From Regional to General: A Vision-Language Model-Based Framework for Corner Cases Comprehension in Autonomous Driving

Xu HAN[†] xhanab@connect.ust.hk
Yehua HUANG[†] {yhuang704}
Song TANG {stang428}@connect.hkust-gz.edu.cn
Xiaowen CHU xwchu@ust.hk
Data Science and Analytics Thrust, Information Hub
The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

Abstract

Large Vision-Language Models (LVLMs) have demonstrated their excellent capabilities in handling multi-modal tasks. However, in the filed of autonomous driving, they still face challenges of handling corner cases in traffic scenes, which often involve complex relationships among the road users and objects. To strengthen LVLMs' abilities in understanding traffic scenes, we propose a prompting-based progressive framework that boosts LVLMs' comprehension of the corner cases. Inspired by the thinking modes of human beings, our framework guides the LVLM to analyze regional factors in the scene first and then comprehend the general situation based on the regional understandings. A significance assessment mechanism is introduced in between to determine the scope of the objects that should be considered by the LVLM. The proposed method significantly outperformed the baselines on the CODA-LM dataset. Our code is available at https://github.com/hyhping2023/ECCV_FNN_Code.

1. Introduction

Large Vision-Language Models (LVLMs) have garnered increasing attention across various fields, including Autonomous Driving [1, 10]. Despite their potential, LVLMs still face significant challenges in handling corner cases in traffic scenarios [3]. These corner cases often involve complex interactions among road users and objects. LVLMs sometimes struggle to comprehend the relationship between obstacles and the ego vehicle and often fail to fully grasp critical details such as the positions and features of the entities on the road. This understanding is crucial for accurately interpreting the overall situation in complex scenarios [8].

[†]Contributed equally to this work

Corner cases typically involve a diverse array of objects, where the general situation is influenced by the key objects within the scene. Humans tend to analyze these scenes by first examining the features of key objects and then discerning the relationships among them to develop an understanding of the overall situation. Inspired by this human approach to analyzing traffic scenes, we propose a progressive framework that prompts LVLMs to comprehend corner cases in autonomous driving in a human-like manner.

In our framework, we begin by guiding LVLMs to analyze the key features of objects in the traffic scene. We then prompt the LVLMs to integrate information from the most significant objects to form a more accurate and comprehensive understanding of the overall situation. To facilitate this process, we have designed a Significance Assessment module that determines the scope of objects that should be considered by the LVLM. After this analysis, the LVLM is prompted to provide driving suggestions for handling the corner case, which is critical for autonomous driving.

Our experiments on the CODA-LM [3] dataset indicate that the proposed framework significantly outperforms the baseline methods on all evaluated metrics, demonstrating its efficacy and reliability of comprehending corner cases.

1.1. CODA-LM Tasks

The CODA-LM dataset, which including over 10,000 images with the corresponding textual descriptions covering global driving scenarios, detailed analyses of corner cases, and future driving recommendation, is adopted in this study for comprehensive evaluation of the LVLMs in corner cases. In the CODA-LM benchmark, LVLMs are required to work on three distinct tasks, namely *general perception*, *region perception* and *driving suggestions*, and their performance will be judged by Large Language Models (LLMs).

In the *general perception* task, LVLMs are required to describe all the potential road obstacles in the traffic scenes

