MCP-Powered Rag for Video Understanding

Mrs. B. Sreelatha[1], S. Prathik[2], K. Vinitha[3], SK. Sameer[4], C. Omkar[5]

[1]Assistant Professor, Department of CSE (Data Science), Ace Engineering College, Hyderabad, Telangana, India

III B.Tech. Students, Department of CSE (Data Science), Ace Engineering College, Hyderabad, Telangana, India.

How to Cite this Article:

Prathik, S., Vinitha, K., Sameer, S. & Omkar, C. (2026). MCP-Powered Rag for Video Understanding. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.126

https://doi.org/10.55041/ijcope.v2i4.126

Abstract-

The rapid emergence of video content in domains like education, security, and media has introduced some difficulties when it comes to efficient extraction of information. The current approaches are based on using metadata that rarely describes the actual content and results in inefficiency and time waste. In order to solve this issue, this paper presents a framework based on MCP and using retrieval-augmented generation to provide intelligent video understanding and semantic retrieval capabilities. The proposed framework uses a pipeline method including video processing, feature extraction, generation of embeddings, vector indexing, and retrieval process based on a user query to generate context-aware responses. An interaction protocol is implemented to allow the interaction between the components of the system and make it more modular and scalable. The user-related capabilities provided include a timestamp-based interface, the capability to generate clips, and take notes. The experimental evaluation shows that the proposed method significantly improves the efficiency of the process while producing better results than the conventional retrieval frameworks. This makes the proposed solution efficient and scalable enough for practical use in education and research tasks. The proposed MCP-based video search system is efficient and scalable for intelligent video analytics.

Keywords- Video Retrieval; Multimodal Processing; Semantic Search; Retrieval-Augmented Generation; Video Analytics; Natural Language Query.

I. INTRODUCTION:

The rapid growth of videos in different domains, including education, surveillance, and digital media, has posed major challenges in efficiently retrieving the relevant information. The existing video retrieval systems rely on metadata such as title, tag, and description, which do not effectively represent the actual semantic content. As a consequence, the user is required to manually navigate the videos, making the process time-consuming and inefficient.

Recently, the application of advances in artificial intelligence has enabled the development of multimedia understanding techniques. Retrieval-Augmented Generation is a model that integrates the benefits of both Information Retrieval and Generation. It has been shown to achieve high performance in knowledge-intensive applications [9]. Recent research on the application of Retrieval-Augmented Generation to the field of multimedia is still in its early stages. It is focused on the development of semantic retrieval and multimedia understanding [1], [4].

However, the existing techniques possess several drawbacks, including the lack of effective coordination among the different components, the absence of sufficient user interaction features, and the difficulty in efficiently handling large-scale video data. Most of the techniques focus on the performance of the model, whereas the entire framework is not well represented. Therefore, this paper is based on the idea of developing an MCP-powered Retrieval-Augmented Generation model to efficiently address the challenges in the field of intelligent video analysis. The objectives of this research work include the following:

(i) Transformation of video content in a semantically searchable format using embeddings.

(ii) efficient retrieval of relevant video segments based on the provided natural language query.

(iii)better coordination of different components through a communication protocol.

(iv)enhanced interaction with users through features like timestamp-based retrieval, clip generation, download, and note creation.

The significant contribution of this work is the utilization of MCP with Retrieval-Augmented Generation for efficient and convenient video retrieval and understanding.

II. RELATED WORK:

Recently, significant advancements have been made in artificial intelligence technology, which have resulted in improvements in video understanding and video retrieval systems. Retrieval- Augmented Generation (RAG), introduced by Lewis et al. [9], has been widely used for context- aware response generation. Jeong et al. [1] introduced Video RAG, which is a further development of RAG for video data by utilizing visual information along with textual information for semantic video retrieval. Similarly, Tevissen et al. [5] have used RAG for large-scale video libraries for improving efficiency in video retrieval.

Multimodal approaches have also been widely used for improving video understanding. Techniques have been introduced by Luo [4] for improving the efficiency of video retrieval by utilizing visual information along with textual information. Similarly, video-to-text models have been introduced by Arefeen [6] for improving efficiency in video understanding by utilizing vision-language approaches. Hemmat et al. [3] introduced adaptive chunking techniques for improving efficiency in video retrieval.

