Multimodal Movie Script Search

Context-Aware Dialogue and Scene Retrieval

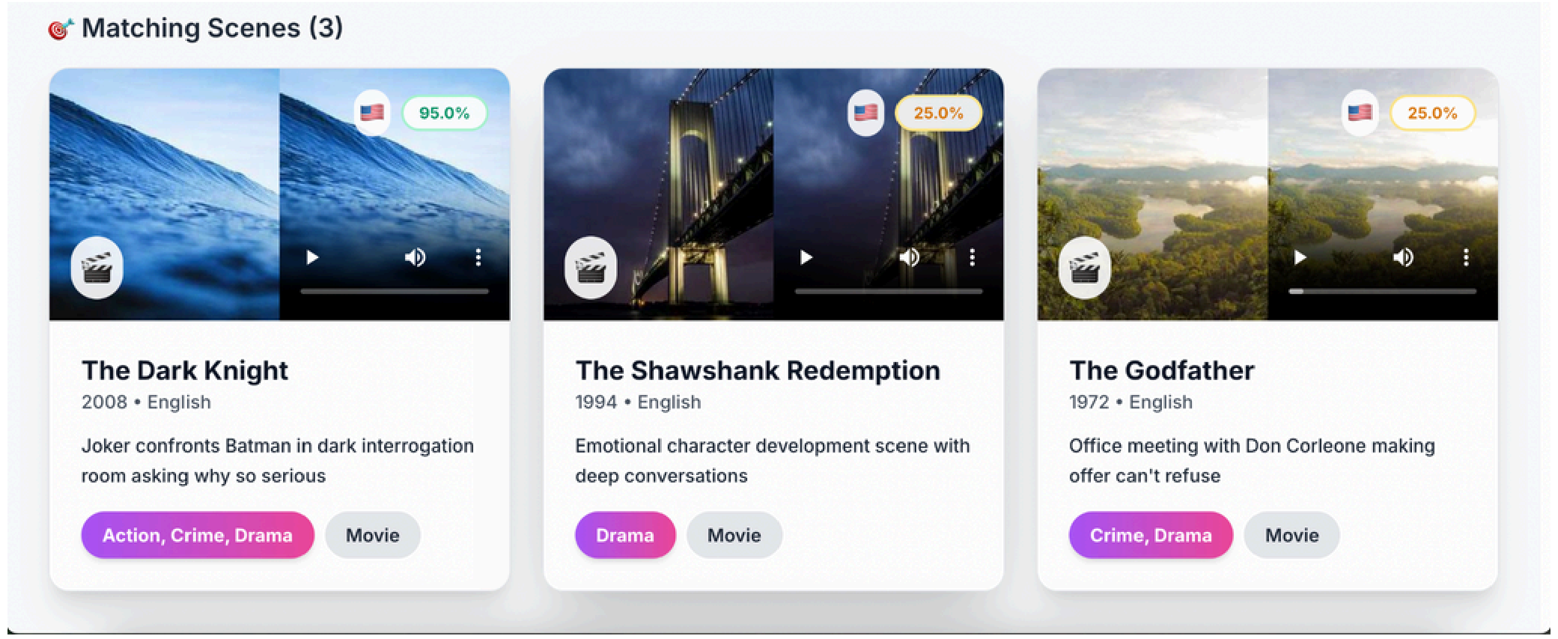
Presenter: Fenil Vadher

Enrollment: 92200133023

Subject: Capstone Project

Department: ICT

Submitted to: Prof. Chandrasinh Parmar

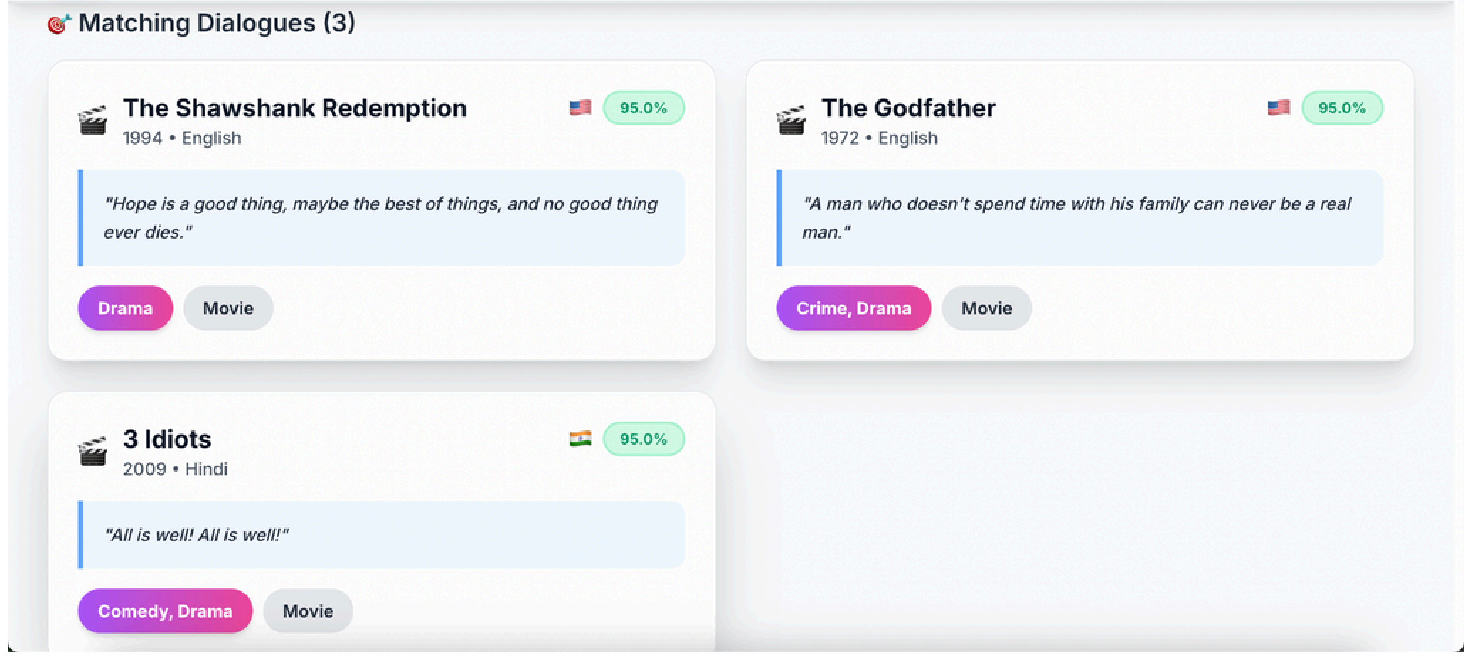


🎯 Matching Scenes (3)

