International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A SURVEY ON MCP-POWERED RAG OVER VIDEOS

[1st]Muhammad Abul Kalam, [2nd]Kollimarla Naga Sai Satya Teja, [3rd] Mudavath Bharath Kumar, [4th]Prabhugari Mallikarjun
[1] *CSE(AI & ML), Assistant Professor at ACE Engineering College, Hyderabad, India*
[2,3,4]*CSE(AI & ML), Student at ACE Engineering College, Hyderabad, India*

Abstract – The current research proposes a novel MCP-powered RAG over videos (video-RAG) model which facilitates interactive engagement with video content using natural language queries. The method ingests, analyzes, and indexes the video dataset so that it can perform semantic search and answer questions accurately through the use of MCP along with the Regie plugin for video retrieval and Cursor for acting as the MCP host in order to facilitate structured and intelligible answers. The retrieval process involves extracting contextually pertinent video clips with specific citations which include timestamp references that correspond to the requested information, thus reducing the need to manually navigate videos to extract information.

Index Terms— Model Context Protocol (MCP), Video Retrieval-Augmented Generation (Video-RAG), Multimodal Retrieval, Video Question Answering, Semantic Video Search, Tool-based Large Language Models, Citation-driven AI, Conversational Video Analytics

I. INTRODUCTION

AI-Powered Meme Generator for Company Advertisements is an innovative web-based tool that increases digital marketing efficiency through intelligent automation. In today's social media era, memes have emerged as a potent tool for brands to reach out to their customers in a humorous and relatable manner. This tool leverages sophisticated AI and NLP algorithms such as Llama 3 8B Instruct to create innovative and relevant captions from the user-input topics, slogans, or product descriptions. It matches these captions with appropriate meme templates or enables the user to choose them, using image processing to produce high-quality output that increases brand visibility and social media presence.

In today's digital age, social media has become one of the most popular and influential platforms for marketing and brand promotion. In today's world, businesses use creative visual content to capture the attention of the audience and communicate their message effectively. Among these, memes have become a popular means of communication because of their humor, relatability, and shareability. Memes not only entertain people but also help brands in a unique way to reach the audience and increase their online visibility.

However, to make such memes on a regular basis, one needs to be quite creative and have a good understanding of trends and preferences. Many organizations face challenges in developing relevant and engaging content that resonates with their brand voice. This gives rise to the need for an intelligent system that can automate the meme development process without compromising the accuracy of context, humor, and aesthetics. Artificial Intelligence (AI) technology has emerged as a promising solution to meet this demand efficiently. The final objective of this project is to assist organizations in enhancing their social media interactions and brand presence. The developed memes can be downloaded or directly shared on different platforms, making marketing faster, more efficient, and more effective. With the increasing need for innovative advertising approaches, this AI-powered meme generator is an exciting solution that combines automation with creativity to help brands reach their audiences in a more relatable and memorable manner.

A. Challenges with Video Content Retrieval

There exist various difficulties associated with video content that make their retrieval rather challenging. Firstly, unlike texts, videos represent a constant flow of data both in terms of sounds and visual content that needs to be processed before extracting valuable information from them. Secondly, traditional techniques of searching for videos based on metadata such as titles and tags may prove insufficient since they do not provide comprehensive coverage of a piece of video content.

Besides, while transcription may provide some benefits in searching, it is also dependent on accurate recognition of speech and thus prone to various errors. Lack of punctuations and inability to convey tone or any visual information makes this approach rather ineffective for retrieval purposes. In addition, keyword-based techniques of searching are likely to miss the point of query due to inability to understand its semantic content, especially in case of large repositories of videos.

The most crucial issue associated with video content, however, seems to be absence of an effective navigation mechanism that would help to find a needed part of a video without watching the whole piece.

B. The Importance of Semantic and Conversational Video Search

Due to the exponential growth in the amount of available video data, the demand for advanced techniques that would allow retrieving information using more advanced methods than simple keyword searches increases. Semantic search is one of the solutions to the problem that enables systems to interpret the meaning of the question and intentions behind it. This means that the systems will be able to retrieve data not based on keywords but on their meanings.

Another important feature that modern information systems possess is the ability to work with conversational interfaces. Thanks to this technology, people can easily communicate with machines by posing natural language questions to them. When used for video retrieval purposes, this feature will enable users to communicate with video data as if talking to an expert.

Semantics and conversational interfaces are two very useful capabilities that can greatly enhance the efficiency of information retrieval and make it more engaging for users.

3. Motivation for MCP-powered Video-RAG System

The main motivation for developing this system is to address the issues that currently exist in video retrieval systems by implementing more advanced approaches and technologies. These systems lack the ability to comprehend the user request, make context-based replies and provide an efficient access to relevant video segments.

With the implementation of Retrieval Augmented Generation (RAG), the developed application will be able to search for relevant video fragments as well as provide adequate answers on the basis of their analysis. In addition, the introduction of the Model Context Protocol (MCP) allows for more effective communication with other services and applications in a structured way.

