Exhibit D – Research Project Requirement Template

Leveraging Vision-Language Models for Efficient Understanding of Vulnerable Roadway Users via a Multimodal Traffic Sensing Approach

Recipient/Grant (Contract) Number: The University of Texas at Austin; The City College of New York/Grant # 69A3552344815 and 69A3552348320

Center Name: National Center for Understanding Future Travel Behavior and Demand (TBD)

Research Priority: Improving Mobility of People and Goods

Principal Investigator(s): Yiqiao Li, Jie Wei, Camille Kamga

Project Partners: N/A

Research Project Funding: $180,000

Project Start and End Date: 06/01/2025 - 05/31/2026

Project Description: The proliferation of 3D and video data from urban intersections offers a unique opportunity to analyze and protect vulnerable road users (VRUs). However, the effectiveness of modern detection models like PointPillar or CenterPoint is limited by the availability of high-quality labeled data. In Year 2, we demonstrated the feasibility of multimodal sensing using LiDAR and cameras. In Year 3, we propose a strategic shift toward **adaptive self-learning traffic monitoring framework** by leveraging multimodal large language models (MLLMs) to serve as both annotators and detectors, thereby minimizing the need for labor-intensive data annotation with a self-improving perception. This approach aims to improve the efficiency of VRU data collection and generate high-quality data to support a deeper understanding of micro-level VRU travel behavior.

This hybrid strategy includes two synergistic phases:

Phase 1 - MLLM-Assisted Annotation: We will explore the use of advanced Vision-Language Models (VLMs) such as Gemini and CLIP2Point to reduce manual annotation costs. Building on recent breakthroughs in Chain-of-Thought (CoT) and few-shot prompting, we will design in-context learning pipelines capable of generating high-fidelity annotations from minimal labeled examples. These pipelines will be applied to both LiDAR projections and camera footage, which produce detailed annotations that capture traffic-related objects (e.g., pedestrian, cyclist, buses, truck) with a specific focus on VRUs as well as their behaviors (e.g., jaywalking, crossing against signal) and interaction events.

Key tasks are: (i) Establish a rendering and preprocessing pipeline for point cloud and image integration; (ii) Design prompt structures and visual-question-answering tasks to guide the MLLM's annotation; (iii) Develop a human-in-the-loop annotation tool to iteratively validate and improve the results with minimal human efforts.

Phase 2: MLLM-Based Detection and Self-Improvement: Leveraging the annotated data generated in Phase 1, we will then implement few-shot learning, RAG or Low-Rank Adaptation (LoRA) based fine-tuning frameworks, and iterative prompting to adapt MLLMs into agentic object detectors and scene interpreters. This phase transitions MLLMs from passive annotators to active detection agents, bypassing traditional labor-intensive supervised training processes.

Key tasks are: (i) Evaluate the performance of zero-shot and few-shot detection; (ii) Use retrieval-based augmentations to enhance context in challenging scenes; (iii) Perform a comparative analysis between our proposed method and established deep learning models such as PointPillars- and VoxelNets-based architectures; (iv) Explore feedback close loops where MLLMs refine their performance using their own corrected outputs.

US DOT Priorities: The project aims to advance the Transformation Research Priority. In addition, the project will focus on the following:

- Research Priorities: Data-Driven Insight (Page 50): Data Science: Harness advanced data collection and data processing capabilities to create timely, accurate, credible, and accessible information to support transportation operation and decision-making.
- New and Novel Technologies: Automation: Harness advanced data collection and data processing capabilities to create timely, accurate, credible, and accessible information to support transportation operations and decision-making.
- Adaptive and Dynamic: Transportation systems can detect and adapt to changing conditions, such as changes in traffic demand, advances in technology, or changes to the environment by reconfiguring capacity and adopting new technologies. Future risks are anticipated, and adaptation strategies are built into the planning, design and operations of infrastructure.

Outputs: The following outputs are anticipated as outputs of this project:

- A two-phase pipeline combining MLLM-assisted annotation and detection.
- A large multimodal dataset with semantic VRU object and trajectory labels, generated with minimal human effort.
- A lightweight visual interface for VRU labeling and model feedback.
- Experimental results comparing MLLM-based detection with traditional 3D perception models.
- At least two peer-reviewed papers submitted to top-tier conference or journals.
- A toolkit demonstrating how to adapt general-purpose MLLMs to traffic scene understanding.

Outcomes/Impacts: This project will develop a novel human-in-the-loop self-learning framework that minimizes dependence on large annotated datasets while improving perception performance in traffic monitoring tasks. MLLMs will learn from their own annotations and iteratively refine their outputs using prompt-based guidance, evolving into effective few-shot object detectors which enable scalable VRU behavior analysis across diverse urban settings. These models will support explainable, context-rich outputs—enhancing safety diagnostics and informing urban design interventions. The outcomes will also provide foundational work for integrating agentic AI systems into real-world transportation infrastructures.

Final Research Report: A URL link to the final report will be provided upon completion of the project.