Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases

Anonymous EMNLP submission

Abstract

Large Vision-Language Models (LVLMs) have received widespread attention in advancing the interpretable self-driving. Existing evaluations of LVLMs primarily focus on the multi-faceted capabilities in natural circumstances, lacking automated and quantifiable assessment for self-driving, let alone the severe road *corner cases*. In this paper, we propose **CODA-LM**, the very first benchmark for the automatic evaluation of LVLMs for self-driving corner cases. We adopt a hierarchical data structure to prompt powerful LVLMs to analyze complex driving scenes and generate high-quality pre-annotation for human annotators, and for LVLM evaluation, we show that using the text-only large language models (LLMs) as judges reveals even better alignment with human preferences than the LVLM judges. Moreover, with CODA-LM, we build **CODA-VLM**, a new driving LVLM surpassing all the open-sourced counterparts on CODA-LM. Our CODA-VLM performs comparably with GPT-4V, even surpassing GPT-4V by **+21.42%** on the regional perception task. We hope CODA-LM can become the catalyst to promote interpretable self-driving empowered by LVLMs.

1 Introduction

The Large Vision-Language Models (LVLMs) (Liu et al., 2023b; OpenAI, 2023b; Gou et al., 2023) have attracted increasing attention, primarily due to their remarkable visual reasoning abilities, which are of paramount importance (Hu et al., 2023; Sima et al., 2023) for autonomous driving. Traditional self-driving systems use a modular design, integrating various modules such as perception, prediction, and planning to handle complicated road scenarios, which, however, are still inadequate to generalize in the open domain, especially for real-world *corner cases* (Li et al., 2022). In this paper, we primarily consider *object-level corner cases*[1], including both *instances of novel categories* and *novel instances of common categories* (Li et al., 2022).

[1]We adopt the definition of object-level corner case in (Breitenstein et al., 2021).

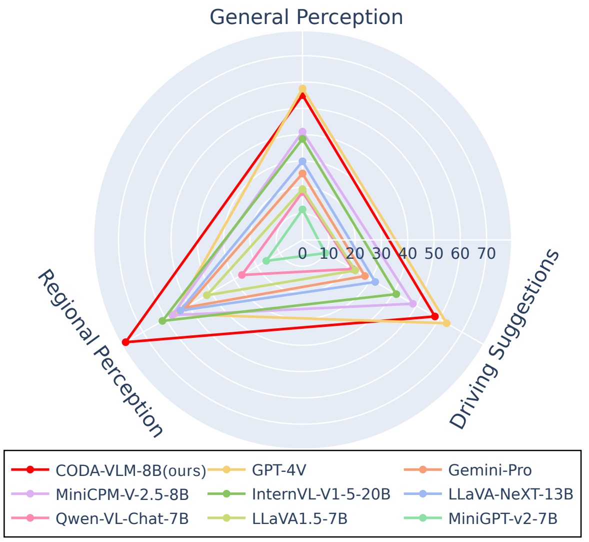


Figure 1: **Comparison among open-sourced and commercial LVLMs on CODA-LM**. CODA-LM provides the very first automated and quantifiable evaluation of LVLMs on road corner cases.

LVLMs, on the other hand, with their extensive world knowledge and reasoning capability, have the potential to overcome these severe challenges. A preliminary study (Wen et al., 2023) has revealed the ability of powerful LVLMs (OpenAI, 2023b) in handling the road corner cases, where samples are selected from CODA (Li et al., 2022), the largest real-world corner case dataset, to prompt GPT-4V. Although effective, their evaluation relies on redundant manual inspections, hindering the scalability of larger-scale LVLM evaluation for self-driving.

In this paper, we propose **CODA-LM**, the very first benchmark for the automated and systematic evaluation of LVLMs on self-driving corner cases. Following Wen et al. (2023), we utilize corner cases from CODA and collect question-answering annotations of three distinct tasks including the *general perception*, the *regional perception*, and the *driving suggestions*. To obtain high-quality pre-annotation, we design a hierarchy data structure to help GPT-

Dataset	Multimodal	Corner	General Per.	Regional Per.	Suggestion
CODA (Li et al., 2022)	✗	✓	✓	✓	✗
StreetHazards (Hendrycks et al., 2019)	✗	✓	✓	✓	✗
nuScenes-QA (Qian et al., 2023)	✓	✗	✓	✗	✗
BDD-X (Kim et al., 2018)	✓	✗	✓	✗	✗
DRAMA (Malla et al., 2023)	✓	✗	✓	✓	✓
DriveLM (Sima et al., 2023)	✓	✗	✓	✓	✓
CODA-LM (ours)	✓	✓	✓	✓	✓

Table 1: **Comparison between CODA-LM and existing datasets**. CODA-LM is the first large-scale multimodal road corner case dataset for autonomous driving with a hierarchical evaluation framework.

4V better analyze complex road scenes and capture all necessary obstacles. The structured responses are then converted to coherent texts, which are then verified by human annotators. Different from the existing LVLM benchmarks (Li et al., 2024), we show the necessity of using text-only LLMs (OpenAI, 2023a) as "judges" for automated evaluation of LVLMs on CODA-LM, which reveals a stronger consistency with humans than LVLM judges (OpenAI, 2023b). Moreover, we propose **CODA-VLM**, a novel driving LVLM achieving the state-of-the-art among all open-sourced LVLMs on CODA-LM, even surpassing GPT-4V on the regional perception task by **+21.42%**. The main contributions of this work contain three parts:

1. We propose **CODA-LM**, the **very first** LVLM benchmark for an automatic and systematic evaluation of LVLMs on road corner cases.

2. We demonstrate that the text-only LLMs can serve as powerful judges to evaluate LVLMs, revealing a stronger consistency with human judgments even compared with LVLM judges.

3. We comprehensively evaluate performance of existing LVLMs on self-driving corner cases, and construct the **CODA-VLM**, a new driving LVLM comparable with GPT-4V on CODA-LM, surpassing all open-sourced counterparts on both driving perception and suggestions.

