AIMS 2K28 Recruitments

Problem Statement: Scene Localization in Dense Images via Natural Language Queries

In many real-world applications such as surveillance, autonomous systems, and contextual visual analytics, dense scenes often contain multiple simultaneous activities. This project aims to build a system that can **identify and localize specific sub-scenes within a single dense image** based on a natural language query describing one of the events occurring in the scene.

Given a high-resolution image depicting multiple activities (e.g. *a street market*, *a park*, *a railway station*), and a textual description such as *"a person snatching a chain"* or *"a vendor selling vegetables to a customer"*, the model must output a **cropped image region that semantically corresponds to the input description.**

Example:
Input Image:

Prompt: Multiple people talking
Output Image examples:

or

Objective

Develop a deep learning model capable of:

- Understanding contextual visual features in dense, multi-activity scenes.
- Parsing textual scene descriptions into semantically meaningful representations.
- Grounding the text in the image by returning accurate bounding box(es) or cropped regions that represent the described scene.

Input

- Image: A single dense image (H x W x 3), possibly containing multiple distinct interactions.
- Query: A free-form natural language description (e.g., "a man snatching a chain").

Output

- A bounding box (x1, y1, x2, y2) or a cropped image patch that corresponds to the described interaction.

Deliverables:

- A working prototype that:
    - Takes a **dense image + scene description** as input.
    - Returns the **relevant cropped region** from the image.
- Documentation
- A short demo video (1–2 mins) showing your system working with at least two queries.

It is your responsibility to collect and organize the **training data** for your project.

Submission Deadline

Submit the deliverables by **15 August, 2025** by **11:59 pm**.

Submission Guidelines

- Submit all code, model weights (if any), and documentation.
- Include a clear README with setup and usage instructions.
- Share the demo video via Google Drive or YouTube.

Note: This assignment is designed to evaluate your approaches and ability to combine vision and language understanding in a practical application. **Creativity, optimization, and intelligent architectural** choices will be valued. Custom modules or techniques over plug-and-play models are highly encouraged.