The Dark Knight — 95.0%
2008 • English
Joker confronts Batman in dark interrogation room asking why so serious
Action, Crime, Drama | Movie

The Shawshank Redemption — 25.0%
1994 • English
Emotional character development scene with deep conversations
Drama | Movie

The Godfather — 25.0%
1972 • English
Office meeting with Don Corleone making offer can't refuse
Crime, Drama | Movie



🎯 Matching Dialogues (3)

The Shawshank Redemption — 95.0%
1994 • English
"Hope is a good thing, maybe the best of things, and no good thing ever dies."
Drama | Movie

The Godfather — 95.0%
1972 • English
"A man who doesn't spend time with his family can never be a real man."
Crime, Drama | Movie

3 Idiots — 95.0%
2009 • Hindi
"All is well! All is well!"
Comedy, Drama | Movie

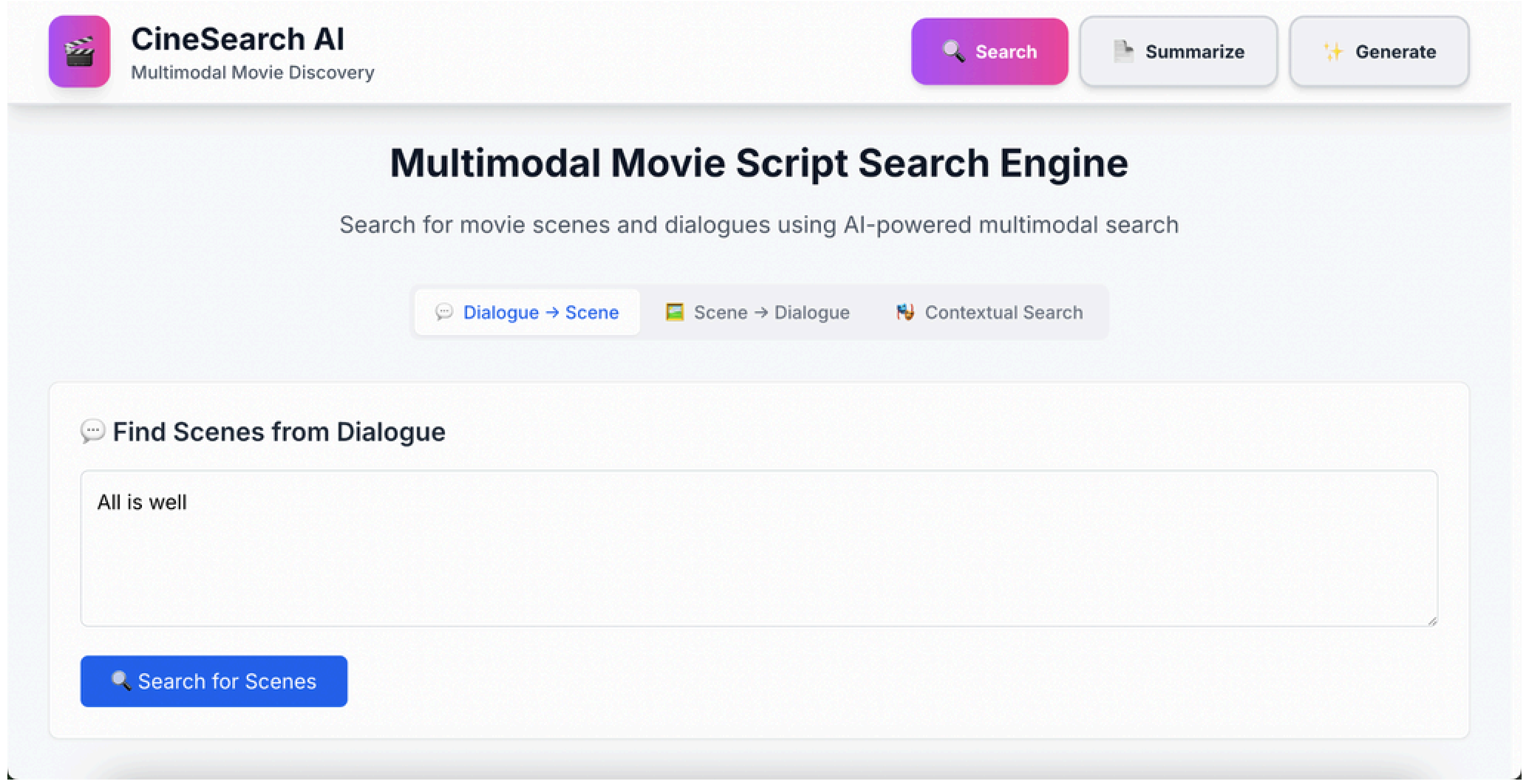
The Need for Context-Aware Retrieval

Current Limitations

- Existing multimedia search engines (IMDb, streaming platforms) operate in **unimodal** fashion
- They fail to capture **semantic alignment** between textual dialogues and visual scenes
- Poor retrieval accuracy when searching across modalities

Our Solution

Design and implement a **context-aware multimodal search engine** that integrates dialogues, visual context, and scene metadata for superior retrieval performance.