2 Related Work

LVLM evaluation primarily focuses on natural image spaces. MME (Fu et al., 2023) introduces manually designed question-answer pairs to measure both perception and cognition abilities on a total of 14 sub-tasks. MMBench (Liu et al., 2023c) employs GPT-4 to transform free-form predictions into the pre-defined multiple-choice questions and introduces the CircularEval strategy for a more robust evaluation. SEED-Bench-2 (Li et al., 2023a) adopts a similar format with MMBench but extends over 27 dimensions, evaluating LVLMs' capabilities in image and text comprehension, interleaved image-text understanding, and generation tasks. Auto-Bench (Ji et al., 2023) generates question-answer-reasoning triplets using LLMs (Touvron et al., 2023; Chen et al., 2023c; Liu et al., 2024b; Gou et al., 2024) as evaluation data. All the evaluation benchmarks above rely on the rigid, manually curated datasets of natural images, and thus, difficult to be applied for complicated driving scenarios.

Autonomous driving datasets. The NuScenes-QA (Qian et al., 2023) manually constructs 460K question-answer pairs based on the object attributes and relationships among objects in scene graphs. BDD-X (Kim et al., 2018) focuses on the behavior of the ego car and provides corresponding reasons. While both datasets concentrate on general perception, DRAMA (Malla et al., 2023) and DriveLM (Sima et al., 2023) further consider regional perception and driving suggestions. DRAMA identifies the most critical targets and offers the corresponding advice, while DriveLM promotes end-to-end autonomous driving understanding through the usage of graph-structured question-answer pairs. Self-driving systems often fail in corner cases, leading to severe accidents. StreetHazards (Hendrycks et al., 2019) is a synthesized dataset where corner cases are simulated via graphics. CODA (Li et al., 2022) is a real-world road corner case dataset with 10K driving scenes, spanning more than 40 classes. As in Tab. 1, the existing corner case datasets lack language modality, while vision-language datasets don't cover road corner cases. Thus, we propose CODA-LM, the first large-scale multimodal road corner case dataset for self-driving with a hierarchical automatic evaluation framework.

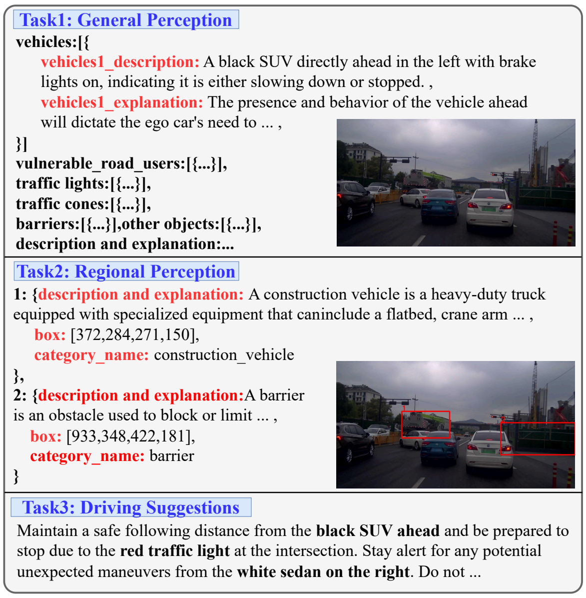


Figure 2: **Task hierarchy of our CODA-LM**, including *general perception* (up), *regional perception* (middle), and *driving suggestions* (bottom), respectively.

3 CODA-LM Dataset

Based on road corner cases from CODA (Li et al., 2022), our CODA-LM comprises 9,768 real-world driving scenarios with 41,722 textual annotations for critical road entities and 21,537 annotations for road corner cases. Critical road entities affecting self-driving decision-making are categorized into seven distinct groups, including *vehicles*, *vulnerable road users (VRUs)*, *traffic signs*, *traffic lights*, *traffic cones*, *barriers*, and *other objects* (*e.g.*, animals and traffic islands). As illustrated in Fig. 2, our CODA-LM involves a task hierarchy with three principal tasks, including the *general perception*, *regional perception*, and *driving suggestion*, as detailed in Sec. 3.1-3.3 separately. Such a systematic task hierarchy requires LVLMs to understand the complex driving environments, providing a comprehensive assessment of **interpretable** self-driving agents empowered by LVLMs.

3.1 General Perception

The foundational aspect of the general perception task lies in a comprehensive understanding of critical road key entities in driving scenarios, including their appearance, location, and reasons why they influence the driving behaviors of our ego car. This task is pivotal in evaluating LVLMs' proficiency in interpreting complex interactive scenes, mirroring the perception process in self-driving. Moreover, to comprehensively evaluate LVLMs' performance in different environments, we classify the images based on the time and weather conditions, including *night* and *daytime* scenes for time conditions, as well as *clear*, *cloudy*, and *rainy* circumstances for the weather conditions.

3.2 Regional Perception

The regional perception task measures LVLMs' capabilities to understand corner case objects when provided with specific bounding boxes, which involves describing objects within the given bounding boxes and explaining why they would influence self-driving behavior. The establishment of regional perception is based on a core realization (Han et al., 2021) that the ability to accurately localize corner cases is crucial for enhancing the overall system's robustness in the practical application of autonomous driving. These scenarios often contain complicated or unusual elements that traditional models might overlook or struggle to interpret correctly, such as *unique traffic signs*, *pedestrians with abnormal behavior*, and *atypical road conditions*. By specifically focusing on these cases, we can gain a comprehensive understanding of the model's capability to comprehend corner case objects.

3.3 Driving Suggestions

The driving suggestions task aims to evaluate the capability of LVLMs in formulating driving advice, a critical component for interpretable self-driving. This task is closely related to the planning process of autonomous driving, requiring the model to provide the optimal driving suggestions for the ego car after correctly perceiving the general and regional aspects of the current driving environment. Via the construction of the driving suggestions task, we can deeply evaluate the performance of LVLMs in formulating effective driving strategies.

