The System Description of CPS Team for Track on Driving with Language of CVPR 2024 Autonomous Grand Challenge

Jinghan Peng*, Jingwen Wang*, Xing Yu*, Dehui Du†
East China Normal University
{51255902115, 51265902037, 51265902017}@stu.ecnu.edu.cn, dhdu@sei.ecnu.edu.cn

Abstract

This report outlines our approach using vision language model systems for the Driving with Language track of the CVPR 2024 Autonomous Grand Challenge. We have exclusively utilized the DriveLM-nuScenes dataset for training our models. Our systems are built on the LLaVA models, which we enhanced through fine-tuning with the LoRA and DoRA methods. Additionally, we have integrated depth information from open-source depth estimation models to enrich the training and inference processes. For inference, particularly with multiple-choice and yes/no questions, we adopted a Chain-of-Thought reasoning approach to improve the accuracy of the results. This comprehensive methodology enabled us to achieve a final score of 0.7799 on the validation set leaderboard.

1. Introduction

Large Language Models (LLMs) and Vision Language Models (VLMs) have achieved remarkable advancements in recent years. While traditional LLMs primarily focus on processing and generating natural language, VLMs extend these capabilities by integrating vision data with language, enabling the processing of text and images simultaneously [4][14]. In autonomous driving, VLMs are particularly advantageous due to their capability to process and integrate visual and language information concurrently. This capability enables the detection of obstacles, recognition of traffic signs, interpretation of road conditions, and informed decision-making based on these visual cues. This combination of vision and language processing enhances the overall performance and safety of autonomous driving systems (ADS) [3]. In summary, the ongoing development of VLMs holds vast potential for autonomous driving, promising more intelligent and reliable systems.

Building on the progress in VLMs, the Driving with Language track of CVPR 2024 Autonomous Grand Challenge serves as a platform to test and demonstrate how this technology can be adapted for practical applications, highlighting the significance of integrating advanced vision-language models into real-world autonomous driving scenarios. This track requires participants to develop models that integrate language modality to address complex driving questions by reasoning and making decisions using multi-view image inputs across various driving scenarios. To address this challenge, we propose a solution aimed at achieving accurate environmental perception, precise motion prediction, and generalizable and explainable driving behavior. We utilize depth estimation models to obtain depth information from images and construct a high-quality dataset through carefully designed prompts. We further fine-tune vision-language models using parameter-efficient fine-tuning methods. These refined models are integrated into a meticulously designed inference pipeline, enhancing the reasoning capabilities by leveraging both vision and language data, thereby significantly improving overall performance. Finally, we integrate the results from multiple individual systems to achieve further enhancements in performance.

The report is structured as follows: Sec. 2 details the training and validation datasets employed, including the preprocessing methods used. Sec. 3 introduces the base model of our VLMs. Sec. 4 describes the fine-tuning process and specifics of the VLMs. Sec. 5 describes the inference methodology and the incorporation of depth information. Sec. 6 presents experimental results and performance analysis. Finally, Sec. 7 provides a summary of our findings and contributions.

2. Dataset

2.1. Training Dataset

For the Driving with Language track, we use the DriveLM-nuScenes [11] dataset as the training dataset. This dataset consists of 4072 sample frames from 696 scenes in the nuScenes [2] dataset, comprising a total of 377,983 ques-

*These authors share equal contribution to this work.
†Corresponding author.

tions. Each scene is composed of a series of sample frames. Each sample frame includes six images, information on several key objects in the current scene, and a series of question-and-answer (QA) pairs. As illustrated in Fig. 1, the six images, each with a resolution of 1600×900, are taken from six cameras mounted at different directions on the vehicle. The key object information defines the crucial objects in the current scene, including the status of specified objects, their visual descriptions, and their 2D object bounding boxes in the images. As shown in Fig. 1, each key object is identified using a unique *KeyObj Tag*. These QA pairs encompass multiple-choice, yes/no, and dialogue-based formats, covering perception, prediction, planning, and behavior tasks related to driving.

To enhance the model's ability to identify key objects accurately, we utilize key object information from the training set to generate additional QA pairs for training. Examples are provided below, where the answer forms the description for the key object.

