]$ close to your planned future trajectory.



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.



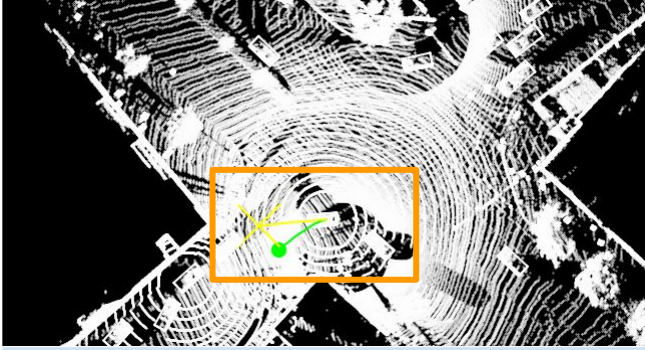
V2V-LLM: Yes, there is a vehicle at $[-0.1, 1.2]$ close to your planned future trajectory.

Figure 29. 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 o: ground-truth notable object locations. Cyan x: model identification outputs.

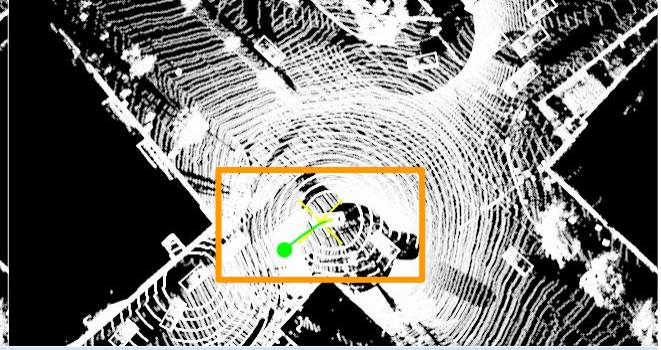
Q5: Planning

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

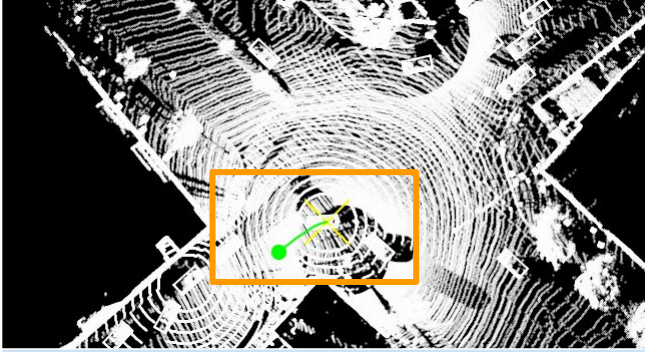
GT: The suggested future trajectory is [(1.0,-0.3),(2.1,-0.8),(3.3,-1.6),(4.7,-2.5),(6.1,-3.6),(7.7,-4.8)].



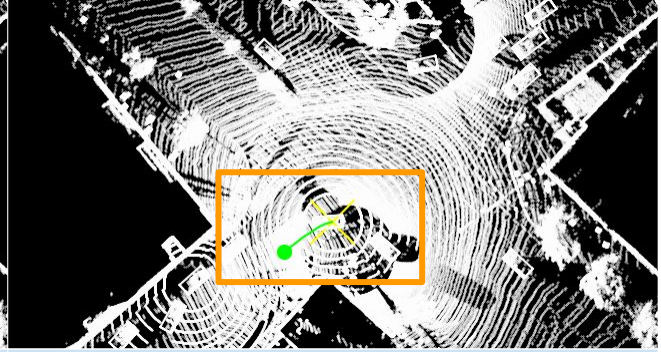
No Fusion: The suggested future trajectory is [(1.5,-0.1),(3.2,-0.3),(5.0,-0.4),(6.9,-0.6),(8.9,-0.8),(11.0,-1.0)].



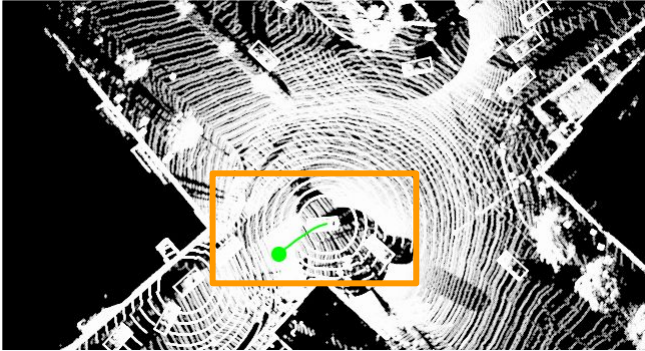
Early Fusion: The suggested future trajectory is [(0.3,-0.0),(0.6,-0.1),(0.9,-0.1),(1.2,-0.1),(1.6,-0.2),(2.0,-0.2)].



