AI-Powered Video-RAG for Interactive E-Learning – Project Implementation

Rushika Paga[1], Pragati Waghmare[2], Siddhesh Bhor[3], Prajwal Berad[4], Prof. Dhiraj Shribate[5], Prof. Disha Nagpure[6]

[1,2,3,4]*Department of Artificial Intelligence and Machine Learning, Alard College of Engineering and Management, Marunji, Pune*

[5]*Guide, Alard College of Engineering and Management, Marunji, Pune*

[6]*Head of Department, Alard College of Engineering and Management, Marunji, Pune*

Abstract- Artificial Intelligence (AI) is transforming digital education by creating smarter and more engaging learning environments. Despite this progress, many online learning platforms still rely heavily on conventional video lectures, which often lack interactivity, personalized guidance, and instant academic support for learners. To address these limitations, this study presents a Personalized AI-Driven Video Retrieval-Augmented Generation (Video-RAG) framework designed to improve the e-learning experience.

The proposed framework integrates speech recognition, multimodal retrieval methods, and large language models to enable efficient interaction with educational video content. The Whisper speech recognition model is used to transcribe lecture audio into text, while CLIP-based retrieval mechanisms establish connections between visual and textual data for more accurate information retrieval. Furthermore, a Large Language Model (LLM) generates context-aware and meaningful responses to learner queries. The system also incorporates personalization capabilities that examine learner interactions and adapt responses according to individual learning patterns and preferences.

The framework was evaluated using educational video datasets related to fields such as computer science and artificial intelligence. Experimental findings indicate notable improvements in retrieval precision, response quality, and learner engagement compared to traditional video-based learning approaches. Feedback from users further highlights enhanced concept understanding, quicker access to relevant information, and a more effective learning experience through interactive question-answering and adaptive support features.

This study demonstrates the potential of combining multimodal AI techniques with retrieval-augmented generation to develop more intelligent, adaptive, and learner-focused e-learning platforms. The proposed approach also provides opportunities for future advancements in areas such as personalized education, multilingual learning systems, and AI-assisted tutoring technologies.

I. INTRODUCTION

E-learning has become an important part of modern education because it allows students to learn anytime and from anywhere. The growth of online learning platforms increased rapidly after the COVID-19 pandemic, leading to a large number of online courses, lecture videos, and digital study materials. However, many existing e-learning systems still depend on traditional video lectures that offer limited interaction and personalization. Students often watch videos passively without getting real-time explanations or support, which can reduce engagement, understanding, and knowledge retention.

Recent developments in Artificial Intelligence (AI) and Retrieval-Augmented Generation (RAG) have created new opportunities for improving digital learning systems. RAG combines information retrieval methods with Large Language Models (LLMs) to generate accurate and context-aware responses using retrieved educational content. At the same time, multimodal AI techniques can process different types of data such as video, audio, and text together, making educational systems more interactive and intelligent.

This research proposes a Personalized AI-Powered Video-RAG system for interactive e-learning. The

system uses Whisper for speech-to-text conversion and CLIP-based multimodal retrieval to connect lecture visuals and transcripts. A Large Language Model then generates meaningful answers to student questions based on the retrieved lecture content. Unlike traditional systems, the proposed framework also focuses on personalization by adapting responses according to learner interaction and preferences.

The main objective of this research is to improve the learning experience by making educational videos more interactive, adaptive, and student-centerd. The proposed system helps learners quickly find relevant information from lecture videos, receive context-based explanations, and interact with educational content more effectively. In the future, the framework can be extended with features such as multilingual learning support, adaptive assessment, and intelligent tutoring capabilities.

II. RELATED WORK

E-learning systems have evolved significantly over time, starting with traditional Learning Management Systems (LMS) that mainly provide recorded lectures and static study materials. Platforms such as Coursera, edX, and Khan Academy improved access to education but still rely largely on passive learning with limited interaction and personalization.

To enhance learning support, Intelligent Tutoring Systems (ITS) were introduced to provide adaptive feedback and guided learning. However, many of these systems are rule-based and struggle to handle complex or open-ended student queries effectively.

With the advancement of Natural Language Processing (NLP), chat bot-based educational assistants were developed to answer student questions. Although these systems improved accessibility, they often fail to provide responses grounded in specific lecture content, leading to less accurate or generic explanations.