Q: *The width and height of the image are 1600 and 900 respectively. <c4,CAM_FRONT,920.8,383.3> represents the key object that the center coordinates of the bounding box in the CAM_FRONT image are (920.8,383.3). What is the object <c4,CAM_FRONT,920.8,383.3>? What is the state of it?*
A: *<c4,CAM_FRONT,920.8,383.3> is a white truck to the front of the ego vehicle. It is moving.*

To enhance the precision of object depth information, we utilize the open-source depth estimation model called Depth Anything [13] to estimate the depth of images, as depicted in Fig. 2. We calculate depth estimates for each pixel within the established bounding boxes of key objects in the training set. From these estimates, we select the 75th percentile value as the representative depth for each object. This depth value is then translated into a textual distance description, such as 'close' or 'far', which is subsequently incorporated into the object's descriptive metadata.

2.2. Validation Dataset

The validation dataset, which mirrors the data distribution of the training dataset, consists of 799 sample frames drawn from 149 scenes in the nuScenes dataset [2], encompassing a total of 15,480 questions. Different questions are evaluated using different metrics, and the final score is calculated by a weighted sum of these various evaluation scores. For key objects in the validation set, we extract their coordinates based on the *KeyObj Tag* and retrieve depth values for each pixel within an 11×11 rectangular frame centered on these coordinates. We then calculate the depth estimates for these objects using the same method applied in the training dataset.

3. Systems

LLaVA (Large Language and Vision Assistant) is an advanced multimodal model that effectively processes both visual and language data by combining a vision encoder with a language model. This integration allows it to perform complex tasks involving both text and images [8]. Our systems utilize LLaVA-1.5-7B and LLaVA-NeXT-7B as foundational models for further fine-tuning.

LLaVA-1.5 integrates a pre-trained visual encoder with the pre-trained language model, allowing for a robust understanding and generation of content across both modalities. The model is particularly innovative in its use of GPT-4 to generate multimodal instruction-following data from image-text pairs, which are then used for instruction tuning on machine-generated data [9]. This approach enables LLaVA-1.5 to handle a wide array of scenarios including conversations, detailed descriptions, and complex reasoning tasks. Its architecture facilitates seamless integration between visual features and textual data through a two-stage training process involving pre-training for feature alignment followed by end-to-end fine-tuning.

Building on the capabilities of LLaVA-1.5, LLaVA-NeXT introduces significant improvements such as enhanced visual reasoning, OCR capabilities, and an expanded understanding of world knowledge. It processes images at higher resolutions and accommodates three different aspect ratios, thus capturing more visual details. These developments ensure that LLaVA-NeXT significantly advances beyond its predecessor in terms of performance and applicability in real-world applications.

4. Training Protocol

We fine-tune the LLaVA model using the training data described in Sec. 2.1. To optimize computational and parameter efficiency, we have opted against using a full fine-tuning approach to train our model. Instead, we employ LoRA [5], an efficient parameter fine-tuning method, to fine-tune all fully connected layers within the language model component of LLaVA. Furthermore, we explore the use of DoRA [10], an advanced version of LoRA, to enhance our fine-tuning process. During the training process, we extract information about the key objects mentioned within each question. We select the corresponding image that contains these key objects to serve as the image input for the model. We then prepend the descriptions of these key objects to the question before inputting it into the model. For questions that solely contain directional information, we select the corresponding image based on the specified direction as the input. We then concatenate descriptions of key objects visible from that direction, prepending this information to the question. For questions lacking both key object details and directional information, we opt for an image facing forward

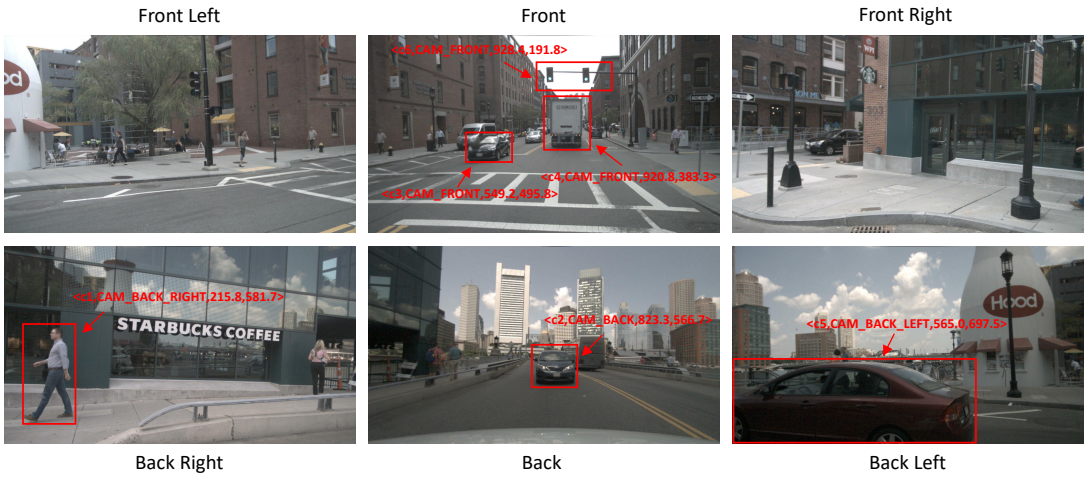

Figure 1. Diagram of six images from a sample frame, captured by cameras mounted on the vehicle in six directions: front, front left, front right, back, back left, and back right. Key objects are marked with IDs and 2D bounding boxes.