4 CODA-LM Construction

4.1 Data Collection

Overview. For each task introduced in Sec. 3, we meticulously design prompts to guide GPT-4V to generate high-quality textual pre-annotations based on visual information, as provided in Figs. 5 and 6. We start by constructing a hierarchical data structure in the JSON format (detailed in the following) to guide GPT-4V for better scene understanding of complex road scenes, categorizing the critical road entities into seven classes. Each entity is detailedly described, explaining how they affect the driving behavior of the ego car. After obtaining the GPT-4V responses for both the general and re-

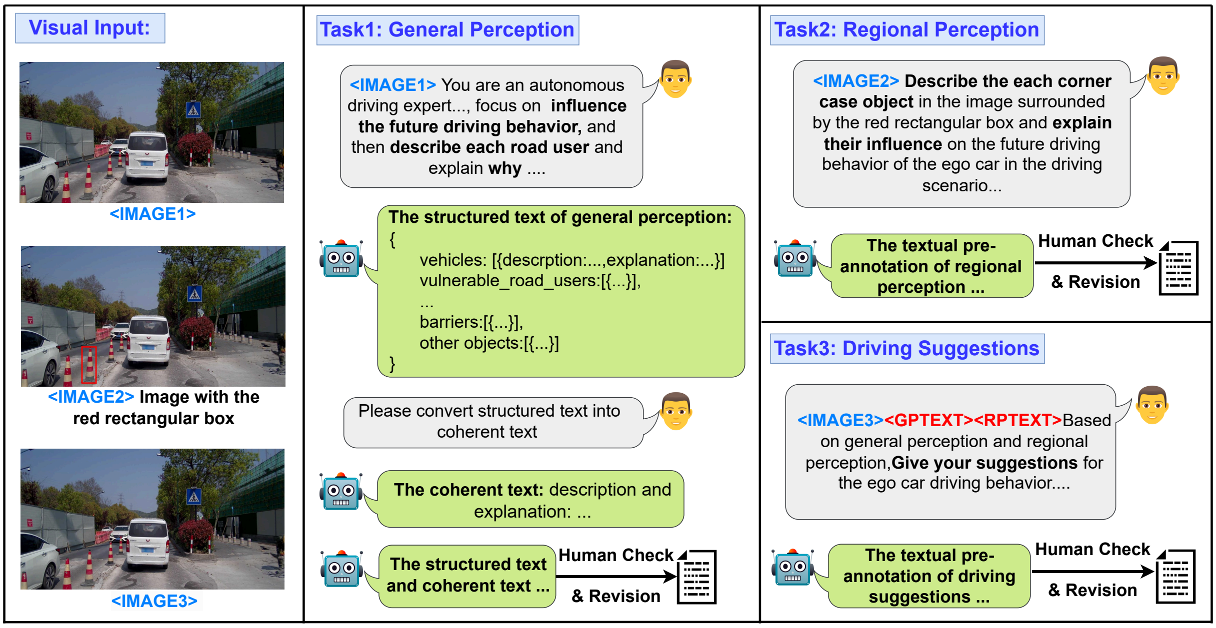


Figure 3: **Overview of CODA-LM construction.** We design a hierarchical data structure in the JSON format to guide GPT-4V to better understand complicated driving scenes and generate high-quality pre-annotations for human annotators to conduct further verification and revision. <GPTEXT> and <RPTEXT> refer to the revised answer from the general perception task and the regional perception task, respectively.

gional perceptions, we combine these with the corresponding road image to form a composite context for the GPT-4V to generate the driving suggestions. Finally, we ask human annotators to verify and revise the pre-annotations. The overall construction pipeline is visualized in Fig. 3.

Hierarchical text structure for general perception. To conduct precise perception and even driving suggestions, it is essential to recognize all road obstacles. However, if directly prompted with plain texts, we notice that GPT-4V suffers from 1) *entity ignorance*: GPT-4V tends to focus on the salient objects while ignoring the insignificant obstacles. 2) *element ignorance*: when prompted with plain texts, GPT-4V might describe road entities without explaining why it affects the ego car or vice versa.

Thus, as in Fig. 3 (middle), we design a hierarchy data structure in the JSON format from *categories* to *objects* and ultimately *data elements*. GPT-4V is guided to first recognize objects of every single *category* separately, and "fill in" *description* and *explanation* of each object. We then prompt GPT-4V again to convert structured texts to coherent natural languages and serve as the final pre-annotations. As in Tab. 4, the **"structure-coherence"** pipeline achieves significant consistency with humans.

Visual prompts for regional perception. We consider two manners to convert bounding boxes as the inputs for LVLMs, 1) *visualization*: suggests marking the targets with red rectangle boxes on the original images, as in Fig. 3 (left). 2) *grounding*: uses normalized coordinates (top-left and bottom-right corners) in text prompts to locate the target, similarly with LLaVA (Liu et al., 2023b). As shown in Tab. 5, visualization with the red rectangle boxes reveals significantly better empirical results, which is then considered as the default vision prompts.

Human verification and revision is ultimately adopted to guarantee the correctness of our CODA-LM annotations. For convenience, we construct a labeling tool GUI based on Gradio (Abid et al., 2019), as in Fig. 9, followed by the ethics review.

Data split. We separate 4,884 scenes as the training set, with 4,384 data samples as the validation set and the remaining 500 samples as the test set to construct the CODA-LM benchmark as in Tab. 2 for a comprehensive comparison among LVLMs on self-driving corner cases.

4.2 Evaluation Framework

Unsatisfactory LVLM judges. LMSYS (Zheng et al., 2023) shows the feasibility of using GPT-4 as judges to evaluate the intelligent chat assistants by giving a 1-10 score, revealing high consistency with human assessment. Inspired by that, we start with a preliminary attempt by using *LVLM judges* (*e.g.*, GPT-4V) to evaluate various LVLMs, which, however, merely obtains a human consistency of around 70% for all three tasks, as shown in Tab. 3.

We assume that this is probably due to the unsatisfactory instruction-following ability of GPT-4V, which cannot always respond in the required format (Bai et al., 2023b). Meanwhile, GPT-4V still lacks the multimodal in-context learning ability, making few-shot evaluation indispensable in complex and varied autonomous driving scenarios.

Text-only LLM as LVLM judges. In this paper, we propose to adopt *text-only LLMs* (*e.g.*, GPT-4) as judges to evaluate LVLMs on driving scenarios. Given the reference ground truths and few-shot ICL examples, GPT-4 is instructed to evaluate the correctness of model responses with a score ranging from 1 to 10. The average score of the whole evaluation set serves as the final `Text-Score`. We provide the evaluation prompts and ICL examples in Figs. 7 to 11. As shown in Tab. 3, the text-only GPT-4 judge evaluates more consistently with human judgments than the GPT-4V judge.

Potential bias and hallucination To revise that, we ask the human annotators to verify and revise the evaluation results given by GPT-4 and finally report the benchmark results in Tab. 2.

