
Supplementary material - STSBench: A Spatio-temporal Scenario Benchmark for Multi-modal Large Language Models in Autonomous Driving

Christian Fruhwirth-Reisinger^{1,2,*} Dušan Malić^{1,2,*} Wei Lin³
David Schinagl¹ Samuel Schulter^{4,†} Horst Possegger^{1,2}

¹Institute of Visual Computing, Graz University of Technology

²Christian Doppler Laboratory for Embedded Machine Learning

³Institute for Machine Learning, Johannes Kepler University Linz

⁴Amazon

Contents

1	Benchmark details	2
1.1	STSnu statistics	2
1.2	Scenario catalog	3
1.3	Scenario definition pseudo code	9
1.4	Scenario verification	9
1.5	Multiple-choice question generation	14
2	Driving expert baselines	14
3	Detailed results & analysis	15
4	Additional experiments	20
5	Prompt examples	23

*Equal contribution. Corresponding authors: {reisinger, dusan.malic}@tugraz.at

†This work is independent of the author's employment at Amazon.

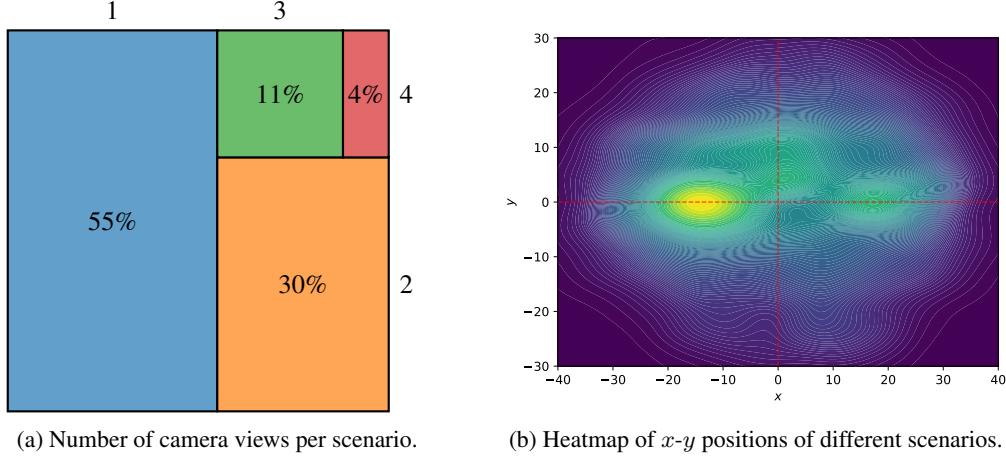


Figure 1: **Scenario statistics.** Distribution of scenarios with agent involvement across camera views and on the x - y plane.

1 Benchmark details

1.1 STSnu statistics

Since driving scenes are dynamic environments in which both the ego vehicle and other traffic participants (agents) are in motion, a single camera view is often insufficient to capture all relevant interactions. As a result, most modern datasets [4, 28, 32] employ multi-camera systems with slightly overlapping fields of view. In our benchmark, we mine scenarios where agents are distributed across varying numbers of camera views, as illustrated in Fig.1a. We observe that approximately half of the scenarios are confined to a single view, while the remaining scenarios span up to four views over the observed time period. The relatively high concentration of actions in a single view can be attributed to two main factors: 1) In the majority of sequences in nuScenes, the ego vehicle either drives straight or remains stationary. 2) Many scenarios occur in the front or rear view of the ego vehicle, as these cameras primarily capture the road it travels on. Each of these factors, individually and in combination, contributes to the skew in this distribution. Nevertheless, the remaining half of the scenarios span multiple views, underscoring the importance of evaluating a model’s ability to generalize across spatially distributed visual inputs.

To maintain a balanced dataset, we sub-sample the mined scenarios and remove samples from overrepresented categories. We define three criteria for agent-related scenarios to guide this selection: 1) occlusion, 2) distance to the ego vehicle, and 3) spatial distribution. Based on these criteria, we retain scenarios that are highly visible, occur in the near surrounding of the ego vehicle, and are spatially well-distributed around it. The first criterion is straightforward since visual systems cannot effectively reason about occluded agents. For STSnu, the underlying dataset nuScenes [4] provides occlusion rates for agent bounding box annotations. We discard all scenarios where the agent is visible less than 30% in the first video frame. The second criterion prioritizes agents that are closer to the ego vehicle, as distant objects are harder to perceive, and the benchmark is not focused on evaluating long-range detection or reasoning about small objects. Therefore, we remove agents located beyond 30 meters from the ego vehicle. The final criterion ensures a diverse set of samples with respect to camera viewpoints and spatial coverage, as illustrated in Fig.1b. We apply farthest point sampling (FPS) in the x - y plane. Afterwards, we still observe an unbalanced distribution, with a disproportionately high number of scenarios occurring in the rear area of the ego vehicle, reflecting biases in the underlying data. However, the remaining scenarios are relatively well distributed across other spatial regions.

1.2 Scenario catalog

The benchmark generation starts with the definition of a scenario catalog. It includes all scenarios that should be covered in the benchmark and assigns negative scenarios for each entry. Negative scenarios do not occur during the assigned scenario and serve later in the benchmark creation as wrong-choice options for the multiple-choice answer generation. In our scenario catalog, we list scenarios for ego-vehicle (Fig. 2), other agents (Fig. 4), interactions between ego-vehicle and agents (Fig. 3), and interactions among other agents (Fig. 5). For the respective categories, we define the scenarios in text form in Tables 3, 2, 1 and, 4.



Figure 2: **Scenario catalog.** Ego scenarios with assigned negative scenarios in red.



Figure 3: **Scenario catalog.** Ego-Agent scenarios with assigned negative scenarios in red.

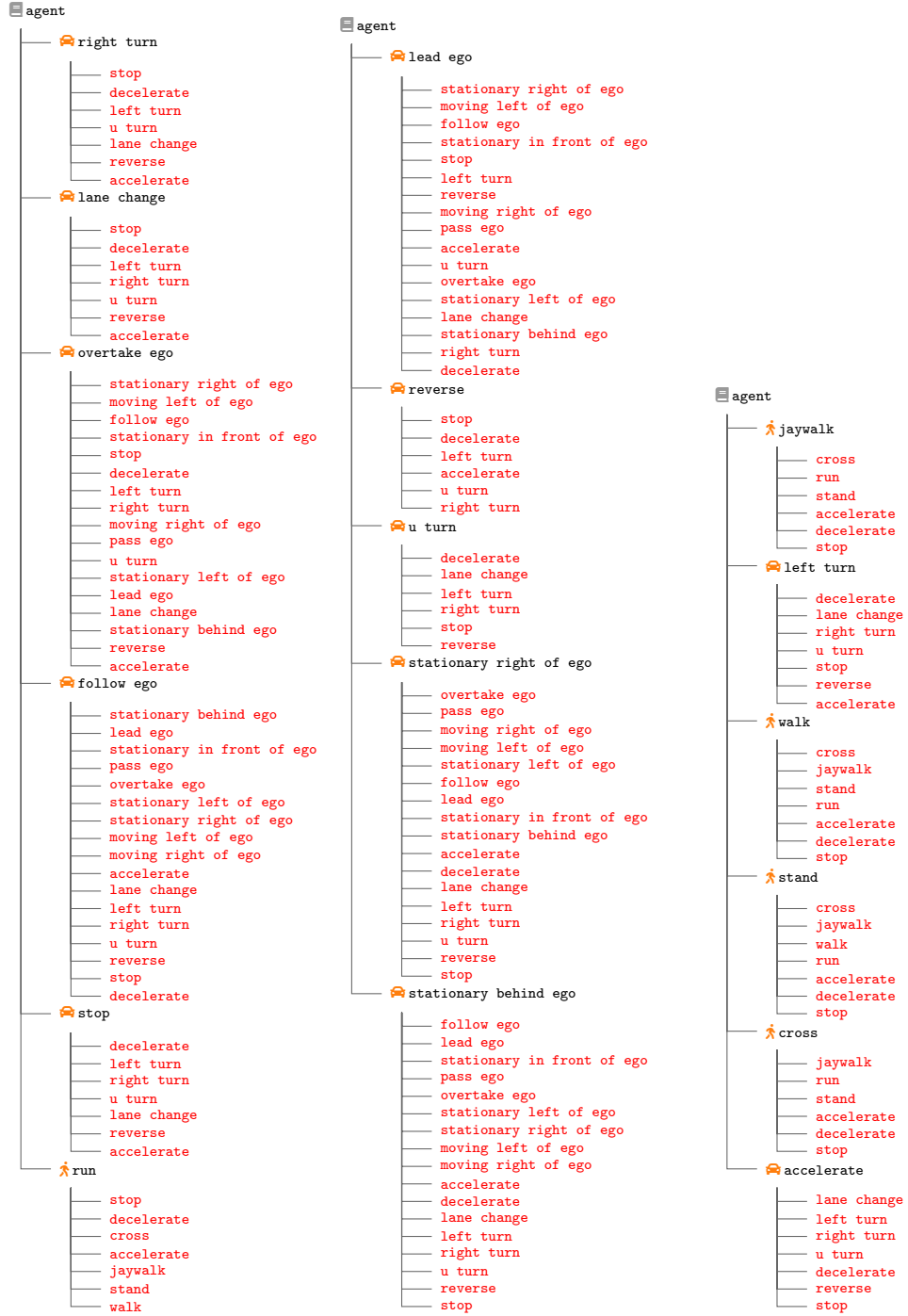


Figure 4: **Scenario catalog.** Agent scenarios with assigned negative scenarios in red.



Figure 5: **Scenario catalog.** Agent-Agent scenarios with assigned negative scenarios in red.

Table 1: **Ego-to-agent scenario definition.** Textual description of scenarios where the ego-vehicle interacts with other agents.

lead agent	Ego travels ahead of agent at a similar speed while maintaining a consistent distance.
follow agent	Ego is driving behind agent at a similar speed while maintaining a consistent distance.
pass agent	Ego in the adjacent lane overtakes the stopped agent.
overtake agent	Ego in the adjacent lane and moves ahead of agent while both are in motion.
stationary left of agent	Agent is fully stopped and remains stationary to the left of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary in front of agent	Agent is fully stopped and remains stationary in front of agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
wait ped cross	Ego comes to a stop or remains stationary, yielding the right-of-way to agent who is crossing or preparing to cross the road, while maintaining awareness of the agent’s movement and ensuring a safe distance until the crossing is complete.

Table 2: **Agent scenario definition.** Textual description of agent scenarios.

stand	Agent (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.
walk	Agent (pedestrian) moves at a steady, moderate pace, typically following designated paths or crosswalks.
jaywalk	Agent (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.
run	Agent (pedestrian) is running and moves rapidly.
cross	Agent (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
accelerate	Agent is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
stop	Agent is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
reverse	Agent is moving in reverse, either to park, navigate a tight space, or adjust its position.
left turn	Agent is executing a left turn at an intersection or junction.
right turn	Agent is executing a right turn at an intersection or junction.
u turn	Agent is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
lane change	Agent is transitioning from its current lane to an adjacent lane.
lead ego	Agent travels ahead of ego at a similar speed while maintaining a consistent distance.
follow ego	Agent is driving behind ego at a similar speed while maintaining a consistent distance.
overtake ego	Agent in the adjacent lane moves ahead of ego while both are in motion.
stationary right of ego	Agent is fully stopped and remains stationary to the right of ego, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary behind ego	Agent is fully stopped and remains stationary behind ego (which is also stopped), such as when waiting at a traffic light, in a parking lot, or in any other queuing scenario.