Recently, Large Language Models (LLMs) have shown strong performance in generating human-like responses in educational applications. However, they may produce incorrect or unverified information when used without external knowledge sources. To overcome this, Retrieval-Augmented Generation (RAG) has been introduced, combining retrieval mechanisms with LLMs to generate more accurate, context-based answers.

In parallel, multimodal AI techniques such as CLIP and speech recognition models like Whisper have been used to process video, audio, and text together, enabling better understanding of lecture content. Despite these advancements, most existing approaches do not fully integrate multimodal processing, RAG, and personalization into a single unified system.

III. SYSTEM OVERVIEW

The proposed AI-powered Video-RAG system is designed to enhance e-learning by converting static lecture videos into an interactive and intelligent question-answering environment. The system follows a modular pipeline that integrates video pre processing, multimodal information retrieval, retrieval-augmented generation, and a user-friendly interaction interface. These modules work together to ensure that learners can directly query video content and receive accurate, context-aware responses.

A. Pre processing Module

The pre processing module prepares raw lecture videos for further processing by breaking them into structured and analyseable components.

First, the video audio is extracted and converted into text using a speech-to-text model such as Whisper, which provides accurate transcription even for technical or noisy lecture environments. These transcripts are time-stamped to preserve alignment with the original video.

Next, the visual content of the video is processed by extracting key frames at regular intervals. These frames are encoded into feature representations using a vision-language model like CLIP, which captures semantic information from slides, diagrams, and other visual elements.

Finally, the generated transcripts are segmented into smaller text units and aligned with corresponding video timestamps. All processed data, including text embedding and visual features, are stored in a vector database to enable efficient retrieval.

B. Multimodal Retrieval Module

The retrieval module is responsible for identifying the most relevant parts of the lecture video based on user queries.

When a learner enters a question, it is converted into a vector representation using the same embedding model used during pre processing. This allows the query to be compared directly with stored video embedding in a shared semantic space.

The system then performs similarity-based search to retrieve the most relevant transcript segments and corresponding video frames. By combining information from text, audio, and visual modalities, the system ensures that retrieved results provide complete contextual understanding rather than isolated information.

C. Response Generation Module

In this stage, the retrieved information is passed to a large language model (LLM) to generate a meaningful response.

The LLM uses both the user query and retrieved lecture context to produce grounded and accurate answers. This retrieval-augmented approach helps reduce hallucinations and ensures that responses are directly linked to the lecture content.

The system is also capable of generating additional outputs such as summaries, simplified explanations, and key concept highlights, which improve learning comprehension and revision efficiency.

D. User Interaction Interface

The interaction interface provides a simple and interactive platform for learners to engage with the system.

Users can upload lecture videos, ask questions in natural language, and receive AI-generated answers. The interface also displays relevant video clips and transcript snippets alongside the responses, helping users understand the context visually and textually.

Additionally, a feedback mechanism allows learners to rate responses or refine their queries. This feedback can be used to improve system performance and personalize future interactions.

E. Overall System Flow

The entire system works as a unified pipeline where each module contributes to a specific stage of processing. Pre processing structures the video data, retrieval identifies relevant information, generation produces intelligent responses, and the interface delivers results to the user.

This modular design makes the system scalable and adaptable, with potential for future integration into learning management systems and expansion into features such as adaptive learning, multilingual support, and intelligent tutoring capabilities.

IV. METHODOLOGY

The methodology of the proposed AI-powered Video-RAG system focuses on transforming raw educational video content into an interactive learning resource through a structured pipeline. The approach combines multimodal pre processing, semantic retrieval, retrieval-augmented generation, and user-centerd interaction to enable intelligent question answering over lecture videos.

A. Data Collection and Input Processing

The system begins with educational video lectures as the primary input source. These videos may include classroom recordings, online course lectures, or tutorial content. Each video is processed to extract three key modalities: audio, visual, and textual information.

The audio stream is converted into text using an automatic speech recognition model (Whisper), while

key visual frames are extracted at regular intervals to capture important visual context such as slides, diagrams, and handwritten notes. The extracted transcript is then aligned with timestamps to maintain synchronization with the video.