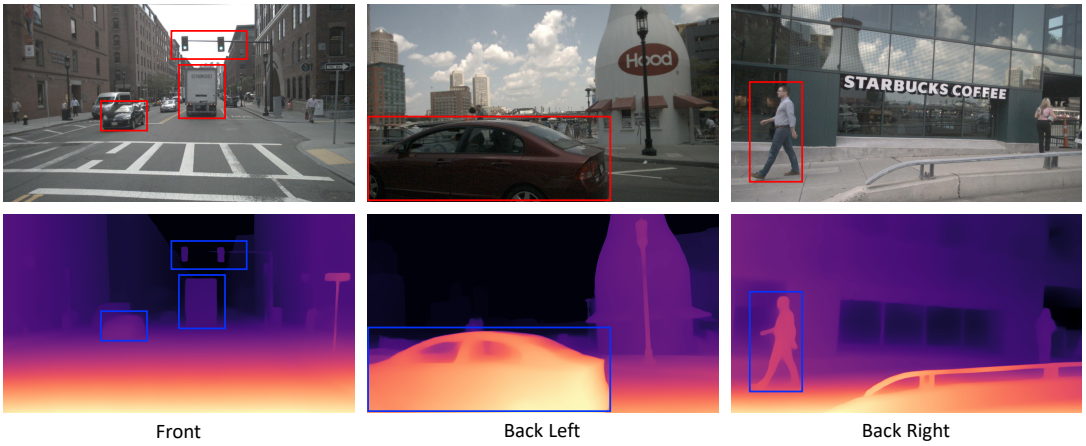

Figure 2. The diagram displays three raw images (Top) and their corresponding depth estimation images (Bottom) from a sample frame. Key objects are highlighted with red bounding boxes in the raw images and blue bounding boxes in the depth estimation images.

as the input. We then concatenate descriptions of all key objects in view and prepend this information to the question. For all our experiments, we employ PyTorch framework on a computation platform with an Intel Xeon Gold 5218R CPU, eight NVIDIA RTX 3090 GPUs, and 256 GB of memory. For the LoRA and DoRA configurations, we set the rank and alpha to 8 and 16, respectively. We implement a cosine learning rate scheduler starting at an initial rate of 2e-5 and incorporate a warm-up phase during the first 3% of the training steps. Each system is fine-tuned on the training set for one epoch to prevent overfitting.

5. Inference

Our proposed inference framework is illustrated in Fig. 3. For a single inference, a text-based question and a scene image are processed through a prompt design module that integrates a depth estimation model with our VLM, creating a prompt enriched with detailed scene information. This prompt and the scene image are then input into the VLM to generate the final answer.

During our preliminary testing, we discovered that the accuracy of the responses was relatively low due to insufficient information available from the initial input. To enhance the accuracy of answers, we designed a pipeline to guide the VLM towards more precise reasoning. For the initial input *Question Text*, heuristic rules are applied to extract *KeyObj Tag* of key objects as the off-line label for subsequent use. For the *Scene Image*, we first employ the depth estimation model to compute the depth information of the image. This depth information, combined with the *KeyObj Tag*, is used to estimate the depth of the key objects, *KeyObj Depth*, which serves as one of the components for constructing the subsequent prompt. Additionally, to im-

