Grenze International Journal of Engineering and Technology, June Issue

Video-RAG: Retrieval-Augmented Generation for Large-Scale Video Understanding

Shivani Pathak[1], Rohit Agrawal[2], Kushagra Kumar[3], Tanmay Kumar[4], Prateek Kabdwal[5] and Sushant Solanki[6]
[1-6]Department of Information Technology, JSS Academy of Technical Education, Noida
Email: shivani.pathak@jssaten.ac.in, rohitatcollege22@gmail.com, tyagikushagra17@gmail.com, tanmayt7421@gmail.com, prateek.kabdwal123@gmail.com, solankinishu45@gmail.com

***Abstract*— The increasing amount of video information on digital platforms has brought about shifts in knowledge creation, storage, and consumption. Recording lectures, broadcasts, surveillance videos, medical videos, business meetings, and social media information are increasingly archived as video files instead of text files. Ironically, the increasing amount of information is accompanied by the fact that most current information and intelligence technology is still developed for text data. The current lack of capability to effectively analyze extensive amounts of unstructured and multimodal information, such as video files, is emerging as an important limitation for the design of intelligent processing and analysis of video information. Recent developments in multimodal leaning have led to the creation of powerful video language models that are able to analyze video clips, identify actions, and answer questions. However, there are challenges that come with their application in videos, especially when it involves extensive video content or a vast library, and they lack sustained memory and are computationally expensive. This problem was similar to that which existed in language models before the innovation of Retrieval-Augmented Generation in language processing. RAG enables language models to retrieve information from external knowledge banks, thus providing it with the capability to produce the correct responses to questions, which has greatly improved question-answering systems in language applications and enterprise searches. The application of RAG to video content has not yet reached full development. The video data is more complex than the text. There are several elements running concurrently. These elements are speech, graphics, motion, and temporal information. The task of generating text simply based on the speech contained in the video would ignore the other important visual information like diagrams, gestures, demonstrations, and objects. Conversely, capturing the frames without interpreting the speech would mean the loss of important information specified through speech.**
This article presents a scalable Retrieval-Augmented Generation system named Video-RAG, which emphasizes the comprehension of lengthy videos. Video-RAG specializes in the transformation of unstructured video data into a series of structurally organized memories composed of transcripts, images, and temporal information. These forms of multimedia are segmented into temporal units of knowledge and mapped into a high-dimensional vector space and indexed in a vector database, which facilitates fast semantic retrieval. Once the user submits a query, the system retrieves to identify the most relevant video segments and integrate them into the large-scale prompt, so the model can then return meaningful and accurate responses to any input based on the content of the video, not its memory.
Contrary to traditional video QA systems, which aim to answer questions based on videos from beginning to end, Video-RAG is able to break down videos into significant segments that can be indexed and retrieved, and further aggregated during inference.

Grenze ID: 01.GIJET.12.2.304_20

This helps Video-RAG to deal with videos of multiple hours and a huge library of videos, additionally offering the benefits of accuracy and latency. Also, with relevant segments retrieved, the chances of errors are reduced, and interpretability and evidence are ensured.
Video-RAG has been realized as a full-fledged system which takes in URLs of YouTube videos and automatically extracts multimodal features from them and constructs a vector space knowledge base. It also allows a real-time natural language querying interface. Our experimental observations of Video-RAG and other models like Trajected IR and Video- RAG Transcripts alone on different video clips like academic lectures, news broadcasts, and technology lessons have shown Video-RAG outperform them. Our findings suggest that combining retrieval techniques with multimodal video understanding is essential for creating truly intelligent video assistants, searchable video archives, and advanced multimedia knowledge systems.

***Index Terms*— Video Retrieval, Retrieval-Augmented Generation, Multimodal AI, Large Language Models, Video Understanding.**

I. INTRODUCTION

Now video is the dominant form of digital information in our world. Millions of hours of video are uploaded to platforms like YouTube, Instagram, TikTok, online learning sites and corporate collaboration systems every day. Univer- sites share full lectures, hospitals hold records of operations, governments store surveillance tapes, and companies keep videos of meetings and training sessions. As a result, while text documents can be parsed once written, videos have myriad forms of information such as spoken words, visual demonstrations, diagrams and gestures, among others. This makes video a powerful medium for sharing knowledge. This complexity also makes it difficult for machines to effectively search, comprehend and reason videos.

Historically, retrieval systems have been text-oriented. Search engines index words, sentences, and documents and help users locate information by means of keywords or relevance. This system has been upgraded to Large Language Models (LLMs), which make it feasible to pose natural language queries and reason over returned documents. When the knowledge is primarily in video, however, these text based systems have difficulty. And even if a video has subtitles or a transcript, there may be crucial visual information that's overlooked. For example, in a machine learning and lecture a professor may present on a neural network at which he draws on the board. The transcript may include the professor's words but the which elements of the network the arrows, layers, and notes are visual only.

