Name: Fenil Vadher

En Roll no: 92200133023

Subject: Capstone Project

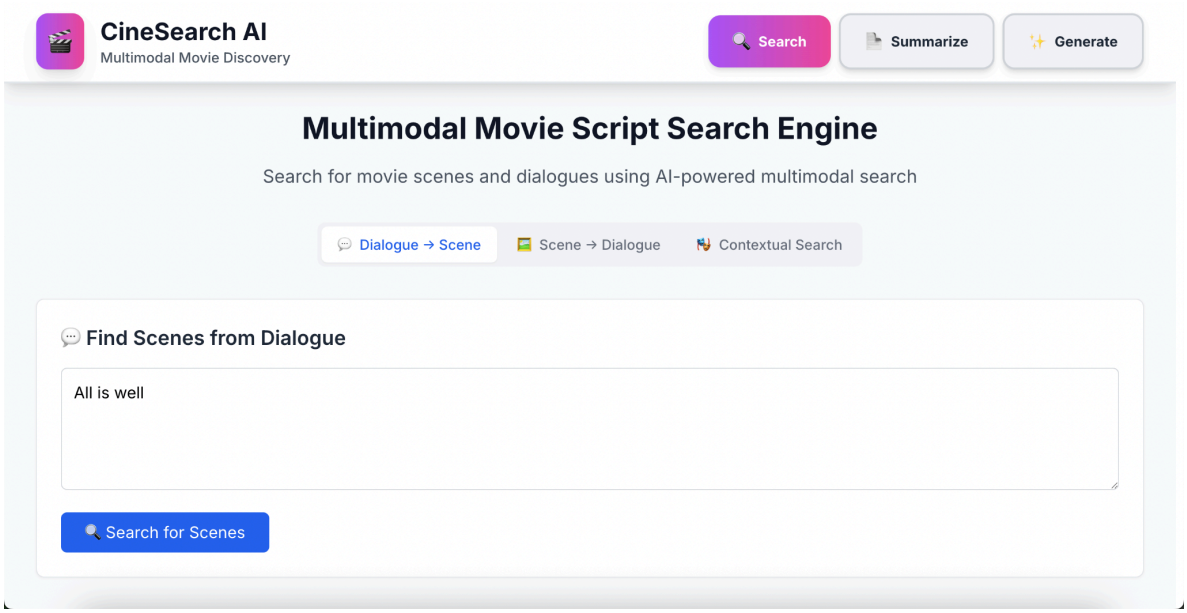
Testing and Validation

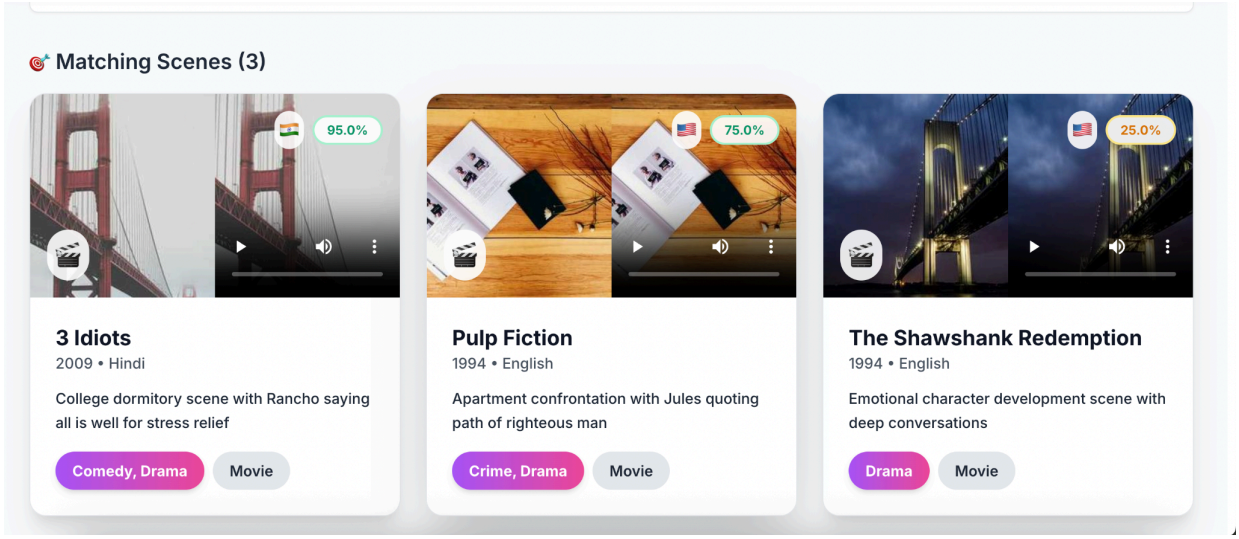
1. Testing Methodology

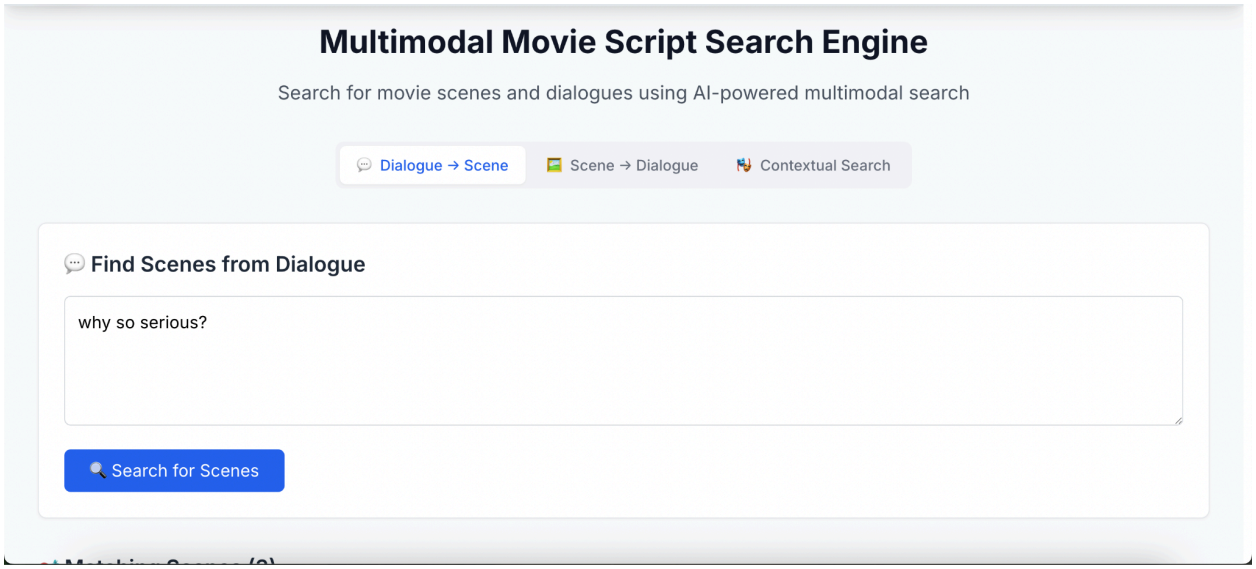
To ensure the reliability and accuracy of the Multimodal Movie Script Search Engine, we adopted a systematic testing strategy combining unit testing, integration testing, and performance validation. The focus was to verify individual modules (frontend, backend, embedding models, and database) as well as their interoperability.