Table 3: **Ego scenario definition.** Textual description of ego-vehicle scenarios.

accelerate	Ego is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
decelerate	Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
stop	Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
left turn	Ego is executing a left turn at an intersection or junction.
right turn	Ego is executing a right turn at an intersection or junction.
lane change	Ego is transitioning from its current lane to an adjacent lane.

Table 4: **Agent-to-agent scenario definition.** Textual description of scenarios where agents interact with other agents.

walk alongside	Agent (pedestrian) and agent (pedestrian) walk side by side at a steady, moderate pace.
walk opposite	Agent (pedestrian) and agent (pedestrian) walk toward each other at a moderate pace, cross paths, and proceed.
lead agent	Agent travels ahead of agent at a similar speed while maintaining a consistent distance.
follow agent	Agent is driving behind agent at a similar speed while maintaining a consistent distance.
pass agent	Agent in the adjacent lane overtakes the stopped agent.
overtake agent	Agent in the adjacent lane and moves ahead of agent while both are in motion.
moving left of agent	Agent is traveling in parallel to the left of agent (e.g., in adjacent lanes or side by side), with one vehicle maintaining a leftward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
moving right of agent	Agent is traveling in parallel to the right of agent (e.g., in adjacent lanes or side by side), with one vehicle maintaining a rightward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
stationary left of agent	Agent is fully stopped and remains stationary to the left of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary right of agent	Agent is fully stopped and remains stationary to the right of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary in front of agent	Agent is fully stopped and remains stationary in front of agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
stationary behind agent	Agent is fully stopped and remains stationary behind agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
wait ped cross	Agent comes to a stop or remains stationary, yielding the right-of-way to a agent who is crossing or preparing to cross the road, while maintaining awareness of the agent’s movement and ensuring a safe distance until the crossing is complete.

Algorithm 1 Pseudo code acceleration.

Input:

- `objs`: A list of Ego or Agent objects, each with:
 - `.velocity`: A numeric velocity attribute
 - `.is_vehicle`: A boolean flag indicating if the object is a vehicle
 - `frames`: Number of consecutive frames to evaluate for acceleration
 - `threshold_ms`: Minimum velocity increase required over the window
-

Output:

- A generator yielding `(start_idx, end_idx)` tuples marking intervals of acceleration
-

Steps

1. If the first object in `objs` is not a vehicle:
Return: `(None, None)`
 2. Extract the velocity of each object into an array `vel`
 3. If length of `vel` < `frames`:
Return: `(None, None)`
 4. Create overlapping sequences of `frames` velocities \rightarrow `vel_w`
 5. Compute first-order velocity differences in each window \rightarrow `acc_w`
 6. Create a boolean mask where:
 - All differences in `acc_w` > 0.1 (positive acceleration)
 - Total velocity change in each window `vel_w` > `threshold_ms`
 7. Merge valid acceleration regions in overlapping masks
 8. For each `start_idx` where mask is True:
Compute end index: `start_idx + frames` \rightarrow `end_idx`
Yield: `(start_idx, end_idx)`
-

1.3 Scenario definition pseudo code

In order to illustrate the logical structure of our mining process, we show the pseudo code for the ego-scenario *acceleration* in Algorithm 1. For full implementation details, we refer the readers to our open-source code repository (<https://github.com/LRP-IVC/STSBench>), where all scenario extraction functions are defined within *annotator/mining/ego.py*.

1.4 Scenario verification

Verification tool. Fast and simple verification requires a clean and versatile visualization interface paired with a simple input mask. In Fig. 7, we show the verification command line tool (a) of STS-Bench, that can be used via mouse or keyboard only. The visualization interface (b) uses Rerun [24] and shows various available input modalities, such as LiDAR point clouds, agent trajectories on a map, velocities, and multi-view camera images for consecutive time frames. The presented validation tool enables fast inspection of mined scenarios in different modalities and effortless acceptance or rejection of samples.

Verification insights. To demonstrate the efficiency of our verification, we provide a detailed time analysis in Fig. 6. We observe an average of 8.3 seconds, 16.0 seconds, and 12.9 seconds for reviewers A, B, and C, respectively. While some scenes require more attention, most of the scenarios are obvious and can therefore be confirmed quickly.

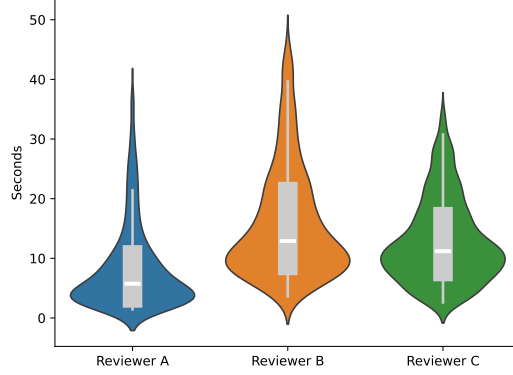


Figure 6: **Validation time.** Comparison of the validation duration for each reviewer.

Table 5: **Cohen’s kappa.** Verification agreement between all reviewer pairs for mined scenarios (Positive) and assigned negative scenarios (Negative).

Reviewer 1	Reviewer 2	Positive	Negative
A	B	0.88940	0.97333
A	C	0.92001	0.97822
B	C	0.88294	0.97767

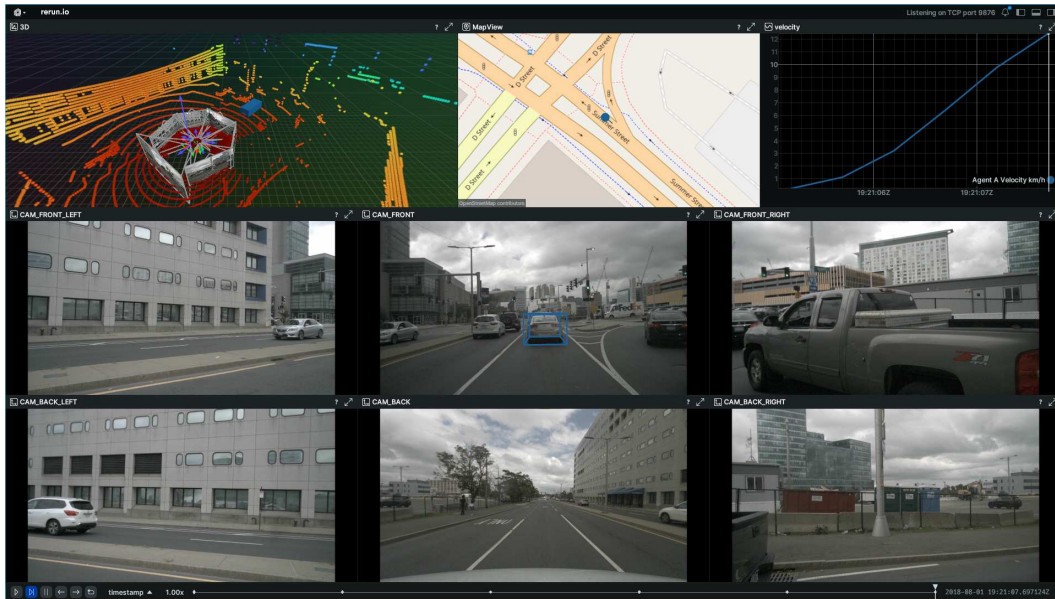
In total, all three reviewers verified 1188 mined scenarios of which 1017 have been confirmed by the majority of them (at least two). Hence, the overall agreement among all reviewers on the mined scenarios was 85.6% confirming consistent verifications reflected by a Fleiss’ kappa of 0.87743. For the same set of scenarios, the agreement regarding suitable predefined negative scenarios has been 79.2%. This lower level of agreement is due to the conservative approach of only considering negative ratings on which all reviewers agree. The actual agreement on negatives is very consistent with a Fleiss’ kappa of 0.97742. This can also be observed by looking at Cohen’s kappa between reviewer pairs as shown in Table 5.

In the following, we show typical verification mistakes and disagreements between reviewers leading to either accepted or rejected scenarios. These verification differences are solved by majority voting. Fig. 8 shows a highly occluded person (blue), which has been considered too difficult by one of the three reviewers. However, although only a small part of the person is visible, this sample can serve as a difficult scenario for the benchmark. The next two scenes in Fig. 9 and Fig. 10 illustrate scenarios that the reviewers have rejected because they do not follow the scenario definition exactly. In Fig. 9, the ego vehicle *passing* the agent (green) is not in the adjacent line. Fig. 10 shows a similar issue, where the agent (green) *following* the ego vehicle has another agent in between, which is not considered following the ego vehicle in our scenario definition. For *jaywalking*, the definition states that a pedestrian crosses the street outside designated crossing areas. However, not all designated areas are properly annotated in map data, as shown in Fig. 11. Construction workers operating in a safe construction space (Fig. 11a) or pedestrians crossing a driveway between two sidewalks (Fig. 11b) are not considered jaywalking.

Disagreements can also occur due to inaccurate visualizations of projected 3D bounding boxes onto the camera image. In Fig. 12, we illustrate an example where bounding boxes indicate two neighboring vehicles (green, blue) in the camera image (a). However, a closer look at the LiDAR scan (b) reveals another agent between them. Consequently, the ambiguous verification of such samples could either stem from the inaccurate projection or a misunderstanding of the scenario description. Finally, subjective perception can also lead to ambiguous verification results. Fig. 13 shows a lane change scenario hardly recognizable from the trajectory in the map view (b). However, the vehicle (blue) is heavily occluded, and lane markings are not visible in the camera image (a). Thus, such hardly recognizable scenarios have been rejected in the manual verification stage.



(a) Verification GUI.

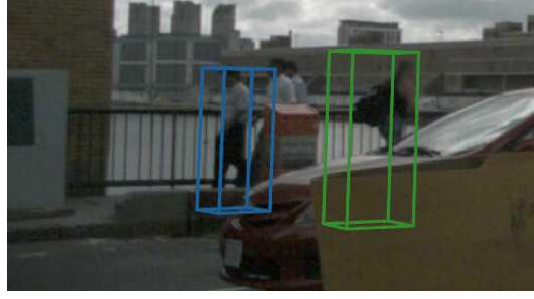


(b) Verification rerun.

Figure 7: **Verification interface.** Screenshots of our tool for fast and simple verification (a) of mined traffic scenarios by inspecting (b) recordings from various available modalities.

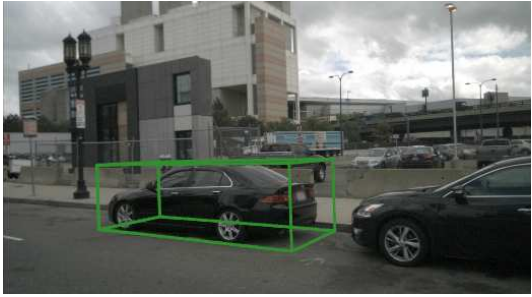


(a) First frame (CAM_FRONT_LEFT).