Evaluation criteria of the general perception include *accuracy*, *hallucination penalty*, and *consistency*. Accuracy evaluates how well LVLMs match with reference ground truths, while hallucination penalty suggests that LVLMs should not mention entities not collected in the reference, which, otherwise, should be penalized when computing scores. Consistency focuses on the relationship between the object description and the explanation of why it affects the ego car. For driving suggestions, the criteria focus on the *rationality*, *relevance*, and *detail level* of driving suggestions generated by LVLMs. Especially for driving suggestions, we require the responses to be specific and actionable, rather than vague or overly broad. Prompts are listed in Fig 7.

Evaluation metrics. As previously introduced, we utilize the `Text-Score` (Zheng et al., 2023) given by text-only GPT-4 judge as the primary evaluation metrics for all three tasks. We further explore the usage of traditional text-generation evaluation metrics as in Tab. 7, which, however, cannot well differentiate the capabilities of various LVLMs under complicated self-driving scenarios.

4.3 CODA-VLM

In this section, we explore improving the performance of LVLM models on road corner cases from the perspectives of both the visual representation and knowledge transfer and construct our **CODA-VLM**, a novel driving LVLM achieving state-of-the-art recognition and planning performance on autonomous driving scenarios.

Knowledge transfer. To acquire more comprehensive pre-training knowledge, we use the LLaVA-Llama-3-8B-v1.1 developed by Xtuner[2] as our baseline, which follows the basic architecture of LLaVA1.5 (Liu et al., 2023a), while replacing the LLM with LLaMA3-8B[3], and performing modality alignment and instruction fine-tuning on a larger dataset. Based on that, we inject knowledge specific to driving scenarios via instruction fine-tuning. Specifically, we organize the image-text pairs from CODA-LM into a dialogue format and employ a rational data sampling strategy to form an instruction-following dataset. Furthermore, to efficiently learn while preserving as much pre-training knowledge as possible, we use LoRA (Hu et al., 2021) to fine-tune both the LLM and the visual encoder[4].

Visual representation. To obtain more effective visual representations and enhance the model's regional perception capabilities, we refer to the dynamic high resolution (*i.e.*, AnyRes) from LLaVA-NeXT (Liu et al., 2024a). While retaining the fixed global image resolution, we split original images into different sub-images, each independently encoded by a shared visual encoder, and finally concatenate all visual tokens together before feeding into LLMs. Moreover, considering context lengths and training costs of LLMs, we observe that a 2×2 MaxPool operation on visual tokens of sub-images can effectively reduce redundancy, achieving a better trade-off between efficiency and performance.

Implementation details. It is worth noting that our approach is simple yet effective. The training of CODA-VLM requires only 3 hours on 8 A800 GPUs. Specifically, we use LoRA with r = 256 and α = 256 for the LLM, and r = 64 and α = 16 for the visual encoder, fine-tuning with a context length of 4096. The learning rate is set to $2e^{-4}$, training for 4 epochs with a batch size of 16 per GPU. We utilize the combination of the train and validation splits of CODA-LM, as discussed in Sec. 4.1. In Sec. 5.2 and 5.3.4, we provide more detailed analysis and empirical ablation results on CODA-VLM.

[2]https://github.com/InternLM/xtuner
[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[4]https://huggingface.co/openai/clip-vit-large-patch14-336

Method	General↑ Text-Score	Regional Perception ↑ ALL	Vehicle	VRU	Sign	Light	Cone	Barrier	Other	Suggestion↑ Text-Score
MiniGPT-v2-7B	11.58	15.93	18.74	13.58	15.71	17.78	15.34	13.02	14.41	10.00
Shikra-7B	12.24	22.94	28.29	17.88	20.00	15.56	21.23	20.00	19.67	10.20
LLaVA1.5-7B	19.30	42.06	46.67	38.47	39.14	48.89	50.83	30.93	33.82	23.16
Qwen-VL-Chat-7B	18.22	26.62	35.48	24.16	20.86	23.33	19.61	17.56	25.86	22.06
MiniCPM-V-2.5-8B	41.12	57.20	61.91	<u>54.82</u>	<u>59.43</u>	46.67	66.57	35.35	<u>58.75</u>	48.48
LLaVA1.5-13B	24.54	42.41	53.62	36.79	33.71	46.67	41.27	30.41	33.82	27.90
LLaVA-NeXT-13B	29.86	53.63	55.51	47.08	54.00	<u>60.00</u>	70.34	40.47	46.45	31.92
InternVL-V1-5-20B	38.38	<u>61.53</u>	<u>63.77</u>	53.14	50.57	57.78	<u>80.34</u>	46.86	57.11	41.18
Gemini-Pro	25.24	51.38	49.03	42.77	37.43	42.22	69.56	45.70	51.32	27.40
GPT-4V	**57.50**	56.26	60.89	40.58	49.43	54.44	66.08	<u>50.17</u>	53.16	**63.30**
CODA-VLM (ours)	<u>55.04</u>	**77.68**	**78.79**	**73.80**	**64.86**	**73.33**	**86.18**	**78.72**	**68.75**	<u>58.14</u>

Table 2: **Comparison among open-sourced and commercial LVLMs on CODA-LM Test set.** All open-sourced LVLMs suffer from the complicated road corner cases, while our **CODA-VLM**, due to its usage of superior vision representation and knowledge transfer, performs the best or second best on all evaluated dimensions, surpassing all open-sourced counterparts. Note that here we re-scale the original 1-10 `Text-Score` to 1-100 for better readability. **Bold** denotes the best results, while <u>underline</u> suggests the second best.

5 CODA-LM Benchmark

In this section, based on the proposed CODA-LM dataset, we start by comparing and analyzing the performance of different LVLMs in Sec. 5.1, followed by an in-depth analysis of model architecture designs in Sec. 5.2. We then conduct an ablation study on critical components of dataset construction and evaluation in Sec. 5.3.

5.1 Main Results

Baselines. In this work, we evaluate a total of 10 LVLMs, including both open-sourced and commercial models. Commercial models consist of the Gemini-Pro (Team et al., 2023) and GPT-4V (OpenAI, 2023b), while the open-sourced LVLMs are categorized based on the parameter sizes of their language models. The 7B/8B variants include the MiniGPT-v2 (Chen et al., 2023a), Shikra (Chen et al., 2023e), LLaVA1.5 (Liu et al., 2023a), Qwen-VL-Chat (Bai et al., 2023a) and MiniCPM-Llama3-V-2.5 (Team, 2024), while the 13B/20B LVLMs consist of LLaVA1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024a) and InternVL-Chat-V1-5 (Chen et al., 2024). Each model is evaluated on the three tasks separately for a comprehensive analysis of their performance on self-driving corner cases.