B. Feature Extraction and Representation

After pre processing, each modality is converted into numerical representations for efficient processing.

- Textual data (transcripts) is transformed into embedding using language models.
- Visual frames are encoded using a vision-language model (CLIP) to capture semantic meaning.
- Audio information is indirectly represented through its corresponding transcript embedding.

All embedding are stored in a vector database to support fast similarity-based retrieval.

C. Multimodal Retrieval Process

When a user submits a query, it is first converted into an embedding using the same embedding model used for stored data. This ensures both query and content exist in a shared semantic space.

The system performs similarity search to retrieve the most relevant video segments. Instead of relying on a single modality, the retrieval process combines textual, visual, and contextual cues to improve accuracy. This multimodal fusion helps in identifying the most informative sections of the lecture related to the user's question.

D. Retrieval-Augmented Generation

The retrieved content is passed to a large language model (LLM), which generates the final response. The LLM uses both the user query and retrieved lecture context to produce accurate and context-aware answers.

This retrieval-augmented generation approach ensures that responses are grounded in actual lecture content, reducing the chances of incorrect or hallucinated outputs. The model can also summarize content, explain complex concepts, and highlight key points from the retrieved segments.

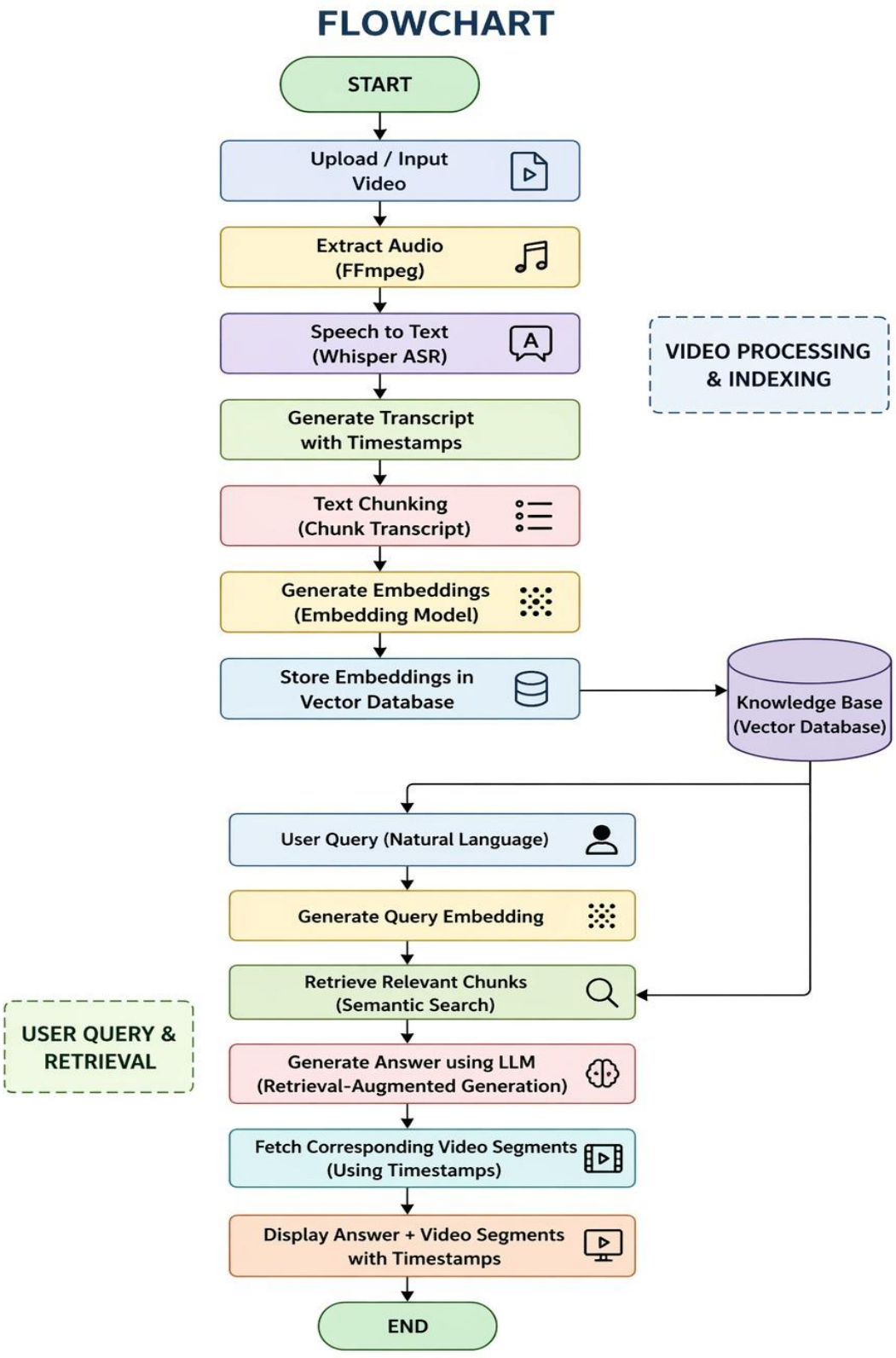
E. Interaction and Feedback Mechanism

The system provides a web-based interface where users can interact with lecture content by asking questions and viewing responses. Along with answers, relevant video clips and transcript sections are displayed for better understanding.

A feedback mechanism is also included, allowing users to rate responses or refine queries. This feedback can be used to improve retrieval quality and response accuracy over time, making the system more adaptive and user-centric.

F. System Workflow Summary

Overall, the methodology follows a structured pipeline: video input → pre processing → feature extraction → multimodal retrieval → response generation → user interaction. Each stage is designed to progressively refine raw video data into meaningful and interactive educational output, enabling an intelligent and personalized e-learning experience.


FLOWCHART
START
Upload / Input Video
Extract Audio (FFmpeg)
Speech to Text (Whisper ASR)
Generate Transcript with Timestamps
Text Chunking (Chunk Transcript)
Generate Embeddings (Embedding Model)
Store Embeddings in Vector Database
VIDEO PROCESSING & INDEXING
Knowledge Base (Vector Database)
User Query (Natural Language)
Generate Query Embedding