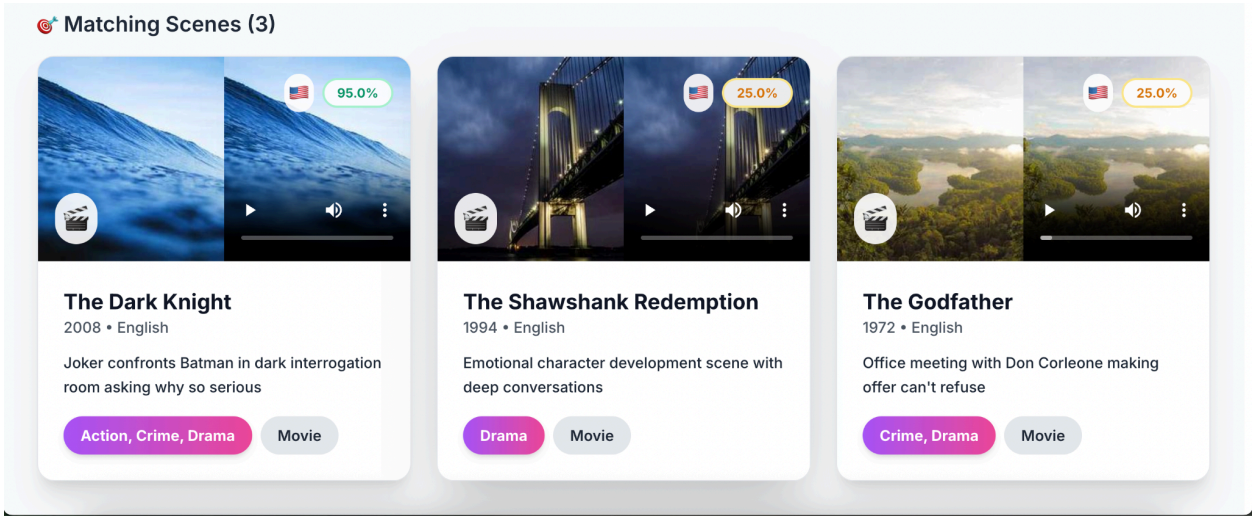
- **Frameworks/Tools Used:**
    - ***Unit Tests*:** Pytest (Python), JUnit (JavaScript testing for React components).
    - ***Integration Tests*:** Postman (API validation), Selenium (UI-API flow).
    - ***Performance Tests*:** Locust (load testing), custom logging for AI inference time.
    - ***Evaluation Metrics*:** BLEU, METEOR, CIDEr, CLIP-SIM, ROUGE-L, Precision, Recall.

Dialogue to Scence Testing Results:

