NexusAD: Exploring the Nexus for Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Mengjingcheng Mo, Jingxin Wang, Like Wang,
Haosheng Chen, Changjun Gu, Jiaxu Leng*, Xinbo Gao*
Chongqing University of Posts and Telecommunications
mo1031@live.com, 212115020@mail.sdufe.edu.cn, s230232032@stu.cqupt.edu.cn
{chenhs, gucj, lengjx, gaoxb}@cqupt.edu.cn

Abstract

*This report presents our approach for the Corner Case Scene Understanding track of the Autonomous Driving Challenge at the ECCV 2024 Workshop. The advent of multimodal large-scale language models (MLLMs) like GPT-4V has showcased remarkable multimodal perception and understanding capabilities, even in dynamic street scenes. However, applying MLLMs to address the corner cases in autonomous driving remains a largely unexplored area. Using the CODA-LM dataset, which features visual images paired with textual descriptions and analyses of corner cases, we adopted InternVL-2.0 as our base model and conducted domain-specific fine-tuning tailored to driving scenes. In this work, we enhance spatial correlation utilization within images by leveraging position and depth information to improve driving scene perception. Additionally, we incorporate chain-of-thought reasoning for greater accuracy and develop a context learning mechanism based on scene-aware retrieval, which further refines the model's understanding. This comprehensive strategy culminated in a final score of **68.97** on the leaderboard. Our code will be released at https://github.com/OpenVisualLab/NexusAD.*

1. Introduction

Large Vision-Language Models (LVLMs) have the potential to greatly enhance autonomous driving by integrating visual and linguistic information, thereby improving system performance, safety, and alignment with human intent through effective perception, prediction, and decision-making. Recent advancements like DriveLM [14], LMDrive [13], DriveVLM [15], OmniDrive [16], ELM [22], CODA-LM [10], among others, have accelerated progress in this field, underscoring LVLMs' ability to create more intelligent and reliable systems. However, even leading models like GPT-4V struggle with corner cases, such as adverse weather or unseen categories, indicating the need for further research to enhance LVLMs' capabilities.

Building on the advancements in LVLMs, the Corner Case Scene Understanding Track of the ECCV 2024 Autonomous Driving Challenge offers a platform to explore the practical application of this technology. This track emphasizes the integration of advanced vision-language models into real-world autonomous driving scenarios. Using the CODA-LM [10] dataset, which is derived from CODA [8] and comprises around 10K images with textual descriptions of global driving scenarios, corner case analyses, and future driving recommendations, participants are tasked with developing models that effectively integrate language modalities. The goal is to address complex driving challenges and foster the creation of more reliable and interpretable autonomous driving systems.

To address the challenge of capturing intrinsic relationships between similar complex scenarios for accurate environmental perception and cognition, as well as achieving generalizable and explainable driving behavior, our proposed solution is summarized as follows. (1) We first utilize object detection and depth estimation models to extract **spatial information** from images, which is then converted into structured textual formats that are more interpretable for large language models. (2) We perform **scene-aware retrieval-augmented** to identify the most relevant samples as contextual examples, which is crucial for enhancing the model's understanding of extreme driving scenarios. (3) We construct a high-quality dataset for fine-tuning using carefully designed **step-by-step prompts** that guide the model in comprehending complex driving scenarios. (4) We employ parameter-efficient fine-tuning to refine the vision-language model and integrate it into a structured inference pipeline, enhancing reasoning and performance through visual and linguistic data.

The report is structured as follows: Section 2 describes

*Corresponding author.

the competition datasets, along with the evaluation methods. Section 3 introduces the base models of our NexusAD, and details the improvements and specifics of our implementation. Section 4 outlines the fine-tuning process and presents the experimental results and performance analysis. Finally, Section 5 summarizes our contributions.

2. Dataset

For the corner case scene understanding track, we use the CODA-LM [10] dataset as the training dataset. CODA-LM is a large-scale multimodal dataset specifically designed for corner cases in autonomous driving. It provides an automated and systematic evaluation framework for assessing the performance of large vision-language models (LVLMs) in handling complex driving scenarios. Carefully crafted prompts guide GPT-4V to generate high-quality text pre-annotations, which are then verified and refined by human annotators. This dataset consists of 4,884 training images, 4,384 validation images, 50 mini-set images, and 500 test images, with most images having a resolution of approximately 1280×720 pixels. And the challenge is designed around three main tasks: general perception(GP), regional perception(RP), and driving suggestions(DS). The performance of the LVLMs is evaluated using text-only GPT-4 as the "judge," an approach that shows stronger alignment with human judgment. The final score is assessed on the test set and is composed of the average scores from the three tasks.

