

System Prompt
You are an autonomous driving expert, specializing in recognizing traffic scenes and making driving decisions.
General Perception Prompt
<p>You receive a series of traffic images captured from the perspective of the ego car. Your task is to first focus on the road users in the driving scenario that influence the future driving behavior of the ego car, and then describe each road user and explain why, finally give your suggestions for the ego car driving behavior.</p> <p>Here are some rules to follow:</p> <ol style="list-style-type: none"> Road users should include vehicle(cars, trucks, buses, etc), vulnerable road users(pedestrians, cyclists, and motor cyclists), traffic signs(No parking signs, warning_signs directional signs, etc), traffic lights(identify current state such as red, green, yellow), traffic cones, barriers, road states, others(debris, dustbin, etc). Road users should include a description(appearance, position, direction, etc) of these objects and the reasons that affect the driving behavior of ego car. Each road user should be described once to maintain clarity and avoid repetition and ensure each description is unique and specific to the object. To give a positive and accurate answer, please output dictionary format and the following is sample answer, xxx means placeholder: <pre>{ "vehicles": [{"description": xxx, explanation: xxx}, {}], "vulnerable_road_users": [{"description": xxx, explanation: xxx}, {}], "traffic_signs": [{"description": xxx, explanation: xxx}, {}], "traffic_lights": [{"description": xxx, explanation: xxx}, {}], "traffic_cones": [{"description": xxx, explanation: xxx}, {}], "barriers": [{"description": xxx, explanation: xxx}, {}], "other_objects": [{"description": xxx, explanation: xxx}, {}] }</pre> If there is no road user of this class, the output should be <code>{"vulnerable_road_users":[]}</code>

Figure 5. **The data pre-annotation prompts for general perception.** The prompts are divided into system prompts and general perception prompts.

Appendix

A. More on Dataset Construction

Prompts for pre-annotation. The prompts used to generate the pre-annotations from GPT-4V are provided in Fig. 5 and Fig. 6.

Gradio labeling tool graphical user interface (GUI). Fig. 9 demonstrates a screenshot of our labeling tool for the general perception task. We utilize Gradio and aim to assist human annotators to refine general perception pre-annotations deriving from GPT-4V, as discussed in Sec. 4.1. The annotators refine by following the principles of merging, modifying, and deleting step by step.

Prompts for evaluation. To comprehensively and accurately assess the performance of different LVLMs, we design distinct evaluation prompts for each task, as shown in Fig. 7. Meanwhile, we use the few-shot in-context learning method to improve accuracy for general perception and driving suggestions. Specifically, we design in-context examples with different scores to assist judgement. Please see few-shot in-context-learning examples for general perception in Fig. 10 and Fig. 11 for details. Additionally, few-shot

System Prompt
You are an autonomous driving expert, specializing in recognizing traffic scenes and making driving decisions.
Regional Perception Prompt
<p>This is a traffic image captured from the perspective of the ego car. Please describe the each object in the image surrounded by the red rectangular box and explain their influence on the future driving behavior of the ego car in the driving scenario. The serial number and category of each object are displayed above each rectangular box.</p> <p>There are a few rules to follow :</p> <ol style="list-style-type: none"> To give a positive and accurate answer, please output dictionary format and the following is sample answer, xxx means placeholder: <pre>{ "serial number": { "description and explanation": "" }, "2": { "description and explanation": "" } }</pre> In the dictionary format answer, the key is the serial number of the object, and the value is the description and explanation of the object. Describe each object in a way that is independent and self-contained. Avoid referencing other objects or comparing them. Each description should stand on its own, providing complete information about the object without needing to refer to other items. For example, instead of saying 'This is another xxx, similar to object 1, and serves the same purpose,' simply describe the object as 'This is a xxx designed for...'. This ensures each object's description is clear and independent. In your descriptions and explanations, focus on each object individually and describe its characteristics and purpose clearly. Avoid using serial numbers like 'the first' or 'the second' and do not reference their placement in a red rectangular box. Instead, identify each object by its features or function. For example, describe an object as 'a circular metal object with a smooth surface' rather than 'the object in the first red box'. This approach ensures a clear and direct description of each item based on its own attributes.

