AIMS-DTU Research Intern Round 2

Project Task: From Picture to Plate

Overview

This intern task focuses on applying Vision-Language Models (VLMs) to generate concise cooking instructions from food images and noisy titles. Applicants will work with a curated subset of recipe data they collect themselves to understand multimodal representation and text generation. The task emphasises creativity, prompt engineering, and an understanding of how vision and language models align.

Objective

Given a food image and a noisy or vague dish title, generate a concise 2–3 step cooking instruction that captures the essence of the dish.

Dataset Setup

Applicants are required to:

Collect or scrape **10–15 food image samples** along with

- A noisy or vague title (e.g., "cheesy bake", "noodly thing")
- Full cooking instructions (can be sourced from websites like Food.com, AllRecipes, etc.)

Applicant's Responsibility:

1. Create 10–15 concise cooking summaries (2–3 steps) manually from the full instructions.
2. Use these as examples to build a model/pipeline that can generalise to new, unseen image-title pairs.

Task Flow

Step 1: Data Collection & Manual Summary Creation

- Scrape or download food images + titles + instructions (10–15 examples)
- Write a 2–3 step natural language cooking summary for each

Step 2: Model Selection & Pipeline Design

- Choose a pretrained VLM
- Develop a pipeline that takes image + title as input and generates a summary
- Use few-shot prompting, template design, or light prompt tuning to guide the model

Step 3: Inference on Test Set

- Generate summaries for a test set of 5–10 new image-title pairs
- Compare generated summaries to original instructions manually or with BLEU/ROUGE (optional)

Step 4: Evaluation and Reflection

Include analysis of:

- When the model got it right/wrong
- Common failure pattern
- Justification for model choice and design decisions

Deliverables

1. Notebook/script to generate summaries
2. 10–15 manually written summaries
3. Model-generated outputs for test data
4. Short video walkthrough (3–5 minutes) explaining:
    - Your pipeline
    - Reasoning behind prompt structure/model selection
    - 2 sample outputs walkthrough

Deadline

Deadline for submission: **29 May 2025, 11:59 pm**
Submission platform: Email your submission to aimsdtu@dtu.ac.in