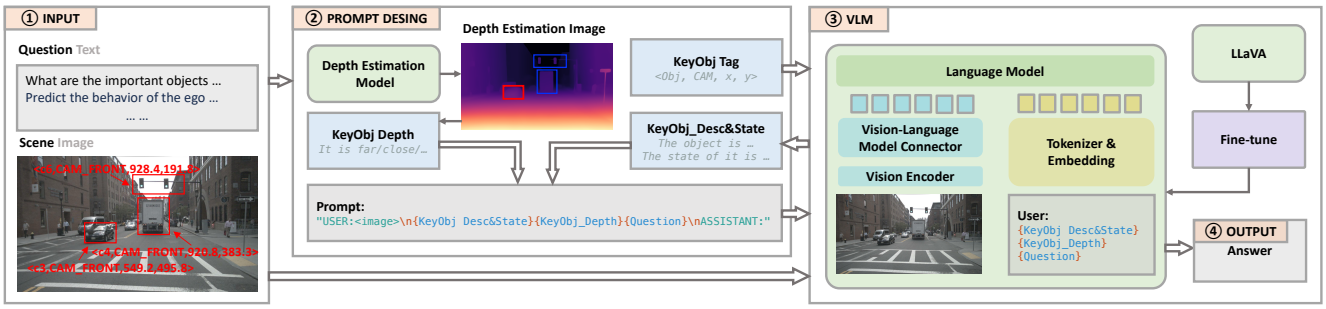

Figure 3. The architecture of our proposed inference framework.

Systems	Accuracy↑	ChatGPT↑	Language↑						Match↑	Final Score↑
			Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE_L	CIDEr		
LLaVA-1.5-7B+LoRA	0.7105	59.9425	0.9267	0.9040	0.8822	0.8602	0.9275	7.3282	90.3968	0.7329
LLaVA-1.5-7B+DoRA	0.6944	61.4596	0.8805	0.8228	0.7728	0.7269	0.8691	5.3107	93.1453	0.7177
LLaVA-NeXT-7B+LoRA	0.7431	64.5203	**0.9411**	**0.9212**	**0.9008**	**0.8798**	**0.9377**	**7.8806**	93.7031	0.7710
Fusion	**0.7861**	**64.8120**	**0.9411**	**0.9212**	**0.9008**	**0.8798**	**0.9377**	**7.8806**	**93.8406**	**0.7799**

Table 1. The table shows the top results for our individual systems and the fusion system on the validation dataset.

prove the quality of the response of VLM, we isolate the process of describing and reasoning about the key objects' states. This isolated query to the VLM includes the key objects' descriptions and states in the scene image *KeyObj Desc&State*, which also forms part of the information for constructing the prompt. The final generated prompt format is as follows:

Prompt: *USER: <image> {KeyObj Desc&State}{KeyObj Depth}{Question} ASSISTANT:*

This prompt includes the initial text information of the question, the depth information of the key objects, and the description and state information of the key objects. It is used along with the scene image as the final input to the VLM, from which the model's inference result is derived. It is important to note that we employ a multi-system fusion approach, utilizing the best-performing model for each question type to organize the final inference results.

Chain-of-Thought prompting (CoT) [12] facilitates complex reasoning through intermediate steps. Our approach combines Zero-shot CoT Prompting [7] and Few-shot CoT prompting [1] to enhance response accuracy for critical question types, including multiple-choice and yes/no questions. Despite our tailored prompt design, we observed a decrease in model performance during evaluations. Further analysis revealed that applying few-shot CoT in the complex context of autonomous driving may have inadvertently constrained the model's inherent reasoning abilities, limiting its capacity to generate diverse responses. Two main strategies for improvement include designing more comprehensive prompts to accommodate a variety of scenarios and fundamentally enhancing the CoT ability of model [6].

6. Result

The evaluation results for our leading individual systems on the validation set, along with the results for the fusion system, are detailed in Tab. 1. While DoRA requires training more parameters and prolongs the training period compared to LoRA, our findings show that LoRA slightly outperforms DoRA. We further fine-tune the LLaVA-NeXT-7B model using the LoRA method, which shows superior performance on both the accuracy and language metrics. In our extensive analysis, we integrate inference results from various individual systems, not just the ones noted as optimal in Tab. 1. For multiple-choice and yes/no questions, we employ a voting method to determine the most commonly selected answer as the final answer. For other types of questions, we choose the answer that achieves the highest evaluation score across the relevant evaluation metrics as the final answer. Ultimately, we achieve an optimal final score of 0.7799.

7. Conclusion

This paper outlines our contributions to the track on Driving with Language of CVPR 2024 Autonomous Grand Challenge. We develop vision language model systems, exclusively training them on the DriveLM-nuScenes dataset. Our systems are based on LLaVA models, which we enhance using the LoRA and DoRA fine-tuning techniques. We also integrate depth information from open-source depth estimation models into both training and inference phases. These efforts culminate in a notable score on the validation set leaderboard.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 1, 2

[3] 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

[4] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization, 2022. Survey Track. 1

[5] 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[6] Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*, 2023. 4

[7] 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. 4

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[10] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 2

[11] 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

[12] 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. 4

[13] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 2

[14] 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