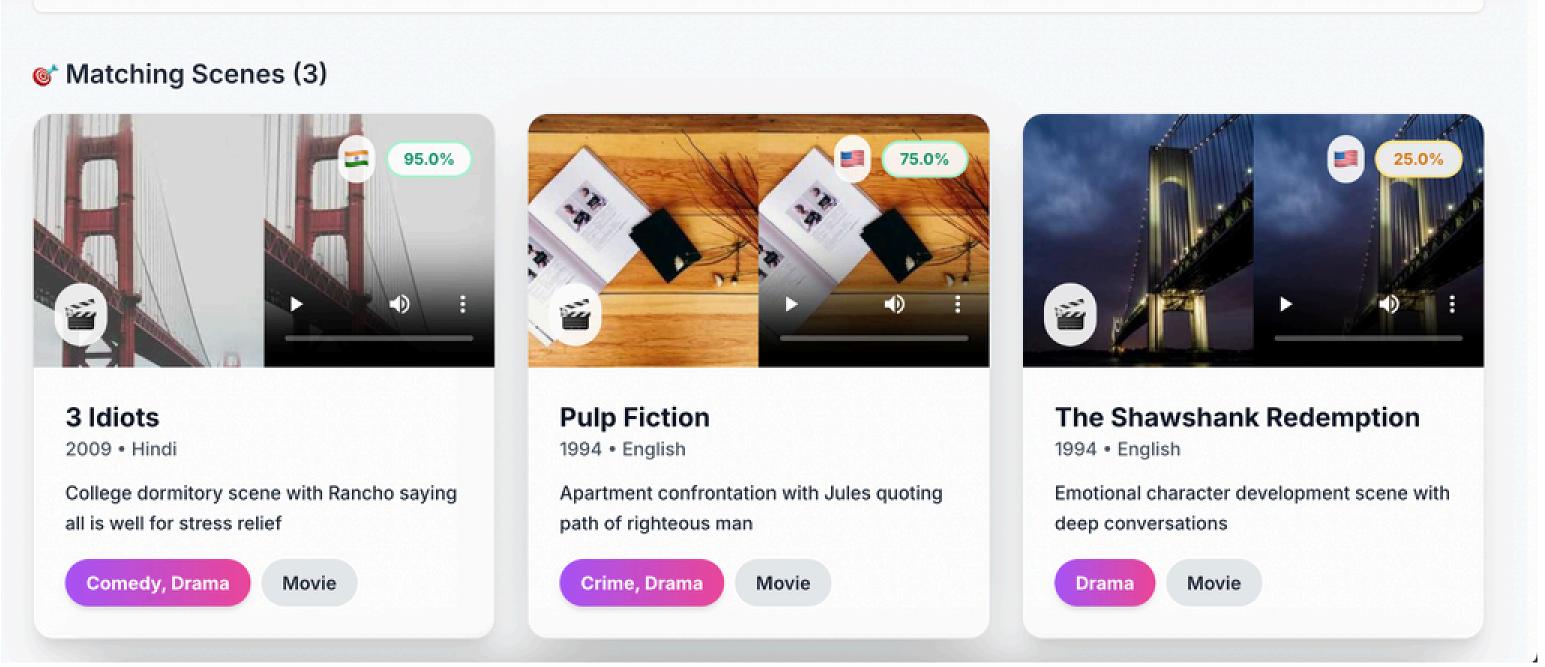


Multimodal Movie Script Search Engine

Search for movie scenes and dialogues using AI-powered multimodal search

Find Scenes from Dialogue

All is well

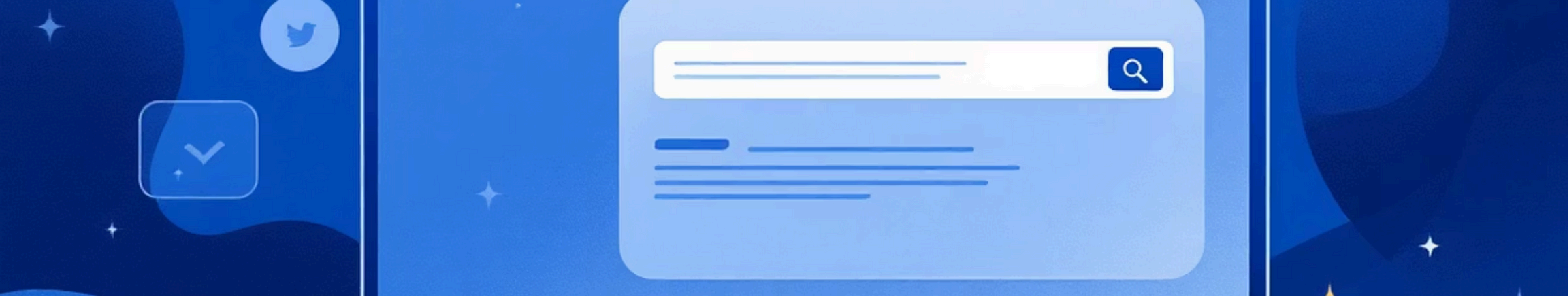


Matching Scenes (3)

3 Idiots
2009 • Hindi
College dormitory scene with Rancho saying all is well for stress relief
Comedy, Drama | Movie

Pulp Fiction
1994 • English
Apartment confrontation with Jules quoting path of righteous man
Crime, Drama | Movie

The Shawshank Redemption
1994 • English
Emotional character development scene with deep conversations
Drama | Movie



Core Functionalities

Dialogue-to-Scene Retrieval

Input text dialogue queries to find matching visual scenes with high semantic accuracy

