ADL Project Proposal

Guide: Dr. A. V. Subramanyam

Ayush Saun (MT24024) **Manogna Pasumarthi** (MT24127)

1. Project Title

EventVision: Context-Aware Image Retrieval from Informative Caption

2. Problem Statement

Traditional image retrieval and generation models, such as **CLIP (77-token limit)**, **ALIGN (256-token limit)**, and **BLIP (128-token limit)**, are optimized for short, visually grounded captions but struggle to process long or multiple captions as contextual input. While these models perform well on object-level queries (e.g., "a dog playing with a ball"), they falter when faced with **complex, event-centric descriptions**, such as "massive protests following the 2023 election in Argentina" or "wildfires forcing evacuations across southern Europe." Such queries require deeper **contextual understanding**, including **causality, temporal dynamics, participant roles, and event-specific semantics**. Consequently, current systems fail to fully exploit **rich textual context**, limiting their ability to retrieve images that truly reflect real-world events. Addressing this challenge requires developing retrieval frameworks capable of understanding **informative, multi-sentence captions** and mapping them to images in a semantically meaningful and context-aware manner[2].

3. Tentative Solution

We plan to experiment with the following methods to understand which could be more suitable:

- Develop an advanced retrieval framework that leverages structured relationships among entities, temporal cues, and contextual signals.
- Implement a flexible model adaptation mechanism that allows selective parameter modulation and expert routing.

4. Work Division / Individual Focus

To ensure clarity and balanced contribution, each member will focus on a specific track:

- **Ayush Saun** Focus on development of advanced retrieval framework.
- **Manogna Pasumarthi** Focus on implementing a flexible model adaptation mechanism.

Both students will collaborate on documentation, integration, and final presentation.

5. Significance / Impact

Unlike conventional captioning and retrieval tasks that focus on individual objects, scenes, or actions, this project emphasizes the integration of contextual elements, including participants, temporal and spatial information, outcomes, and overall significance. It promotes the creation of models capable of producing narrative-rich, semantically meaningful outputs.

References

[1] Thien-Phuc Tran, Minh-Quang Nguyen, Minh-Triet Tran, Tam V. Nguyen, Trong-Le Do, Duy-Nam Ly, Viet-Tham Huynh, Khanh-Duy Le, Mai-Khiem Tran, and Trung-Nghia Le, *Event-Enriched Image Analysis Grand Challenge at ACM Multimedia 2025*, arXiv preprint arXiv:2508.18904, 2025. [arXiv]

[2] Dinh-Khoi Vo, Van-Loc Nguyen, Minh-Triet Tran, and Trung-Nghia Le, *EVENT-Retriever: Event-Aware Multimodal Image Retrieval for Realistic Captions*, arXiv preprint arXiv:2509.00751, 2025. [arXiv]

[3] *A Hybrid Dense-Sparse Multi-Stage Re-ranking Framework for Event-Based Image Retrieval*, 2025. [PDF]

[4] *Hierarchical Article-to-Image: Leveraging Multi-Granularity Text Representations for Article Ranking and Text-Visual Similarity for Image Retrieval*, 2025. [PDF]