**LIST OF GROUP MEMBERS**

| **NAME** | **NETID** |
| --- | --- |
| Sridharan Subramanian | ss233 |
| Brijesh Muthumanickam | brijesh2 |
| Murali Karteek Gandiboyina | mkg7 |
| Sai Rohit Muralikrishnan | srm17 |

**PROJECT TITLE:** From Vision to Understanding: Leveraging VLMs to enable autonomous driving decisions.

**PROJECT DESCRIPTION AND GOALS:**

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.

**MEMBER ROLES:**

1. Sridharan Subramanian - Start with Data preparation + Exploratory Data Analysis
2. Brijesh Muthumanickam - Explore LLAMA
3. Murali Karteek Gandiboyina - Explore DriveGPT4
4. Sai Rohit Muralikrishnan - Explore Graph Visual Question Answers(Graph VQA)

**RESOURCES:**

* [DriveLM: Driving with Graph Visual Question Answering](https://arxiv.org/abs/2312.14150)
* [Embodied Understanding of Driving Scenarios](https://arxiv.org/abs/2403.04593)
* [LLM4Drive: A Survey of Large Language Models for Autonomous Driving](https://arxiv.org/abs/2311.01043)
* [A Survey of Reasoning with Foundation Models](https://arxiv.org/abs/2312.11562)

**Reservations**

Potential difficult part:

Other vehicle Prediction and related QA - second level of goal

Planning

Minimum goal for now - Accurate Perception of important objects like signs, nearby vehicles front & back.

**Relationship to your background:**

All the team members are pursuing MEng in Autonomy and Robotics. Autonomous systems is a field we are exploring and learning new things everyday. This project is directly related to the team’s interests as it combines core principles of computer vision and large language models (LLMs). Incorporating both offers an opportunity to further develop skills in these key areas. While some core computer vision techniques like object detection, classification and segmentation are familiar to us, combining these with LLMs in prediction & planning will be a new avenue.

**Datasets**

We plan to first explore the open source datasets given below from nuScenes:

Training Dataset

[**https://drive.google.com/file/d/1DeosPGYeM2gXSChjMODGsQChZyYDmaUz/view**](https://drive.google.com/file/d/1DeosPGYeM2gXSChjMODGsQChZyYDmaUz/view)

Validation dataset

[**https://drive.google.com/file/d/18f8ygNxGZWat-crUjroYuQbd39Sk9xCo/view**](https://drive.google.com/file/d/18f8ygNxGZWat-crUjroYuQbd39Sk9xCo/view)

**Evaluation Metrics**

In order to evaluate this effectiveness of this Vision-Language Model implementation, we try to score it based on following metrics:

* ***Object Score*** - Accuracy of correctly predicted important objects vs total number of objects in the scene.
* ***Natural Language Answers Score -*** Scoring the difference between predicted natural language answers vs ground truth answers.