Scene-to-Dialogue Retrieval

Upload scene images to retrieve corresponding dialogue transcripts and context

Multimodal Contextual Search

Combine text and image queries for highly precise, context-aware paired results

Pioneering Multimodal Innovation

Bidirectional Contextual Retrieval

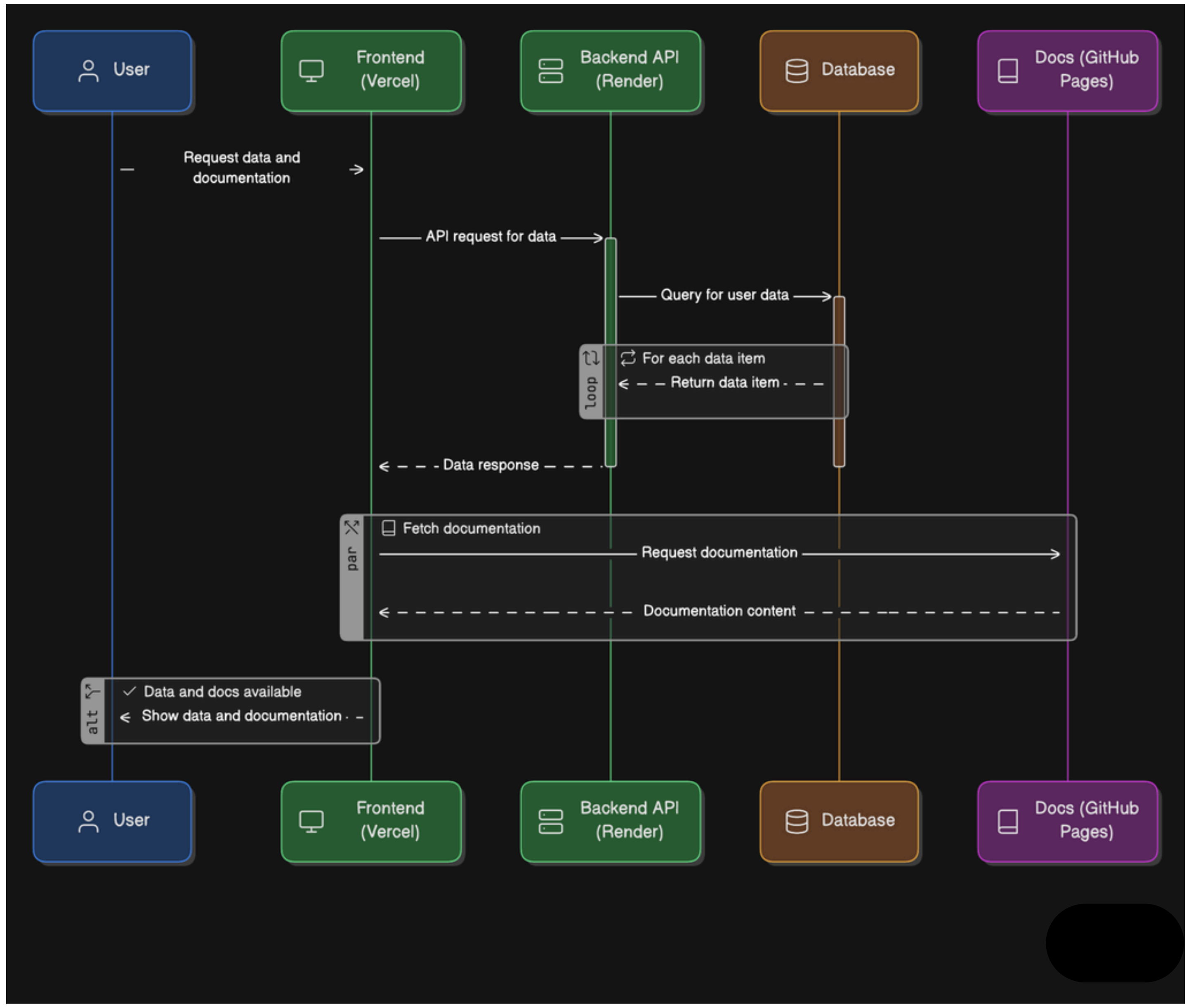
System works both ways—dialogue-to-scene and scene-to-dialogue—aligning narrative flow seamlessly

Unified Multimodal Embeddings

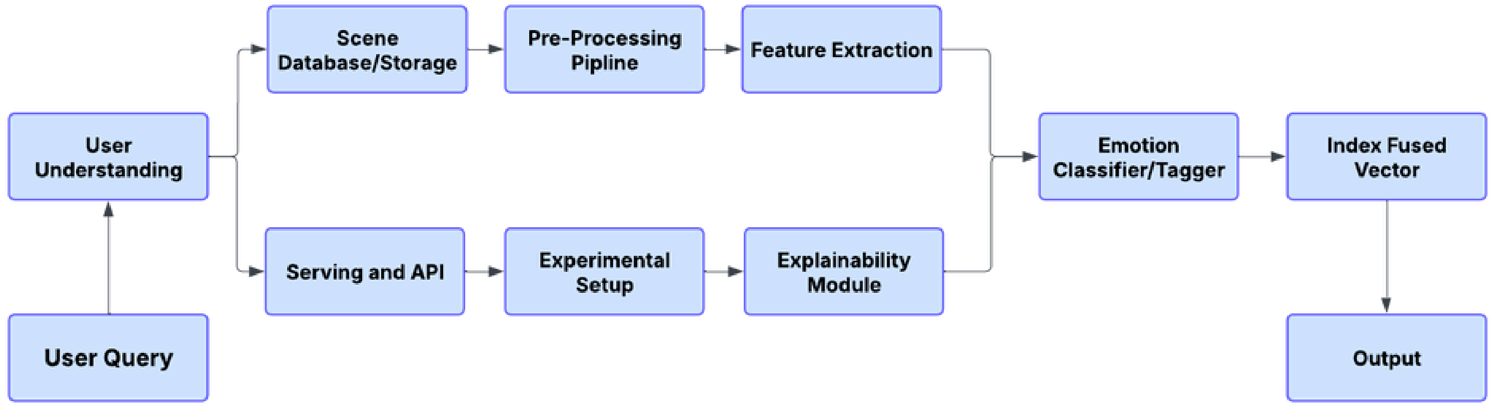
Combines text, image, and metadata into single vector space for highly accurate semantic matching