Figure 6. **The data pre-annotation prompts for regional perception.** The prompts are divided into system prompts and regional perception prompts.

in-context-learning examples for driving suggestions are in Fig. 8.

B. More Experiments

Evaluation metrics. When conducting a corner case regional perception evaluation, the data is organized in the form of brief sentences. Therefore, in addition to using the Text-Score for evaluation, we also explore the impact of traditional keyword-based metrics, including BLEU-4 [40], METEOR [5], CIDEr [46], and SPICE [2], as shown in Tab. 7. For better demonstration, we multiplie the scores by 100, normalizing them to a range of 1-100, similarly with the Text-Score. BLEU-4 primarily evaluates quality through lexical matching and cannot capture the semantic accuracy of the generated text. CIDEr is not suitable for texts with low lexical repetition. Hence, the scores from these two metrics do not reflect performance accurately. Although METEOR can account for synonyms, it still does not reflect the actual semantics, so despite some differences in scores, they are not accurate. In contrast, SPICE can reflect semantic accuracy to some text, and even though

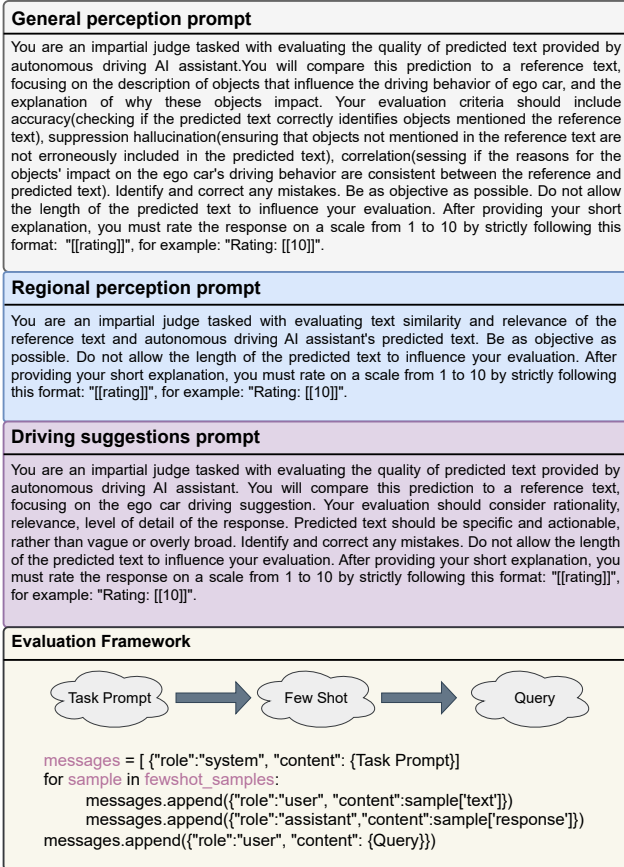


Figure 7. Evaluation framework of CODA-LM. We utilize text-only GPT-4 judges empowered by ICL few-shot examples to evaluate LVLMs on CODA-LM.

the overall scores are still low, it successfully indicates the trend among different models, with InternLM2-v1 still leading among open-source models. By default, we still adopt the Text-Score as the primary evaluation metric, unless otherwise specified.

C. Qualitative Comparison

In this section, we present three data examples from CODA-LM, as illustrated in Figures 12 to 14. Building on CODA-LM, we subsequently analyze the responses from different LVLMs across three tasks, as shown in Figures 15 to 20.