Furthermore, the ability to cite sources used to generate an answer can be considered as one of the main motivations for developing this system. This is very useful for applications requiring a high level of reliability and accuracy of generated information, especially for such fields as education and enterprise environments.

In general, this paper proposes an innovative approach to creating a more powerful video-based interface for generating and retrieving responses to user requests.

D. Technology Stack and System Overview

Integration with the Model Context Protocol (MCP) makes it possible for the system to be used as a tool in AI-powered platforms like Cursor, which enables

developers and users to work with video data without leaving their environment. In addition, tools like MoviePy are utilized for extracting timestamped video clips, thereby ensuring that users get access to specific segments without putting in any extra effort.

The entire system design is based on a structured pipeline involving video ingestion, semantic indexing, retrieval, response generation, and video clip extraction.

The integration of the system with the Model Context Protocol (MCP) enables its operation as a tool in AI environments like Cursor, making it possible for developers and users to work with video content within their processes. Further, software like MoviePy is employed for generating timestamped video clips, allowing the users to gain access to specific video segments without any extra effort.

In terms of the system architecture, the entire process flow involves video ingestion, semantic indexing, retrieval, response generation, and clipping video segments.

II. LITERATURE SURVEY

The development of Artificial Intelligence in recent years has greatly enhanced the capacity of multimedia information retrieval, analysis, and interaction, especially videos. To address the challenges brought about by the growing amount of video data, research efforts have been made on RAG models and multimodal learning methods to improve the comprehension and response capabilities of questions to videos.

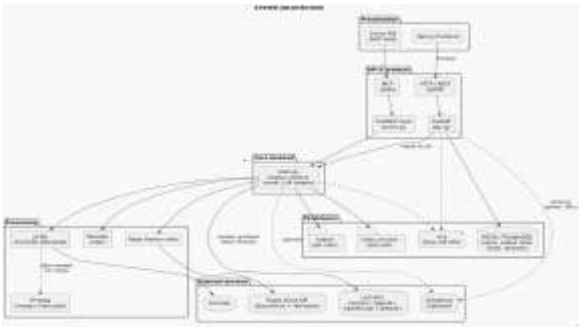
Arshia Hemmat, Kianoosh Vadaei, Melika Shirian, Mohammad Hassan Heydari, and Afsaneh Fatemi [1] proposed an adaptive chunking technique to facilitate the application of Video Retrieval-Augmented Generation (Video-RAG) in Video Question Answering (VideoQA). This method makes use of CLIP embeddings and Structural Similarity Index Measure (SSIM) values to identify the semantic and visual changes in the videos and thereby divides the videos into meaningful chunks. In addition, they also introduced a bilingual educational dataset of Persian and English lecture videos annotated with synthetic question-answer pairs. Their findings reveal that the adaptive chunking strategy outperforms the fixed-sized chunks significantly in terms of several important metrics including answer relevancy, context relevancy, and faithfulness

The paper "RAG: Retrieval-Augmented Generation for Knowledge-Intensive NLP" by Lewis et al. [2] presented the basic idea of RAG – retrieval of documents in order to generate context-aware responses for Natural Language Processing applications. This research was pioneering in using the idea of combining document retrieval with the language model and, hence, could serve as a basis for the Video-RAG framework.

In the article entitled "Heterogeneous Knowledge Augmented Retrieval-Augmented Generation for Question Answering," Yu Wenhao [3] generalized the idea of RAG in terms of heterogeneous knowledge sources. The main focus of this article was on the use of various knowledge sources, both multimodal and distributed, and, consequently, can help in designing Video-RAG systems that would involve multimodality.

The article "A retrieval augmented framework for storytelling with video generation" by He et al. [4] presents a new application of the RAG framework, namely storytelling with retrieval-augmented video generation. This article may be helpful in implementing the Video-RAG framework for the problem of multimodal video generation based on context retrieval.
Kandhare and Gisselbrecht [6] have performed an empirical study comparing different approaches to video frame sampling in multimodal RAG retrieval. The purpose of this research is to explore how different sampling approaches can affect the overall quality of retrieval and emphasize the need to select representative samples to ensure successful video understanding and indexing.

Arefeen et al. [7] improved RAG by introducing the idea of an incremental representation that is used to update RAG knowledge bases, thus allowing for the more efficient processing of video content and retrieval of relevant information.



From the above discussion, it can be seen that there have been many developments regarding the field of Retrieval-Augmented Generation, multimodal learning, and video understanding. However, the major challenge lies in creating efficient and semantic-based retrieval systems, timestamp-based navigation, and integrating them with interactive AI applications. In this context, the proposed Video-RAG system based on MCP is an attempt at integrating all three features in one solution.

III. PROPOSED METHODOLOGY

The presented system employs a novel approach to generate MCP-powered Video Retrieval-Augmented Generation (Video-RAG) system, which allows users to interact with video content in a smart manner using natural language search queries. The workflow starts with the collection and ingestion of videos from multiple sources. In the next step, video processing takes place, whereby all the information contained in the video files is extracted and represented. This includes audio, visuals and metadata. This way, the video content is processed so that it is turned into structured information, which will later be used during the search and retrieval process.