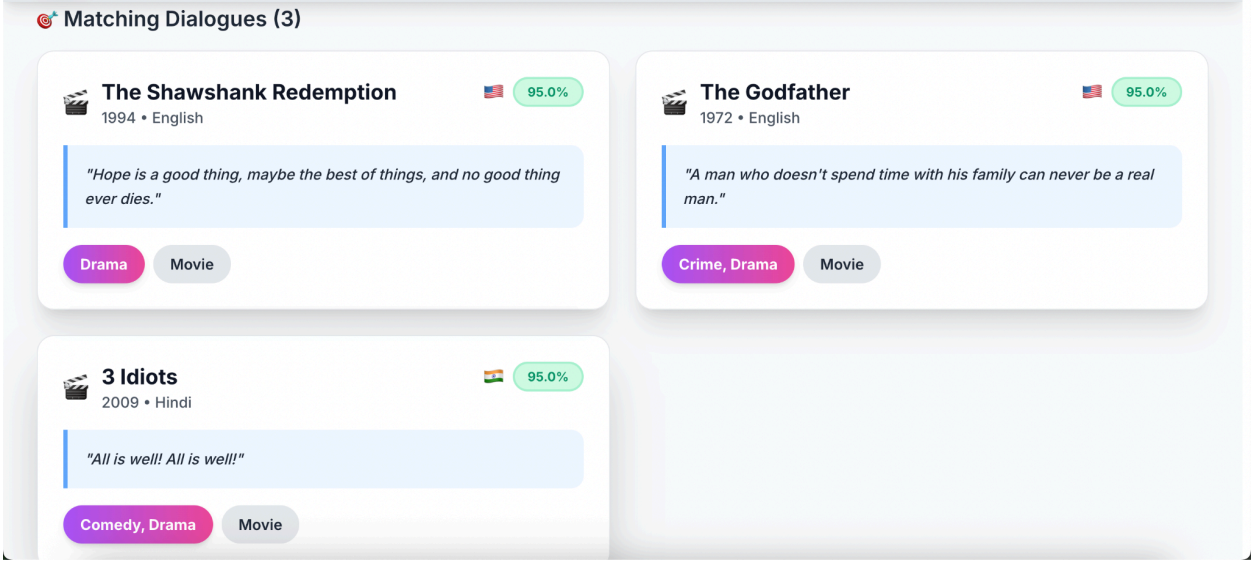




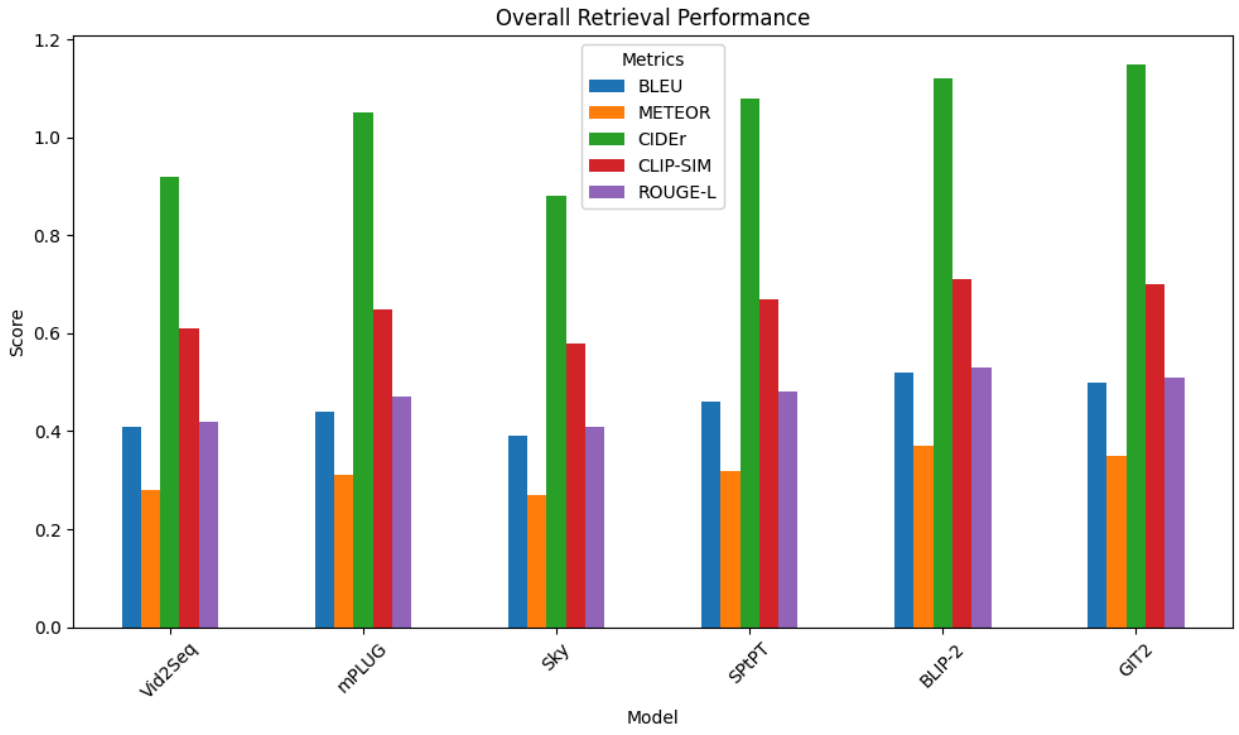


