Name: Fenil Vadher

En Roll no: 92200133023

Subject: Capstone Project

Innovation and Originality

1. Novelty in Approach

The proposed **Multimodal Movie Script Search Framework** introduces a **context-aware dialogue and scene retrieval system** that leverages **multimodal embeddings** (text + vision) to enable fine-grained movie understanding. Unlike traditional unimodal search systems that operate on **text-only (script-based)** or **video-only (content-based retrieval)**, our approach integrates **dialogues, visual context, and scene metadata** into a unified search engine.

Key innovative aspects include:

1. **Context-Aware Retrieval Across Modalities**
    - Existing methods often limit retrieval to **keyword-based dialogue search** or **frame-based video retrieval**.
    - Our framework enables **dialogue-to-scene, scene-to-dialogue, and multimodal contextual queries**, providing richer narrative alignment.
    - Example: Searching with the phrase *"Show me the fight scene where the villain says 'This ends now'"* retrieves the exact **visual scene + dialogue context** instead of unrelated textual matches.
2. **Novel Use of Multimodal Transformer Models**
    - We integrate **Vid2Seq, BLIP-2, mPLUG, GIT2, Sky, and SPtPT**, each excelling in visual-language alignment, dialogue modeling, and sequence-to-sequence

learning.

- The **ensemble integration** of these models provides a **robust multimodal embedding space**, reducing semantic mismatches across modalities.

3. **Custom JSON-Based Dataset for Reproducibility**
   - Instead of relying on generic benchmark datasets (e.g., MS-COCO, MovieQA), we constructed a **custom dataset of movies and dialogues**, formatted in **JSON** for easy reproducibility and extension.
   - This structured dataset enables **scene–dialogue alignment research**, which is currently underexplored.
4. **Hybrid Evaluation with NLP + Vision Metrics**
   - While prior work mostly reports **textual overlap metrics (BLEU, ROUGE)**, our evaluation includes **vision–language similarity metrics (CLIP-Sim)** alongside **precision/recall**, creating a more **holistic benchmark**.
   - This dual evaluation ensures fairness across modalities.

Comparison to Existing Systems:

Aspect	Existing Solutions	Proposed Framework
Search Modality	Text-only or Video-only	Cross-modal: Dialogue ↔ Scene ↔ Multimodal
Models Used	Rule-based or unimodal DL models	Multimodal Transformers (Vid2Seq, BLIP-2, etc.)

Dataset Availability	Limited, non-reproducible	Custom JSON dataset, openly reusable
Evaluation Metrics	BLEU, ROUGE (text-only)	BLEU, ROUGE, CIDEr, CLIP-Sim, Precision/Recall
Application Scope	Script retrieval or video indexing	Movie understanding, recommendations, script summarization, media archiving

Thus, the originality lies in **fusing multiple modalities**, **introducing structured datasets**, and **establishing a new evaluation pipeline** that goes beyond unimodal limitations.

2. Contribution to the ICT Domain

The project makes contributions to both **research** and **practical applications** within ICT, particularly in **AI/ML, Natural Language Processing, Computer Vision, and Multimedia Information Retrieval**.

1. **Advancement in Multimodal AI Research**
    - By unifying text, dialogue, and visual embeddings, our work **extends multimodal retrieval capabilities** in the ICT domain.
    - Provides a **benchmark pipeline** for future researchers working on media understanding.
2. **Practical Applications for ICT Systems**
    - **Content Indexing:** Enables production houses to archive and search massive script–scene datasets efficiently.
    - **Recommendation Systems:** Enhances streaming platforms (e.g., Netflix, Amazon Prime) by retrieving **contextually relevant clips** instead of

metadata-only filtering.

- **Script Summarization & Storyboarding:** Assists directors, editors, and writers in **narrative planning**.
- **Intelligent Video Assistants:** Can serve as the foundation for AI-powered **film study tools** or **personalized learning systems** for media students.

3. **Contribution to Data Reproducibility in ICT Research**
   - The introduction of a **custom JSON-formatted multimodal dataset** bridges a **critical gap in reproducible research** for dialogue–scene alignment.
   - By making the dataset **reusable and extensible**, the project contributes to the ICT community's open research culture.
4. **Alignment with ICT Trends**
   - Multimodal AI (vision + language fusion) is one of the fastest-growing domains in ICT (as highlighted by IEEE and ACM reports).
   - The system aligns with **real-world industry needs** such as video analytics, intelligent media indexing, and **generative AI applications** in film.

3. Evidence of Novelty

- **Stakeholder Relevance:** Media students, filmmakers, OTT platforms, and AI researchers express the need for **fine-grained scene/dialogue search tools** beyond basic keyword search.
- **Technical Uniqueness:** Integration of **state-of-the-art multimodal transformers** in a **production-ready framework (Flask + React + Vector DB)** is underexplored in literature.
- **Supporting References:**

1. IEEE Xplore – Recent works on multimodal search highlight the **gap in dialogue–scene retrieval**.
2. ACM Digital Library – Reports emphasize **need for reproducible datasets** in multimodal AI.
3. Industry Reports (Gartner, 2024) – OTT platforms face **increasing demand for content personalization**, requiring multimodal solutions.

4. Summary of Innovation and Originality

The **Multimodal Movie Script Search Framework** is innovative because it:

- Enables **bidirectional and multimodal contextual retrieval** (dialogue ↔ scene).
- Uses an **ensemble of advanced multimodal transformer models** for embedding alignment.
- Introduces a **novel dataset structure** that is reusable and reproducible.
- Combines **NLP + vision metrics** for a holistic evaluation methodology.
- Contributes to **real-world ICT applications** (media indexing, recommendations, script summarization).

Thus, this work goes beyond existing approaches by **pioneering a new research direction** in **intelligent movie understanding** and offering practical benefits to both academia and industry.