Two-Stage LVLM system: 1st Place Solution for ECCV 2024 Corner Case Scene Understanding Challenge

Ying Xue[1,2] Haiming Zhang[1,2]
Yiyao Zhu[3] Wending Zhou[1,2] Shuguang Cui[2,1] Zhen Li[2,1]
[1] FNii-Shenzhen
[2] School of Seience and Engineering, CUHK-Shenzhen
[3] Hong Kong University of Science and Technology

Abstract

This technical report outlines the methods we employed for Track 1 Corner Case Scene Understanding of ECCV 2024 Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving Challenge. The challenge is to generate general perception, region perception description, and driving suggestions for the corner case driving scene. We propose a two-stage method consisting of preliminary output and refinement. We first fine-tune the LLaVA-Next model with LoRA to get a coarse output, then utilize GPT-4 to refine the result. This system combines the learning ability of LLaVA-Next and the strong reasoning ability of GPT-4. As a result, our system achieved the top 1 score of 72.12 on the final leaderboard. The code and checkpoints are released on https://github.com/Chloe-gra/ECCV2024_Challenge_llmforad_solution

1. Introduction

In the field of autonomous driving, corner cases refer to situations that occur infrequently in the dataset but can significantly impact the system's performance. For example, sudden changes in road conditions and rare traffic signs or signal patterns are representative corner cases. They are characterized by high rarity, high complexity, and high risk. Tackling corner cases is a challenging task in autonomous driving.

Multimodal Large Language Models (MLLMs) or Large Vision Language Models (LVLMs) have developed rapidly in recent years [6]. Their application to autonomous driving areas, including perception, prediction, and planning, has helped improve the performance of autonomous driving tasks [7, 8].

The Corner Case Scene Understanding Track is aimed at improving global scene understanding, region reasoning, and actionable navigation for autonomous vehicles based on the CODA-LM dataset. This track calls for participants to develop a vision language model to address the corner case autonomous driving tasks including general perception, region perception, and driving suggestions.

Targeting this corner case dataset, we proposed a solution that aims to enhance the perception and reasoning performance of MLLM. We design a two-stage model, which first fine-tunes the LLaVA-Next model to generate a coarse result, and then utilizes GPT-4 to refine the preliminary output with designed prompts to get final results.

2. Dataset

CODA-LM [4] is constructed from the CODA dataset [3] and consists of around 10K images with the corresponding textual descriptions of general perception, regional perception, and driving suggestions. Different from other autonomous driving datasets. CODA-LM focuses on corner case analysis. Specifically, the train set contains 4884 samples sourced from the CODA2022 validation set. Its validation set has 4384 samples sourced from the CODA2022 test set optionally for training. The test dataset has 500 samples drawn from the CODA2022 test set. Each sample features a front-view, single-frame image, providing a comprehensive basis for analyzing and addressing corner cases in autonomous driving. There are three tasks to address:

General Perception. This task measures the comprehensive understanding ability of the key entities in a driving scene, including their appearance, location, and the reasons why they affect driving. It focuses on seven types of road users, including vehicles, vulnerable road users (VRUs), traffic cones, traffic signals, traffic signs, barriers, and miscellaneous.

Region Perception. Given a specific bounding box, the MLLM should describe the specific object in the bounding box and explain why it affects self-driving behavior. The participants can access to ground truth category names by comparing the annotations of CODA-LM and CODA

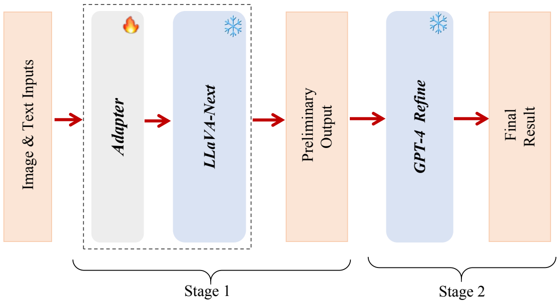


Figure 1. The overall pipeline of our method.

datasets, which are allowed to be used.

Driving Suggestions. This task is to make driving recommendations in the area of autonomous driving. It evaluates the MLLM's ability to provide optimal driving recommendations based on the general and regional perception.

3. Method

The pipeline of our method is illustrated in Fig. 1, consisting of two stages.

The stage 1 takes the original images along with three types of questions as inputs, to obtain the preliminary descriptions that offer coarse but comprehensive details.

Afterward, the stage 2 is utilized to refine the preliminary results to achieve satisfying results.

In the subsequent sections, we will first illustrate the foundation model we used, and then elaborate on the two stages in detail.

3.1. Foundation Model.

LLaVA-NeXT [5] is an LVLM with impressive capabilities in multimodal understanding and reasoning. As an advanced version of LLaVA (Large Language and Vision Assistant), it has stronger abilities including multimodal understanding, OCR, and an expanded understanding of world knowledge. Therefore, we utilize LLaVA-NeXT-7B as the foundation model for further fine-tuning.

GPT-4 [1] is a stronger LVLM that can handle more complex language tasks with greater accuracy, coherence, and creativity. We chose GPT-4o for our model, which has a faster reasoning speed.

