Enhancing Vision Language Models for Autonomous Driving with Multi-view Multi-dataset Fusion

Anonymous CVPR submission

Paper ID a78b83c1

Abstract

*In this report, we present our MMAD , which enhances Vision Language Models for **A**utonomous **D**riving with **M**ulti-view **M**ulti-dataset Fusion. We introduce an end-to-end approach to solving question-answering (QA) tasks related to autonomous driving by leveraging vision language models(VLMs) with multi-perspective images and multi-dataset fusion. We enhance VLM by improving both the model input and training methodologies. For model input, our MMAD accepts multiple images as input and incorporates text prompts to direct the model's attention to different perspectives within the images. In terms of model training, we adopt a pretrain-finetuning paradigm. The model undergoes pretraining on publicly available multi-modal datasets and is subsequently fine-tuned with multi-modal data specific to autonomous driving scenarios. This process enables the VLM to comprehend and respond to questions within the context of autonomous driving. MMAD achieves an accuracy of 75% on the DriveLM validation set and secures a ChatGPT score of 65.6, placing it within the top 5 performers on the leaderboard. This demonstrates the effectiveness of our approach in advancing the capabilities of autonomous driving systems through the integration of multi-modal inputs and robust training strategies.*

1. Introduction

In the rapidly evolving landscape of autonomous driving technology, the integration of advanced computational models is pivotal to achieving safe, efficient, and intelligent transportation systems. Previous works are mainly developed on singular task-oriented model and the system is separated into three sub-modules, including perception[11, 14], prediction[5, 7], and planning[6, 18]. Recenet approaches[9, 10] have made strides in simplifying autonomous driving (AD) through end-to-end unified models. These models process raw sensor data directly to predict the results. While these efforts have achieved notable success, they also present challenges regarding the models' interpretability and robustness against various conditions.

One of the most promising developments in this field is the incorporation of Vision Language Models[13] (VLMs), which marks a significant leap in the evolution of intelligent transportation. VLMs, with their ability to process visual and textual data, offer a new dimension in how autonomous vehicles perceive and interpret their surroundings. By incorporating the reasoning capabilities of Large Language Models (LLMs), these systems can make more informed decisions, leading to safer and more efficient driving behaviors. LMDrive[20] utilizes the CARLA simulator to create a closed-loop dataset that includes navigation instructions. Simultaneously, DriveLM[22] has expanded the scope by developing a dataset that covers a range of tasks from perception to prediction and decision-making, leveraging the nuScenes dataset. GPT-Driver[15]refines the GPT-3.5 model to function as a motion planner by translating detection and prediction outcomes into textual data. Additionally, [24] has designed an interpretable, end-to-end autonomous driving system that processes multimodal inputs. Approaches such as DME-Driver[8] and Reason2drive[16] incorporate logical reasoning into decision-making tasks. This integration aims to provide models with capabilities akin to human reasoning and decision-making in driving scenarios.

In this technical report, we describe the details of our MMAD . To be more specific, we have made enhancements to the existing Vision Language Model by introducing **multi-image inputs** tailored for multi-perspective image and **conducting pretraining and instruction tuning** focused on scenarios pertinent to autonomous driving.

2. Method

2.1. Multi image input

In our pursuit to enhance VLMs for autonomous driving applications, we undertake a significant modification to the input mechanism of the model. Traditionally, VLMs have been designed to process single images, which, while ef-

CVPR #a78b83c1

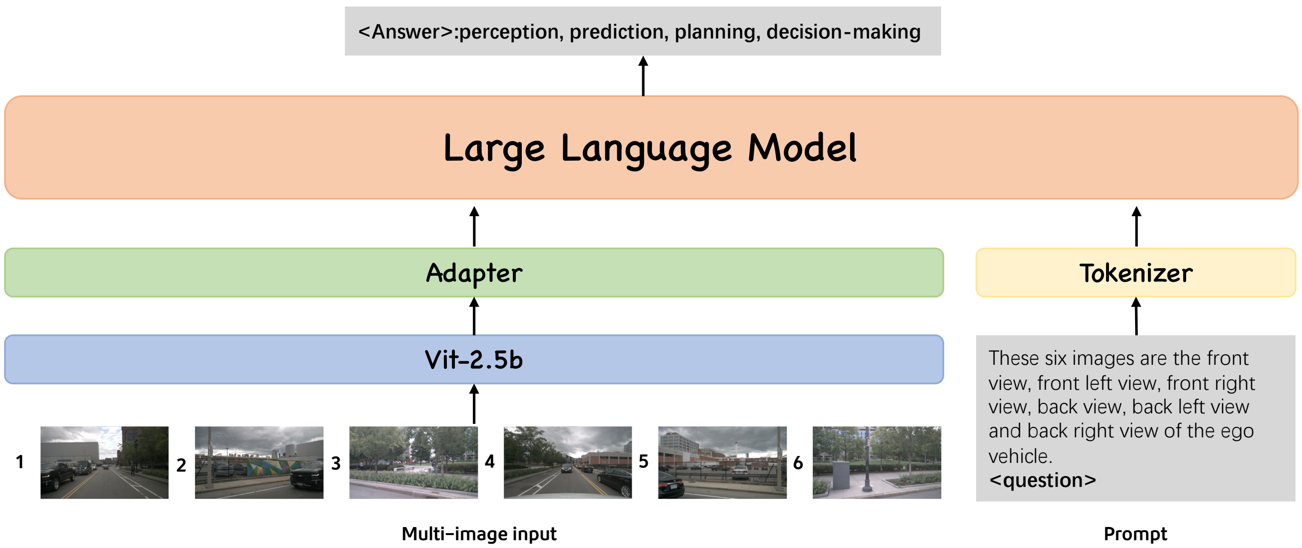

Figure 1. **Framework of MMAD .** MMAD takes the multi-view images and prompt as input, and predict the results.