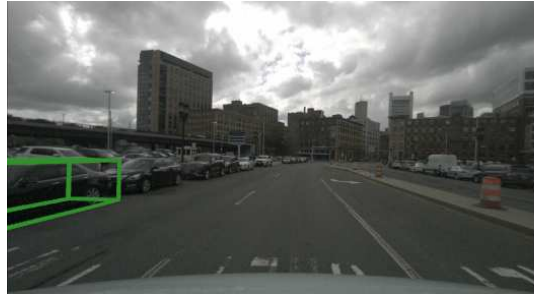


(b) Fourth frame (CAM_FRONT_LEFT).

Figure 8: **Agent walk opposite.** Accepted scenario sample after review, since the occlusion level has been rated difficult but reasonable. The bounding box is not visible for the tested models.



(a) First frame (CAM_FRONT_RIGHT).



(b) Last frame (CAM_BACK).

Figure 9: **Ego-vehicle passes agent.** The scenario sample was rejected after review since the ego-vehicle is not in the adjacent lane of the referenced agent. The bounding box is not visible for the tested models.



(a) First frame (CAM_BACK).



(b) Last frame (CAM_BACK).

Figure 10: **Agent following ego-vehicle.** Rejected scenario sample after review since there is another agent between the ego-vehicle and the referenced agent. The bounding box is not visible for the tested models.

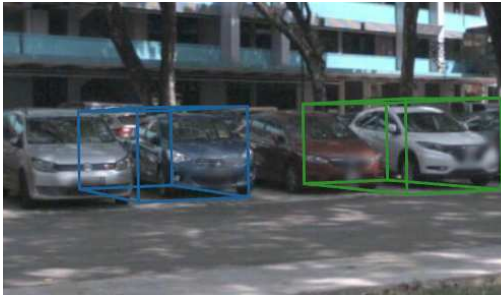


(a) Construction worker in a safe space.

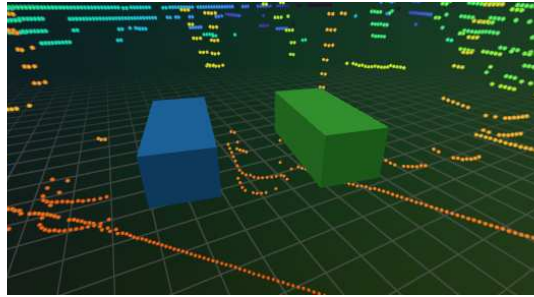


(b) Pedestrian on sidewalk.

Figure 11: **Agent jaywalking.** Rejected scenario sample after review since there is no jaywalking. The bounding box is not visible for the tested models.



(a) First frame (CAM_FRONT_LEFT).



(b) 3D LiDAR scene.

Figure 12: **Agent right of agent.** The scenario sample was rejected after review since there is another agent between the two involved agents. The bounding box is not visible for the tested models.



(a) First frame (CAM_BACK_RIGHT).



(b) Trajectory on map.

Figure 13: **Agent lane change.** The scenario sample was rejected after review since no clear lane change was visible. The bounding box is not visible for the tested models.

Table 6: **End-to-end driving models.** Driving expert VLMs for end-to-end driving, *i.e.*, open-loop planning (OLP), closed-loop driving (CLD), and control signal prediction (CSP). Additional proxy tasks are scene generation (SG), dense captioning (CAP), and counterfactual reasoning (CR). Rows marked in gray provide explicit reasoning, code, and pre-trained model weights. Code marked with "~" does not include evaluation scripts or configurations for nuScenes.

Method	Evaluation Dataset	Environment	Task	Explicit reasoning	Model Weights	Code
ORION [10]	nuScenes, CARLA [9]	Real, Sim	CLD, OLP	✓	✓	~
DriveMM [11]	nuScenes	Real	VQA	✓	✓	✓
RDA-Driver [12]	nuScenes	Real	OLP	✓	✗	✗
EMMA [13]	WOD, nuScenes	Real	OLP	✓	✗	✗
Senna-VLM [15]	nuScenes, DriveX [15]	Real	CSP	✓	✓	✓
GPVL [17]	nuScenes	Real	OLP	✓	✗	✓
VLP [22]	nuScenes	Real	OLP	✗	✗	✗
LMDrive [25]	CARLA [9]	Sim	CLD	✗	✓	✓
DriveVLM-Dual [26]	nuScenes	Real	OLP	✓	✗	✗
InsightDrive [27]	nuScenes	Real	OLP	✓	✗	✗
OmniDrive [30]	nuScenes	Real	OLP, CR	✓	✓	✓
BEVDriver [33]	CARLA [9]	Sim	CLD	✗	✗	✗
Sce2DriveX [35]	nuScenes	Real	OLP, CSP	✓	✗	✗
HERMES [37]	nuScenes	Real	SG, OLP	✓	✗	✗
OpenDriveVLA [38]	nuScenes	Real	OLP, CAP	✓	✗	✗

1.5 Multiple-choice question generation

The evaluation on our benchmark requires the generation of input prompts that contain the task description and a question. In addition to the preamble (setting the context) and scenario descriptions (for all answer choices), we ask the models one of the following questions, depending on the scenario category:

- **Ego:** Which of the following options best describes ego driving maneuver?
- **Agent:** Which of the following options best describes the driving behavior of the <reference to agent>?
- **Ego-to-agent:** Which of the following options best describes the ego driving behavior with respect to the <reference to agent>?
- **Agent-to-agent:** Which of the following options best describes <reference to agent 1> maneuver with respect to the <reference to agent 2>?

Notice that this example is only valid for DriveMM [11] and requires adaptation for other models in order to get better performance. For instance, the fine-tuning of Senna-VLM [15] always refers in a first-person manner to the ego vehicle, which requires altering questions respectively. Detailed prompting examples for all evaluated models are provided in Sec.5.

2 Driving expert baselines

In this section, we discuss the choice of driving expert models from recent publications listed in Table 6. To be considered for evaluation, the model must have explicit reasoning capabilities in addition to its primary task. Furthermore, the code and model weights must be publicly available at the time of submission to guarantee fair comparison. Therefore, we opted for DriveMM [11], Senna-VLM [15], and OmniDrive [30].

DriveMM. Not particularly developed for the end-to-end driving task, DriveMM [11] performs joint training on multiple driving datasets. The model architecture is inspired by LLaVa [19] and comprises a SigLIP [34] vision encoder and Llama-3.1 [1] as LLM. In the first step, the model gets pre-trained on a multi-modal dataset that includes text-image pairs for images and videos. For the final fine-tuning, the model learns from single images, multi-view images, and videos, leveraging various driving datasets, *i.e.*, DriveLM [26], LingoQA [20], NuInstruct [8], OmniDrive [30], MAPLM [5], and CODA-LM [18]. Hence, the model learns to process various input modalities from single images to multi-view videos.

Table 7: **Ego scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego scenarios. Accuracy grouped by scenario categories: Left Turn (**LT**), Accelerate (**Acc**), Right Turn (**RT**), Lane Change (**LC**), Decelerate (**Dec**), Stop and Average (**Avg**). The best results are highlighted in bold.

	LT	Acc	RT	LC	Dec	Stop	Avg
Llama 3.2 [1]	9.1	24.3	0.0	18.8	29.4	18.2	16.6
DeepSeek V3 [7]	27.3	86.5	10.0	12.5	41.2	9.1	31.1
GPT-4o [21]	100.0	91.9	50.0	68.8	52.9	18.2	63.6
InternVL 2.5 1B [6]	0.0	29.7	20.0	37.5	23.5	9.1	20.0
Qwen2.5-VL 7B [3]	36.4	2.7	70.0	6.3	52.9	45.5	35.6
InternVL 2.5 8B [6]	18.2	21.6	50.0	6.3	70.6	63.6	38.4
Senna-VLM [15]	18.2	10.8	0.0	6.3	17.7	0.0	8.8
OmniDrive [30]	25.0	25.0	23.8	35.3	27.8	25.0	24.4
DriveMM [11]	63.6	51.4	70.0	18.8	11.8	45.5	43.5

OmniDrive. This model is designed to provide driving decisions and waypoints for planning. It encodes multi-view images into BEV features leveraging the StreamPETR [29] encoder and maps this representation into language space, incorporating a Q-Former [16]. The training contains a series of objectives, including 3D grounding, open-loop planning, counterfactual reasoning, and scene understanding. Annotations for training on the nuScenes [4] dataset are automatically generated leveraging ground truth information and VLMs.

Senna. This end-to-end driving model consists of two modules, Senna-VLM and SennaE2E. The idea is to predict high-level planning decisions from multi-view video inputs with Senna-VLM and encode them into meta-action features, enhancing the final planning trajectories of SennaE2E, which is a state-of-the-art end-to-end driving planner [14]. The images are encoded with CLIP [23] and compressed in order to reduce the number of input tokens. The model is trained in multiple steps, starting with pre-training on multiple sources, including the instruction following data from LLaVa [19]. Two fine-tuning steps follow on private data with automatically generated VLM annotations, leveraging ground truth information. First, the model learns to describe the scene, detect traffic signals, identify vulnerable road users, and so on. Second, the model learns to predict meta-actions such as accelerate, decelerate, left, right, *etc.*

3 Detailed results & analysis

We provide detailed evaluations for ego, agent, ego-to-agent, and agent-to-agent scenarios in Table 7, 8, 9, and 10, respectively. Therefore, we list the accuracy for each scenario and the weighted average for the corresponding category. In addition to evaluating the performance of the individual models, we also gain the following insights:

Ego scenarios. Large language models (LLMs) perform exceptionally well for ego scenarios, shown in Table 7, as they have the ground truth trajectory of the ego vehicle at their disposal. Especially GPT-4o [21] with advanced reasoning can successfully derive scenarios from given trajectories. We illustrate the model’s reasoning after correctly identifying a *left turn* in Fig. 15. An exception is the differentiation between the closely related scenarios *decelerate* and *stop*. Since the vehicle may not be fully stationary in the trajectory data, such as when moving at less than 1 km/h, the language model typically interprets the situation as deceleration. In contrast, powerful vision-language models (VLMs) often recognize it as a potential stop, based on contextual cues like a vehicle ahead, a red traffic light, or a stop sign. However, end-to-end driving models fail to perform well across different scenarios. Only DriveMM [11] provides reasonable results for more obvious scenarios such as *left turn* or *right turn*.

Agent scenarios. Table 8 shows the detailed evaluation of various models on agent scenarios. These scenarios that involve a single agent are more difficult for LLMs. A simple reason is that some of these scenarios, *i.e.*, *jaywalking*, *crossing*, require road markings to be detected correctly. Surprisingly,

Table 8: **Agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego-to-agent scenarios. Accuracy grouped by scenario categories: Jaywalk (**JW**), Left Turn (**LT**), Walk (**W**), Stand (**S**), Cross (**C**), Accelerate (**Acc**), Right Turn (**RT**), Lane Change (**LC**), Overtake Ego (**OE**), Follow Ego (**FE**), Stop, Run, Lead Ego (**LE**), Reverse (**Rev**), U Turn (**UT**), Stationary Right Of Ego (**SRoE**), Stationary Behind Ego (**SBE**) and Average (**Avg**). The best results are highlighted in bold.