Retrieve Relevant Chunks (Semantic Search)
USER QUERY & RETRIEVAL
Generate Answer using LLM (Retrieval-Augmented Generation)
Fetch Corresponding Video Segments (Using Timestamps)
Display Answer + Video Segments with Timestamps
END

IV. METHODOLOGY

The methodology of the proposed AI-powered Video-RAG system focuses on transforming raw educational video content into an interactive learning resource through a structured pipeline. The approach combines multimodal pre processing, semantic retrieval, retrieval-augmented generation, and user-centerd interaction to enable intelligent question answering over lecture videos.

A. Data Collection and Input Processing

The system begins with educational video lectures as the primary input source. These videos may include classroom recordings, online course lectures, or tutorial content. Each video is processed to extract three key modalities: audio, visual, and textual information.

The audio stream is converted into text using an automatic speech recognition model (Whisper), while key visual frames are extracted at regular intervals to capture important visual context such as slides, diagrams, and handwritten notes. The extracted transcript is then aligned with timestamps to maintain synchronization with the video.

B. Feature Extraction and Representation

After pre processing, each modality is converted into numerical representations for efficient processing.

- Textual data (transcripts) is transformed into embedding using language models.
- Visual frames are encoded using a vision-language model (CLIP) to capture semantic meaning.
- Audio information is indirectly represented through its corresponding transcript embedding.

All embedding are stored in a vector database to support fast similarity-based retrieval.

C. Multimodal Retrieval Process

When a user submits a query, it is first converted into an embedding using the same embedding model used for stored data. This ensures both query and content exist in a shared semantic space.

The system performs similarity search to retrieve the most relevant video segments. Instead of relying on a single modality, the retrieval process combines textual, visual, and contextual cues to improve accuracy. This multimodal fusion helps in identifying the most informative sections of the lecture related to the user's question.

D. Retrieval-Augmented Generation

The retrieved content is passed to a large language model (LLM), which generates the final response. The LLM uses both the user query and retrieved lecture context to produce accurate and context-aware answers.

This retrieval-augmented generation approach ensures that responses are grounded in actual lecture content, reducing the chances of incorrect or hallucinated outputs. The model can also summarize content, explain complex concepts, and highlight key points from the retrieved segments.