fective, limits their ability to capture the full context of dynamic driving environments. To address this limitation, we re-engineer the VLMs to accept multiple images simultaneously. This adaptation allows the model to analyze a scene from various perspectives, thereby providing a more comprehensive and nuanced understanding of the driving context. As shown in Fig. 1, we input six images, separated by numbers between each pair of images. The vision encoder, such as a Vision Transformer (ViT), extracts features from each image. These extracted visual features are then transformed by an adapter module composed of attention and multi-layer perceptron (MLP) components into visual tokens for input into the large language model. By incorporating multi-view inputs, our model can better interpret complex traffic situations, recognize multiple objects and their spatial relationships, and make more informed decisions. This advancement in input methodology is crucial for the development of autonomous driving systems that can operate safely and efficiently in real-world conditions.

2.2. Training

The training process of MMAD consists of two stages: generalized pre-training and instruction tuning for autonomous driving. Specifically, the generalized pre-training involves mapping the image features to the input space of the language model, enabling the language model to comprehending the given image, *e.g.* generating a description for a single image. The instruction fine-tuning further enhances the pre-trained models for autonomous driving tasks such as perception, prediction, planning and decision-making on multi-view images, through fine-tuning on multiple autonomous driving instruction datasets. We introduce these two training steps in the following.

2.2.1 Generalized pre-training

In the generalized pre-training step, our goal is to enable a pre-trained large language model to comprehending the images. To achieve that, we training the VL adapter on large-scale set of image-text pairs, similar to existing multi-modal large-scale model [1, 12]. The training dataset is composed of several publicly accessible sources, including CC3M [21], CC12M [3], LAION-en [19], LAION-COCO [19], SBU [17] and COCO Caption [4]. In the training, we freeze the large language model and the vision encoder, and only optimize the VL adapter. The training objects is to minimize the cross-entropy of the text tokens.

2.2.2 Instruction tuning for autonomous driving

Till here, we have obtained a vision language model that can comprehend the image context, for example, generating a description for a single image. However, this simple understanding of images is far from sufficient for autonomous driving, which requires in-depth perception of the scene, accurate prediction, and complex logical reasoning. To address this issue, we propose to finetune the pre-trained vision language model on multiply autonomous driving instruction datasets. Specifically, we fuse the recent DriveLM [22] and OmniDrive [23] dataset to enhance the vision language model for autonomous driving:

DriveLM. The dataset comprises 696 scenes from nuScenes [2], with 4072 samples and approximately 0.3 million image-question pairs. The questions cover various aspects such as perception, prediction, planning, and behavior. We utilize the entire dataset for model fine-tuning. Furthermore, to enhance the model's predictive performance on multiple-choice questions, we convert certain questions into a question-and-answer format specifically designed for multiple-choice questions.

OmniDrive. The dataset consists of 28,130 samples, which comprise approximately 0.4 million image-question pairs. It covers various aspects, including scene description, attention, counterfactual reasoning, decision making, planning, and general conversation. This dataset includes more complex tasks that can significantly improve the model's comprehension and reasoning skills for intricate driving scenarios.

In addition to the two autonomous driving datasets, we also incorporate the instruction tuning dataset from LLaVA [12] to enhance and sustain the model's general reasoning abilities.

3. Experiments

We use a learning rate of 1e-5 with a cosine annealing schedule during instruction tuning. The model is trained for 1 epoch by using 16 NVIDIA A100 GPUs and the total batch size is 16. As shown Tab. 1, we integrate multiple autonomous driving datasets under the condition of multi-image input, ultimately achieving an accuracy of 75% on the DriveLM validation set and secures a ChatGPT score of 65.6, placing it within the top 5 performers on the leaderboard.

Table 1. The final score on validation set.

Dataset	Acc.	CG.	B1	B2	B3	B4	RL.	CI.	Mat.	Score
DriveLM	0.59	0.65	0.74	0.68	0.62	0.56	0.74	0.12	0.38	0.55
+LLaVA665k	0.72	0.64	0.76	0.70	0.64	0.59	0.74	0.17	0.44	0.58
+OmniDrive	0.75	0.66	0.76	0.70	0.64	0.59	0.74	0.18	0.45	0.60

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 2

[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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[5] Fang Da and Yu Zhang. Path-aware graph attention for hd maps in motion prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6430–6436. IEEE, 2022. 1

[6] Lingping Gao, Ziqing Gu, Cong Qiu, Lanxin Lei, Shengbo Eben Li, Sifa Zheng, Wei Jing, and Junbo Chen. Cola-hrl: Continuous-lattice hierarchical reinforcement learning for autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13143–13150. IEEE, 2022. 1

[7] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1

[8] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. *arXiv preprint arXiv:2401.03641*, 2024. 1

[9] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[10] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. 1

[11] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 1

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 3

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

[14] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1

[15] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1

[16] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. *arXiv preprint arXiv:2312.03661*, 2023. 1

[17] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 2

[18] Oliver Scheel, Luca Bergamini, Maciej Wolczyk, Błażej Osiński, and Peter Ondruska. Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *Conference on Robot Learning*, pages 718–728. PMLR, 2022. 1

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[20] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023. 1

[21] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[22] 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, 2

[23] 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. 2

[24] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 1