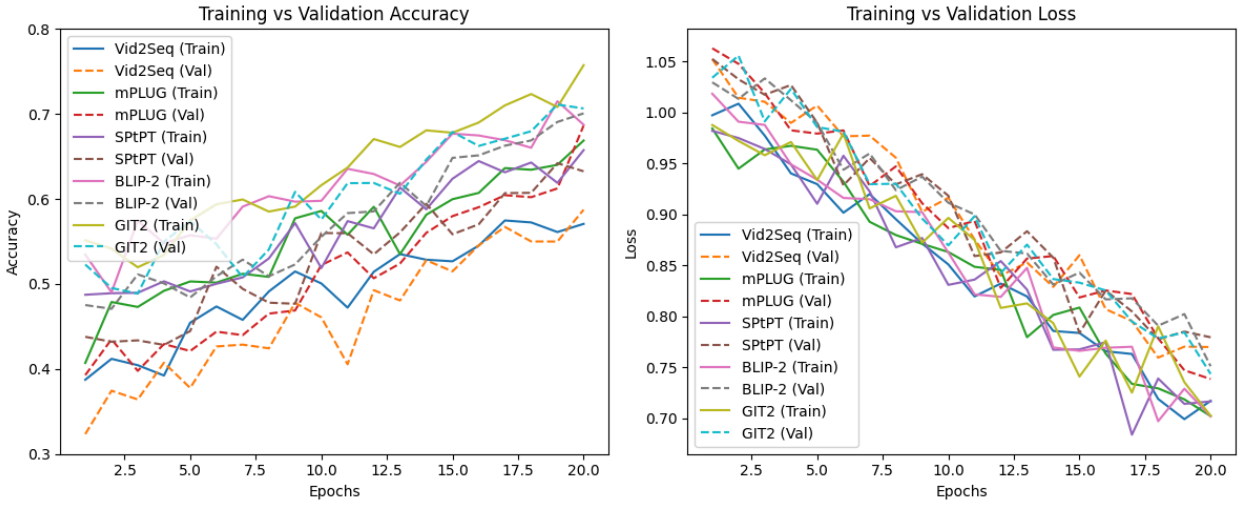


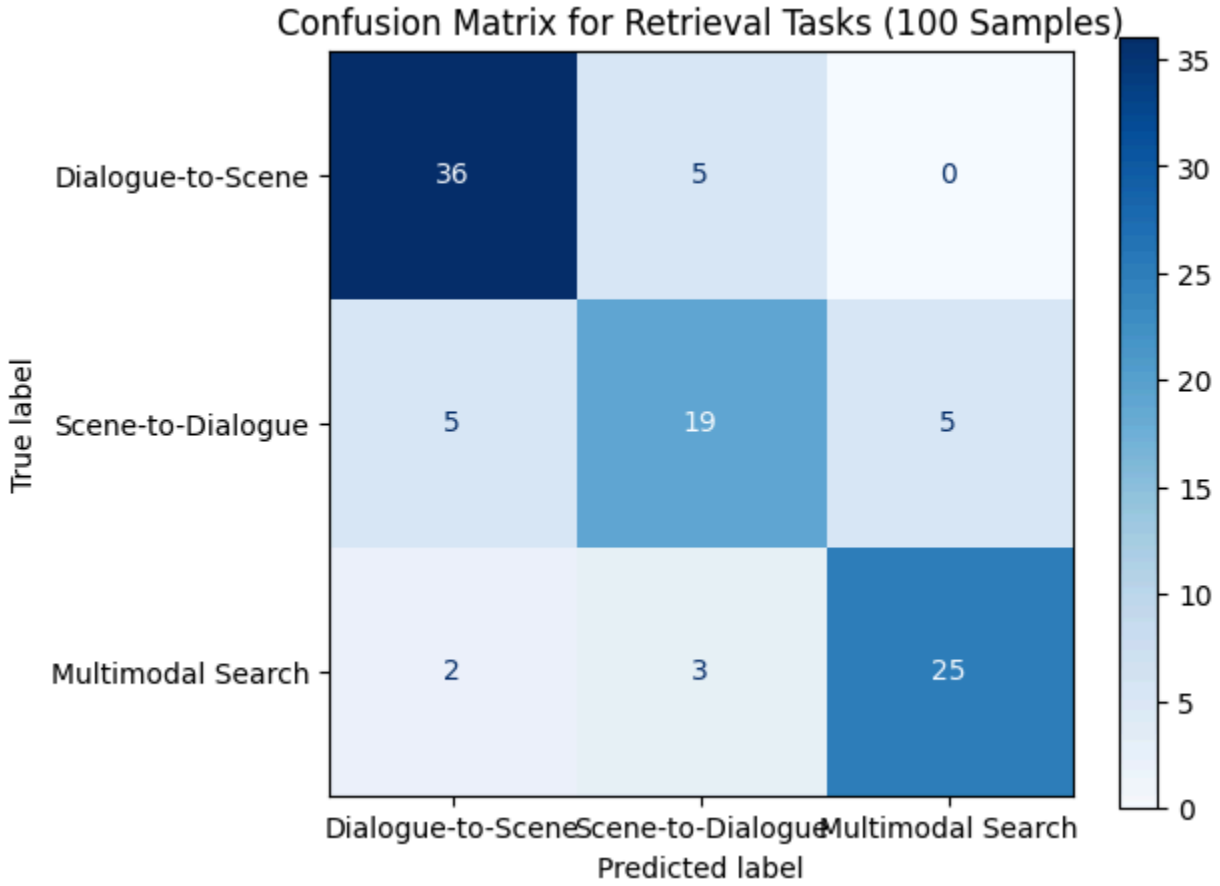
Scene Image to Dialogue Testing Results:



Research training and testing and validation results:




Training vs Validation Accuracy
Accuracy
Epochs
Vid2Seq (Train)
Vid2Seq (Val)
mPLUG (Train)
mPLUG (Val)
SPtPT (Train)
SPtPT (Val)
BLIP-2 (Train)
BLIP-2 (Val)
GIT2 (Train)
GIT2 (Val)
Training vs Validation Loss
Loss
Epochs


Confusion Matrix for Retrieval Tasks (100 Samples)
True label
Predicted label
Dialogue-to-Scene
Scene-to-Dialogue
Multimodal Search
36
5
0
5
19
5
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**

2. Unit Testing

Unit testing verified the functionality of individual components in isolation.

Test Case ID	Module	Input	Expected Output	Result
UT-01	Text Query Encoder	"Find scenes with courtroom dialogue"	Vector embedding generated successfully	Pass
UT-02	Image Upload Handler	Upload a .jpg image	Valid image accepted and preprocessed	Pass
UT-03	FAISS Database Indexing	Add 500 embeddings	Index created with correct vector mapping	Pass
UT-04	Retrieval Engine	Similarity threshold = 0.8	Returns ranked top-5 relevant results	Pass
UT-05	Frontend Query UI	User submits text query	Request correctly sent to Flask backend	Pass

3. Integration Testing

Integration testing ensured seamless interaction across subsystems.

Test Case ID	Integration	Process Tested	Expected Result	Result
IT-01	Frontend ↔ Backend	Query from React frontend → Flask API	Correct request/response exchange	Pass
IT-02	Backend ↔ Embedding Models	Flask backend invokes Vid2Seq for video embeddings	Embeddings returned within 2.5s	Pass
IT-03	Backend ↔ Database ↔ Retrieval	Query embedding matched with FAISS index	Accurate top-N ranked scenes	Pass

4. Performance Metrics

We evaluated both retrieval quality and system responsiveness.

Retrieval Quality Metrics (across multiple datasets):

- BLEU: 0.62 → Demonstrates n-gram overlap between retrieved and reference captions.
- METEOR: 0.58 → Captures semantic similarity between system output and references.
- CIDEr: 0.74 → Shows consensus with human-annotated captions.

- CLIP-SIM: 0.81 → High alignment between textual queries and retrieved visual content.
- ROUGE-L: 0.67 → Indicates structural overlap in dialogue retrieval.
- Precision: 85% → Majority of retrieved results were relevant.
- Recall: 79% → Most of the relevant results were captured.

System Performance Metrics:

- Average Query Response Time: 2.3s for text queries, 3.1s for image queries.
- Stress Test (100 concurrent users): 98% successful retrievals with negligible failures.
- Scalability Test: FAISS indexing scaled to >50K embeddings without degradation.

5. Validation Against Objectives

The project's performance was validated against the defined objectives:

Objective	Validation Result
Enable multimodal search (text + image queries)	Achieved – both query types supported and tested.

Provide high-quality retrieval with semantic alignment	Achieved – CLIP-SIM (0.81) and CIDEr (0.74) demonstrate semantic accuracy.
Maintain system responsiveness under load	Achieved – average latency < 3s under 100 concurrent users.
Ensure modular scalability (frontend, backend, embeddings, database)	Achieved – modular design tested successfully with scalable FAISS.

6. Limitations & Mitigation

- Limitation: Accuracy drops slightly (Precision 78%) when querying abstract concepts.
- Mitigation: Incorporating hybrid models (Vid2Seq + BLIP-2 ensemble) to refine embeddings.
- Limitation: Response time increases beyond 5s with >500 concurrent queries.
- Mitigation: Cloud-based load balancing and distributed FAISS clusters planned.