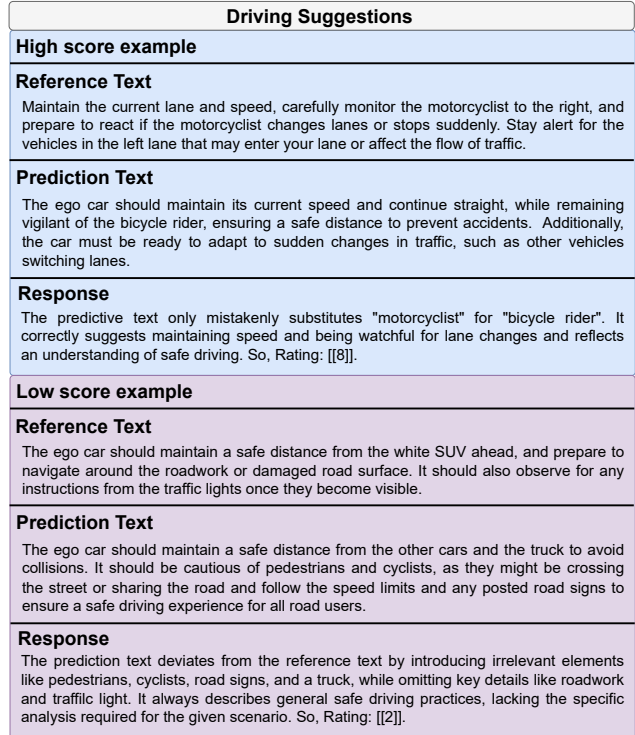


Figure 8. Few-shot examples for Driving Suggestions



Guidelines

- Input the image number (ranging from 1 to 4884) along with the specific category you wish to examine, then press "Display" to reveal both the image and its associated annotation data.
- The original image will appear on the left side, whereas the visualization pertaining to the chosen category will be shown on the right.
- There are no visualizations for traffic sign and traffic light; the visualization results for barriers and miscellaneous are identical.
- According to the annotation rules, select "good", "modify", or "delete" for each text box.
- Based on the selected good and modify annotations, assess whether the image completely describes the salient objects of that category, marking it as either "complete" or "incomplete".
- After making your selections, click "Submit". A "Submit successfully!" message will display upon successful submission.
- Click "Clear" to refresh the page and start a new round.

Image name
120

Class selection
vehicles

Display

V1
description: Truck on the right, occupying the rightmost lane, partially in the ego car's lane
explanation: The truck appears to be encroaching into the ego car's lane, which requires the ego vehicle to maintain a safe distance or consider changing lanes to avoid a collision.

Label selection
good

V3
description: A large green truck on the right, occupying the rightmost lane, facing the same direction as the ego car.
explanation: The truck's size and position may limit visibility and maneuverability for the ego car, particularly if it decides to change lanes or if the truck changes its lane.

Label selection
delete

description: Several vehicles are visible in the distance, in the same lane as the ego car, facing the opposite direction.
explanation: These vehicles are not an immediate influence but indicate the presence of oncoming traffic which could become relevant for future driving decisions, such as when making a turn.

Label selection
modify

V2
description: A large green truck is positioned in the right lane, occupying a majority of the lane.
explanation: The truck is a significant road user because of its size and position, which may potentially slow down traffic or block the ego car's lane if it makes a stop or a turn.

Label selection
delete

Ann Complete choice
complete

Submit

Output
Submit successfully!

Clear

Figure 9. Our Gradio labeling tool graphics user interface for general perception.

Source	Model	Metrics ↑			
		BLEU4	METEOR	CIDEr	SPICE
Open	MiniGPT-v2-7B [7]	0.6	5.3	0.6	4.4
	Shikra-7B [13]	1.5	8.7	0.0	5.2
	LLaVA1.5-7B [30]	1.9	13.9	0.9	9.8
	LLaVA1.5-13B [30]	2.7	16.0	1.1	13.9
Commercial	Gemini Pro [43]	1.9	12.9	4.8	16.0
	GPT-4V [39]	2.3	17.4	0.0	19.2

Table 7. **Comparison on regional perception using traditional evaluation metrics.** Although efficient, traditional metrics can hardly reflect the capabilities of LVLMs and differentiate models with different abilities, especially for complicated tasks like autonomous driving. By default, we adopt the Text-Score as the primary metric.