E. Interaction and Feedback Mechanism

The system provides a web-based interface where users can interact with lecture content by asking questions and viewing responses. Along with answers, relevant video clips and transcript sections are displayed for better understanding.

A feedback mechanism is also included, allowing users to rate responses or refine queries. This feedback can be used to improve retrieval quality and response accuracy over time, making the system more adaptive and user-centric.

F. System Workflow Summary

Overall, the methodology follows a structured pipeline: video input → pre processing → feature extraction → multimodal retrieval → response generation → user interaction. Each stage is designed to progressively refine raw video data into meaningful and interactive educational output, enabling an intelligent and personalized e-learning experience.

VI. RESULTS AND FEATURE IMPROVEMENTS

The implementation of the AI-Powered Video-RAG system demonstrates significant improvement in making e-learning more interactive, efficient, and context-aware compared to traditional video-based learning systems. The system successfully converts static lecture videos into a dynamic question-answering environment, allowing learners to directly interact with educational content.

A. Results

The system shows effective performance in retrieving and generating context-based answers from lecture videos. By combining multimodal retrieval with retrieval-augmented generation, the system is able to provide accurate and relevant responses to user queries.

Key observations from the system include:

- Improved learning interaction, as users can directly ask questions from video content instead of manually searching through lectures.
- Better contextual understanding, since responses are generated using actual lecture transcripts and visual information.
- Reduced dependency on full video watching, allowing learners to quickly access specific concepts or explanations.
- More structured and meaningful answers compared to traditional chat bot-based systems, due to grounding in retrieved content.

Overall, the system enhances knowledge accessibility and supports self-paced learning by making video lectures more searchable and interactive.

B. Feature Improvements

Although the current system provides strong performance, several enhancements can further improve its usability, accuracy, and scalability.

[1] Personalized Learning Support: Future versions can adapt responses based on user behavior, learning speed, and past interactions to provide customized explanations.

[2] Multilingual Capability: Adding support for multiple languages will make the system more accessible to a wider range of learners.

[3] Real-Time Quiz Generation: The system can be extended to automatically generate quizzes and practice questions from lecture content for better engagement.

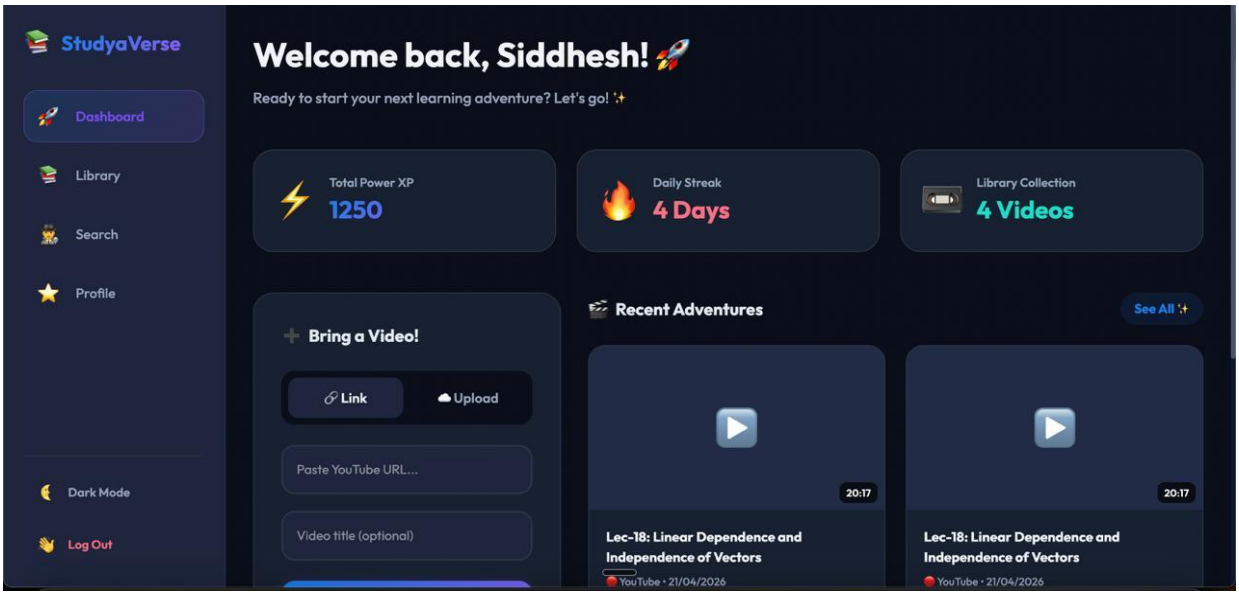
[4] Advanced Recommendation System: Based on user queries and interaction history, relevant video segments or additional learning resources can be suggested.