	JW	LT	W	S	C	Acc	RT	LC	OE	FE	Stop	Run	LE	Rev	UT	SRoE	SBE	Avg
Llama 3.2 [1]	16.0	16.7	20.8	9.1	10.4	32.0	15.4	29.2	0.0	20.5	10.3	20.0	35.3	0.0	0.0	0.0	0.0	13.8
DeepSeek V3 [7]	28.0	45.8	33.3	13.6	41.7	20.0	30.8	12.5	83.3	89.7	20.7	0.0	94.1	0.0	0.0	100.0	100.0	42.0
GPT-4o [21]	36.0	62.5	25.0	13.6	33.3	52.0	38.5	16.7	33.3	94.9	6.9	80.0	82.4	0.0	0.0	100.0	100.0	45.6
InternVL 2.5 1B [6]	16.0	29.2	35.4	0.0	70.8	22.0	26.9	16.7	66.7	15.4	17.2	20.0	11.7	0.0	0.0	0.0	0.0	20.5
Qwen2.5-VL 7B [3]	4.0	29.2	68.8	72.7	79.2	2.0	42.3	20.8	33.3	61.5	10.3	80.0	17.7	0.0	0.0	50.0	50.0	36.6
InternVL 2.5 8B [6]	8.0	16.7	85.4	75.0	89.6	28.0	65.4	33.3	33.3	64.1	20.7	40.0	47.1	100.0	0.0	100.0	100.0	53.3
Senna-VLM [15]	32.0	0.0	58.3	31.8	35.4	30.0	30.8	12.5	33.3	61.5	6.9	20.0	41.2	0.0	0.0	50.0	0.0	26.1
OmniDrive [30]	26.3	44.8	32.4	14.5	50.6	28.2	45.9	3.2	66.7	20.6	27.3	28.6	15.2	0.0	0.0	0.0	0.0	23.8
DriveMM [11]	0.0	16.7	75.0	22.7	50.0	82.0	30.8	8.3	16.7	79.5	37.9	0.0	76.5	0.0	0.0	50.0	100.0	38.0

Table 9: **Ego-to-agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego-to-agent scenarios. Accuracy grouped by scenario categories: Lead Agent (**LA**), Pass Agent (**PA**), Wait Pedestrian Cross (**WPC**), Follow Agent (**FA**), Overtake Agent (**OA**), Stationary Left Of Agent (**SLoA**), Stationary In Front Of Agent (**SiFoA**) and Average (**Avg**). The best results are highlighted in bold.

	LA	PA	WPC	FA	OA	SLoA	SiFoA	Avg
Llama 3.2 [1]	27.5	18.2	57.1	35.3	0.0	0.0	0.0	19.7
DeepSeek V3 [7]	50.0	72.7	57.1	70.6	100.0	50.0	100.0	71.5
GPT-4o [21]	77.5	25.0	57.1	70.6	100.0	100.0	100.0	75.8
InternVL 2.5 1B [6]	17.5	68.2	78.6	52.9	0.0	0.0	50.0	38.2
Qwen2.5-VL 7B [3]	10.0	0.0	100.0	76.5	0.0	0.0	0.0	26.6
InternVL 2.5 8B [6]	32.5	9.1	100.0	94.1	0.0	0.0	50.0	40.8
Senna-VLM [15]	27.5	27.3	85.7	70.6	100.0	0.0	0.0	44.4
OmniDrive [30]	32.4	44.1	80.0	19.4	100.0	0.0	25.0	43.0
DriveMM [11]	77.5	29.6	92.9	88.2	0.0	0.0	50.0	48.3

GPT-4o [21] performs above random guessing. However, this outcome is less indicative of true scenario understanding and more a reflection of the model’s strong general reasoning capabilities. We illustrate this in Fig. 14, where GPT-4o [21] arrives at the correct answer through a process of elimination. The best overall performance for this category reaches InternVL 2.5 8B [6]. The VLM has difficulties with challenging scenarios like *jaywalking* and *lane change*, requiring in-depth visual understanding, but also with scenarios that involve the ego vehicle, such as *overtaking ego*, *lead ego*, and *follow ego*. The relatively bad performance of driving expert models highlights the need to enhance spatio-temporal understanding for these models.

Ego-agent scenarios. The detailed results for all ego-to-agent scenarios are listed in Table 9. We observe that LLMs perform very well at this task. However, we also notice a respectable performance of driving expert models. This is unsurprising since these scenarios are also part of existing benchmarks and datasets. The objective of driving experts mostly includes reasoning about agents in their close vicinity. The most challenging cases are *pass agent*, and both scenarios of identifying *stationary* objects in a certain position, *i.e.*, *left* and *right* of the ego vehicle. The performance on these relatively unknown but simple scenarios highlights a strong bias toward previously seen scenarios and underscores the limited zero-shot generalization capabilities of these models.

Agent-to-agent scenarios. The interactions between two agents not involving the ego vehicle are the most difficult to reason about for the assessed models. In Table 10, InternVL 2.5 8B [6] reaches the best average score across all scenarios. We can observe that OmniDrive [30] is surprisingly good

Table 10: **Agent-to-agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for Agent-to-agent scenarios. Accuracy grouped by scenario categories: Follow Agent (**FA**), Stationary Left Of Agent (**SLoA**), Stationary Right Of Agent (**SRoA**), Walk Alongside (**WA**), Stationary In Front Of Agent (**SiFoA**), Walk Opposite (**WO**), Lead Agent (**LA**), Wait Pedestrian Cross (**WPC**), Overtake Agent (**OA**), Stationary Behind Agent (**SBA**), Pass Agent (**PA**), Moving Left Of Agent (**MLoA**), Moving Right Of Agent (**MRoA**) and Average (**Avg**). The best results are highlighted in bold.

	FA	SLoA	SRoA	WA	SiFoA	WO	LA	WPC	OA	SBA	PA	MLoA	MRoA	Avg
Llama 3.2 [1]	43.2	5.9	16.7	25.0	12.5	32.5	37.8	20.0	25.0	11.1	20.9	33.3	0.0	21.8
DeepSeek V3 [7]	52.3	5.9	8.3	65.0	6.3	95.0	26.7	96.0	25.0	5.6	55.8	33.3	100.0	44.2
GPT-4o [21]	65.9	5.9	5.6	75.0	6.3	95.0	71.1	84.0	41.7	0.0	30.2	33.3	50.0	43.4
InternVL 2.5 1B [6]	6.8	2.9	2.8	17.5	25.0	27.5	24.4	40.0	50.0	16.7	58.1	66.7	50.0	29.9
Qwen2.5-VL 7B [3]	63.6	23.5	36.1	55.0	25.0	45.0	20.0	84.0	25.0	38.9	20.9	0.0	50.0	37.5
InternVL 2.5 8B [6]	43.2	61.8	63.9	72.5	0.0	72.5	26.7	100.0	0.0	50.0	32.6	100.0	50.0	51.8
Senna-VLM [15]	63.6	23.5	25.0	30.0	25.0	27.5	28.9	68.0	25.0	16.7	25.6	0.0	50.0	31.5
OmniDrive [30]	54.7	12.7	13.5	1.8	9.4	4.5	44.3	57.1	52.9	32.6	56.5	0.0	0.0	26.2
DriveMM [11]	34.1	35.3	55.6	32.5	6.3	17.5	31.1	24.0	8.3	27.8	11.6	33.3	50.0	28.3

at scenarios like *overtaking* and *passing agent*. However, the model does not recognize pedestrian scenarios, *i.e.* *walk opposite* and *walk alongside*, and is not able to identify positions of objects, *i.e.*, *left* and *right* for both motion states *stationary* and *moving*. In contrast, InternVL 2.5 8B [6] is exceptionally good at scenarios that can be identified by visual cues, but does not work well for scenarios spanned across different camera viewpoints, such as *overtaking* or *passing agent*.

System Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.

User Prompt:

An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car. The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters. In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK). The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame. Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment. The following are driving maneuvers and actions along with their respective descriptions:
Agent 1 crossing street: Agent 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
Agent 1 jaywalking: Agent 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.
Agent 1 running: Agent 1 (pedestrian) is running and moves rapidly.
Agent 1 decelerating: Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Agent 1 stopping: Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number: 0
x: 0.00
y: 0.00
rotation: 0.00
...
Frame number: 5
x: 7.95
y: 0.33
rotation: 0.03
Agent 1:
Class: pedestrian
Frame number: 0
LiDAR x: 6.45
LiDAR y: -22.25
LiDAR rotation: 2.93
CAM x: 580.66
CAM y: 509.22
CAM: CAM_BACK
...
Frame number: 5
LiDAR x: 4.63
LiDAR y: -29.92
LiDAR rotation: 2.90
CAM x: 696.33
CAM y: 513.36
CAM: CAM_BACK
Which of the following options best describes agent 1 driving maneuver?
Options:
A. Agent 1 crossing street
B. Agent 1 jaywalking
C. Agent 1 running
D. Agent 1 decelerating
E. Agent 1 stopping
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

GPT-4o [21]: B

Correct: B

User Prompt:

Can you elaborate your decision?

Answer:

GPT-4o [21]: Agent 1 is moving consistently and progressively away from the ego vehicle without any indication of stopping, running rapidly, or interacting with specific crossing points. The movement doesn't suggest crossing at a designated area, so the best interpretation given the information is that Agent 1 is likely crossing outside designated areas, which aligns with jaywalking behavior.

Figure 14: GPT-4o [21] reasoning for a jaywalking scenario.

System Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.

User Prompt:

An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.

The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.

In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK).

The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame.

Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Ego stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Ego turning left: Ego is executing a left turn at an intersection or junction.

Ego performing u-turn: Ego is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.

Ego turning right: Ego is executing a right turn at an intersection or junction.

Ego decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.

Ego:
Frame number: 0
x: 0.00
y: 0.00
rotation: -0.00
:
:
:
Frame number: 5
x: 7.75
y: 3.50
rotation: 0.82

Which of the following options best describes ego driving maneuver?
Options:
A. Ego stopping
B. Ego turning left
C. Ego performing u-turn
D. Ego turning right
E. Ego decelerating
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

GPT-4o [21]: B

Correct: B

User Prompt:

Can you elaborate your decision?

Answer:

GPT-4o [21]: The ego vehicle's motion data shows increasing x and y coordinates along with a gradually increasing rotation angle from frames 0 to 5. The trajectory and rotation suggest the vehicle is moving in an arc to the left. This pattern aligns with the maneuver described as "Ego turning left," hence the choice B.

Figure 15: GPT-4o [21] reasoning for an ego left turning scenario.