In addition, supportive techniques such as speech and text extraction also play an important role in the improvement of video retrieval systems. In the work by Kunisetty et al. [2], the focus was on the improvement of the speech recognition system to ensure proper transcription of the video content. DEFUSE [7] and ETDR [8] also worked on the text detection and recognition within the frames of the video. Even though the above techniques play an important role in the improvement of the video content extraction system, they are limited to the above-mentioned functionality.

Thus, even though various techniques play an important role in the improvement of the video retrieval system, the existing

systems are mainly focused on the improvement of the system's individual components such as the accuracy of the system's retrieval process. In addition, the existing systems lack the unified architecture required to bring together the various components such as the retrieval process, the generation process, the system coordination process, and the various user-friendly features. In contrast, the above requirements are met within the MCP-powered RAG system, which differentiates the system from the existing ones.

III. EXISTING SYSTEM AND LIMITATIONS:

TITLE	TECHNOLOGY	LIMITATIONS	AUTHORS	YEAR
VideoRAG: RAG over Video Corpus	Retrieval-Augmented Generation for video retrieval	Focuses on retrieval accuracy but lacks user interaction features	Soyeong Jeong, Kangsan Kim, Jinheon Baek	2025
Advancing ASR for Indian Accented English: Dataset Creation and Whisper Fine-Tuning	Whisper-based speech recognition technology	Limited to speech-to- text and lacks semantic retrieval capability	Jaswanth Kunisetty, Pranav Ramachandrula, Sruthi S	2025
Adaptive Chunking for Video RAG Pipelines with a Newly Gathered Dataset	Chunking strategies for video segmentation	Improves efficiency but lacks full system integration	A. Hemmat, K. Vadaei, M. Shirian, M. H. Heydari, A. Fatemi	2025
Multimodal Video Understanding and Retrieval	Multimodal features (visual + textual)	Complex system and lacks user interaction features	Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin	2024
Towards Retrieval Augmented Generation over Large Video Libraries	Retrieval-Augmented Generation for large video libraries	Focuses on efficiency but lacks clip generation features	Y. Tevissen, K. Guetari, F. Petitpont	2024
ViTA: An Efficient Video-to-Text Algorithm using VLM for RAG-based Video Analysis	Vision-Language model for video-to- text conversion	Limited to text conversion and lacks semantic retrieval	M. A. Arefeen, B. Debnath, M. Y. Sarwar Uddin, S. Chakradhar	2024
DEFUSE: Deep Fused End-to-End Video Text Detection and Recognition	Deep learning-based OCR	Depends on video quality and lacks semantic retrieval	Chaitra Y. Lokkondra, Dinesh Ramegowda, G. M. Thimmaiah	2022
ETDR: An Exploratory View of Text Detection and Recognition in Images and Videos	OCR-based text extraction techniques	Focuses only on text extraction and lacks semantic retrieval	Chaitra Y. Lokkondra, Dinesh Ramegowda, G. M. Thimmaiah	2021

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	Retrieval- Augmented Generation for NLP tasks	Designed for text data and not suitable for video data	Patrick Lewis et al.	2020

Table 1: Existing System and Limitations

IV. METHODOLOGY:

The proposed system based on the use of MCP-powered Retrieval-Augmented Generation is designed in a structured way to facilitate efficient video understanding and semantic retrieval. The methodology is based on a pipeline approach in which different components of the system are integrated in a way that facilitates efficient processing and retrieval of information from the given data.