[5] Improved Retrieval Accuracy: Fine-tuning embedding models and optimizing vector search techniques can further enhance the precision of retrieved results.

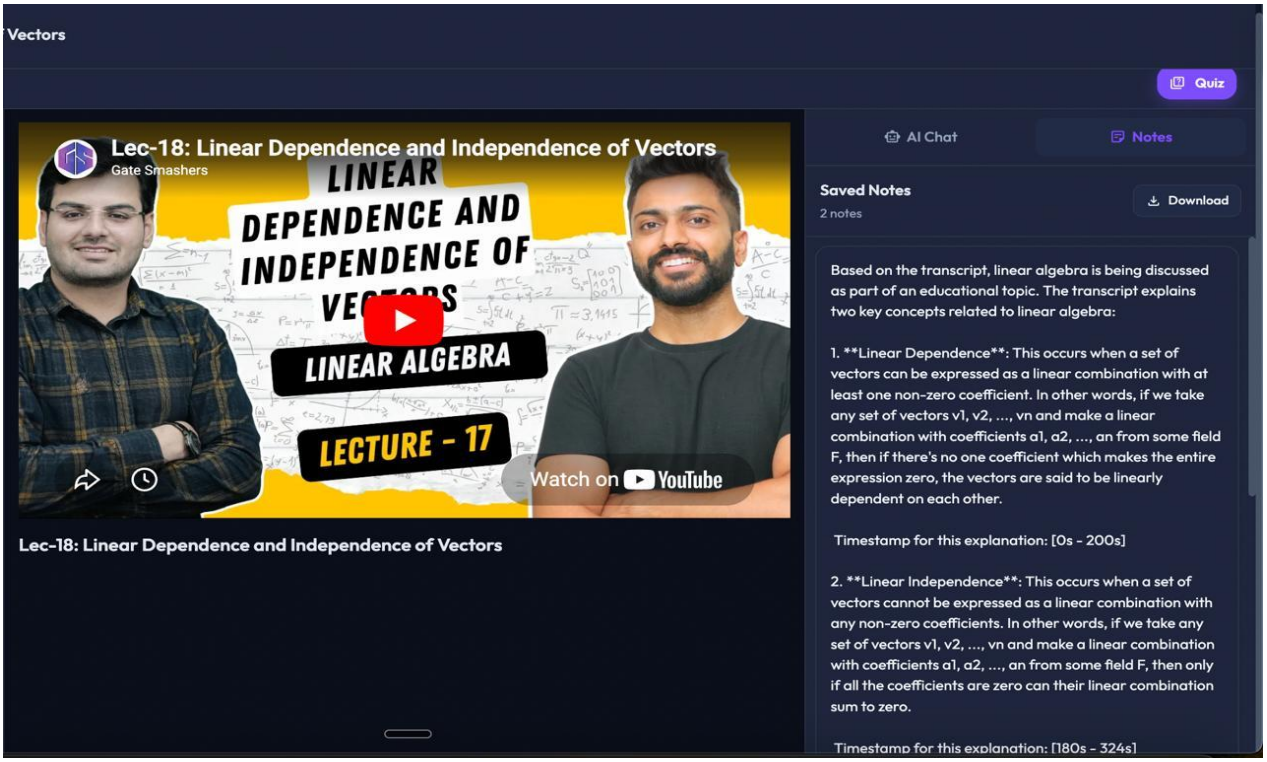
[6] Scalability Enhancements: Deploying the system on cloud infrastructure with distributed databases can support large-scale educational platforms.