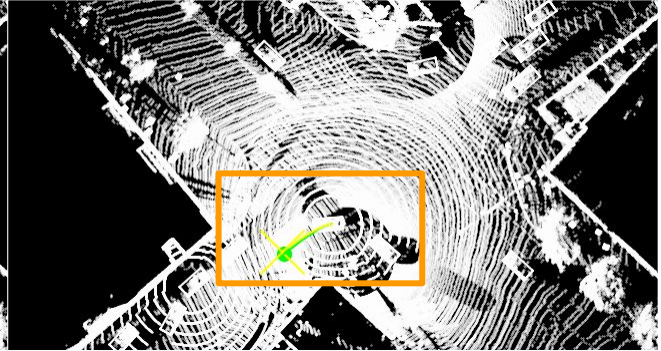
AttFuse: The suggested future trajectory is [(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0),(0.0,-0.0)].



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)].



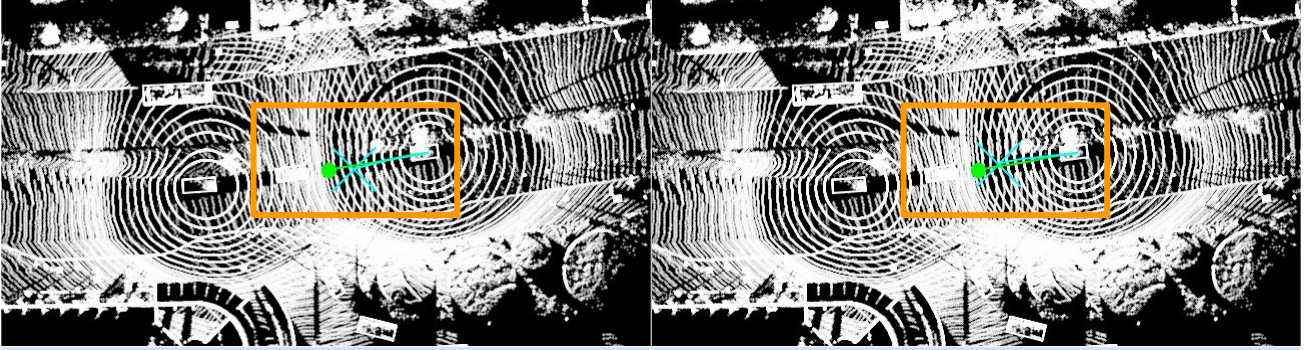
V2V-LLM: The suggested future trajectory is [(1.1,-0.3),(2.3,-0.8),(3.6,-1.5),(5.0,-2.4),(6.5,-3.4),(8.1,-4.5)].

Figure 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 o:** ending waypoints in ground-truth answers. **Yellow curve:** model planning outputs. **Yellow x:** ending waypoints in model outputs.

Q5: Planning

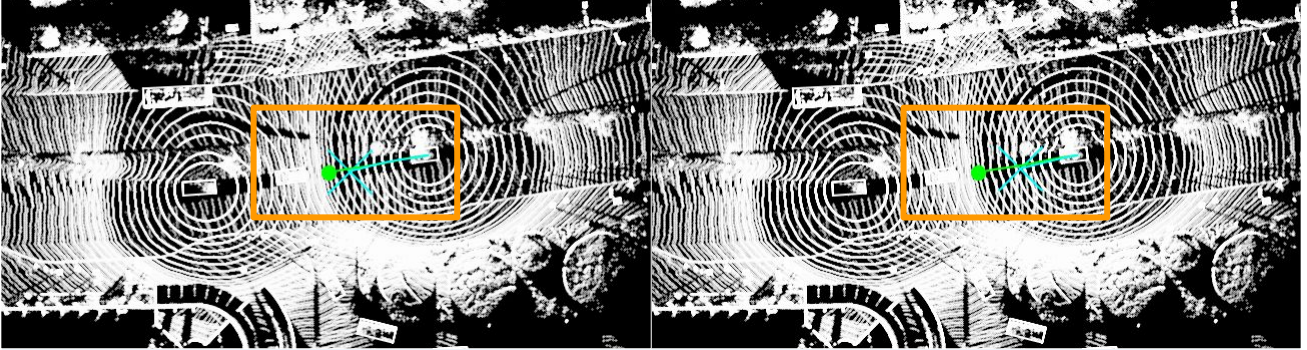
Q: I am CAV_1. What is the suggested future trajectory to avoid collision with nearby objects?