Setting. To ensure the reproducibility of our evaluation results, we use the same prompt for generating responses for all evaluated LVLMs and employ greedy decoding during inference, which generates the next token with the highest probability at each step as output, thus eliminating randomness during inference. As discussed in Sec. 4.2, GPT-4 is used as the judge for evaluation, with the temperature coefficient set to 0 and a fixed random seed, to ensure consistency when scoring different models.

Results. The comparison results on the CODA-LM Test set are reported in Tab. 2. Among the open-sourced baselines, MiniCPM-V-2.5-8B achieves the best performance, probably due to the usage of the powerful LLaMA3 base model, only ranking second to Intern-VL-1.5-20B on regional perception. Among the commercial models, GPT-4V continues to demonstrate a leadership position, ranking first on general perception and driving suggestions. Interestingly, Gemini-Pro is polarized, showing poor results in general perception and driving suggestions while excelling in regional perception. **CODA-VLM**, instead, achieves the best or second best on all the evaluated dimensions, surpassing all open-sourced counterparts. CODA-VLM obtains comparable performance with GPT-4V, even exceeding GPT-4V by **+21.42%** on regional perception. A qualitative comparison is given in Fig. 4.

5.2 Analysis

Visual represention. Recent works (Liu et al., 2024a; Chen et al., 2024) have revealed the significant benefit of utilizing high-resolution images as input for LVLMs. For regional perception, simply increasing the image resolution from 224 to 336 enables LLaVA1.5-7B to outperform Shikra-7B by

Judge	Reference	General	Regional	Suggestion
GPT-4	GT	**83.67**	**85.71**	**89.80**
GPT-4V	Image	69.39	75.51	69.39
GPT-4V	Img & GT	79.59	79.59	87.76

Table 3: **Consistency between different judges and human judgments.** Text-only GPT-4 judges reveal superior consistency for all tasks. GT denotes ground truth answers. Default settings are marked in gray.

Judge	Reference	Consistency (%)
GPT-4	Plain	71.43
GPT-4	Structured & Concat	77.55
GPT-4	Structured & Coherent	**83.67**

Table 4: **Consistency among human judgments and GPT-4 judges with different references.** The *structured coherence* manner reveals significant superiority.

20%. By further increasing the effective resolution with the AnyRes, LLaVA-NeXT-13B surpasses the LLaVA1.5-13B by over 11%. The compression of visual tokens is another factor. Even with a 448 image resolution, Qwen-VL-Chat-7B is 16% lower than LLaVA1.5-7B with 336 image inputs, largely due to the usage of Q-former for token compression. In contrast, InternVL-V1-5-20B merges four adjacent tokens, while MiniCPM-LLaMA3-V-2.5 resamples each sub-image individually, both effectively reducing redundant tokens while maximizing performance retention. The same tendency can be observed in general perception and driving suggestions tasks. Therefore, in CODA-VLM, we adopt AnyRes with a 2×2 MaxPool to achieve the balance between performance and efficiency.

Knowledge transfer. The knowledge embedded in LVLMs significantly influences the performance, which, on the one hand, comes from pre-trained visual encoders and LLMs, while on the other hand, also arises from high-quality visual instruction fine-tuning. As reported in Tab. 2, MiniCPM-V-2.5-8B surpasses LLaVA-NeXT-13B by 12% and 17% in general perception and driving suggestions, despite having smaller LLMs, revealing the significance of LLaMA3-8B. Moreover, we observe that GPT-4V exceeds open-sourced LVLMs by a significant margin on general perception and driving suggestions, indicating that current open-sourced LVLMs still lack the domain-specific knowledge of self-driving. Therefore, in CODA-VLM, we adopt LLaMA3-8B as our base model and conduct the domain-specific fine-tuning with driving scenes in CODA-LM.

Method	Grounding	Visualization
Shikra-7B	20.39	**22.94**$^{+2.55}$
LLaVA1.5-13B	18.41	**42.41**$^{+24.0}$
GPT-4V	12.85	**56.26**$^{+43.41}$

Table 5: **Ablation on visual prompts for regional perception.** Visualization with red rectangle boxes shows consistent improvements among all evaluated models.

Model	Training Time	General Perception	Driving Suggestion
LLaVA-1.5	-	15.84	29.24
+ Drive SFT Data	1.5h	53.35	60.83
+ CLIP LoRA	1.6h	53.65	61.17
+ AnyRes	6h	**57.46**	**61.83**
+ 2×2 MaxPool	3h	56.04	61.42

Table 6: **Ablation on our CODA-VLM components.** Training time (hours) is estimated with 8× A800 GPUs.

5.3 Ablation Study

5.3.1 Human Consistency of Judges

Following LMSYS (Zheng et al., 2023), we adopt the ranking-based manner to calculate the consistency of the GPT-4 and GPT-4V judges with human judgments. We randomly sample 50 samples from the CODA-LM Test set, and for each sample, we further sample two model responses from Tab. 2, followed by random shuffling. We then ask judges to determine the ranking (with ties) of the two candidate responses and human consistency is calculated as the probability of the GPT judge agreeing on the ranking with human judgments.

As reported in Tab. 3, the text-only GPT-4 judge with the reference answers achieves more than 80% consistency for all three tasks, surpassing the GPT-4V variants by a large margin. The GPT-4V judge suffers when only images are provided as the reference, which is relieved when reference answers are provided, but still inferior to the text-only GPT-4 judge, even with a higher expense.

5.3.2 Hierarchical Data Structure for General Perception

We ablate the necessity of using the "structured-coherence" pipeline in Tab. 4. Following Sec. 5.3.1, we evaluate the quality of pre-annotations by using them as the reference for the GPT-4 judge and then calculate the consistency with human judgments. We compare with 1) plain text prompting and 2) structured prompting followed by concatenating annotations of each category to consecutive texts. As shown in Tab. 4, generating structured responses followed by coherence obtains the best consistency.