To maximize the performance of our model, we combine the two strong LVLMs as our foundation model. The LLaVA-NeXT model is responsible for learning the novel dataset. The GPT-4 model, with a higher level of common sense and reasoning ability, is utilized to refine the preliminary results learned by LLaVA-NeXT.

3.2. Stage 1: Preliminary Output

Stage 1 aims to obtain preliminary results for the three tasks for the subsequent usage. During the training phase, we finetune the LLaVA-NeXT-7B model on the CODA-LM dataset. As LLaVA-NeXT-7B has 7 billion trainable parameters, this will considerably impact memory usage to finetune it directly. To address this problem, we use LoRA [2] an efficient parameter fine-tuning method, to finetune the LLaVA-NeXT model. It allows the freeze of the existing weights and only trains a couple of adapter layers on top of the base model. We add adapters to all linear layers of the model, except for the ones present in the vision encoder and multimodal projector. The prompt is nearly the same as the given question for each task in the dataset. We fine-tuned three models with different percentages of the three tasks' data samples.

When inference, we first utilize GPT-4 to choose the most suitable model among the three fine-tuned models for each task. Then we ensemble the selected models to get the preliminary output of Stage 1.

We tend to select the model that learns the targeting task fully and is less affected by the other tasks that output irrelevant content. For the general perception task, GPT-4 should choose the answer that can generally and accurately describe the objects in the image. For the region perception task, GPT-4 is required to choose a concise answer that only describes the object in the red rectangle. For the driving suggestion task, it should select an answer without unnecessary description but focus on the suggestions.

Given the test samples, the three candidate models first output a small part of test data for all three tasks to select the best model cost-efficiently. For each task, we use GPT-4 to choose the most suitable answer for each sample according to the requirements above. The most frequently selected model will be the one to output preliminary results for each task.

3.3. Stage 2: Refinement

Stage 2 is designed to refine the results from Stage 1 and solely performs inference. We use GPT-4 as LVLM in this stage. The stronger reasoning ability of GPT-4 can help to modify the inaccurate content of the preliminary output from Stage 1. For each task, GPT-4 processes the draft from Stage 1 referring to the image to obtain the final output. Next, we will introduce how we design prompts targeting each task.

For the general perception task, we note that the preliminary output of LLaVA-Next has some problems like hallucination and erroneous description, especially for distant and small objects. Therefore, we choose to delete the inaccurate content detected by GPT-4 to alleviate hallucination and inconsistency. Besides, by experiments, we discovered that adding a comment on the level at which each road user af-

Method	General Perception	Region Perception	Driving Suggestion	**Final Score**
LLaVA-Next baseline	25.88	52.06	30.90	36.28
Preliminary Output (Stage 1)	52.84	78.33	61.62	64.26
Final Output (Stage 1 + Stage 2)	**58.70**	**83.41**	**74.26**	**72.12**

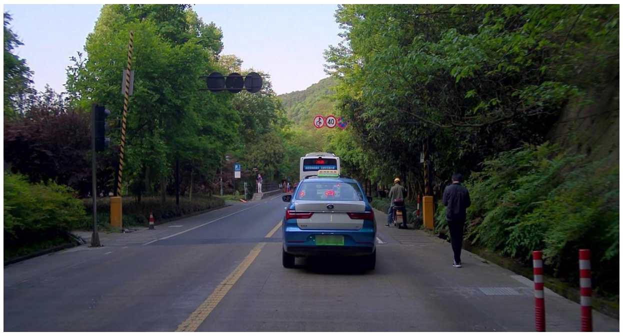
Table 1. The results on test set.

General Perception

Answer: In the traffic image, there is a blue car directly in front of the ego car. This vehicle is significant as it directly impacts the ego car's ability to proceed. ……On the right side of the road, there is a pedestrian walking away from the ego car, wearing dark clothing. The pedestrian's presence requires the ego car to monitor them in case they decide to cross the road or enter the driving lane.\n\nAhead, there are multiple traffic signs indicating a maximum speed limit of 40 km/h. ……n\nAdditionally, red and white traffic cones are placed on the right side of the road, ……\n\nImpact on the ego car:\n\n- **Blue car ahead**: Immediate and significant impact, as the ego car must maintain a safe distance and be prepared to stop if the car in front stops.\n\n- **Bus**: Moderate impact……

GT: The scene depicts a variety of road users and objects, each affecting the driving dynamics of the ego car.\n\nA blue car with a taxi sign is directly in the ego car's path, dictating the pace and necessary following distance. ……\n\nOn the sidewalk to the right, a pedestrian is walking in the same direction as traffic, necessitating monitoring in case they step onto the road.……\n\nAbove, traffic signs are in place, indicating no overtaking, and setting a speed limit of 40 km/h, requiring compliance from the ego car. However, the traffic light's state is not visible in the image, prompting the ego car to proceed with caution and be ready to react to possible changes based on the other vehicles' behavior.\n\nAdditionally, two red and white traffic cones on the right side of the road mark the driving area boundary and suggest a narrowing of the road, possibly hinting at nearby roadwork or a temporary obstacle. ……

Scene Image



Driving Suggestion

Answer: ### Revised Draft Driving Suggestion:\n\n1. **Maintain a Safe Following Distance**: The ego car should maintain a safe distance from the blue car directly ahead. ……\n\n3. **Pedestrian Awareness**: ……\n\n4. **Stay Within the Speed Limit**: Multiple traffic signs indicate a maximum speed limit of 40 km/h. …….\n\n5. **Traffic Light Caution**: The traffic light is currently not illuminated. ……\n\n6. **Traffic Cones and Lane Restrictions**: The cones on the right side of the road signify a closed lane or road work. …….