3. Method

3.1. Foundation Model

In this paper, we utilize the InternVL2-26B [1] model as the foundation for corner case perception and understanding in autonomous driving. The InternVL2 family is a series of state-of-the-art vision-language models known for their superior semantic understanding and cross-modal reasoning capabilities. It comprises a vision encoder that converts images into feature representations and a language decoder that integrates these features with natural language to produce outputs. This model excels in handling diverse scenarios, including dialogue, detailed descriptions, and complex reasoning tasks, due to its dynamic high-resolution visual encoding and robust semantic comprehension. For this challenge, we enhance corner case perception by dividing images from the ego vehicle into six 448×448 sub-images, each processed into 256 image tokens using a vision transformer multi-layer perceptron (ViT-MLP) and pixel shuffle.

3.2. Preliminary Visual Perception

For visual perception in autonomous driving scenarios, our initial approach considered classic methods such as BEVFormer [11, 18], VoxFormer [9], or TPVFormer [5] for 3D object detection or semantic occupancy prediction. However, during practical implementation, we found that starting with 2D detection on the image and then estimating the depth of the detected objects yielded more accurate results. This approach also better aligns with the object categories in the competition dataset.

Visual Grounding. The InternVL2 model [1] supports grounding tasks, enabling queries like "Find `<bbox>`" to locate `<ref-object>` or "Find `<ref-object>`" to obtain coordinates `<bbox>`." However, when tested on the CODA-LM dataset, which emphasizes extreme scenarios in autonomous driving, the model's performance was suboptimal. While it excelled in object recognition, its localization precision was lacking. Trained on general object datasets like ref-COCO [20], the model tends to prioritize common objects, often missing long-tail targets in extreme scenarios. To address this, we selected Grounding DINO [12] as the initial object detector for the base perception module, ensuring more comprehensive detection results. We used the predefined categories in the CODA dataset as input to the detection model, retrieving the locations of all relevant objects in the image.

Depth Estimation. Accurate distance estimation of road objects, particularly obstacles, is vital for autonomous driving safety. To enhance object depth accuracy, we utilize the open-source DepthAnything v2 [19] to estimate image depth (see Figure 1). Depth estimates are obtained for each target based on their positions in the detection results. These depth values are categorized into four levels: "long range," "mid range," "short range," and "immediate," and integrated into the object's descriptive metadata. The detection and depth information are then formatted according to the seven major categories defined by CODA-LM, with details provided in section 6.1 of the supplementary materials.

3.3. Scene-aware Retrieval-Augmented Generation

Retrieval Augmented Generation (RAG) [7] effectively enhances large language models by combining pre-trained parameters with non-parametric memory, making it well-suited for handling knowledge-intensive tasks. In autonomous driving, accurate perception and scene understanding are crucial for ensuring safe operation. A promising approach to achieve these capabilities is through a context-guided model that integrates retrieved expert demonstrations, enabling high-performance, interpretable, and generalizable autonomous driving [21]. Initially, we considered using GraphRAG [2], but its computational demands for local deployment proved prohibitive. Additionally, direct index query methods were generally ineffective, making it difficult for the model to retrieve content that is truly relevant for decision-making. To address these challenges, we designed a retrieval method grounded in image semantics and scene similarity. This method selects the