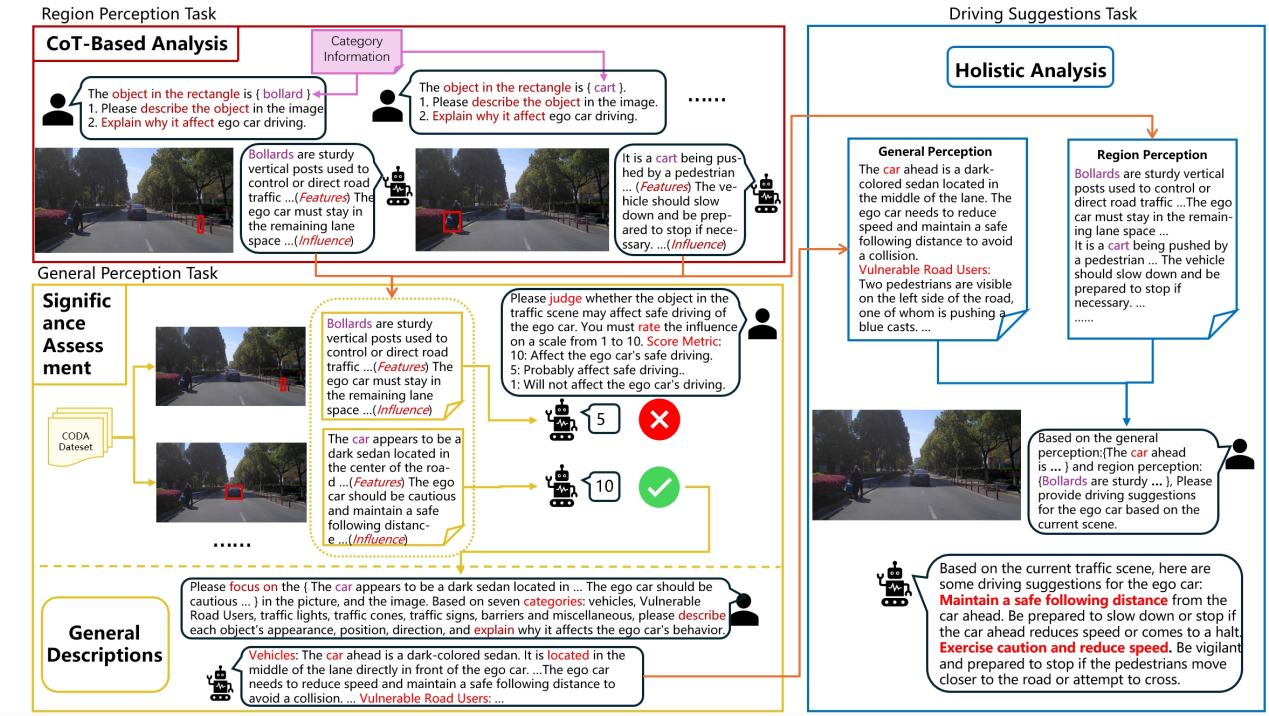

Figure 1. **Overview of the progressive framework.** It starts by requiring the LVLM to analyze regional entities separately. Then we guide LVLMs to score the significance of the entities and select more important objects, and their information is provided for constructing general descriptions. Finally, based on these general and regional perception, we prompt LVLMs to provide driving suggestions.

and provide explanations about why they would affect the driving decisions. The *region perception* task measures the LVLMs' capability to understand corner objects when provided with specific bounding boxes. The *driving suggestions* task aims to evaluate LVLMs' capability in formulating driving advice in the autonomous driving domain..

2. Methods

In the proposed progressive framework, we start by guiding LVLMs to analyse the annotated objects in the image separately, where the category information of the objects were acquired from the CODA dataset[2]. We prompt the LVLM to describe core features of the object in bounding box and consider its influence to the ego vehicle based on its descriptions. To better perceive the whole image, we expand the range of analyzed entities by identifying all the annotated traffic participants in the image from the CODA dataset. After receiving the analysis of all annotated objects, we instruct the LVLM to assess the significance of the objects in the current scene by judging their influence on ego vehicle. A score of 1 to 10 is rated by the LVLM and the information of high influencing objects is reserved for the LVLM to generate general perception by only focusing on these pieces of information. Finally, the LVLM is prompted to provide driving suggestions based on the integration of the general perception descriptions and the region perception descriptions obtained in former steps.

2.1. Region Perception

Chain-of-Thought (CoT)-Based Analysis[9]. In this method, the LVLM is required to categorize the targeted object at first. In practice, the category of the object can be obtained from the CODA dataset or applying object detection methods, such as YOLO[1]. With the category information, the LVLM is prompted to figure out the features of the object in the bounding box, e.g., appearance and position. Grounded in the descriptions, the LVLM is then instructed to analyze the potential influence on the ego vehicle.

2.2. General Perception

Object Level Analysis (OLA) Since corner cases consist of a variety of traffic objects, to better perceive corner cases, LVLMs should firstly recognize primary objects in the traffic scenes and identify their categories accurately. At first, we applied YOLO to detect key entities in the traffic scenes, which is out of our satisfactory due to its lower accuracy of category recognition compared with using annotations in CODA dataset. In addition, limited by the resolution of the image, YOLO fell short of recognize relatively small ob-

[1]https://docs.ultralytics.com/zh/yolov5/quickstart_tutorial/

Score ↑	Significance Level
10	The object will definitely affect the ego vehicle's safe driving, the situation is urgent to be taken into account and there is a large probability that the accident will happened.
5	The object will probably affect the ego vehicle's safe driving but the accident can be avoided.
0	The object will not affect the ego vehicle's driving.