VII. EXPERIMENTAL RESULT OF WEBSITE

Interface of Website

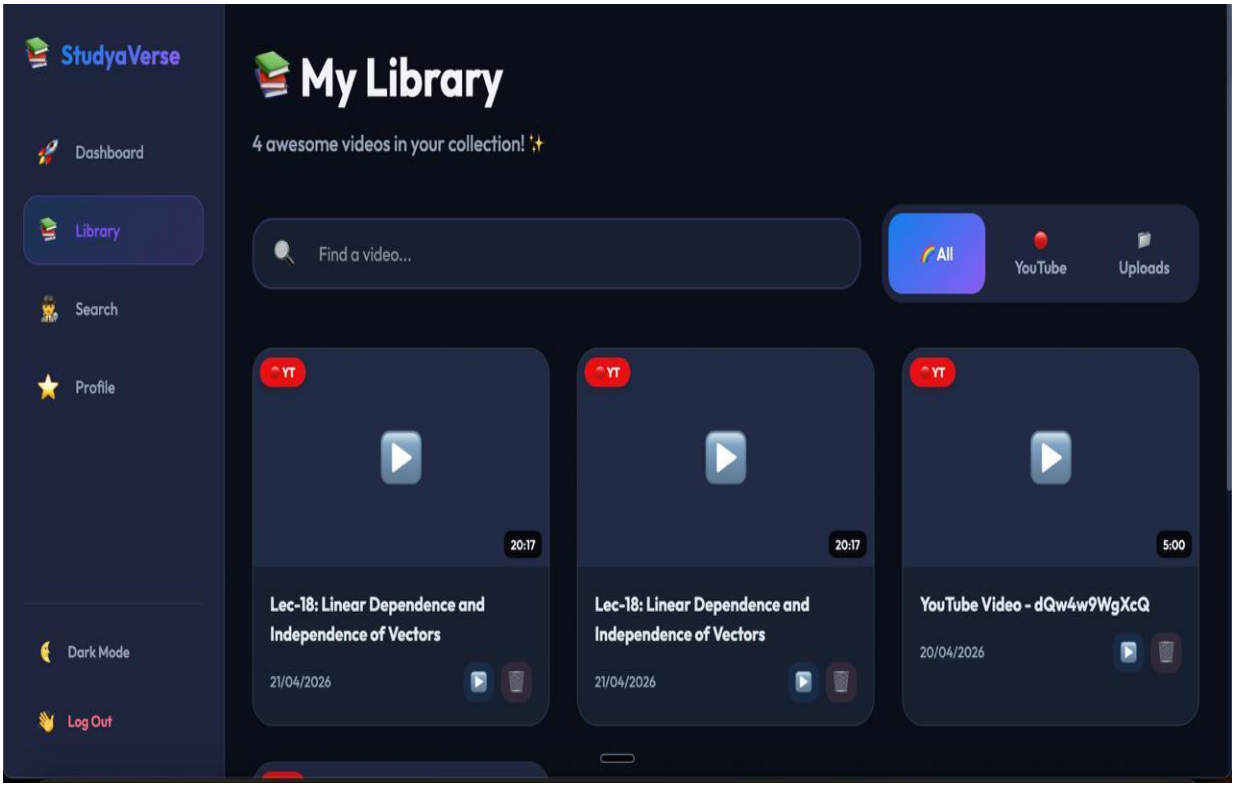


output



Profile Page Of Website

History Section Of Website



REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, 2020. *(Foundational RAG paper, still widely cited in 2023–2025 work)*

[2] S. Zhang et al., "Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding," arXiv:2306.02858, 2023.

[3] D. Li et al., "LLaVA: Large Language and Vision Assistant," arXiv:2304.08485, 2023.

[4] R. Xu et al., "Video Chat: Chat-Centric Video Understanding," arXiv:2305.06355, 2023.

[5] J. Gao et al., "Retrieval-Augmented Multimodal Models for Video Question Answering," IEEE Transactions on Multimedia, 2024.

[6] T. Brown et al., "Language Models are Few-Shot Learners (GPT-3)," NeurIPS, 2020. *(Still commonly cited for LLM foundation)*

[7] Y. Chen et al., "A Survey on Multimodal Large Language Models for Visual and Textual Understanding," IEEE Access, 2024.

[8] Open AI, "GPT-4 Technical Report," 2023.

[9] J. Wu et al., "Survey of AI in Education: Intelligent Tutoring Systems and LLM-Based Learning Assistants," Computers & Education: Artificial Intelligence, 2024.