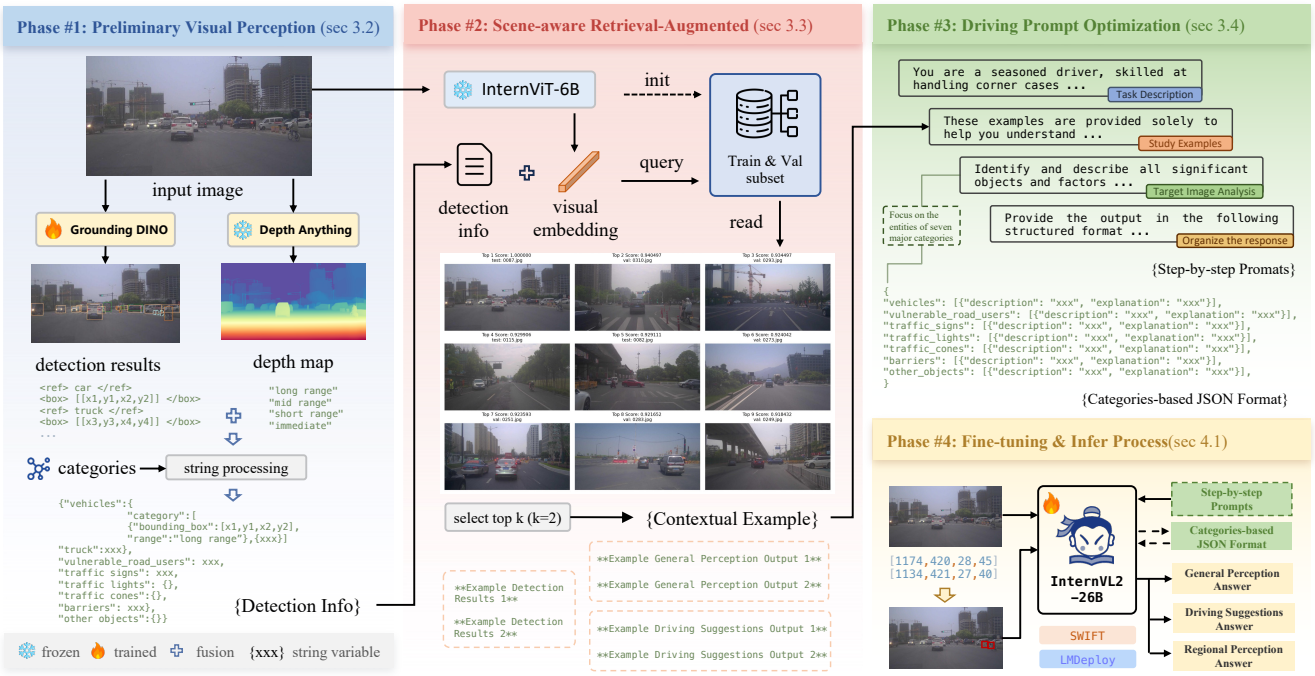

Figure 1. Overview of the proposed NexusAD framework. It mainly consists of four phases: preliminary visual perception, scene-aware retrieval enhancement, driving prompt optimization, and fine-tuning/inference process.

most relevant samples for the current scene, providing essential contextual information. Specifically, we first extract image features via the pre-trained InternViT-6B [1], which produces semantically aligned vector features. Recognizing that the distribution of road users significantly impacts perception and understanding, particularly in corner cases, we developed a scene-matching algorithm that emphasizes the category, orientation, and distance of targets relative to the vehicle. As demonstrated in Figure 1, our method effectively identifies the most pertinent samples for the current scene, enabling the model to assimilate contextual information more effectively and produce higher-quality responses. *More examples of scene-aware retrieval-augmented visualizations are provided in section 8 of the supplementary materials.*

3.4. Driving Prompt Optimization

Chain of Thought Prompting (CoT) [6, 17] has demonstrated significant improvements in the performance of large models on arithmetic, common sense, and symbolic reasoning tasks. By guiding models through intermediate reasoning steps, CoT enables them to achieve superior results compared to standard prompts. [14] In this section, we present our approach to constructing step-by-step prompts [3] designed to enhance the quality of model-generated answers. We first describe the formulation of a structured, category-aware output and then discuss the critical considerations for each of the three tasks. In addition, we have considered guidance on the content of concern for extreme scenarios of autonomous driving in the task prompts.

Format Output. The CODA-LM paper employs a labeling strategy that categorizes perceptual information into seven key areas, covering target descriptions and their implications for autonomous vehicles. This strategy utilizes GPT-4V to generate globally perceived responses, following a phased, step-by-step approach to mitigate the illusion problem in large models and improve answer quality. In line with this, we adopted a similar two-stage pipeline during the fine-tuning phase, instructing the model to generate output in a JSON format that aligns with the seven major categories. *The output format for the intermediate stage is detailed in section 6.2 of the supplementary materials.*

Prompt of General Perception. General perception evaluates the LVLM's ability to comprehensively understand key road entities. In this section, we designed a description of the perception task based on the original prompt provided by the official guidelines and utilized samples obtained through scene similarity retrieval as context for guiding the model. Our approach directs the model to analyze the given driving scene from the perspective of the autonomous vehicle, focusing on aspects such as the appearance, position, orientation, and impact of road entities on the vehicle. Additionally, we incorporated specific guidance for the model to handle special situations, such as obstacles, requiring heightened attention to these elements. To mitigate potential issues of contextual hallucination, we in-

cluded instructions in the prompts to suppress illusions, ensuring that the model validates its responses against the image content. Specific prompts can be found in the supplementary materials 7.1.

