CS 543 PROJECT PROPOSAL: FROM VISION TO UNDERSTANDING: LEVERAGING VLMS TO ENABLE AUTONOMOUS DRIVING DECISIONS

Sridharan Subramanian
ss233@illinois.edu

Brijesh Muthumanickam
brijesh2@illinois.edu

Karteek Gandiboyina
mkg7@illinois.edu

Sai Rohit Muralikrishnan
srm17@illinois.edu

1 PROJECT DESCRIPTION AND GOALS

1.1 WHAT IS THE PROBLEM?

Our project focuses on integrating Vision Language Models (VLMs) with autonomous driving systems to enhance their decision-making capabilities and improve overall driving performance. VLMs, which are trained on large-scale web data, combine visual and language understanding, making them powerful tools for interpreting complex driving environments. By using these models, we aim to introduce advanced reasoning into autonomous systems, enabling them to make more informed and generalizable driving decisions.

The core idea behind our approach, called Drive Language Model, is to take visual input from multiple camera views and use the model's reasoning ability to answer critical questions about driving situations, such as traffic conditions, potential hazards, or navigation choices. This integration not only boosts the system's ability to generalize across different driving environments but also makes the driving process more interactive and explainable for human users. The ultimate goal is to create an end-to-end driving system that can respond intelligently to complex real-world scenarios, improving both safety and communication with users.

1.2 MEMBER ROLES

- Sridharan Subramanian: Start with Data preparation + Exploratory Data Analysis
- Brijesh Muthumanickam: Explore LLAMA
- Karteek Gandiboyina: Explore DriveGPT4
- Sai Rohit Muralikrishnan: Explore Graph Visual Question Answers (Graph VQA)

1.3 POTENTIAL DIFFICULT PART AND PLANNING

One of the challenging aspects of this project is predicting the behavior of other vehicles and related question-answering (QA), which is a second-level goal for us.

Our minimum goal for now is to achieve accurate perception of important objects, such as traffic signs and nearby vehicles, both in front and behind.

1.4 RELATIONSHIP TO YOUR BACKGROUND

All the team members are pursuing an MEng in Autonomy and Robotics. Autonomous systems are a field we are exploring and learning new things about every day. This project aligns directly with the team's interests as it combines core principles of computer vision and large language models (LLMs). Incorporating both allows us to further develop skills in these critical areas.

While we are familiar with core computer vision techniques such as object detection, classification, and segmentation, combining these with LLMs for tasks like prediction and planning will be a new area for us to explore.

2 DATASETS

We plan to first explore the following open-source datasets from nuScenes:

- **Training Dataset**: Dataset
- **Validation Dataset**: Dataset

3 EVALUATION METRICS

To evaluate the effectiveness of the Vision-Language Model implementation, we will score it based on the following metrics:

- **Object Score**: This metric will measure the accuracy of correctly predicted important objects versus the total number of objects in the scene.
- **Natural Language Answers Score**: This will score the difference between the predicted natural language answers and the ground truth answers.

REFERENCES

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.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Jingwei Wen, Xipeng Qiu, Yi-Chen Guo, Hui Xiong, Qun Liu, and Zhenguo Li. A survey of reasoning with foundation models. *arXiv*, abs/2312.11562, 2023. URL https://api.semanticscholar.org/CorpusID:266362535.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv*, abs/2311.01043, 2023. URL https://api.semanticscholar.org/CorpusID:264935408.

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.

Sima et al. (2023); Zhou et al. (2024); Yang et al. (2023); Sun et al. (2023)