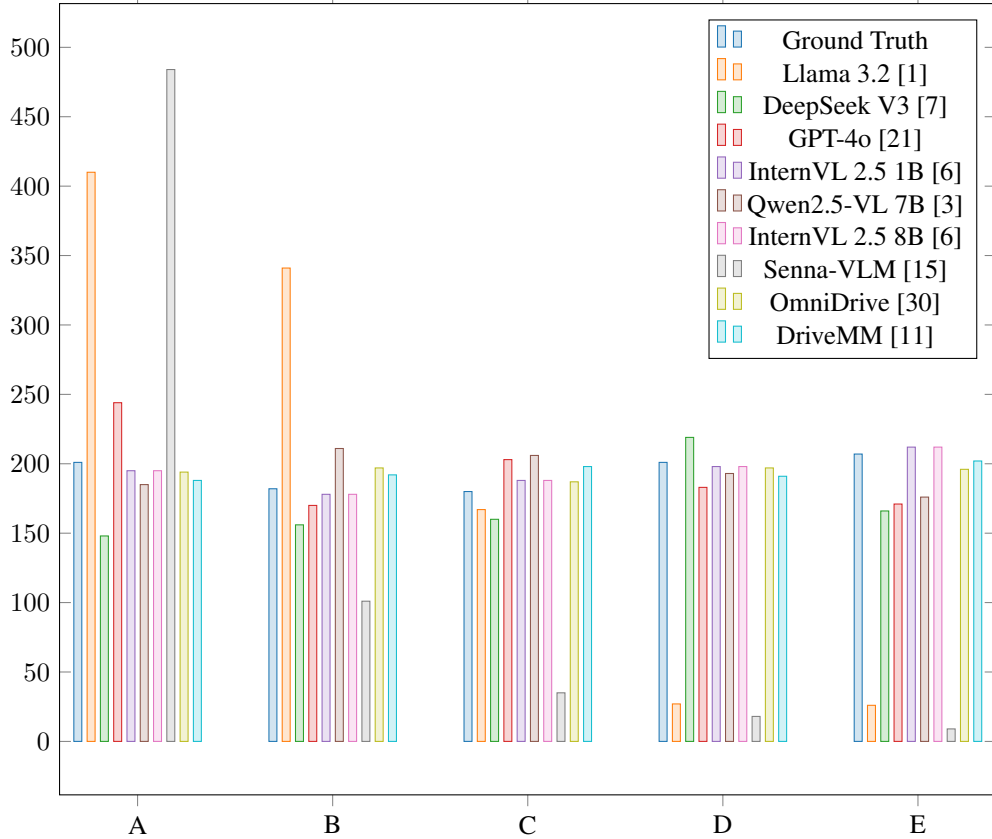


Figure 16: **Model choice preferences.** Distribution of correct multiple-choice letters (ground truth) compared to predictions of all tested models.

4 Additional experiments

Multiple-choice letter distribution. To mitigate potential biases arising from uneven ground truth letter distributions, our experimental design employs a uniform distribution across all multiple-choice options. This critical aspect, visualized in Fig. 16, addresses the known tendency of some LLMs to favor specific letter choices [36]. While methods such as GPT-4o [21], InternVL 2.5 8B [6], and DriveMM [11] appear to be free of this bias, our analysis reveals that Llama 3.2 [1] exhibits a preference for alphabetically earlier letters, a pattern similarly observed in the Senna-VLM [15] expert model. Furthermore, Senna-VLM [15] demonstrates a tendency to disregard prompt instructions, such as the restriction to single-letter outputs, and occasionally includes irrelevant information (most probably from its training data), as illustrated in Fig. 33. The overall output vocabulary of Senna-VLM is depicted in the word cloud in Fig. 17c.

Vision-language models augmented with additional ground-truth information. To gain further insights on off-the-shelf VLMs and driving expert models, we provide these models with additional ground-truth trajectories of involved agents (text form) alongside the visual input in the prompt, mirroring the evaluation of LLM-only models. This experiment evaluates whether combining structured 3D motion data with visual context would enhance the reasoning performance of vision-language models. As shown in Table 11, this additional information significantly degraded the performance of Senna-VLM [15], suggesting that its training compromises the LLM’s generalization capabilities, consistent with our findings in Fig. 17c. The performance of OmniDrive [30] remained consistent, indicating an inability to effectively utilize information outside its training data. Notably, DriveMM [11]’s performance increased by 4% on the average score, which brings it close to the performance of DeepSeek V3 [7]. Given that DriveMM is trained on six diverse datasets, this highlights the importance of diverse training data in preserving the generalization abilities of LLMs.

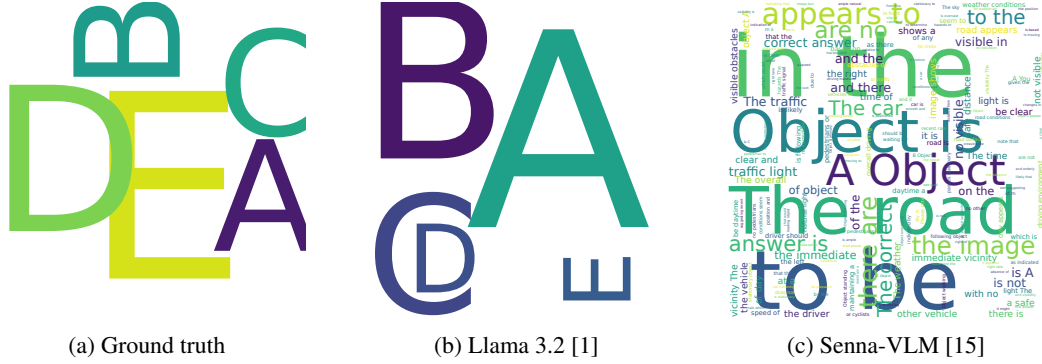


Figure 17: **Ambiguous model responses.** Distribution of individual letters that were the correct multiple-choice option (a), and the outputs generated by the Llama 3.2 [1] (b) and Senna-VLM [15] (c) models.

Table 11: **Ground-truth augmented VLM evaluation.** Evaluation of VLMs and driving expert models that receive ground-truth trajectories in text form alongside to the visual prompts.

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
InternVL 2.5 1B [6]	22.87	42.06	23.05	21.13	27.28
Qwen2.5-VL 7B [3]	37.90	49.05	42.71	27.28	39.24
InternVL 2.5 8B [6]	27.72	49.87	51.18	37.63	41.60
Senna-VLM [15]	3.46	0.00	4.35	4.81	3.16
OmniDrive [30]	27.02	34.61	22.75	29.27	28.41
DriveMM [11]	53.11	71.41	39.54	29.91	48.49

Surprisingly, we observe no significant performance gains among the majority of models, suggesting that the VLMs struggle to effectively fuse the visual and textual modalities. We hypothesize that the models either fail to align the trajectory descriptions with visual cues, or that the added input increased complexity without meaningful gain, possibly confusing the model. This result highlights an important limitation: current VLMs may not yet be equipped to jointly reason over multi-modal inputs that combine structured 3D representations and rich visual context.

Ablation of the query frame. The default model setting is to provide the initial frame of the sequence as the query frame. We evaluate how the performance is affected by providing a later query frame as the initial input. Specifically, we evaluate OmniDrive [30] by querying with the first (default), middle, and last frames in each scenario. Table 12 shows that querying with the middle frame achieves the highest performance, presumably because it integrates more comprehensive temporal information than the alternatives. Conversely, using the last frame as the query leads to a minor decrease in performance, which we hypothesize is due to capacity limitations of OmniDrive for long-range temporal dependency modeling.

OmniDrive [30] chain-of-thought (CoT). OmniDrive leverages the chain-of-thought (CoT) [31] capabilities of vision-language models (VLMs) to infer trajectories in a sequential, step-by-step manner, starting from scene descriptions, 3D grounding, and other relevant contextual information. In our evaluation, we mirrored this approach by prompting the model at the conclusion of its internal CoT process. This design choice aims to ensure the model encapsulates crucial spatio-temporal information necessary for answering our queries. To ablate the impact of CoT, we conducted an evaluation of OmniDrive while disabling the CoT mechanism (Table 12). As anticipated, the model’s performance experienced a slight degradation without CoT, underscoring its positive contribution to overall performance.

DriveMM [11] confusion matrices. To gain further insights into the best performing driving expert model DriveMM, we provide the confusion matrix for each scenario category in Fig. 18.

Table 12: **Reference input frame evaluation.** OmniDrive [30] ablation by querying the model using the n^{th} frame, with or without the Chain of Thought (CoT) capabilities enabled.

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
1 st w/ CoT	24.40	42.97	23.78	26.15	29.33
3 rd w/ CoT	20.89	46.92	32.93	27.24	32.00
6 th w/ CoT	21.70	43.65	23.60	26.72	28.92
1 st w/o CoT	29.91	35.87	21.76	25.39	28.23

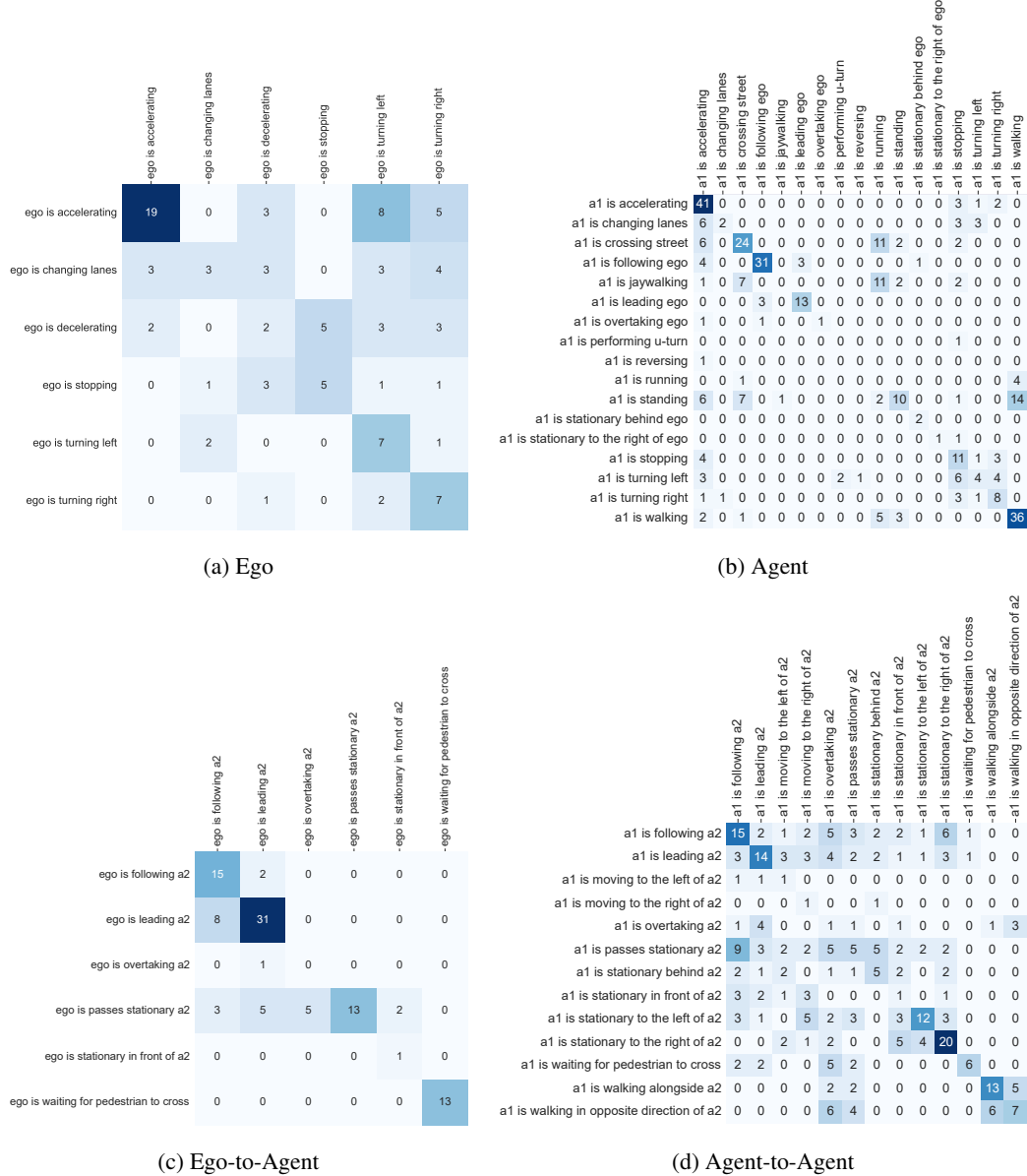


Figure 18: **Confusion matrix for DriveMM [11].** Model decisions for all four scenario categories, *i.e.*, ego (a), agent (b), ego-to-agent (c), and agent-to-agent (d) where *ego* denotes the ego vehicle, *a1* the first (active) agent and *a2* the second (passive) agent.