Prompt of Regional Perception. Regional perception measures the LVLM's ability to understand objects within specific bounding boxes and explains how these objects impact autonomous driving behavior. To avoid issues of contextual hallucination caused by example-based learning, we did not utilize context learning in regional perception. Instead, we designed specific steps and guidance based on the competition's focus categories to obtain responses more aligned with autonomous driving scenarios. *Detailed prompts can be found in the supplementary materials 7.2.*

Prompt of Driving Suggestion. Evaluating the ability of LVLM to generate driving suggestions is a critical aspect of the autonomous driving planning process. Compared to perception tasks, driving suggestions require a deeper understanding of driving scene rules and stronger reasoning capabilities. Therefore, while optimizing the task description and providing context examples for learning, we placed particular emphasis on driving-related rules in the prompts, such as maintaining a safe distance, adjusting speed, and identifying and following lane markings. Additionally, to better align with the driving advice annotations in the dataset, we instructed the model to output only the final suggestions, omitting the intermediate reasoning steps. *Detailed prompts can be found in the supplementary materials 7.3.*

4. Experiments

4.1. Fine-tuning & Inference Process

We used the SWIFT (Scalable lightWeight Infrastructure for Fine-Tuning) to fine-tune the InternVL-2.0 [1] model based on the training and validation subsets of the CODA-LM [10] dataset. Due to considerations of computational and parameter efficiency, as well as limitations imposed by the available gpu memory, we opted not to use full parameter fine-tuning, which typically offers better theoretical performance. Instead, we employed LoRA [4], an efficient parameter fine-tuning method, to adjust all fully connected layers in the components of the InternVL2-26B model. For the LoRA configuration, we set the rank to 8 and alpha to 32. We implemented a cosine learning rate scheduler, starting with an initial rate of 2e-4 and including a warm-up phase in the first 5% of the training steps. During the experiment, the random seed was set to 42, the context length to 4096, the batch size to 1, the gradient accumulation steps to 16, and the maximum number of epochs to 10. For the inference process, we deployed the model using LMDeploy, with the system prompt identical to that used during training: *"You are a seasoned driver, skilled at handling corner cases."* All of our experiments were conducted using the PyTorch framework on a computing platform equipped with an Intel Xeon Platinum 8350 CPU, 4 NVIDIA A100 GPUs, and 1024GB of memory.

Method	FS	GP	RP	DS
GPT-4V	59.02	57.50	56.26	63.30
CODA-VLM	63.62	55.04	77.68	58.14
InternVL-2.0-26B	52.11	43.39	64.91	48.04
NexusAD (Ours)	**68.97**	**57.58**	**84.31**	**65.02**

Table 1. The table shows the best results for our system on the test set. (FS: Final Score; GP: General Perception; RP: Regional Perception; DS: Driving Suggestion.)

4.2. Results on the Leaderboard

Table 1 presents the score results on the test set, ranked according to the competition standings. CODA-VLM [10] serves as the organizer baseline, while InternVL-2.0-26B [1] represents the baseline adopted in our approach. The table demonstrates that our improved methods and ideas have led to significant enhancements across all three tasks compared to the baseline model. Notably, our approach surpasses the performance of the powerful closed-source model GPT-4V. By leveraging structured spatial perception information, enhanced contextual learning for retrieval, and meticulously designed step-by-step guidance prompts, we achieved improvements of 14.19, 19.40, and 16.98 in general perception, region perception, and driving suggestion tasks, respectively, relative to the baseline model. Ultimately, we attained a final score of 68.97.

5. Conclusion

This report details our approach for ECCV 2024 workshop on multimodal perception and comprehension of corner cases in autonomous driving. We utilized the open-source multimodal model InternVL-2.0 and applied parameter efficient fine-tuning on the CODA-LM dataset to address corner cases in autonomous driving. Our methodology integrates detection and depth information, converting it into language patterns to enhance the model's performance on scene perception. We employed scene-aware retrieval to select the most relevant samples from the training set, providing the model with richer contextual information. Additionally, we designed prompts to guide the model through step-by-step reasoning, resulting in improved answer quality. The proposed NexusAD effectively captures complex associations in corner case driving scene, achieving a notable score on the final leaderboard.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 3, 4

[2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 2

[3] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[4] 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. In *International Conference on Learning Representations*, 2022. 4

[5] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 2

[6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 3

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2

[8] 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. 1

[9] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 2

[10] 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, 2, 4

[11] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[13] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 1

[14] 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, 3

[15] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1

[16] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 1

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[18] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 2

[19] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2

[20] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2

[21] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 2

[22] 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