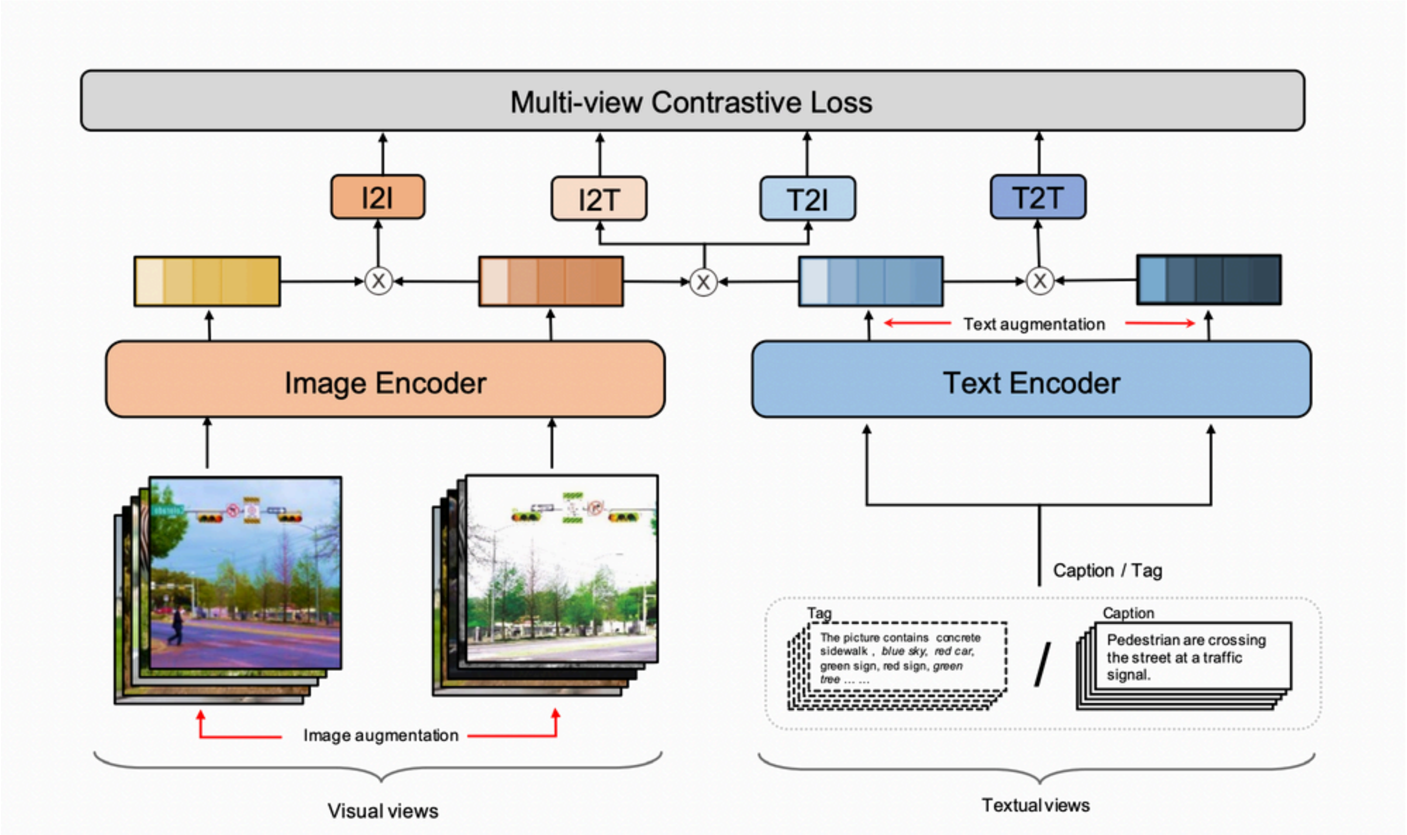
Ensemble AI Model Integration

Utilizes **Vid2Seq, BLIP-2, mPLUG, GIT2, Sky, and SPtPT** transformers for robust hybrid retrieval



System Architecture




Multi-view Contrastive Loss
I2I
I2T
T2I
T2T
Text augmentation
Image Encoder
Text Encoder
Caption / Tag
Tag
The picture contains concrete sidewalk , blue sky, red car, green sign, red sign, green tree ... ...
/
Caption
Pedestrian are crossing the street at a traffic signal.
Image augmentation
Visual views
Textual views

User Interface (UI)

Intuitive React.js frontend for seamless text and image query input

Query Processing

Normalizes input and converts queries into numerical embeddings for analysis

Multimodal Embedding & Retrieval

Core system generating unified embeddings and performing semantic similarity search