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)].



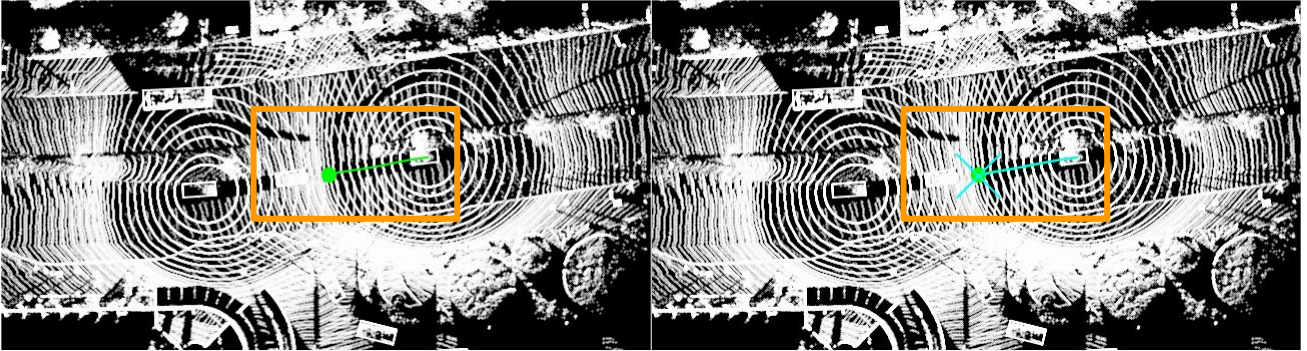
No Fusion: The suggested future trajectory is [(2.2,0.0),(4.4,0.0),(6.5,-0.0),(8.5,-0.2),(10.5,-0.3),(12.3,-0.5)].

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)].

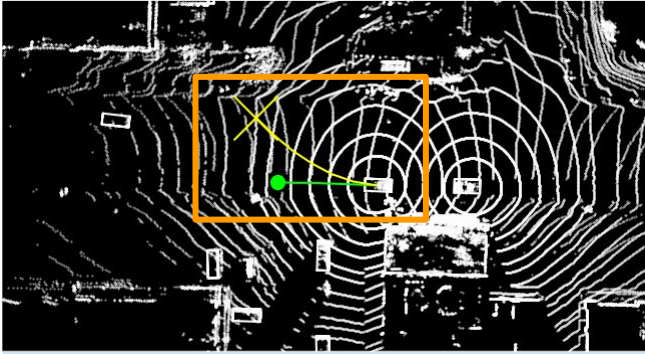
Figure 31. V2V-LLM and baseline methods' *planning* results on V2V-QA's V2X-split testing set. **Green curve:** future trajectories in ground-truth answers. **Green o:** ending waypoints in ground-truth answers. **Cyan curve:** model planning outputs. **Cyan x:** ending waypoints in model outputs.

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

[[2.5,0.1),(5.2,0.2),(7.9,0.3),(10.6,0.4),(13.4,0.4),(16.1,0.5)].

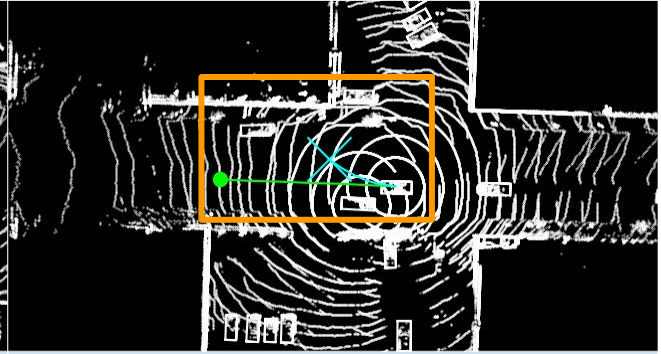


V2V-LLM: The suggested future trajectory is
[[3.7,0.6),(7.2,1.8),(10.4,3.5),(13.6,5.6),(16.6,8.1),(19.6,11.0)].

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

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