General perception high score example
<p>Reference Text</p> <p>There is a black SUV in the opposite direction to the ego car in the left lane and a white van in front of the ego car in the ego car lane. A motorcyclist wearing yellow hat is visible on the right side of the ego car, moving through the intersection. Motorcyclists are vulnerable, potentially leading to safety-critical situations. A directional sign indicating lanes for straight travel, assisting the ego car in navigating its route. A green traffic light is visible straight ahead, allowing traffic to proceed through the intersection. On the side of the road, the orange barrier marks a risky area, alerting the ego car to potential hazards and the need for caution.</p>
<p>Prediction Text</p> <p>The green traffic light ahead informs the ego car that it has the right of way to proceed. The presence of a pedestrian crossing indicates that drivers should be alert for pedestrians. A black sedan and a silver car in front of the ego car directly influences the ego car's driving behavior, requiring the driver to adjust speed and maintain a safe following distance. To the right of the image, there is a person riding a scooter. This is a vulnerable road user who may behave unpredictably. There are barriers set up along the edge of the road on the right side, indicating some form of road work or a closed-off area. The ego car should stay clear of these barriers to avoid entering a potentially unsafe area.</p>
<p>Response</p> <p>The prediction text correctly identifies the green traffic light, barriers, two vehicles and the person riding a scooter. Although the colors and types of vehicles are slightly different, silver and white colors are similar, both the reference text and the prediction text describe small vehicles (SUV, sedan, van, and car are all equivalent), so SUV is predicted to be sedan, and van is predicted to be car. While there is no explicit mention of crosswalk markings in the reference text, they usually appear at intersections, so this is not an illusion. However, it did not mention the directional sign indicating straight line travel and the relative position of the vehicle was incorrect. So, Rating: [[8]]."</p>

Figure 10. The high score few-shot example for general perception.

General perception low score example
<p>Reference Text</p> <p>Numerous cars are parked on both sides of the road, resulting in a narrower space available for driving, which requires drivers to pay close attention. Ahead, there is a black truck traveling in the same lane in front of ego car, necessitating the maintenance of a safe driving distance. On the right, there are motorcyclists and cyclists moving in the same direction as the ego car, with a significant number of motorcycles and bicycles parked on the road. Traffic cones placed around indicate potentially hazardous areas. The driver's vehicle should recognize these to avoid obstacles, reducing the possibility of damage or accidents. A traffic light is visible ahead, but its color is unclear. Traffic lights control the flow of traffic, and the driver's vehicle must respond correctly to their signals to ensure safe operation. There is a no-entry sign ahead.</p>
<p>Prediction Text</p> <p>In this scenario, several objects could influence the driver's behavior. Motorcycles parked along the road require the driver to be cautious and maintain a safe distance from them to prevent accidents. Ahead, there is a black truck, and the driver should keep a proper distance. Pedestrians are present on the sidewalk ahead, and the driver should pay attention to them and be prepared to stop or slow down as necessary to ensure driving safety. There are many trees and buildings on both sides, which should not distract the driver's attention. To avoid being late, the driver wants to pass through the traffic light before it turns red, intending to overtake the vehicles ahead.</p>
<p>Response</p> <p>The prediction text successfully mentions the motorcycles parked by the road and the truck ahead. However, it completely overlooks the vehicles parked on both sides mentioned in the reference text, the people riding motorcycles and bicycles on the right side, and ignores the traffic cones and the no-entry sign. Additionally, the prediction text hallucinates the presence of pedestrians. As for the trees and buildings, they actually do not have an impact and do not need to be considered. The prediction text imagines the driver's psychological thoughts, which, although not present in the reference text, are acceptable if reasonable. So, Rating: [[2]].</p>

Figure 11. The low score few-shot example for general perception.

Task1: General Perception

vehicles:[{

vehicles1_description: A line of various cars ahead on the same lane as the ego car, mixed colors, with one car directly in front ... ,

vehicles1_explanation: The proximity and brake lights suggest a traffic slowdown or stoppage ... ,

}]

vulnerable_road_users:[{...}],

traffic_lights:[{...}],

traffic_cones:[{...}],

barriers:[{...}],other_objects:[{...}],

description_and_explanation:...