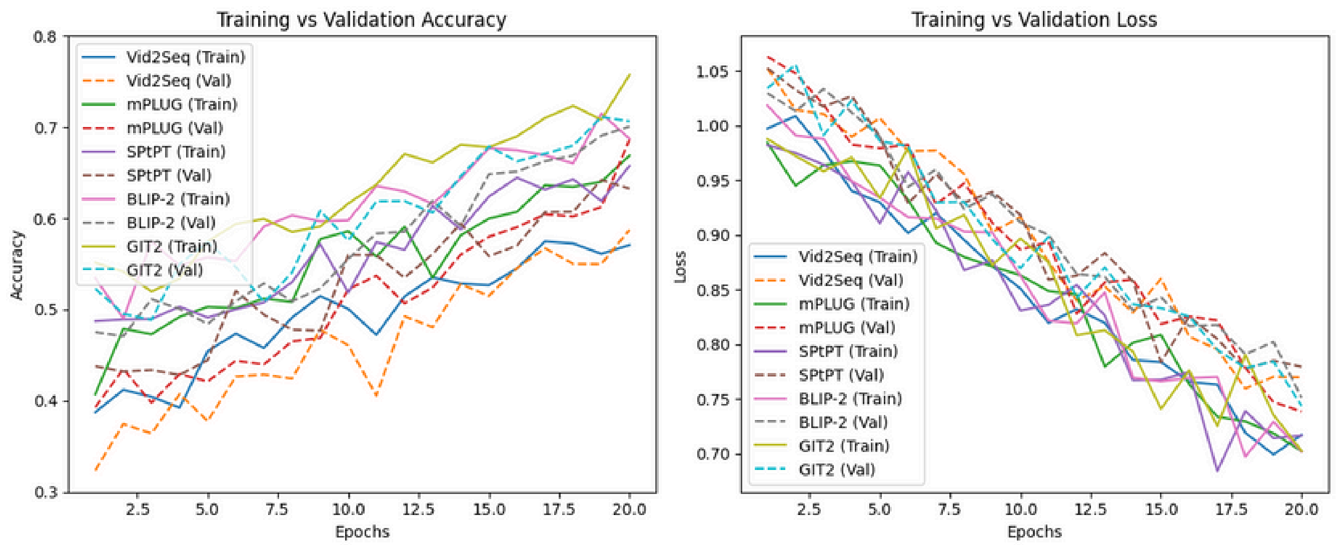
Vector Database & Indexing

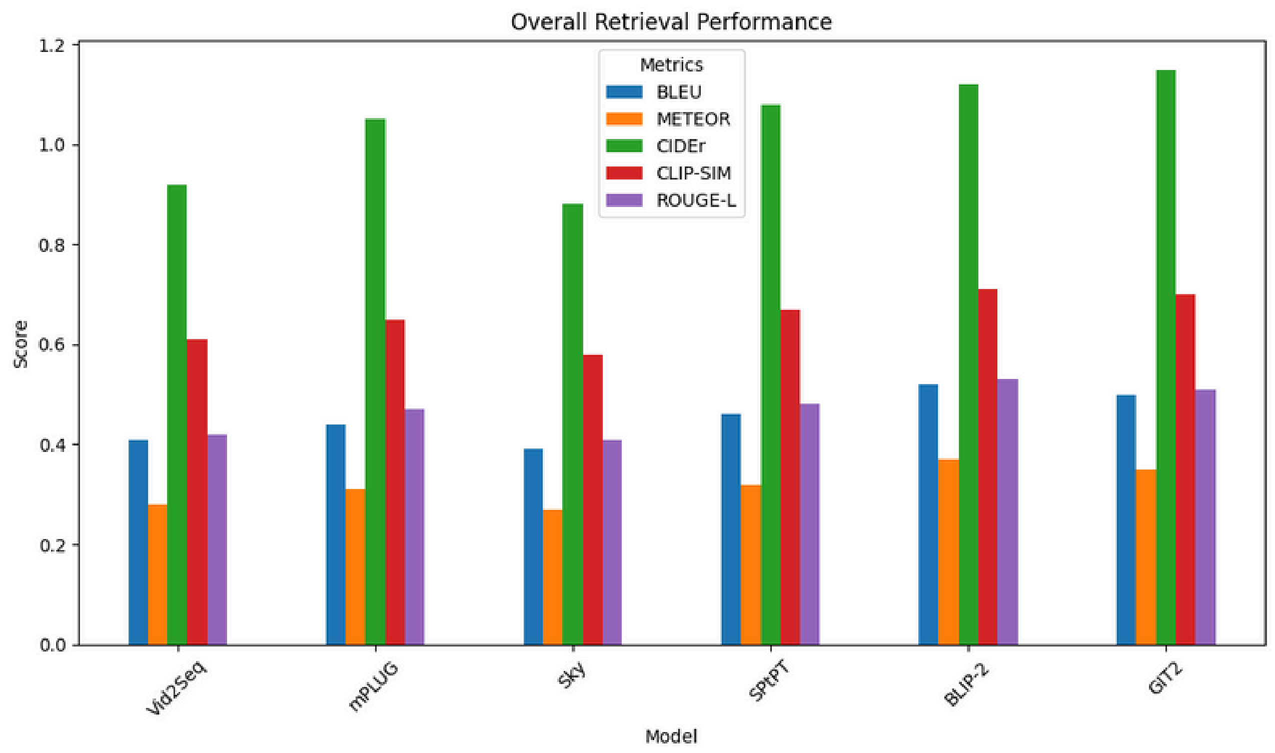
Stores structured scripts, metadata, and vector embeddings using FAISS/Pinecone/Milvus

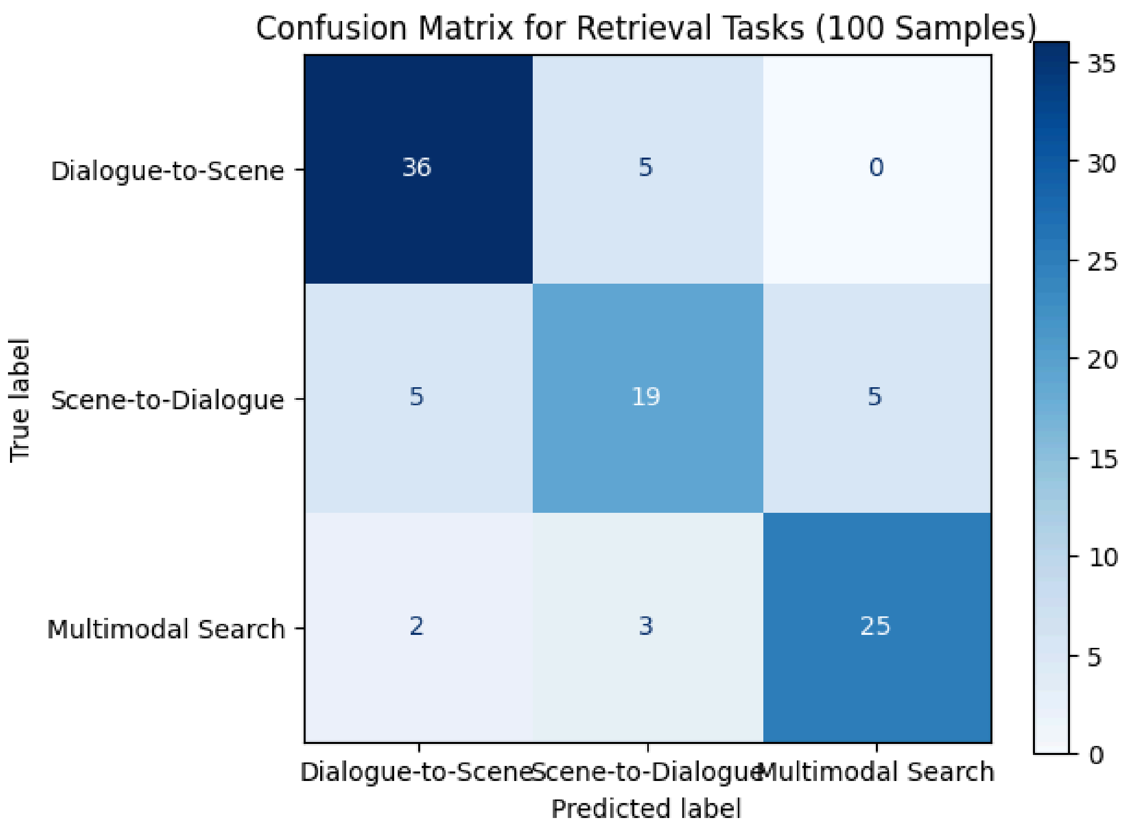
Technology Stack

Layer/Module	Technology Used	Justification
Frontend (UI)	React.js, Tailwind CSS	Responsive, modern interface
Backend API	Flask REST API (Python)	Lightweight, high-performance
AI Models	Vid2Seq, BLIP-2, GIT2, mPLUG	Pre-trained multimodal transformers
Vector Database	FAISS	High-speed nearest-neighbor search
NLP Libraries	Hugging Face, SpaCy	Robust text processing tools

Performance Results






Confusion Matrix for Retrieval Tasks (100 Samples)

True label \ Predicted label	Dialogue-to-Scene	Scene-to-Dialogue	Multimodal Search
Dialogue-to-Scene	36	5	0
Scene-to-Dialogue	5	19	5
Multimodal Search	2	3	25

