CS 766 Mid-Term Project Report

Chen Wei

1 Project Overview

This report outlines the mid-term progress of the final project, which aims to address the persistent challenge of open-vocabulary semantic segmentation in complex urban streetscapes. The core objective remains the application of Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA), to adapt a pre-trained Vision-Language Model (VLM) for specialized spatial feature extraction. The efficacy of this optimized computer vision pipeline will ultimately be evaluated through an interpretable driving risk assessment framework.

Since the submission of the initial proposal, our primary focus has been on establishing the computational environment, formulating the dataset pipeline, and conducting a rigorous empirical evaluation of the zero-shot baseline capabilities of the selected VLM. Key progress milestones include:

- **Data Pipeline Engineering:** We have successfully acquired, filtered, and structured the multi-view Street View Imagery (SVI) dataset sampled from the foundational study. To handle the high-resolution nature of the imagery, a robust pre-processing pipeline has been implemented to manage image resizing, normalization, and tensor conversion for batch processing. The manual annotation process for a representative ground-truth subset, essential for the supervised fine-tuning phase, is currently in progress.
- **Baseline Deployment and Inference:** The pre-trained VLM architecture has been successfully deployed within our local computing environment. Initial zero-shot inference tests have been executed across a subset of the data, establishing a critical qualitative baseline for spatial feature extraction prior to the application of LoRA weight injection.

2 Intermediate Results and Analysis

At this mid-term stage, intermediate results are predominantly qualitative, designed to expose the performance bottlenecks of the unoptimized VLM baseline. As hypothesized in the initial motivation, relying solely on zero-shot inference in foundational VLMs yields overly homogenized outputs and severe hallucinations when analyzing localized, domain-specific urban factors.

A systematic review of the baseline semantic masks reveals significant limitations in zero-shot spatial reasoning. For instance, the model consistently struggles with delineating boundaries under complex lighting conditions, frequently misclassifying severe shadows as physical obstacles. Furthermore, when evaluating "Sight Obstruction Risk" and "Traffic Sign Integrity", the zero-shot VLM often fails to detect heavily occluded or non-standard infrastructural elements. These intermediate findings strongly corroborate our core motivation: while the prohibitive computational costs of full-parameter training must be avoided, targeted network weight optimization via LoRA is strictly necessary to force the model to learn the specific inductive biases required for accurate streetscape segmentation.

3 Difficulties and Revisions

As the software implementation has advanced, practical and architectural limitations have necessitated strategic adjustments to the project scope to ensure feasibility and scientific rigor:

- **Data Constraints and Confounding Variables:** Initial considerations to expand the environmental complexity of the model by integrating severe weather conditions were re-evaluated.

Acquiring, temporally aligning, and annotating high-quality "bad weather" street view imagery proved unfeasible within the project's timeframe. More importantly, introducing disparate weather conditions introduces uncontrollable confounding variables into the spatial analysis. Consequently, the project scope has been strictly refined to utilize the existing, validated dataset from the base semantic segmentation paper. This ensures methodological consistency and directs computational focus entirely toward architectural optimization rather than data collection.

- **Abandonment of RAG Architecture:** Early exploratory research into incorporating Retrieval-Augmented Generation (RAG) to provide external textual context for the VLM revealed structural incompatibilities. Our preliminary probing indicated that while RAG architectures excel in providing macroscopic semantic context for text generation, they fundamentally lack the spatial granularity and rigid localization mechanics required for dense pixel prediction and mask generation. Therefore, the RAG component has been formally abandoned, as it adds unnecessary computational overhead and is practically useless for this specific spatial segmentation task.
- **Computational Overhead Mitigation:** Processing high-resolution SVI through a foundation VLM resulted in unexpected out-of-memory (OOM) errors during early batch processing. We have subsequently optimized our PyTorch data loaders, calibrated batch sizes, and implemented gradient checkpointing to ensure memory stability for the upcoming fine-tuning phase.

4 Revised Timeline

To ensure rapid iteration once the LoRA fine-tuning is complete, the programmatic scaffolding for the downstream evaluation has already been constructed. The data structures for the ten spatial indicators have been defined, and the boilerplate code for the eXtreme Gradient Boosting (XGBoost) classifier and TreeSHAP explainer is in place, currently awaiting the populated semantic mask outputs.

In light of the refined data scope and the architectural decision to bypass the RAG module, the project timeline has been updated:

- **Week 9–10 (Current):** Finalize manual annotations for the ground-truth subset. Inject trainable rank decomposition matrices and execute the full LoRA fine-tuning pipeline on the frozen VLM.
- **Week 11–12:** Process the complete SVI dataset to generate robust pixel-level semantic masks and compute the extracted spatial indicators. Feed these features into the pre-configured XGBoost classifier.
- **Week 13–14:** Apply TreeSHAP to extract feature attributions and generate local/global dependency plots. Finalize the comparative evaluation against the zero-shot baseline and synthesize the final report.