GT: Proceed with caution given the potential unpredictability of the pedestrian and person standing by the bicycle. Maintain a safe distance from the blue car ahead, and prepare to reduce speed if needed to safely navigate past the parked white vehicle and traffic cones on the right. Observe the behavior of vehicles ahead for cues about the traffic light state, and comply with the posted speed limit and no-honking rule.

Figure 2. Detailed qualitative results on CODA-LM dataset. Given the scene image, our method could generate comprehensive answers. We use the different colors to emphasize the corresponding descriptions between ours and ground truth.

fects the ego car can effectively enhance the score. For each object, GPT-4 will estimate its impact as high/moderate/low with an explanation. It emphasizes the importance of the key objects, improving the correlation when compared with reference text. Meanwhile, it lowers the significance of the trivial or irrelevant objects, thus decreasing the hallucination to a certain extent. Thus the prompt is as follows:

System prompt: ”You are an AI assistant tasked with integrating image descriptions related to autonomous driving. You will receive an image captured from autonomous driving and a draft description. Only include visible objects in the scene. Accurately describe the objects and their impact on the ego vehicle. Delete the inaccurate content. Add a comment on how much each road user affects the driving of the ego car.”

For region perception, the ground truth category of the object is known. We perceive that the region perception output from Stage 1 is relatively accurate if the category is correct since it is much easier for LLaVA-Next to learn a single object. Also, the dataset of region perception is much larger. Therefore, to cost-saving refine the result, we only modify the output with incorrect categories. We determine if the category is correct by determining if the category name appears in the description. To more efficiently refine the result, we first re-assign the draft based on the category. For the misclassified sample (The ground truth category is Category A, but the description classifies it to Category B), we randomly select another output of Category A that correctly classified from Stage 1 as the draft of the misclassified sample. If no output of Category A is correctly classified, the draft only contains the ground truth category. Then we use the following prompt to refine the preliminary output:

System prompt: ”You are an AI assistant tasked with integrating image descriptions related to autonomous driving. You will receive an image with red rectangle captured from autonomous driving and a draft description of the object in the rectangle. The task is to describe the object inside the red rectangle in the image and explain why it affects ego car. The category of the object is correct. Based on the image, modify the draft regional description. Start with ’This object is ’”

The driving suggestion task requires making driving recommendations in the area of autonomous driving. The suggestion is based on the results of perception by LVLM. We noticed that the preliminary suggestion output from Stage 1 doesn’t detect the key objects thoroughly, thus impairing the completeness and specificity of the driving suggestion. Therefore, we add the general perception results to the prompt as a reference for the modification of driving suggestions. The refined driving suggestion is more detailed and precise. Each key object has a targeting driving suggestion, which enhances the final score largely. The prompt is shown as follows.

System prompt: ”You are an AI assistant tasked with modify the driving suggestion related to autonomous driving. You will receive an image captured from autonomous driving, a general perception description and a draft driving suggestion. Accurately modify the driving suggestion based on the image and general perception.”

4. Results

Evaluation. CODA-LM utilizes GPT-score as the evaluation protocol. The evaluation criteria should include accuracy (checking if the predicted text correctly identifies objects mentioned in the reference text), suppression of hallucination(ensuring that objects not mentioned in the reference text are not erroneously included in the predicted text), correlation(assessing if the reasons for the objects’ impact on the ego car’s driving behavior are consistent between the reference and predicted text).

Result analysis. The results of the test set are shown in Tab. 1. Compared with the LLaVA-Next baseline model, our two-stage approach obtained significant improvements in all three tasks. We also showcase a detailed example of general perception and driving suggestion results in Fig. 2.

5. Conclusion

This report introduces our approach to tackling the corner case understanding task based on the CODA-LM dataset. We develop a two-stage vision language model system. By first finetuning the LLaVA-Next model to learn the CODA-LM dataset, then using the stronger common sense and reasoning ability of GPT-4 to refine the output, we achieved the best score on the leaderboard.

6. Acknowledgments

This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the National Key R&D Program of China with grant No. 2018YFB1800800, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by the NSFC 61931024&12326610, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), and the Key Area R&D Program of Guangdong Province with grant No. 2018B03033800, by Tencent&Huawei Open Fund.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[3] 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. In *European Conference on Computer Vision*, pages 406–423. Springer, 2022. 1

[4] 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

[5] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[7] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 1

[8] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024. 1