Title: Proposal for a Multi-Modal Transformer Model

Problem Statement:

Current models struggle to efficiently combine text and image features for contextual understanding.

Objective:

Develop a lightweight transformer model integrating visual and textual embeddings.

Proposed Method:

Use pre-trained visual encoder for image features.

Use transformer-based text encoder.

Fuse embeddings using cross-attention layers.

Expected Outcomes:

Improved contextual understanding across modalities.

Efficient representation learning for multi-modal tasks.

Future Work:

Extend to video-text tasks.

Optimize for real-time deployment.