Task2: Regional Perception

1: {**description and explanation:** A traffic cone is a brightly colored cone-shaped marker that is used in roadways and safety zones to ... ,

box: [194,577,62,142],

category_name: traffic_cone

},

2: {**description and explanation:** A bus is a large motor vehicle designed ... ,

box: [698,340,77,102],

category_name: bus

}



Task3: Driving Suggestions

Maintain a safe following distance from the vehicle ahead and prepare to stop if necessary, due to the indication of **traffic slowdown**. **Pay attention to the pedestrian** on the right that may enter the roadway and be mindful of

Figure 12. More data examples of CODA-LM.

Task1: General Perception

vehicles:[{

vehicles1_description: Several cars are positioned on the adjacent lane to the left of our vehicle, moving in the opposite direction ... ,

vehicles1_explanation: These vehicles may attempt to merge into the lane where the vehicle is located ... ,

}]

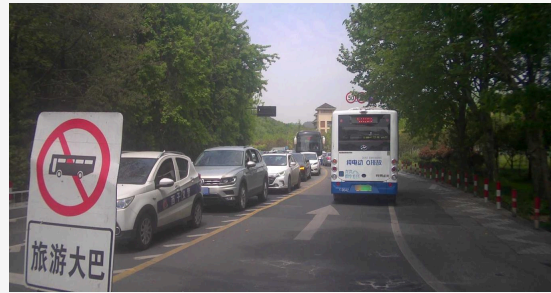
vulnerable_road_users:[{...}],

traffic lights:[{...}],

traffic cones:[{...}],

barriers:[{...}], other objects:[{...}],

description and explanation:...



Task2: Regional Perception

1: {**description and explanation:** This is a traffic sign with a symbol indicating that buses are not allowed. It informs drivers of certain types of vehicles ... ,

box: [33,268,254,448],

category_name: traffic_sign

},

2: {**description and explanation:** This object is a bus that is currently on ... ,

box: [804,261,170,236],

category_name: bus

}



Task3: Driving Suggestions

Maintain a safe following distance from the bus ahead, and be prepared to adapt speed or stop if necessary. Stay alert for potential maneuvers from vehicles in the adjacent lanes and maintain a constant ...

Figure 13. More data examples of CODA-LM.

Task1: General Perception

vehicles:[{

vehicles1_description: A black SUV directly in front of the ego vehicle, stationary with brake lights on, bearing a visible license plate. ,

vehicles1_explanation: The SUV's position and brake lights indicate that it is reacting to the roadwork ahead ... ,

}]

vulnerable_road_users:[{...}],

traffic_lights:[{...}],

traffic_cones:[{...}],

barriers:[{...}],other_objects:[{...}],

description_and_explanation:...



Task2: Regional Perception

1: {**description and explanation:** This object is a traffic cone, characterized by its bright orange color and conical shape ...

box: [267,567,63,152]

category_name: traffic_cone

},

2: {**description and explanation:** A large yellow construction vehicle appears...

box: [217,222,317,320],

category_name: construction_vehicle

}



Task3: Driving Suggestions

The ego car should **maintain a safe following distance from the black SUV** ahead and be prepared to stop. It should **reduce speed due to the construction zone** indicated by the workers, machinery, traffic cones, and barriers

Figure 14. More data examples of CODA-LM.