System Prompt:			
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.			
User Prompt:			
<p>An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.</p> <p>The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.</p> <p>In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK).</p> <p>The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame.</p> <p>Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment.</p> <p>The following are driving maneuvers and actions along with their respective descriptions:</p> <p>Ego stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.</p> <p>Ego turning left: Ego is executing a left turn at an intersection or junction.</p> <p>Ego performing u-turn: Ego is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.</p> <p>Ego turning right: Ego is executing a right turn at an intersection or junction.</p> <p>Ego decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.</p> <p>Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.</p> <p>Ego:</p> <pre>Frame number: 0 x: 0.00 y: 0.00 rotation: 0.00 . . . Frame number: 5 x: 7.75 y: 3.50 rotation: 0.82</pre> <p>Which of the following options best describes ego driving maneuver? Options:</p> <ul style="list-style-type: none"> A. Ego stopping B. Ego turning left C. Ego performing u-turn D. Ego turning right E. Ego decelerating <p>Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.</p>			
Answer:			
GPT-4o [21]: B	DeepSeek V3 [7]: D	Llama 3.2 [1]: A	Correct: B

Figure 19: LLM Ego scenario prompt

5 Prompt examples

To demonstrate the required prompt adaptations, we include examples for all model types and scenario types (Ego, agent, ego-to-agent, and agent-to-agent) in the following. In particular, Fig. 19–22 show the LLM prompts. For the representative off-the-shelf VLMs, Fig. 23–26 provide prompt examples for InternVL 2.5 [6], while Fig. 27–30 provide prompt examples for Qwen2.5-VL 7B [2]. For the driving expert models, we provide exemplary prompts for each scenario type for Senna-VLM [15] in Fig. 31–34, OmniDrive [30] in Fig. 35–38, and DriveMM [11] in Fig. 39–42.

System Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.

User Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios. (An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.

The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.

In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK).

The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame.

Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Agent 1 leading ego: Agent 1 travels ahead of ego at a similar speed while maintaining a consistent distance.

Ego overtaking agent 2: Ego in the adjacent lane and moves ahead of agent 2 while both are in motion.

Ego decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Ego waiting for pedestrian to cross: Ego comes to a stop or remains stationary, yielding the right-of-way to agent 2 who is crossing or preparing to cross the road, while maintaining awareness of the agent 2's movement and ensuring a safe distance until the crossing is complete.

Ego passes stationary agent 2: Ego in the adjacent lane overtakes the stopped agent 2.

Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.

Ego:
Frame number: 0
x: 0.00
y: 0.00
rotation: 0.00
.
.
.
Frame number: 5
x: 0.79
y: 0.03
rotation: 0.04
Agent 2:
Frame number: 0
LiDAR x: -2.97
LiDAR y: 7.03
LiDAR rotation: -2.48
CAM x: 221.21
CAM y: 630.17
CAM: CAM_FRONT
.
.
.
Frame number: 5
LiDAR x: -5.41
LiDAR y: 4.08
LiDAR rotation: -2.30
CAM x: 828.79
CAM y: 578.67
CAM: CAM_FRONT_LEFT
Which of the following options best describes ego driving behavior with respect to agent 2?
Options:
A. Agent 1 leading ego
B. Ego overtaking agent 2
C. Ego decelerating
D. Ego waiting for pedestrian to cross
E. Ego passes stationary agent 2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

GPT-4o [21]: C

DeepSeek V3 [7]: C

Llama 3.2 [1]: B

Correct: D

Figure 20: LLM Ego-to-Agent scenario prompt

System Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.

User Prompt:

An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car. The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters. In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK).

The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame.

Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Agent 1 turning left: Agent 1 is executing a left turn at an intersection or junction.

Agent 1 changing lanes: Agent 1 is transitioning from its current lane to an adjacent lane.

Agent 1 reversing: Agent 1 is moving in reverse, either to park, navigate a tight space, or adjust its position.

Agent 1 decelerating: Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Agent 1 performing u-turn: Agent 1 is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.

Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.

Ego:

Frame number: 0
x: 0.00
y: 0.00
rotation: 0.00
.
.
.
Frame number: 5
x: 15.50
y: -0.17
rotation: -0.01
Agent 1:
Class: car
Frame number: 0
LiDAR x: 5.17
LiDAR y: -5.98
LiDAR rotation: 1.47
CAM x: 1326.79
CAM y: 643.39
CAM: CAM_BACK_RIGHT
.
.
.
Frame number: 5
LiDAR x: 5.91
LiDAR y: -11.37
LiDAR rotation: 1.52
CAM x: 1599.80
CAM y: 480.51
CAM: CAM_BACK_RIGHT

Which of the following options best describes agent 1 driving maneuver?

Options:

A. Agent 1 turning left
B. Agent 1 changing lanes
C. Agent 1 reversing
D. Agent 1 decelerating
E. Agent 1 performing u-turn

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

GPT-4o [21]: C.

DeepSeek V3 [7]: B

Llama 3.2 [1]: B

Correct: B

Figure 21: LLM Agent scenario prompt

System Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents in diverse driving scenarios.

User Prompt:

An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car. The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car. This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent, the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters. In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK). The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians relative to the global frame. Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent within the traffic environment. The following are driving maneuvers and actions along with their respective descriptions: Agent 1 stationary to the right of agent 2: Agent 1 is fully stopped and remains stationary to the right of agent 2, which is also stationary, such as when waiting at a traffic light or in a parking lot. Agent 1 overtaking agent 2: Agent 1 in the adjacent lane and moves ahead of agent 2 while both are in motion. Agent 1 stationary in front of agent 2: Agent 1 is fully stopped and remains stationary in front of agent 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing. Agent 1 stationary to the left of agent 2: Agent 1 is fully stopped and remains stationary to the left of agent 2, which is also stationary, such as when waiting at a traffic light or in a parking lot. Agent 1 moving to the right of agent 2: Agent 1 is traveling in parallel to the right of agent 2 (e.g., in adjacent lanes or side by side), with one vehicle maintaining a rightward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light. Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite. Ego: . . . Agent 1: Class: car Frame number: 0 LiDAR x: -11.89 LiDAR y: 28.12 LiDAR rotation: -1.58 CAM x: 291.60 CAM y: 546.14 CAM: CAM_FRONT . . . Frame number: 5 LiDAR x: -12.35 LiDAR y: -27.44 LiDAR rotation: -1.57 CAM x: 1223.86 CAM y: 520.56 CAM: CAM_BACK Agent 2: Class: car Frame number: 0 LiDAR x: -15.65 LiDAR y: 26.08 LiDAR rotation: -1.58 CAM x: 90.13 CAM y: 564.20 CAM: CAM_FRONT . . . Frame number: 5 LiDAR x: -15.99 LiDAR y: -26.44 LiDAR rotation: -1.52 CAM x: 1351.97 CAM y: 527.46 CAM: CAM_BACK Which of the following options best describes agent 1 driving behaviour with respect to agent 2? Options: A. Agent 1 stationary to the right of agent 2 B. Agent 1 overtaking agent 2 C. Agent 1 stationary in front of agent 2 D. Agent 1 stationary to the left of agent 2 E. Agent 1 moving to the right of agent 2 Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

GPT-4o [21]: B

DeepSeek V3 [7]: E

Llama 3.2 [1]: E

Correct: B

Figure 22: LLM Agent-to-Agent scenario prompt

Video:



Prompt:

Frame-1: IMAGE_TOKEN

Frame-2: IMAGE_TOKEN

Frame-3: IMAGE_TOKEN

Frame-4: IMAGE_TOKEN

Frame-5: IMAGE_TOKEN

Frame-6: IMAGE_TOKEN

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:

You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

You are changing lanes: You are transitioning from your current lane to an adjacent lane.

You are turning right: You are executing a right turn at an intersection or junction.

You are stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

You are turning left: You are executing a left turn at an intersection or junction.

Given that, Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT, which of the following options best describes your driving maneuver?

Options:

A. You are decelerating

B. You are changing lanes

C. You are turning right

D. You are stopping

E. You are turning left

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

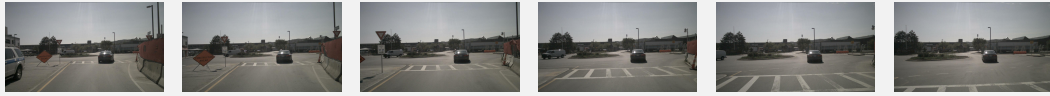
InternVL 2.5 1B [6]: A. You are decelerating

InternVL 2.5 8B [6]: E

Correct: E

Figure 23: VLM InternVL 2.5 8B/1B Ego scenario prompt

Video:



Prompt:

Frame-1: IMAGE_TOKEN
Frame-2: IMAGE_TOKEN
Frame-3: IMAGE_TOKEN
Frame-4: IMAGE_TOKEN
Frame-5: IMAGE_TOKEN
Frame-6: IMAGE_TOKEN

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:

You are stationary to the right of object 2: You are fully stopped and remain stationary to the right of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.

You are overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.

You are stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

You are following object 2: You are driving behind object 2 at a similar speed while maintaining a consistent distance.

You are stationary behind object 2: Object 1 are fully stopped and remain stationary behind object 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.

Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider object 2, `<ref>car</ref><box>[578, 523, 688, 682]</box>` in Frame-1. Which of the following options best describes your driving behavior with respect to the object 2?

Options:

- A. You are stationary to the right of object 2
- B. You are overtaking object 2
- C. You are stopping
- D. You are following object 2
- E. You are stationary behind object 2


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

InternVL 2.5 1B [6]: E. You are stationary behind object 2	InternVL 2.5 8B [6]: D	Correct: D
------------------------------------------------------------	------------------------	------------

Figure 24: VLM InternVL 2.5 8B/1B Ego-to-Agent scenario prompt

Video:



Prompt:

```

Frame-1: IMAGE_TOKEN
Frame-2: IMAGE_TOKEN
Frame-3: IMAGE_TOKEN
Frame-4: IMAGE_TOKEN
Frame-5: IMAGE_TOKEN
Frame-6: IMAGE_TOKEN

```

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:

Object 1 is decelerating: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Object 1 is accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.

Object 1 is running: Object 1 (pedestrian) is running and moves rapidly.

Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.

Object 1 is jaywalking: Object 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.

Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider object 1, `<ref>pedestrian</ref><box>[443, 510, 519, 701]</box>` in Frame-1. Which of the following options best describes object 1 maneuver?

