ABSTRACT

Name of student: **Shashwat Rajvanshi** Roll no: **231030053**

Degree for which submitted: **Master of Technology**

Department: **Civil Engineering**

Thesis title: **Efficient Language-based Description Generation of Dashcam Videos for Vision-Language Models**

Name of Thesis Supervisor: **Prof. Pranamesh Chakraborty**

Month and year of thesis submission: **January, 2026**

High-quality language-based description generation of traffic videos has emerged as a key area for advancing research in autonomous driving, yet producing reliable ground truth for video datasets remains a time-consuming and resource-intensive process. Traditional manual annotation requires human experts to parse complex scenes, describe vehicle behaviors, interpret traffic context, and reason about pedestrian and surrounding-vehicle interactions. As modern driving datasets continue to expand in size and complexity, the limitations of purely manual approaches—in terms of effort, scalability, and consistency—have become increasingly evident.

This thesis investigates whether structured, checklist-based annotations can be produced more efficiently by combining Vision-Language Model (VLM) capabilities with minimal temporal supervision. A 35-item checklist has been designed to capture verifiable aspects of each scene, including ego-vehicle motion, road layout, traffic-control elements, pedestrian attributes, and interactions with surrounding vehicles. Four different benchmark description pipelines have been implemented and evaluated along with a proposed *Single-Frame Temporal Augmentation* strategy. The latter generates per-frame captions at 1 fps and enriches them with human-based temporal metadata available in benchmark dataset, allowing static descriptions to be combined into a coherent scene-level annotation.

All outputs were converted into structured checklist responses and compared against human-annotated ground truth using item-wise Accuracy and F1-scores. The results show that while full video models capture global structure, they often miss localized or fine-grained details. In contrast, the proposed Single-Frame strategy consistently achieves competitive or superior factual agreement in most categories, particularly for spatial and attribute-based items, despite not relying on explicit motion modeling. Overall, the study demonstrates that reliable ground truth for driving clips can be generated without relying on expensive manual annotation or complex video models.