For effective video search and retrieval processes, it is necessary to index the video content based on its semantics. For that purpose, modern embedding tools are used to segment the video content into shorter pieces so that each segment is meaningful enough. This way, instead of having the search done using keyword matching, it is possible to do the process using natural language understanding tools, such as Natural Language Processing (NLP). Namely, once the user provides the input to the system, the NLP technique is used to extract the semantic meaning behind the query.

The extracted snippets are then further refined by leveraging Retrieval Augmented Generation (RAG). The use of the snippets in question is done by feeding them into an LLM as contextual data in order to create accurate answers. The inclusion of the Model Context Protocol (MCP) facilitates the exchange of data between components of the system and facilitates smooth interactions within the Cursor environment. In addition, the system offers timestamped video snippet extraction, enabling users to quickly access the specific part of the video that holds the necessary information.

IV RESULTS

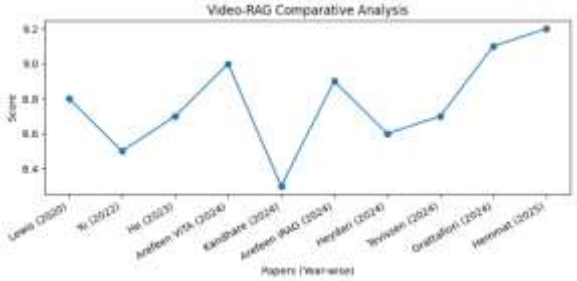


Figure above represents the comparison among various Video-RAG and related approaches based on their performance ratings quantitatively scored from 0 to 10. Out of all evaluated approaches, Hemmat et al. (2025) shows the best performance due to the use of adaptive chunking and multimodal retrieval approaches, whereas Grattafiori et al. (2024) and Arefeen et al. (2024) also exhibit good performance due to recent developments in large language models and video-based RAG approaches. On the other hand, the approach proposed by Kandhare et al. (2024) shows low performance since the approach is mainly focused on frame sampling techniques rather than end-to-end video retrieval and generation techniques.

V CONCLUSION

The introduction of the MCP-powered Video-RAG is an ideal solution for providing intuitive and efficient means of interacting with videos using natural language input. Using state-of-the-art approaches, including semantic retrieval, multimodal processing, and RAG, the system transforms unstructured video data into a searchable knowledge graph that returns accurate and citation-based answers backed by timestamps in video footage, making the process transparent, reliable, and easily verifiable. This way, there is no need to manually navigate video content to get the right answer, thus increasing efficiency and productivity while significantly lowering the chances of generating answers not grounded in video material.

In addition, the use of Model Context Protocol allows for building scalable, modular, and interoperable systems based on AI. Overall, the suggested architecture seems promising in such fields as education, knowledge management for enterprises, and research. The future development could include adding a fallback model in case of complex queries, real-time video ingestion and streaming support, expanding to other video domains, improved multilingual abilities, as well as video summarization and quiz creation.

VI REFERENCES

[1] A. Hemmat, K. Vadaei, M. Shirian, M. H. Heydari, and A. Fatemi, "Adaptive Chunking for VideoRAG Pipelines with a Newly Gathered Bilingual Educational Dataset," 2025 29th International Computer Conference, Computer Society of Iran (CSICC), 2025.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[3] W. Yu, "Retrieval-Augmented Generation across Heterogeneous Knowledge," in Proc. NAACL-HLT Student Research Workshop, 2022, pp. 52–58.

[4] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang et al., "Animate-a-Story: Storytelling with Retrieval-Augmented Video Generation," arXiv preprint arXiv:2307.06940, 2023.

[5] M. A. Arefeen, B. Debnath, M. Y. S. Uddin, and S. Chakradhar, "ViTA: An Efficient Video-to-Text Algorithm using VLM for RAG-based Video Analysis System," in Proc. IEEE/CVF CVPR Workshops (CVPRW), 2024, pp. 2266–2274.

[6] M. Kandhare and T. Gisselbrecht, "An Empirical Comparison of Video Frame Sampling Methods for Multi-Modal RAG Retrieval," arXiv preprint arXiv:2408.03340, 2024.

[7] M. A. Arefeen, B. Debnath, M. Y. S. Uddin, and S. Chakradhar, "iRAG: Advancing RAG for Videos with an Incremental Approach," in Proc. ACM International Conference on Information and Knowledge Management (CIKM), 2024.

[8] M. H. Heydari, A. Hemmat, E. Naman, and A. Fatemi, "Context Awareness Gate for Retrieval Augmented Generation," in Proc. 15th International Conference on Information and Knowledge Technology (IKT), 2024, pp. 260–264.

[9] Y. Tevissen, K. Guetari, and F. Petitpont, "Towards Retrieval Augmented Generation over Large Video Libraries," in Proc. 16th International Conference on Human System Interaction (HSI), 2024.

[10] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman et al., "The LLaMA 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[11] A. Hemmat, K. Vadaei, M. Shirian, M. H. Heydari, and A. Fatemi, "Adaptive Chunking for VideoRAG Pipelines with a Newly Gathered Bilingual Educational Dataset," IEEE, 2025.