Options:

- A. Object 1 is decelerating
- B. Object 1 is accelerating
- C. Object 1 is running
- D. Object 1 is crossing street
- E. Object 1 is jaywalking


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

InternVL 2.5 1B [6]: D	InternVL 2.5 8B [6]: D	Correct: E
------------------------	------------------------	------------

Figure 25: VLM InternVL 2.5 8B/1B Agent scenario prompt

Video:



Prompt:

Frame-1: IMAGE_TOKEN
 Frame-2: IMAGE_TOKEN
 Frame-3: IMAGE_TOKEN
 Frame-4: IMAGE_TOKEN
 Frame-5: IMAGE_TOKEN
 Frame-6: IMAGE_TOKEN

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:
 Object 1 is following object 2: Object 1 is driving behind object 2 at a similar speed while maintaining a consistent distance.
 Object 1 is stationary to the left of object 2: Object 1 is fully stopped and remains stationary to the left of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.
 Object 1 is accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
 Object 1 is moving to the right of object 2: Object 1 is traveling in parallel to the right of object 2 (e.g., in adjacent lanes or side by side), with one vehicle maintaining a rightward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
 Object 1 is leading object 2: Object 1 travels ahead of object 2 at a similar speed while maintaining a consistent distance.

Consider that the Frame-1 is captured with CAM_BACK_LEFT, Frame-2 is captured with CAM_BACK_LEFT, Frame-3 is captured with CAM_BACK, Frame-4 is captured with CAM_BACK, Frame-5 is captured with CAM_BACK, Frame-6 is captured with CAM_BACK. Also, consider object 1 `<ref>car</ref><box>[71, 467, 251, 608]</box>` in Frame-1 and object 2 `<ref>car</ref><box>[150, 491, 321, 615]</box>` in Frame-1. Which of the following options best describes object 1 maneuver with respect to the object 2?

Options:

- A. Object 1 is following object 2
- B. Object 1 is stationary to the left of object 2
- C. Object 1 is accelerating
- D. Object 1 is moving to the right of object 2
- E. Object 1 is leading object 2


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

InternVL 2.5 1B [6]: C. Object 1 is accelerating	InternVL 2.5 8B [6]: B	Correct: B
--------------------------------------------------	------------------------	------------

Figure 26: VLM InternVL 2.5 8B/1B Agent-to-Agent scenario prompt

Video:



Prompt:

Frame-1: {IMAGE_TOKEN}
 Frame-2: {IMAGE_TOKEN}
 Frame-3: {IMAGE_TOKEN}
 Frame-4: {IMAGE_TOKEN}
 Frame-5: {IMAGE_TOKEN}
 Frame-6: {IMAGE_TOKEN}

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:
 You are stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
 You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
 You are performing u-turn: You are performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
 You are turning left: You are executing a left turn at an intersection or junction.
 You are turning right: You are executing a right turn at an intersection or junction.

Given that, Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT, which of the following options best describes your driving maneuver?

Options:

- A. You are stopping
- B. You are accelerating
- C. You are performing u-turn
- D. You are turning left
- E. You are turning right

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

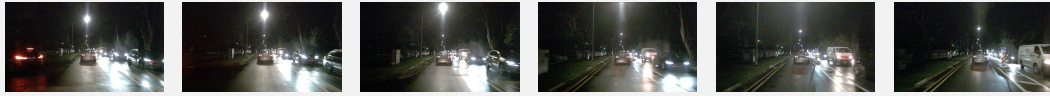
Answer:

Qwen2.5-VL 7B [3]: D

Correct: B

Figure 27: VLM Qwen 2.5 7B Ego scenario prompt

Video:



Prompt:

Frame-1: {IMAGE_TOKEN}
Frame-2: {IMAGE_TOKEN}
Frame-3: {IMAGE_TOKEN}
Frame-4: {IMAGE_TOKEN}
Frame-5: {IMAGE_TOKEN}
Frame-6: {IMAGE_TOKEN}

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:

You are stationary behind object 2: Object 1 are fully stopped and remain stationary behind object 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.

You are following object 2: You are driving behind object 2 at a similar speed while maintaining a consistent distance.

You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.

You are passing stationary object 2: You in the adjacent lane overtakes the stopped object 2.

Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider object 2, which is car inside region [370, 233, 459, 307] in Frame-1. Which of the following options best describes your driving behavior with respect to the object 2?

Options:

- A. You are stationary behind object 2
- B. You are following object 2
- C. You are decelerating
- D. You are accelerating
- E. You are passing stationary object 2

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

Qwen2.5-VL 7B [3]: B

Correct: B

Figure 28: VLM Qwen 2.5 7B Ego-to-Agent scenario prompt

Video:



Prompt:

Frame-1: {IMAGE_TOKEN}
 Frame-2: {IMAGE_TOKEN}
 Frame-3: {IMAGE_TOKEN}
 Frame-4: {IMAGE_TOKEN}
 Frame-5: {IMAGE_TOKEN}
 Frame-6: {IMAGE_TOKEN}

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:
 Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
 Object 1 is accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
 Object 1 is decelerating: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
 Object 1 is stationary: Object 1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.
 Object 1 is jaywalking: Object 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.

Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is captured with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider object 1, which is pedestrian inside region [296, 232, 313, 265] in Frame-1. Which of the following options best describes object 1 maneuver?

Options:

- A. Object 1 is crossing street
- B. Object 1 is accelerating
- C. Object 1 is decelerating
- D. Object 1 is stationary
- E. Object 1 is jaywalking

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

Qwen2.5-VL 7B [3]: A

Correct: A

Figure 29: VLM Qwen 2.5 7B Agent scenario prompt

Video:



Prompt:

Frame-1: {IMAGE_TOKEN}
 Frame-2: {IMAGE_TOKEN}
 Frame-3: {IMAGE_TOKEN}
 Frame-4: {IMAGE_TOKEN}
 Frame-5: {IMAGE_TOKEN}
 Frame-6: {IMAGE_TOKEN}

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.

The following are driving maneuvers and actions along with their respective descriptions:
 Object 1 is running: Object 1 (pedestrian) is running and moves rapidly.
 Object 1 is overtaking object 2: Object 1 in the adjacent lane and moves ahead of object 2 while both are in motion.
 Object 1 is passing stationary object 2: Object 1 in the adjacent lane overtakes the stopped object 2.
 Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
 Object 1 is walking alongside object 2: Object 1 (pedestrian) and object 2 (pedestrian) walk side by side at a steady, moderate pace.

Consider that the Frame-1 is captured with CAM_BACK, Frame-2 is captured with CAM_BACK, Frame-3 is captured with CAM_BACK, Frame-4 is captured with CAM_BACK, Frame-5 is captured with CAM_BACK, Frame-6 is captured with CAM_BACK. Also, consider object 1, which is pedestrian inside region [510, 239, 536, 283] in Frame-1 and object 2, which is pedestrian inside region [534, 238, 556, 280] in Frame-1. Which of the following options best describes object 1 maneuver with respect to the object 2?

Options:

- A. Object 1 is running
- B. Object 1 is overtaking object 2
- C. Object 1 is passing stationary object 2
- D. Object 1 is crossing street
- E. Object 1 is walking alongside object 2

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

Qwen2.5-VL 7B [3]: E

Correct: E

Figure 30: VLM Qwen 2.5 7B Agent-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistant is specilized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.

The following are driving maneuvers and actions along with their respective descriptions:

Decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Reversing: You are moving in reverse, either to park, navigate a tight space, or adjust your position.

Turning left: You are executing a left turn at an intersection or junction.

Turning right: You are executing a right turn at an intersection or junction.

The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:

<image>

<FRONT LEFT VIEW>:

<image>

<FRONT RIGHT VIEW>:

<image>

<BACK LEFT VIEW>:

<image>

<BACK RIGHT VIEW>:

<image>

<BACK VIEW>:

<image>

You are driving, which of the following options best describes your driving maneuver?

Options:

- A. Decelerating
- B. Stopping
- C. Reversing
- D. Turning left
- E. Turning right

For example, a correct answer format is like 'A'. ASSISTANT:


Answer:

Senna-VLM [15]: A. Decelerating

Correct: E

Figure 31: Senna-VLM [15] Ego scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistant is specialized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.

The following are driving maneuvers and actions along with their respective descriptions:

You overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.

Decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

You passes stationary object 2: You in the adjacent lane overtakes the stopped object 2.

You stationary in front of object 2: You are fully stopped and remain stationary in front of object 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.

Reversing: You are moving in reverse, either to park, navigate a tight space, or adjust your position.

The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:

<image>

<FRONT LEFT VIEW>:

<image>

<FRONT RIGHT VIEW>:

<image>

<BACK LEFT VIEW>:

<image>

<BACK RIGHT VIEW>:

<image>

<BACK VIEW>:

<image>

I will now provide you with the position and velocity information of the dynamic objects:

Object 2: car, 7 meters ahead, 6 meters right, speed of 0 m/s.

Please predict which of the following options best describes your driving behavior with respect to Object 2.

Options:

A. You overtaking object 2

B. Decelerating

C. You passes stationary object 2

D. You stationary in front of object 2

E. Reversing

For example, a correct answer format is like 'A'. ASSISTANT:

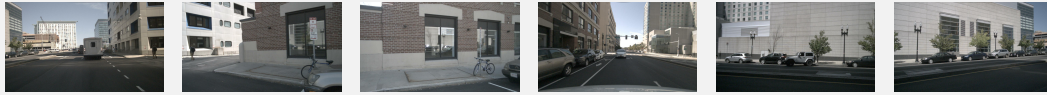
Answer:

Senna-VLM [15]: A. You overtaking object 2

Correct: C

Figure 32: Senna-VLM [15] Ego-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistant is specialized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.

The following are driving maneuvers and actions along with their respective descriptions:

U-turn: Object 1 is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.

Stopping: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Turning right: Object 1 is executing a right turn at an intersection or junction.

Reversing: Object 1 is moving in reverse, either to park, navigate a tight space, or adjust its position.

Accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.

The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:

<image>

<FRONT LEFT VIEW>:

<image>

<FRONT RIGHT VIEW>:

<image>

<BACK LEFT VIEW>:

<image>

<BACK RIGHT VIEW>:

<image>

<BACK VIEW>:

<image>

I will now provide you with the position and velocity information of the dynamic objects:

Object 1: car, 0 meters behind, 14 meters left, speed of 1 m/s.

Please predict which of the following options best describes Object 1 driving behavior.

Options:

A. U-turn

B. Stopping

C. Turning right

D. Reversing

E. Accelerating

For example, a correct answer format is like 'A'. ASSISTANT:

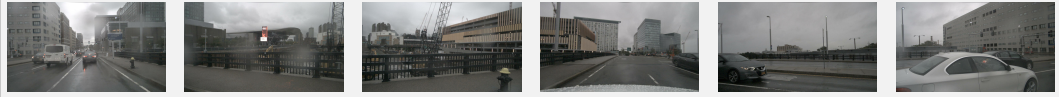
Answer:

Senna-VLM [15]: Object 1 is moving in the same direction as the camera, and there is no indication of a U-turn or a stop. The car is maintaining a safe distance from the vehicle in front, which is also moving in the same direction. The road appears to be clear, and there are no visible obstacles or traffic signals requiring immediate attention. The weather conditions seem to be clear and sunny, which is favorable for driving. The time of day appears to be daytime, as there is ample natural light. The overall driving environment is safe and conducive to maintaining a steady speed.

Correct: D

Figure 33: Senna-VLM [15] Agent scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistant is specialized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.

