**Final Project Proposal Update**

**Topic: Improving Fine-Grained Image Classification with Prompt Engineering Using CLIP**

1. Introduction

Fine-Grained Image Classification is a task focused on distinguishing between subcategories that are visually alike, for instance, classifying different car models or dog species. Pre-trained models like CLIP have shown good performance in zero-shot classification for general tasks, but their effectiveness diminishes when applied to fine-grained scenarios. This project aims to explore how prompt engineering - structuring domain-specific textual inputs - can improve CLIP's ability to classify fine-grained categories. Different approaches of prompt refinement will be taken until CLIP's alignment with subtle visual distinctions is enhanced.

2. Problem Definition and challenges

In zero-shot classification tasks, CLIP performs well with general categories like comparing cats and dogs, but it struggles with fine-grained categories (e.g. specific species of animals or car models). The limitation could arise from prompts being too generic to capture specific and subtle distinctions. Fine-grained categories like bird species in the CUB-200-2011 dataset, require prompts that highlight subtle differences in visual and contextual features.

This project considers to confront three main challenges. First, crafting prompts that capture domain-specific features without overfitting. Second, Balancing prompt design simplicity with effectiveness. Third, analyzing failure cases to iteratively refine prompts for improved accuracy.

3. Related Works

**Fine-Grained Visual Classification**

**CLIP and Vision-Language Alignment**

**Prompt Engineering for Vision-Language Models**

4. Datasets

5. SOTA methods and baselines