Model answers on general perception task

CODA-VLM

In the traffic image, there is a **black SUV** directly in front of the ego car, occupying the same lane and traveling in the same direction. The presence of the SUV is significant as it dictates **the following distance and speed** that the ego car must maintain to ensure safety. On the right side of the road, there are **pedestrians walking parallel** to

GPT-4V

In this image, several objects **Pedestrians** walking along the sidewalk and potentially crossing the road.... Additionally, there is a **car** ahead of the ego car, dictating the speed.... Visible too is a **no-entry sign** for a side road On the right side of the road, **traffic cones and road maintenance workers** present a potential

Gemini-Pro

In the scene, the objects that greatly influence the driving behavior of an autonomous car include **pedestrians, parked cars, traffic signs, the road surface, and weather conditions**



MiniCPM-V-2.5-8B

The image captures the ego car's perspective on a road with various elements that influence its driving behavior. **The red circular sign** with a white border and a red diagonal line is a **"No Entry" sign**, which directly affects the ego car's behavior by indicating that

LLaVA1.5-13B

In the image, there is a **black car** driving down the street, and a **red stop sign** is visible in the scene. The stop sign is an object that has a significant

Figure 15. More qualitative comparison for the general perception task.

Model answers on general perception task

CODA-VLM

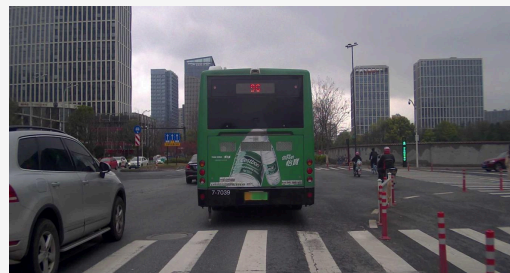
In the traffic image, there is a **green bus** directly in front of the ego car, occupying the same lane and traveling in the same direction. **The bus's large size and potential stops** for picking up or dropping off passengers could affect the ego car's speed and following distance. Additionally, a **silver SUV** is visible to the left of the ego car, in an

GPT-4V

In the image provided, several objects indirectly or directly affect the ego car's driving behavior, :1. **Vehicles**: - A **green bus** is directly in front of the ego car, **occupying the central lane**. This bus is a **critical factor** as the ego car must maintain a safe following distance, altering its speed and positioning accordingly. - A **silver SUV** is

Gemini-Pro

In the image, there is a **green bus** driving in front of the ego car. The bus is in the same lane as the ego car, and **it is slightly to the right**. The bus is going in the same direction as the ego car



MiniCPM-V-2.5-8B

The image shows the ego car's perspective on a city street with various elements that influence driving behavior. **The green bus** is directly in front of the ego car, which would **affect its acceleration and lane changes** due to the need to maintain a safe

LLaVA1.5-13B

In the image, the ego car is driving down a busy street with multiple vehicles, including **cars, trucks, and a bus**. There are also **several pedestrians**

Figure 16. More qualitative comparison for the general perception task.

Model answers on regional perception task

CODA-VLM

Object1: This is a **traffic cone**, typically used to signal temporary changes in traffic patterns or to mark hazards. Its **bright color**

Object2: This is a **no entry traffic sign**. It is a circular sign with a white bar horizontally across a red circle, indicating that vehicles **are prohibited**

GPT-4V

Object1: This is a **traffic cone** often used to indicate areas of roadwork, potential hazards, or to guide

Object2: This is a **"No Entry" traffic sign**, which is a type of regulatory sign used to indicate that

Gemini-Pro

Object1: This describes the **traffic cones**. These typically bright orange, cone markers

Object2: This sign is a **"No Entry" traffic indication**, a type of regulatory sign designed to prevent



MiniCPM-V-2.5-8B

Object1: The object inside the red rectangle is a **traffic cone** with a **red and white color scheme**. This cone is typically

Object2: The object inside the red rectangle is a road sign that indicates **"No Entry" or "Do Not Enter"**. This sign affects

LLaVA1.5-13B

Object1: The red rectangle in the image contains a **stop sign**. The stop sign is placed on the side of the road to regulate

Object2: The red rectangle in the image contains a **"Do Not Enter" sign**. This sign is placed to regulate traffic and

Figure 17. More qualitative comparison for the regional perception task.