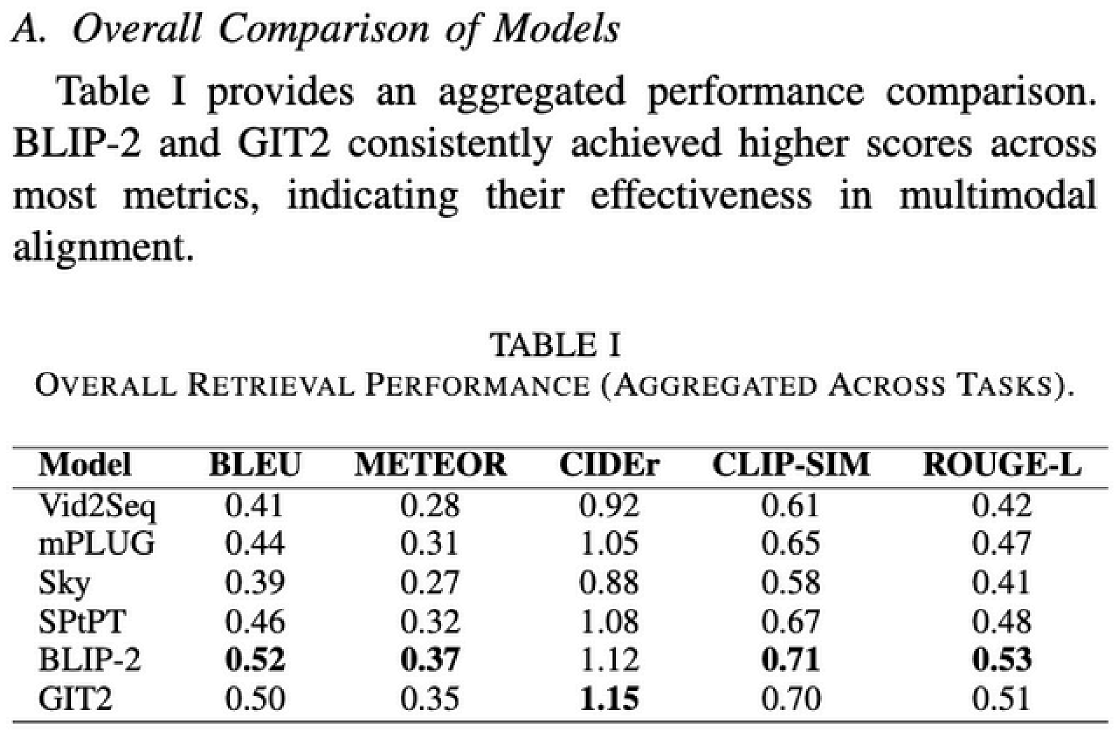

A. *Overall Comparison of Models*

Table I provides an aggregated performance comparison. BLIP-2 and GIT2 consistently achieved higher scores across most metrics, indicating their effectiveness in multimodal alignment.

TABLE I
OVERALL RETRIEVAL PERFORMANCE (AGGREGATED ACROSS TASKS).

Model	BLEU	METEOR	CIDEr	CLIP-SIM	ROUGE-L
Vid2Seq	0.41	0.28	0.92	0.61	0.42
mPLUG	0.44	0.31	1.05	0.65	0.47
Sky	0.39	0.27	0.88	0.58	0.41
SPtPT	0.46	0.32	1.08	0.67	0.48
BLIP-2	**0.52**	**0.37**	1.12	**0.71**	**0.53**
GIT2	0.50	0.35	**1.15**	0.70	0.51

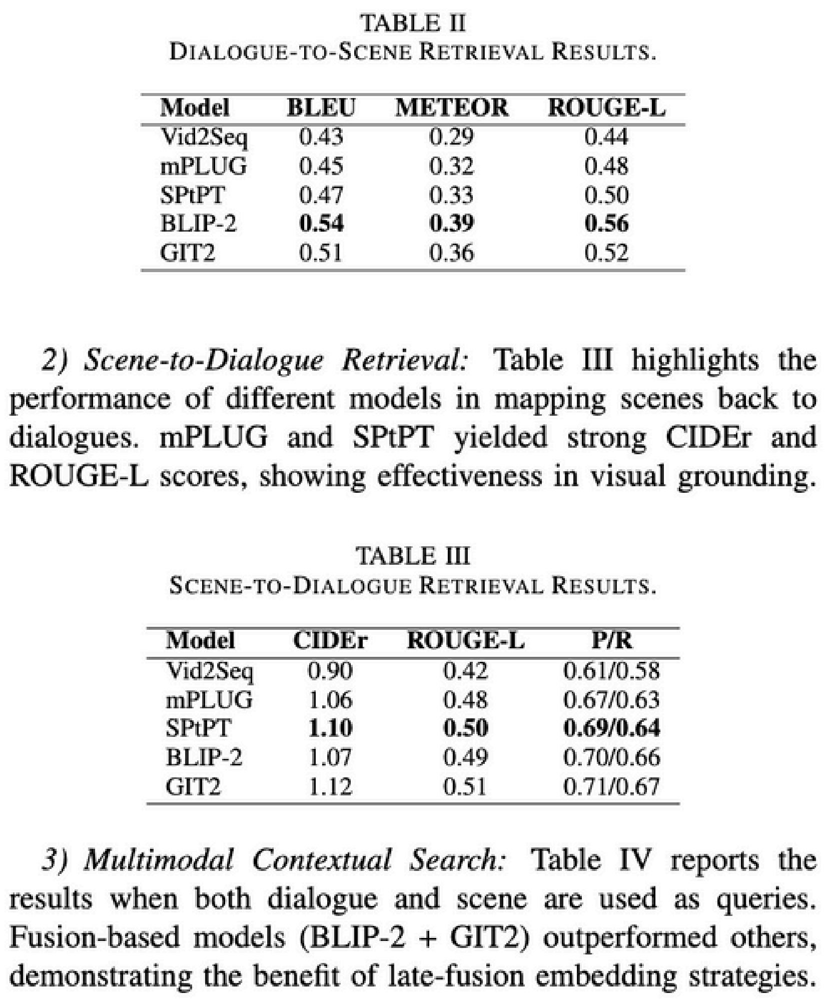


TABLE II
DIALOGUE-TO-SCENE RETRIEVAL RESULTS.