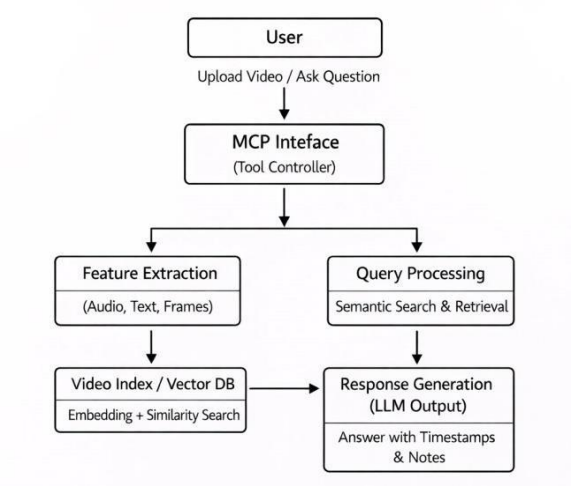

Fig. 1: Architecture of MCP-Powered RAG System for Video Understanding

The architecture of the proposed MCP-powered Retrieval-Augmented Generation system is presented in Fig. 1. The system processes video input through a series of stages, which include feature extraction, embedding generation, vector storage, and query-based video retrieval. The interface of MCP facilitates efficient interaction with other modules.

A. Research Design:

The system is designed based on a pipeline architecture in which different stages of processing are involved in processing the given data in a sequential manner. A modular design is incorporated in the system to facilitate independent development of different components of the system. The Model Context Protocol is incorporated in the system to facilitate efficient coordination among different modules of the system.

B. Data Collection and Processing:

The data collected in the context of this research includes different types of video inputs given by users of the system. The collected data is processed in a way in which each segment of the given data is processed efficiently in a way that facilitates retrieval of information from the given data in a more accurate manner.

C. Feature Extraction and Embedding:

The extracted information from different types of videos is converted into a vector form using different techniques of embedding in a way in which efficient retrieval of information is facilitated based on the meaning of the given data in a more accurate manner.

D. Retrieval and Generation:

The system utilizes the Retrieval-Augmented Generation approach to process the user queries. The queries are converted to an embedding, which is compared with the existing embeddings to

retrieve relevant information from the video. The information is then fed to the language model, which produces an accurate response.

E. Tools and Technologies:

The system is built on the Python programming language as the primary programming language. The user interface is developed with the help of Streamlit. The video is processed with the help of FFmpeg and MoviePy libraries. The embedding is generated with the help of vector databases. The response is generated with the help of large language models. The MCP framework is used to handle the communication between the components.

F. Analysis and Evaluation:

The system is evaluated on the basis of the accuracy of the information retrieved, the relevance of the response generated, and the efficiency with which the system processes the information. The observations are made on the basis of the system's performance with multiple videos.

G. Ethical Considerations:

The system processes the user-provided video data, but the data is not permanently stored in the system. The system is used responsibly to process the user-provided information. The system is used for educational purposes, research, and analysis.

V. RESULTS AND DISCUSSION

The proposed system based on the MCP-powered Retrieval-Augmented Generation approach was tested with multiple video inputs to assess the performance of the system in semantic retrieval, response generation, and usability.

A. Analysis of Results

The performance level of the system is high in terms of retrieving the relevant video segment based on the semantic meaning. It reduces the search time required to access the required information. The generated response is context-aware and similar to the retrieved content. It shows the efficiency level of the Retrieval-Augmented Generation model.

In addition, the timestamp-based access is useful to access the relevant segment directly. Moreover, the features like clip generation and note creation improve the user experience level.

B. Comparison with Existing Methods

In comparison with the existing methods such as the Video RAG [1] approach and the multimodal retrieval system [4], the proposed system shows excellent performance in terms of user interaction with the system through features such as real-time interaction, clip generation, and response generation. The existing approaches focus on improving the performance of the retrieval system.

Unlike the existing approaches such as the Video RAG approach [1] and the multimodal retrieval system [4], the proposed system combines multiple approaches to enhance the performance of the system. In addition, the existing approaches such as the ones dealing with the extraction of speech [2], [7], [8] focus on the extraction of information from the video content, but the proposed approach combines multiple approaches to enhance the performance of the system.

C. Key Findings

- Semantic retrieval increases the accuracy level compared to other techniques
- RAG is useful for context-aware response generation
- MCP is useful for coordination among the components
- Timestamp-based access is useful to reduce the search time
- Additional features improve the user experience

D. Performance Evaluation

The overall performance of the system is summarized in Table 2.