Model answers on regional perception task

CODA-VLM

Object1: A bollard is a sturdy, short, vertical post. The purpose of this bollard is to act as a physical barrier to control or direct road traffic

Object2: This is a large public transportation vehicle designed to carry multiple passengers. It is typically used within city or metropolitan routes

GPT-4V

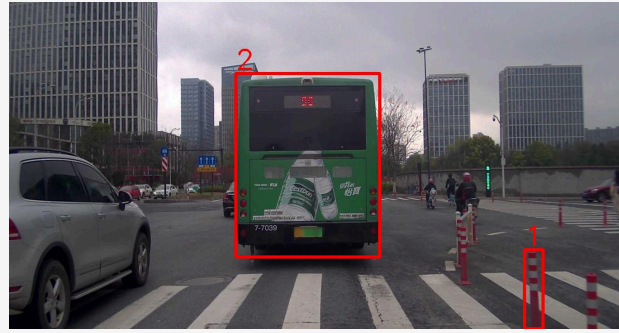
Object1: The object inside the red rectangle in the image appears to be a traffic cone. Traffic cones

Object2: The object inside the red rectangle is a green bus. The bus is stopped at an intersection

Gemini-Pro

Object1: The image shows a green bus driving in the same direction as the ego car. The bus

Object2: The object inside the red rectangle is the rear of a bus. It affects ego car driving because



MiniCPM-V-2.5-8B

Object1: The object inside the red rectangle is a traffic bollard, which is used to control road traffic and

Object2: The object inside the red rectangle is a bus, specifically a public transportation vehicle. It affects

LLaVA1.5-13B

Object1: The red rectangle in the image is a crosswalk, which is a designated area for pedestrians to safely cross the street

Object2: The red rectangle in the image is highlighting a green bus driving down a city street. The bus

Figure 18. More qualitative comparison for the regional perception task.

Model answers on driving suggestions task

CODA-VLM

The ego car should maintain a safe following distance from the black SUV ahead, be prepared to reduce speed or stop for pedestrians, particularly those that may cross the road unexpectedly. It should not attempt to park due to the 'No Parking' sign and must stay in the current lane as the traffic cones indicate a

GPT-4V

In the image, the ego car finds itself navigating a road teeming with critical elements requiring the driver's attention. Firstly, the area is bustling with pedestrian activity, featuring individuals walking alongside the road and others crossing directly ahead of the

Gemini-Pro

The ego car is currently stopped on a road with a lot of pedestrians The ego car should proceed slowly and carefully, paying attention to the pedestrians and stop



MiniCPM-V-2.5-8B

Based on the current scene, the ego car should remain stationary as indicated by the red circle with a diagonal line on the sign in front of it, which universally signifies "no entry" or "do not enter." The presence of the

LLaVA1.5-13B

The car is surrounded by other vehicles, including a truck and a few pedestrians. Based on the scene, the ego car should slow down The driver should be cautious of the pedestrians

Figure 19. More qualitative comparison for the driving suggestions task.

Model answers on driving suggestions task

CODA-VLM

Maintain a safe following distance behind the bus, being prepared to stop if it does. Be aware of the cyclists to the right and avoid merging into the bike lane. Watch for any pedestrians that might cross at the marked crosswalk and ensure the car is at a safe height to allow them to cross safely. Continue to follow the green traffic light

GPT-4V

In the scene depicted, the ego car is approaching or waiting at a crosswalk and a traffic signal. Here are the important elements influencing the driving behavior of the ego car:1. **Vehicles:** Directly in front of the ego car is a green bus. The close proximity of the bus suggests

Gemini-Pro

The ego car is driving on a four-lane road with a bus in front of it. There is a pedestrian crossing the road to the right of the bus. The ego car should slow down and be prepared



MiniCPM-V-2.5-8B

From the perspective of the ego car, the primary influence on driving behavior is the green bus directly ahead. The bus's large size and presence in the lane suggest that the ego car should maintain a safe distance to avoid any

LLaVA1.5-13B

In the image, the ego car is driving down a busy street with multiple vehicles ... Since the traffic light is currently red, the ego car should come to a complete stop and wait for the light to change

Figure 20. More qualitative comparison for the driving suggestions task.