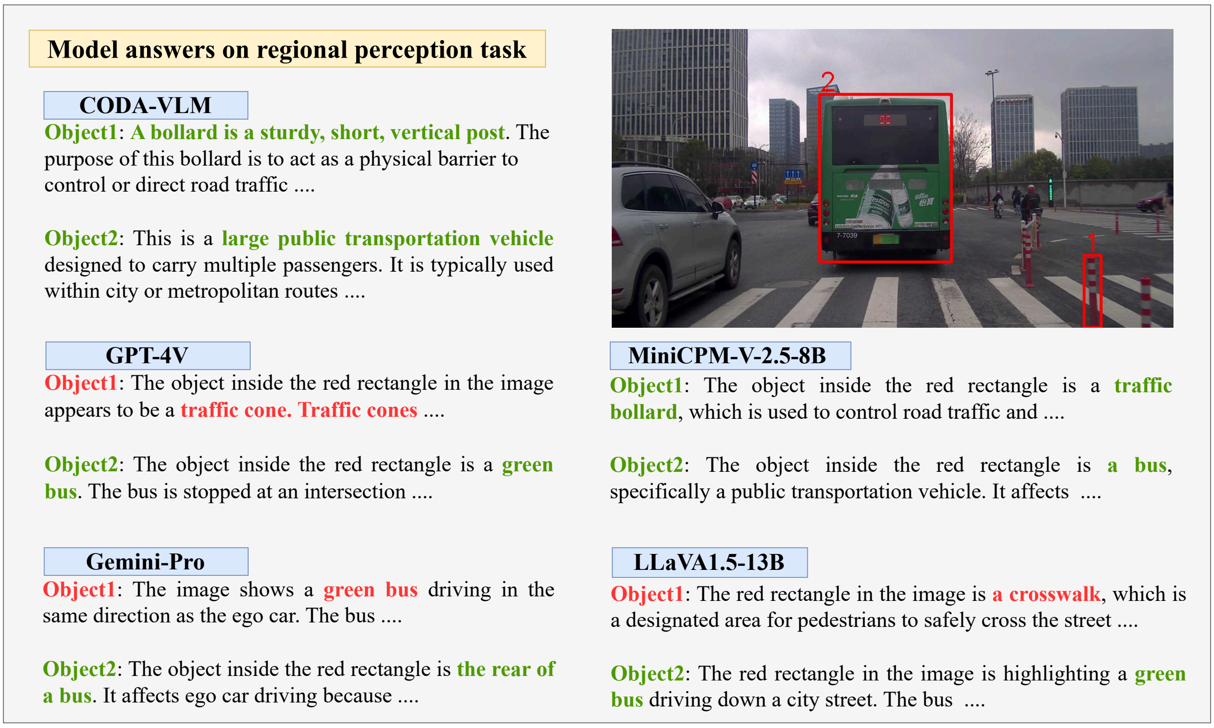


Figure 4: **Qualitative comparison among different LVLMs on the regional perception task.** Mistakes within the model responce are highlighted in **red**, whereas the accurate parts are emphasized in **green**.

5.3.3 Visual Prompts for Regional Perception

We ablate the advantage of using visualization over grounding as visual prompts for regional perception. The prompt for visualization is "*Please describe the object inside the red rectangle in the image and explain why it affects ego car driving*", while the prompt for grounding is "*Please provide a description for this object and explain why this object affects ego car driving: [x1, y1, x2, y2]*". As reported in Tab. 5, visualization demonstrates consistent improvement for all evaluated LVLMs, even for Shikra-7B which has been pre-trained with grounding data specifically.

5.3.4 CODA-VLM Components

We ablate the usage of different components of CODA-VLM on a 200-image subset of the CODA-LM Test set. Starting from a pre-trained LLaMA3-8B-based LLaVA1.5 checkpoint, we ablate the usage of 1) domain-specific fine-tuning, 2) training CLIP encoder with LoRA, 3) adopting AnyRes and 4) conducting 2×2 MaxPool step by step. As shown in Tab. 6, our CODA-VLM achieves a better trade-off among efficiency and performance.

6 Conclusion

In this paper, we propose CODA-LM, a novel real-world multimodality road corner case dataset for autonomous driving with a hierarchy task framework, spanning from general and regional perception to driving suggestions, to support automated evaluation of Large Vision-language Models (LVLMs) on self-driving corner cases. We conduct a comprehensive evaluation of representative LVLMs on road corner cases and propose CODA-VLM, a novel driving LVLM specialized in driving perception and suggestions. However, we are still far from a fully intelligent driving agent and we hope our CODA-LM can serve as the catalyst to promote the development of reliable and interpretable autonomous driving systems.

Limitations

CODA-LM is built on corner cases from CODA, which might not cover all possible unexpected conditions in driving scenarios, and we opt to explore controllable generation (Chen et al., 2023d; Gao et al., 2023, 2024; Li et al., 2023b; Liu et al., 2023d; Wang et al., 2024) to generate corner cases in the future. CODA-LM focuses on interpretable self-driving, and we will explore collecting action-level annotations. The current data collection pipeline relies on human verification and revision to ensure the quality of annotations, and an automatic data calibration method is also appealing. How to better incorporate visual pre-trained prior (*e.g.*, self-supervised learning (Chen et al., 2021, 2023b; Liu et al., 2022; Zhili et al., 2023)) is also open.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Jasmin Breitenstein, Jan-Aike Termöhlen, Daniel Lipinski, and Tim Fingscheidt. 2021. Corner cases for visual perception in automated driving: Some guidance on detection approaches. *arXiv preprint arXiv:2102.05897*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. 2021. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*.

Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. 2023b. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*.

Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. 2023c. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*.

Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. 2023d. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023e. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. 2024. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv preprint arXiv:2403.09572*.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. 2021. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2019. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. 2023. Planning-oriented autonomous driving. In *CVPR*.

Yuanfeng Ji, Chongjian Ge, Weikai Kong, Enze Xie, Zhengying Liu, Zhengguo Li, and Ping Luo. 2023. Large language models as automated aligners for benchmarking vision-language models. *arXiv preprint arXiv:2311.14580*.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *ECCV*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*.

Pengxiang Li, Zhili Liu, Kai Chen, Lanqing Hong, Yunzhi Zhuge, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. 2023b. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arxiv preprint arxiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. 2023d. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *arXiv preprint arXiv:2310.05873*.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. 2024b. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*.

Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. 2022. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. Drama: Joint risk localization and captioning in driving. In *WACV*.

OpenAI. 2023a. ChatGPT.

OpenAI. 2023b. ChatGPT-4V System Card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2023. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

MiniCPM-V Team. 2024. MiniCPM-V.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, and Kai Zhang. 2024. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *arXiv preprint arXiv:2403.13304*.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. 2023. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeruIPS*.

LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. 2023. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*.

System Prompt

You are an autonomous driving expert, specializing in recognizing traffic scenes and making driving decisions.

General Perception Prompt

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.
Here are some rules to follow:
1. 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).
2. Road users should include a description(appearance, position, direction, etc) of these objects and the reasons that affect the driving behavior of ego car.
3. Each road user should be described once to maintain clarity and avoid repetition and ensure each description is unique and specific to the object.
4. To give a positive and accurate answer, please output dictionary format and the following is sample answer, xxx means placeholder:
{"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}, {}]
}
5.If there is no road user of this class, the output should be {"vulnerable_road_users":[]}

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

System Prompt

You are an autonomous driving expert, specializing in recognizing traffic scenes and making driving decisions.

Regional Perception Prompt

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.
There are a few rules to follow :
1.To give a positive and accurate answer, please output dictionary format and the following is sample answer, xxx means placeholder:
{
"serial number": {
"description and explanation": ""
},
"2": {
"description and explanation": ""
}
}
2. 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.
3. 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.
4. 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.

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 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 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), 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 the overall scores are still low, it successfully indicates the trend among different models, with InternLM2-vl still leading

General perception prompt

You are an impartial judge tasked with evaluating the quality of predicted text provided by autonomous driving AI assistant.You will compare this prediction to a reference text, focusing on the description of objects that influence the driving behavior of ego car, and the explanation of why these objects impact. Your evaluation criteria should include accuracy(checking if the predicted text correctly identifies objects mentioned the reference text), suppression hallucination(ensuring that objects not mentioned in the reference text are not erroneously included in the predicted text), correlation(sessing if the reasons for the objects' impact on the ego car's driving behavior are consistent between the reference and predicted text). Identify and correct any mistakes. Be as objective as possible. Do not allow the length of the predicted text to influence your evaluation. After providing your short explanation, you must rate the response on a scale from 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[10]]".

Regional perception prompt

You are an impartial judge tasked with evaluating text similarity and relevance of the reference text and autonomous driving AI assistant's predicted text. Be as objective as possible. Do not allow the length of the predicted text to influence your evaluation. After providing your short explanation, you must rate on a scale from 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[10]]".

Driving suggestions prompt

You are an impartial judge tasked with evaluating the quality of predicted text provided by autonomous driving AI assistant. You will compare this prediction to a reference text, focusing on the ego car driving suggestion. Your evaluation should consider rationality, relevance, level of detail of the response. Predicted text should be specific and actionable, rather than vague or overly broad. Identify and correct any mistakes. Do not allow the length of the predicted text to influence your evaluation. After providing your short explanation, you must rate the response on a scale from 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[10]]".

Evaluation Framework


```
messages = [ {"role":"system", "content": {Task Prompt}]
for sample in fewshot_samples:
        messages.append({"role":"user", "content":sample['text']})
        messages.append({"role":"assistant","content":sample['response']})
messages.append({"role":"user", "content": {Query}})
```

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.

among open-source models. By default, we still adopt the Text-Score as the primary evaluation metric, unless otherwise specified.

C More Discussion

Potential risks and ethical considerations are not a problem for us, since we only use the open-sourced resources for academic usage. For all the open-sourced datasets and models, we follow the intended usage under the official license. We do not use data containing personally identifying information or offensive content.

Human annotators. We recruit human annotators from university graduate students with advanced English reading and writing capabilities. Annotators are mainly coming from Asia, and we pay them the standard part-time hourly wage.

Before annotation, we clearly explain the annotation tasks (check details in Sec. 3 and 4.1) and the annotation GUI, as in Fig. 9 and inform them that their annotations will only be utilized to construct this dataset for academic usage to obtain their agreements.

Driving Suggestions

High score example

Reference Text

Maintain the current lane and speed, carefully monitor the motorcyclist to the right, and prepare to react if the motorcyclist changes lanes or stops suddenly. Stay alert for the vehicles in the left lane that may enter your lane or affect the flow of traffic.

Prediction Text

The ego car should maintain its current speed and continue straight, while remaining vigilant of the bicycle rider, ensuring a safe distance to prevent accidents. Additionally, the car must be ready to adapt to sudden changes in traffic, such as other vehicles switching lanes.

Response

The predictive text only mistakenly substitutes "motorcyclist" for "bicycle rider". It correctly suggests maintaining speed and being watchful for lane changes and reflects an understanding of safe driving. So, Rating: [[8]].

Low score example

Reference Text

The ego car should maintain a safe distance from the white SUV ahead, and prepare to navigate around the roadwork or damaged road surface. It should also observe for any instructions from the traffic lights once they become visible.

Prediction Text

The ego car should maintain a safe distance from the other cars and the truck to avoid collisions. It should be cautious of pedestrians and cyclists, as they might be crossing the street or sharing the road and follow the speed limits and any posted road signs to ensure a safe driving experience for all road users.

Response

The prediction text deviates from the reference text by introducing irrelevant elements like pedestrians, cyclists, road signs, and a truck, while omitting key details like roadwork and traffic light. It always describes general safe driving practices, lacking the specific analysis required for the given scenario. So, Rating: [[2]].

Figure 8: **Few-shot examples for Driving Suggestions**

D 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.