The following are driving maneuvers and actions along with their respective descriptions:

Object 1 overtaking object 2: Object 1 in the adjacent lane and moves ahead of object 2 while both are in motion.

Turning right: Object 1 is executing a right turn at an intersection or junction.

Object 1 waiting for pedestrian to cross: Object 1 comes to a stop or remains stationary, yielding the right-of-way to a object 2 who is crossing or preparing to cross the road, while maintaining awareness of the object 2's movement and ensuring a safe distance until the crossing is complete.

Decelerating: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Object 1 stationary to the right of object 2: Object 1 is fully stopped and remains stationary to the right of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.

The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:

<image>

<FRONT LEFT VIEW>:

<image>

<FRONT RIGHT VIEW>:

<image>

<BACK LEFT VIEW>:

<image>

<BACK RIGHT VIEW>:

<image>

<BACK VIEW>:

<image>

I will now provide you with the position and velocity information of the dynamic objects:

Object 1: car, 15 meters ahead, 3 meters left, speed of 6 m/s.

Object 2: car, 18 meters ahead, 6 meters left, speed of 5 m/s.

Please predict which of the following options best describes Object 1 driving behavior with respect to Object 2.

Options:

A. Object 1 overtaking object 2

B. Turning right

C. Object 1 waiting for pedestrian to cross

D. Decelerating

E. Object 1 stationary to the right of object 2

For example, a correct answer format is like 'A'. ASSISTANT:

Answer:

Senna-VLM [15]: Object 1 is overtaking Object 2.

Correct: A

Figure 34: Senna-VLM [15] Agent-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

You are changing lanes: You are transitioning from your current lane to an adjacent lane.

You are turning left: You are executing a left turn at an intersection or junction.

You are performing u-turn: You are performing a 180-degree turn at an intersection or junction, reversing its direction of travel.

You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

You are turning right: You are executing a right turn at an intersection or junction.

Which of the following options best describes your driving maneuver?

Options:

- A. You are changing lanes
- B. You are turning left
- C. You are performing u-turn
- D. You are decelerating
- E. You are turning right

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [30]: C

Correct: A

Figure 35: OmniDrive [30] Ego scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

You are moving to the left of object 2: You are traveling in parallel to the left of object 2 (e.g., in adjacent lanes or side by side), with one vehicle maintaining a leftward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.

You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.

You are overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.

You are passing stationary object 2: You in the adjacent lane overtakes the stopped object 2.

You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Consider the object 2, which is a car located at coordinates (+4.4, +7.5) and moving at a velocity of 0.1 m/s. Which of the following options best describes your driving behavior with respect to the object 2?

Options:

- A. You are moving to the left of object 2
- B. You are accelerating
- C. You are overtaking object 2
- D. You are passing stationary object 2
- E. You are decelerating

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.


Answer:

OmniDrive [30]: A

Correct: D

Figure 36: OmniDrive [30] Ego-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Object 1 is stopping: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Object 1 are changing lanes: Object 1 is transitioning from its current lane to an adjacent lane.

Object 1 is performing u-turn: Object 1 is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.

Object 1 is turning right: Object 1 is executing a right turn at an intersection or junction.

Object 1 is turning left: Object 1 is executing a left turn at an intersection or junction.

Consider the object 1, which is a car located at coordinates (+0.3, -29.0) and moving at a velocity of 6.8 m/s. Which of the following options best describes object 1 maneuver?

Options:

- A. Object 1 is stopping
- B. Object 1 are changing lanes
- C. Object 1 is performing u-turn
- D. Object 1 is turning right
- E. Object 1 is turning left


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [30]: A	Correct: B
-------------------	------------

Figure 37: OmniDrive [30] Agent scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Object 1 is walking in opposite direction of object 2: Object 1 (pedestrian) and object 2 (pedestrian) walk toward each other at a moderate pace, cross paths, and proceed.

Object 1 is stationary: Object 1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.

Object 1 is jaywalking: Object 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.

Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.

Object 1 is passing stationary object 2: Object 1 in the adjacent lane overtakes the stopped object 2.

Consider the object 1, which is a pedestrian located at coordinates (-18.7, -2.5) and moving at a velocity of 1.8 m/s, and the object 2, which is a pedestrian located at coordinates (-20.4, +1.1) and moving at a velocity of 1.3 m/s. Which of the following options best describes object 1 maneuver with respect to the object 2?

Options:

- A. Object 1 is walking in opposite direction of object 2
- B. Object 1 is stationary
- C. Object 1 is jaywalking
- D. Object 1 is crossing street
- E. Object 1 is passing stationary object 2

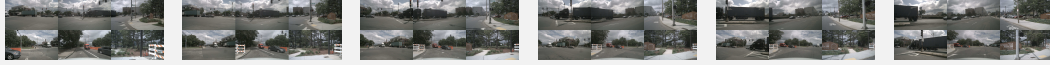
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [30]: E	Correct: A
-------------------	------------

Figure 38: OmniDrive [30] Agent-to-Agent scenario prompt

Multi-view Video:



Prompt:

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Ego is stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Ego is turning right: Ego is executing a right turn at an intersection or junction.

Ego is reversing: Ego is moving in reverse, either to park, navigate a tight space, or adjust its position.

Ego is decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Ego is turning left: Ego is executing a left turn at an intersection or junction.

Which of the following options best describes ego driving maneuver?

Options:

- A. Ego is stopping
- B. Ego is turning right
- C. Ego is reversing
- D. Ego is decelerating
- E. Ego is turning left

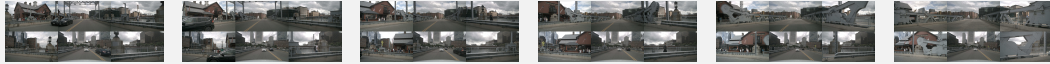
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [11]: B Correct: B

Figure 39: DriveMM [11] Ego scenario prompt

Multi-view Video:



Prompt:

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

Ego is decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.

Ego is following c2: Ego is driving behind c2 at a similar speed while maintaining a consistent distance.

Ego is stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

Ego is changing lanes: Ego is transitioning from its current lane to an adjacent lane.

Ego is leading c2: Ego travels ahead of c2 at a similar speed while maintaining a consistent distance.

Which of the following options best describes the ego driving behavior with respect to the <c2,CAM_BACK,54,59>?

Options:

- A. Ego is decelerating
- B. Ego is following c2
- C. Ego is stopping
- D. Ego is changing lanes
- E. Ego is leading c2


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [11]: E Correct: E

Figure 40: DriveMM [11] Ego-to-Agent scenario prompt

Multi-view Video:



Prompt:

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

c1 is standing: c1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.

c1 is jaywalking: c1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.

c1 is crossing street: c1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.

c1 is running: c1 (pedestrian) is running and moves rapidly.

c1 is walking: c1 (pedestrian) moves at a steady, moderate pace, typically following designated paths or crosswalks.

Which of the following options best describes the driving behavior of the <c1,CAM_FRONT,45,56>?

Options:

- A. c1 is standing
- B. c1 is jaywalking
- C. c1 is crossing street
- D. c1 is running
- E. c1 is walking


Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [11]: C Correct: A

Figure 41: DriveMM [11] Agent scenario prompt

Multi-view Video:



Prompt:

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.

The following are driving maneuvers and actions along with their respective descriptions:

c1 is passes stationary c2: c1 in the adjacent lane overtakes the stopped c2.

c1 is stopping: c1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.

c1 is leading c2: c1 travels ahead of c2 at a similar speed while maintaining a consistent distance.

c1 is stationary to the right of c2: c1 is fully stopped and remains stationary to the right of c2, which is also stationary, such as when waiting at a traffic light or in a parking lot.

c1 is accelerating: c1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.

Which of the following options best describes <c1,CAM_FRONT_RIGHT,95,60> maneuver with respect to the <c2,CAM_FRONT_RIGHT,95,60>?

Options:

- A. c1 is passes stationary c2
- B. c1 is stopping
- C. c1 is leading c2
- D. c1 is stationary to the right of c2
- E. c1 is accelerating

Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [11]: D Correct: C

Figure 42: DriveMM [11] Agent-to-Agent scenario prompt

References

- [1] Meta AI. LLaMA 3.2: Open Foundation and Instruction Models, 2024.
- [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] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.
- [5] 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 *CVPR*, 2024.
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.
- [7] DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [8] Xinpeng Ding, Jinahua Han, Hang Xu, Xiaodan Laing, Xu Hang, Wei Zhang, and Xiaomeng Li. Holistic Autonomous Driving Understanding by Bird’s-Eye-View Injected Multi-Modal Large Models. In *CVPR*, 2024.
- [9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *CoRL*, 2017.
- [10] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkan Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. *arXiv preprint arXiv:2503.19755*, 2025.
- [11] Zhijian Huang, Chengjian Fen, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Drivemm: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024.
- [12] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making Large Language Models Better Planners with Reasoning-Decision Alignment. In *ECCV*, 2024.
- [13] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.
- [14] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *ICCV*, 2023.
- [15] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313*, 2024.
- [16] 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 *ICML*, 2023.

- [17] Tengpeng Li, Hanli Wang, Xianfei Li, Wenlong Liao, Tao He, and Pai Peng. Generative Planning with 3D-vision Language Pre-training for End-to-End Autonomous Driving. *arXiv preprint arXiv:2501.08861*, 2025.
- [18] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases. In *WACV*, 2025.
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023.
- [20] Ana-Maria Marcu, Long Chen, Jan Hünemann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. LingoQA: Video Question Answering for Autonomous Driving. In *ECCV*, 2024.
- [21] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
cia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and
et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024.
- [22] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi,
Senem Velipasalar, and Liu Ren. VLP: Vision Language Planning for Autonomous Driving. In
CVPR, 2024.
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*,
2021.
- [24] Rerun Development Team. Rerun: A Visualization SDK for Multimodal Data, 2024. URL
<https://www.rerun.io>. Available from <https://www.rerun.io/> and [https://github.com/rerun-
io/rerun](https://github.com/rerun-io/rerun).
- [25] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, and Hong-
sheng Li. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *CVPR*,
2024.
- [26] 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 *ECCV*, 2024.
- [27] Ruiqi Song, Xianda Guo, Hangbin Wu, Qinggong Wei, and Long Chen. InsightDrive: Insight
Scene Representation for End-to-End Autonomous Driving. *arXiv preprint arXiv:2503.13047*,
2025.
- [28] Pei Sun, Henrik Kretschmar, 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 *CVPR*, 2020.
- [29] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring Object-
Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *ICCV*, 2023.
- [30] 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 preprint arXiv:2405.01533*, 2024.
- [31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le,
Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models.
NeurIPS, 2022.
- [32] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh
Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva
Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving
Perception and Forecasting. In *NeurIPS*, 2021.

- [33] Katharina Winter, Mark Azer, and Fabian B. Flohr. BEVDriver: Leveraging BEV Maps in LLMs for Robust Closed-Loop Driving. *arXiv preprint arXiv:2503.03074*, 2025.
- [34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, October 2023.
- [35] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2DriveX: A Generalized MLLM Framework for Scene-to-Drive Learning, 2025.
- [36] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *ICLR*, 2024.
- [37] Xin Zhou, Dingkan Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation. *arXiv preprint arXiv:2501.14729*, 2025.
- [38] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463*, 2025.