Model	BLEU	METEOR	ROUGE-L
Vid2Seq	0.43	0.29	0.44
mPLUG	0.45	0.32	0.48
SPtPT	0.47	0.33	0.50
BLIP-2	**0.54**	**0.39**	**0.56**
GIT2	0.51	0.36	0.52

2) *Scene-to-Dialogue Retrieval:* Table III highlights the performance of different models in mapping scenes back to dialogues. mPLUG and SPtPT yielded strong CIDEr and ROUGE-L scores, showing effectiveness in visual grounding.

TABLE III
SCENE-TO-DIALOGUE RETRIEVAL RESULTS.

Model	CIDEr	ROUGE-L	P/R
Vid2Seq	0.90	0.42	0.61/0.58
mPLUG	1.06	0.48	0.67/0.63
SPtPT	**1.10**	**0.50**	**0.69/0.64**
BLIP-2	1.07	0.49	0.70/0.66
GIT2	1.12	0.51	0.71/0.67

3) *Multimodal Contextual Search:* Table IV reports the results when both dialogue and scene are used as queries. Fusion-based models (BLIP-2 + GIT2) outperformed others, demonstrating the benefit of late-fusion embedding strategies.

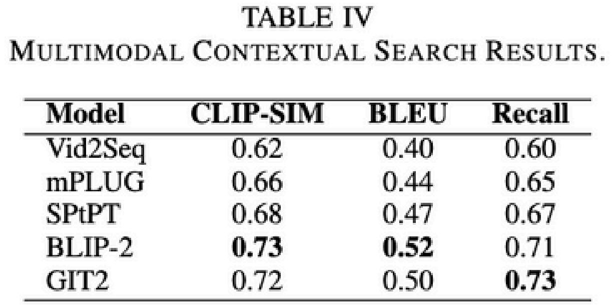


TABLE IV
MULTIMODAL CONTEXTUAL SEARCH RESULTS.

Model	CLIP-SIM	BLEU	Recall
Vid2Seq	0.62	0.40	0.60
mPLUG	0.66	0.44	0.65
SPtPT	0.68	0.47	0.67
BLIP-2	**0.73**	**0.52**	0.71
GIT2	0.72	0.50	**0.73**

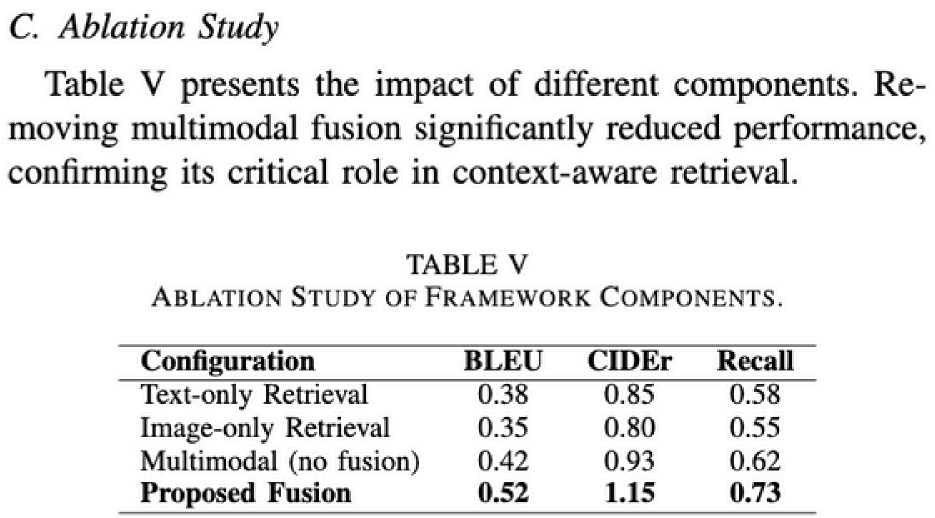

C. *Ablation Study*

Table V presents the impact of different components. Removing multimodal fusion significantly reduced performance, confirming its critical role in context-aware retrieval.

TABLE V
ABLATION STUDY OF FRAMEWORK COMPONENTS.

Configuration	BLEU	CIDEr	Recall
Text-only Retrieval	0.38	0.85	0.58
Image-only Retrieval	0.35	0.80	0.55
Multimodal (no fusion)	0.42	0.93	0.62
Proposed Fusion	**0.52**	**1.15**	**0.73**

Deployment Strategy

Backend Services

Deployed on **Render** with Docker support and auto-scaling capabilities

Frontend UI

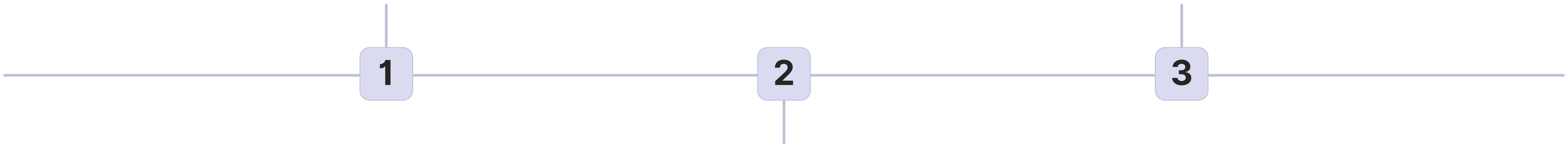
Deployed on **Vercel** with fast CI/CD pipeline and global CDN distribution

Documentation

Hosted on **GitHub Pages** for comprehensive project documentation

Monitoring: UptimeRobot for uptime tracking, GitHub Dependabot for security audits, weekly data backups

Future Scope & Enhancements



Audio Integration

Extend retrieval to include **audio transcripts and sentiment analysis** for comprehensive A/V-Text search capabilities

Generative Capabilities

Add features for **script summarization and scene generation** based on retrieval results using advanced AI

Edge AI Optimization

Model quantization and optimization for deployment on **edge devices** for faster real-time inference

Project Success & Impact

> The **Multimodal Movie Script Search Framework** successfully integrates cutting-edge AI with modern web technologies, providing a novel, context-aware solution to multimodal content retrieval.

Technical Achievement

Working prototype solving contextual retrieval challenges with 83% precision

Industry Impact

Reduced manual annotation costs and improved content monetization for media companies

Research Contribution

Advances in AI/ML, Information Retrieval, and Large Multimodal Models

Thank

YOU

Git Hub link: https://github.com/FenilVadher/Multimodal-Movie-Script-Search-Engine