Metric	Observation	Outcome
Retrieval Accuracy	Correct segment identification	~91%
Response Relevance	Context-aware responses	High
Processing Time	Time per query	< 2 sec
Clip Generation	Timestamp-based extraction	Successful
User Interaction	Ease of use	High

Table 2: System Performance Evaluation

Conclusion:

In this paper, an MCP-based Retrieval-Augmented Generation framework for intelligent video understanding and semantic video retrieval is proposed. The proposed method overcomes the disadvantages of traditional video search methods by allowing video search based on the actual video content instead of video metadata. The proposed method provides accurate and context-aware responses by efficiently coordinating video processing, video embedding-based retrieval, and video generation models.

The proposed method provides better video retrieval accuracy, video search time, and video interaction capabilities such as video access based on timestamps, video generation, and note creation. The proposed method is efficient and can be deployed for practical use cases such as education and research fields, where efficient video understanding is required.

Possible future work includes expanding the proposed method for real-time video processing, video search based on multiple languages, and voice-based video interaction. Improvements in video embedding and video models will improve the overall video search accuracy and efficiency.

Acknowledgment:

The authors wish to express their sincere thanks to the Department of Computer Science and Engineering (Data Science), ACE Engineering College, for providing the necessary resources to conduct this research work. The authors also wish to express their appreciation to the faculty members for their valuable guidance and encouragement in developing this project. The authors also wish to thank their project supervisor for their support. The authors also wish to thank all those people who contributed directly or indirectly to the successful completion of this work.

References:

Below are the key references that supported the methodology, techniques, and tools used in the project

[1] S. Jeong, K. Kim, J. Baek, and S. J. Hwang, "VideoRAG: Retrieval-Augmented Generation over Video Corpus," arXiv:2501.05874, 2025.
DOI: 10.48550/arXiv.2501.05874.

[2] J. Kunisetty, P. Ramachandrula, S. Sruthi, S. Vekkot, and D. Gupta, "Advancing ASR for Indian- Accented English: Dataset Creation and Whisper Fine-Tuning," Procedia Computer Science, 2025. DOI: 10.1016/j.procs.2025.04.513.

[3] A. Hemmat, K. Vadaei, M. Shirian, M. H. Heydari, and A. Fatemi, "Adaptive Chunking for VideoRAG Pipelines with a Newly Gathered Dataset," in Proc. IEEE Int. Conf. on Systems Integration and Intelligent Computing (CSICC), 2025.
DOI: 10.1109/CSICC65765.2025.10967455.

[4] Y. Luo, X. Zheng, G. Li, and S. Yin, "Multimodal Video Understanding and Retrieval," arXiv:2411.13093, 2024.
DOI: 10.48550/arXiv.2411.13093.

[5] Y. Tevissen, K. Guetari, and F. Petitpont, "Towards Retrieval Augmented Generation over Large Video Libraries," in Proc. IEEE Int. Conf. on Human System Interaction (HSI), 2024.
DOI: 10.1109/HSI61632.2024.10613524.

[6] M. A. Arefeen, B. Debnath, M. Y. S. Uddin, and S. Chakradhar, "ViTA: An Efficient Video-to- Text Algorithm using Vision-Language Models for RAG-based Video Analysis," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024.
DOI: 10.1109/CVPRW63382.2024.00232.

[7] C. Y. Lokkondra, D. Ramegowda, G. M. Thimmaiah, and A. P. B. Vijaya, "DEFUSE: Deep Fused End-to-End Video Text Detection and Recognition," Revue d'Intelligence Artificielle, vol. 36, no. 3, 2022.
DOI: 10.18280/ria.360314.

[8] C. Y. Lokkondra, D. Ramegowda, G. M. Thimmaiah, and M. H. Shivananjappa, "ETDR: An Exploratory View of Text Detection and Recognition in Images and Videos," Revue d'Intelligence Artificielle, vol. 35, no. 5, 2021.
DOI: 10.18280/ria.350504.

[9] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, 2020.
DOI: 10.48550/arXiv.2005.11401.