Table 1. **The score metric of the Significance Assessment(SA) module.** We provided three metrics for the consideration of brevity.

jects in the image. With all things considered, we finally decided to recognize the objects in the image based on the annotations in the CODA dataset. With the recognition and categorization task finished, to analyze the influence of the objects, we apply CoT-Based Analysis method and acquire corresponding descriptions about features and influence of an expanded range of entities.

Significance Assessment (SA) Traffic scenes always consists of plenty of objects but only part of them may truly affect the ego car driving. LVLMs may fall into confusion of selecting and describing entities without detailed guidance. For example, with all the objects information generated in the OLA module, they tend to fall into hallucination[5] and contains much useless information in their answers. To filter out unimportant entities and reserve core information, we introduce the Significance Assessment module to measure the urgency and the significance of the object based on our metric in Table 1. Grounded in the significance scores given by the LVLM judge, we select high importance objects and prompt their information about features, categories and influence on the ego car to the LVLM, which helps the LVLM grasp the key entities in the image and contains more details of the important entities. With these actions, the LVLM perceives the general situations with focusing on the key information and provides a more accurate and comprehensive description on general situation.

Category Level Analysis (CLA) Following the chain-of-thought method[9], we tried to split the task of analyzing all objects at once into smaller tasks of analyzing only one kind of objects at a time and integrating them later. However, in our practice, this method performs out of our satisfactory because LVLMs can't describe the objects as detailed and accurate as OLA does.

2.3. Driving Suggestions

Holistic Analysis (HA) In most cases, giving driving suggestion contains a hidden step, which is analyzing the general situation in the traffic scene in advance. To provide more accurate and targeted suggestions, we humans tend to understand the general situation in advance and advise based on the general situation. Consequently, we combine the general perception information with region perception information generated in advance to construct a comprehensive general situation description. After that, we instruct the LVLMs to provide suggestions referring to the corresponding general situation description.

Local Analysis (LA) Typically, driving suggestions contains corresponding driving suggestions for primary entities in the traffic scene. So we tried to instruct the LVLMs to provide driving suggestions for each key entities in the image and we also tried to apply SA method to reserve more important entities. However, both of them perform poorly compared with HA method due to hallucination of the LVLM during the integration period.

Retrieval Augmented Generation (RAG) In our practical application, RAG was also taken into consideration to reduce hallucination[7]. However, due to the lack of diversity and high variance in traffic scenarios, the retrieval differs significantly with the targeted answer, which leads to the deprecation of the RAG method.

3. Experimental Setup and Results

3.1. Experimental Setup

In our practice, we set the temperature to 0.2 for the large models based on the consideration that more deterministic and less varied responses are desired in these tasks. We follow the official evaluation protocols of CODA-LM [3] and calculate a GPT-Score for General Perception, Region Perception, and Driving Suggestion separately. GPT-4o-2024-05-13[2] is adopted as judge and the Final-Score is averaged.

3.2. Main Results

Model	General Perception Score ↑	Region Perception Score ↑	Driving Suggestions Score ↑	Final Score ↑
LLaVA-1.5-7B[4]	19.30	42.06	23.16	28.17
CODA-VLM[3]	55.04	77.68	58.14	63.62
GPT-4V[6]	57.50	56.26	63.30	59.02
Ours	**59.00**	**84.37**	**70.80**	**71.39**

Table 2. **The comparison of the evaluation scores among our framework and other methods on CODA-LM test set**. The CoT-Based method, Object Level Analysis, Significance Assessment module, and Holistic Analysis are applied in our framework.

As Table 2 shows, the proposed framework surpasses the baseline methods from CODA-LM in all evaluated tasks, indicating that our framework is capable of comprehending corner cases effectively and giving reasonable driving suggestions in various kinds of traffic situations.

[2]https://openai.com/index/hello-gpt-4o/

Model	Categories							All ↑
	Vehicles ↑	Traffic Lights ↑	Traffic Signs ↑	Traffic Cones ↑	Vulnerable Road Users ↑	Barriers ↑	Miscellaneous ↑	
LlaVA-v1.5-7B[4]	68.95	67.50	50.00	82.17	75.00	77.06	70.95	72.52
LlaVA-v1.5-13B[4]	71.93	75.00	51.05	79.28	80.00	73.53	71.43	72.08
LlaVA-v1.6-34B[4]	76.67	80.00	56.84	83.19	83.75	82.47	75.71	77.85
Claude-3.5-Sonnet[3]	67.72	80.00	55.26	83.19	81.67	81.76	76.67	75.64
GPT-4o-mini[4]	**80.70**	77.50	**60.26**	83.04	82.50	83.41	76.67	79.23
GPT-4o-2024-05-13	75.79	**87.50**	56.58	**84.93**	**87.50**	**84.82**	**78.57**	**79.33**

Table 3. **Test scores of different models in analyzing objects of different categories.** GPT-4o performs best in most of the categories.