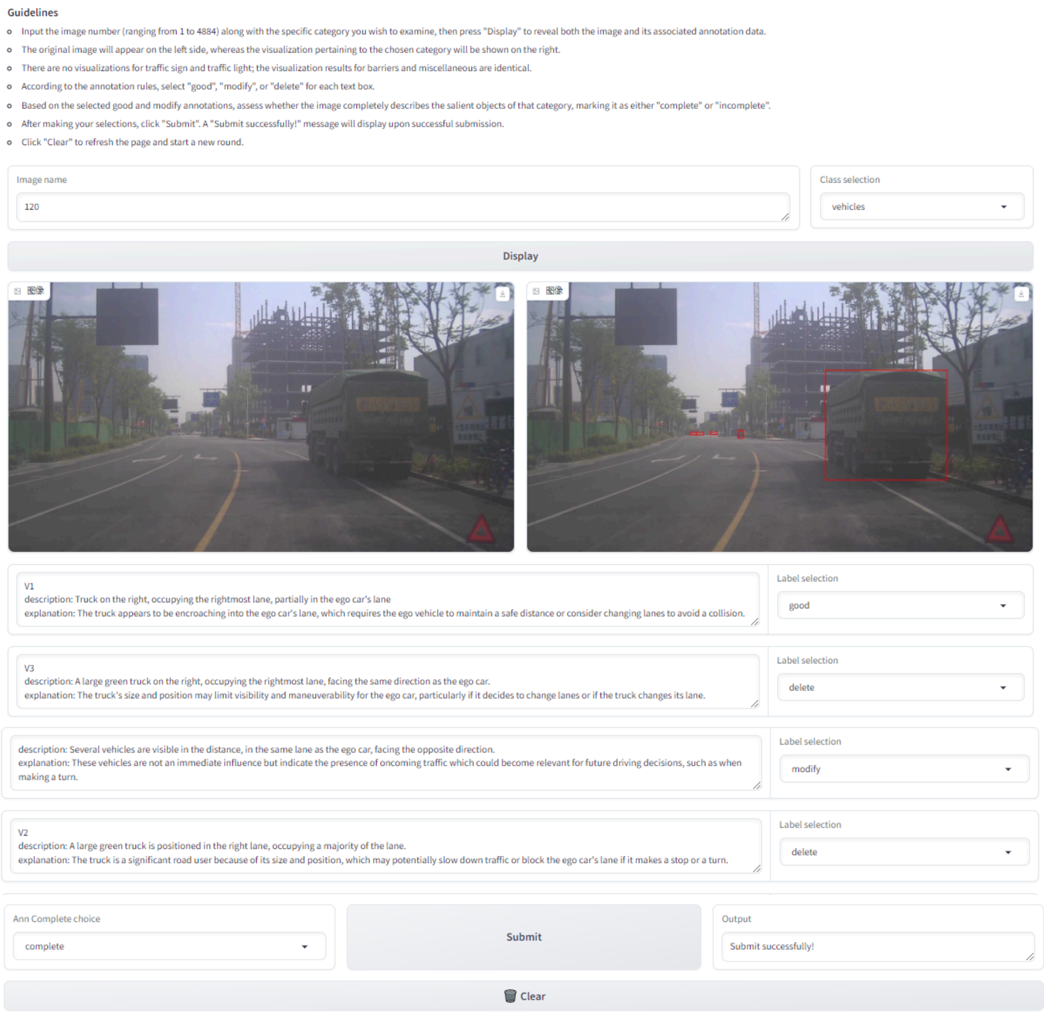


Figure 9: **Our Gradio labeling tool graphics user interface** for general perception.

General perception high score example

Reference Text

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 yelow 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.

Prediction Text

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 sliver car in front of th 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.

Response

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]]."

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

General perception low score example

Reference Text

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.

Prediction Text

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.

Response

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]].

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

Source	Model	Metrics ↑			
		BLEU4	**METEOR**	**CIDEr**	**SPICE**
Open	MiniGPT-v2-7B (Chen et al., 2023a)	0.6	5.3	0.6	4.4
	Shikra-7B (Chen et al., 2023e)	1.5	8.7	0.0	5.2
	LLaVA1.5-7B (Liu et al., 2023a)	1.9	13.9	0.9	9.8
	LLaVA1.5-13B (Liu et al., 2023a)	**2.7**	16.0	1.1	13.9
Commercial	Gemini Pro (Team et al., 2023)	1.9	12.9	**4.8**	16.0
	GPT-4V (OpenAI, 2023b)	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.

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.**

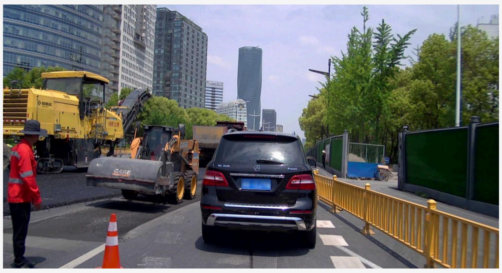
Task1: General Perception

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 shap ...

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

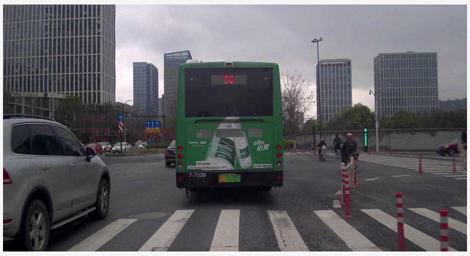
The image captures the ego car's perspective on a road with various elements that influence its driving behavior. **The red circular sign** with a white border and a red diagonal line is a **"No Entry" sign**, which directly affects the ego car's behavior by indicating that ....

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 ....

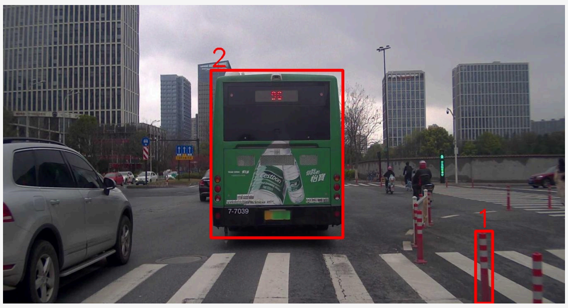
LLaVA1.5-13B

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 ....

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 ....

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 ....

LLaVA1.5-13B

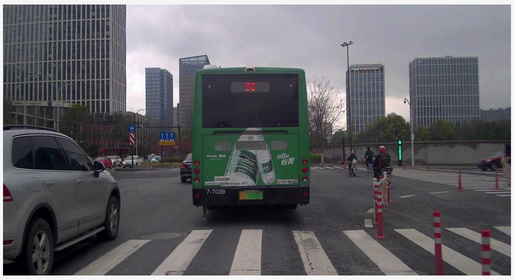
The car is surrounded by other vehicles, including a truck and a few pedestrians. Based on the scene, the ego car should **slow dow** .... The driver should **be cautious of the pedestrians** ....

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** ....

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** ....

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** ....

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.**