3.3. Region Perception

In the region perception task, there are 7 categories of objects (Vehicles, Traffic Lights, Traffic Signs, Traffic Cones, Vulnerable Road Users, Barriers, Miscellaneous) to be analyzed. We have tested a range of models, including both commercial and open-sourced, on validation set for this task. As Table 3 shows, GPT-4o performs best in most categories. Consequently, we select GPT-4o-2024-05-13 as the base model for all three tasks.

3.4. General Perception

We designed a Significance Assessment (SA) module in the general perception task to select the objects that may affect the normal driving of the ego vehicle. The distributions of SA score in the validation set and the test set are similar, as shown in Figure 2, which implies the consistency and reliability of our proposed module. Additionally, to select an appropriate threshold for the SA method, we tested a threshold from 3 to 8, as shown in Figure 3. According to the results, a threshold of either 6 or 8 performs the best, and we chose 8.

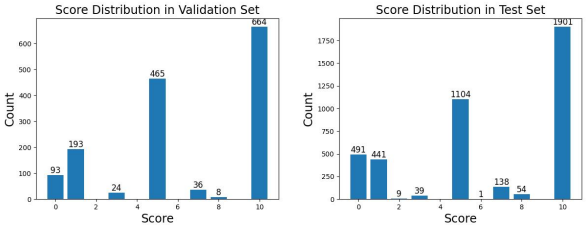

Figure 2. The distributions of SA score in the validation set and the test set are similar.

Besides the SA module, several methods, such as Object Level Analysis (OLA) and Category Level Analysis (CLA), have been explored. The results on the validation set are shown in Table 4 and indicate that combining OLA with SA outperforms all other methods.

3.5. Driving Suggestions

For the driving suggestions task, several methods, including Holistic Analysis (HA), Local Analysis (LA), and Retrieval Augmented Generation (RAG), have been tried. The results in Table 5 show that the HA approach performs the best.

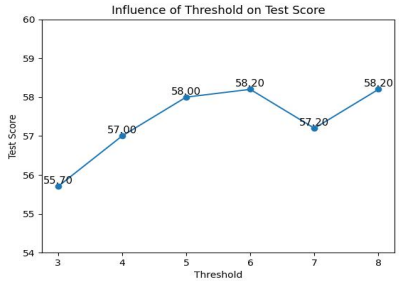

Figure 3. **Influence of setting different thresholds for SA.** A threshold of 6 or 8 performs best and we choose 8 finally.

Method	GPT-4o	OLA	CLA	OLA + SA
General Perception Score ↑	40.30	54.30	51.30	**58.20**

Table 4. **General perception scores with different methods.** The method combing Object Level Analysis with Significance Analysis surpasses all other methods.

Method	GPT-4o	RAG	LA	HA
Test Score ↑	67.27	61.20	58.20	**67.80**

Table 5. **Driving suggestions score with different methods.** The Holistic Analysis method performs the best.

4. Conclusion

We propose a progressive framework that guide LVLMs to comprehend general situations of the corner cases based on their regional understanding. In the framework, we introduce CoT-Based Analysis for better regional comprehension and propose Object Level Analysis and Significance Assessment modules to prompt LVLMs integrating regional information more accurately and perceiving traffic scenes more comprehensively. Additionally, we conduct the Holistic Analysis for LVLMs to provide driving suggestions based on former analysis. Our framework enables LVLMs to analyze traffic scenes in a human way, which improves the performance of LVLMS in handling corner cases.

References

[1] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 1

[2] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 2

[3] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 1, 3

[4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 4

[5] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rong-Zhi Li, and Wei Peng. A survey on hallucination in large vision-language models. *ArXiv*, abs/2402.00253, 2024. 3

[6] OpenAI. Gpt-4v(ision) system card. 2023. 3

[7] Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation. 2024. 3

[8] Haowei Sun, Shuo Feng, Xintao Yan, and Henry X Liu. Corner case generation and analysis for safety assessment of autonomous vehicles. *Transportation research record*, 2675 (11):587–600, 2021. 1

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 2, 3

[10] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1