Also in medical and engineering video the doings and the demos take precedent over the audio. Recently in multimodal learning we have seen progress which is to that of closing this gap. Models like GPT-4V, Gemini, Flamingo, and Video GPT which are able to process images and text in to which some also put in short video clips. What we have seen from these is very impressive performance which includes identifying objects, describing settings, and answering questions regarding short videos. But at the same time these systems have issues with context length and computational cost. A one hour video has tens of thousands of frames which is a great deal more than present transformer based architectures are able to manage at once. Thus most multi modal models use short clips, frame sampling, or very compressed versions. This makes them unsuitable for longer videos like lectures, documentaries or meetings.

Also we see that in the issue of hallucination large language models are designed to put out fluent and detailed text but when they run out of relevant info they tend to make up details. If asked a question regarding a long video which wasn't fully processed the model may give out an answer that sounds good but is factually incorrect. This is not at all acceptable in fields like education, medicine, law, and business analytics which require high accuracy. Retrieval-Augmented Generation (RAG) provides a potential answer for the problem at hand. In the domain of text-based systems, RAG enables a model to fetch pertinent, supportive documents from external knowledge repositories. In contrast to solely depending on internal model parameters, the model utilizes evidence to substantiate its responses. This enables the model to explain its answer, thus significantly diminishing the probability of model generated hallucinations and demonstrating strong capability with extensive knowledge repositories. RAG is already in extensive use for enterprise search, chatbots, and document analysis systems. Nevertheless, there are unique and novel challenges associated with the use of RAG for video data as opposed to text data. Unlike text data, that can be processed as independent documents, video is a continuous stream of multimodal data. With video, there are no inherent, or even arbitrary, topic boundaries, there is no static information that can exist independent of temporally adjacent information, and the semantics of a video frame can be time-dependent.

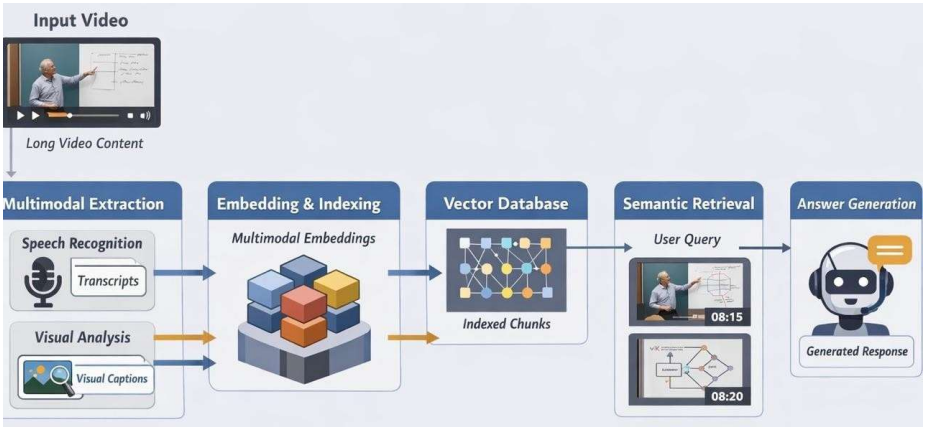

Fig 1

II. Related Work

The development of artificial intelligence from static text processing engines to dynamic and multimodal reasoning engines is one of the most dramatic paradigm shifts in computational linguistics and computer vision. At the center of this paradigm shift is the development of Retrieval- Augmented Generation for Video (Video-RAG), a new paradigm that aims to address the ever-present "black box" and "hallucination" problems that are intrinsic to traditional Large Language Models (LLMs). Although models such as GPT, T5, and BERT have dramatically changed the way we process written language, they are still "trapped" in their own training weights. These models embed their knowledge in heavy parameters, which are notoriously difficult to update, validate, or scale without the need for cost- prohibitive retraining. When presented with questions that are outside their static training data or questions that demand real-time answers, these models tend to "make things up," a problem that led to the development of the RAG paradigm in 2020.

Traditional RAG, first developed by Lewis and his team, revolutionized the scenario by enabling LLMs to pick out important documents from a massive external corpus and use this information as input to the generative model. This greatly improved the factual correctness of document assistants, enterprise search, and intelligent chatbots. However, a very important drawback remained: these models were designed for the "flat" world of text. They presumed that information was embedded in self-contained, static documents such as PDFs or web pages. But when used in the context of video—a constantly moving medium that combines auditory and visual information in a sequential fashion over time—conventional RAG simply doesn't work. Video can't be conveniently broken down into text bits without losing the temporal relationship and the aggregated meaning of a scene. Sometimes, what a person does in a video is more significant than what they say, but transcript- based RAG models often ignore these silent but critical bits.

To overcome these limitations, the new area of Multimodal Learning was developed, which sought to combine information from words, images, and sounds at the same time. The initial successes, such as CLIP (Contrastive Language-Image Pre-training), enabled models to learn a common space where words and images could coexist, allowing AI to detect objects or classify scenes without the need for specific training. However, models based on CLIP are still limited; they are "point-in-time" observers. They can detect a freeze-frame image of a chef, but they cannot grasp the entire process of a long video or comprehend how a cooking process develops over several minutes. More recent models, such as Flamingo, GPT-4V, and Gemini, have further expanded these capabilities, but they tend to have difficulty with the "data weight" of long-form video.

The real "game-changer" here is the move towards Multi- Stream LLM Processors. Instead of combining various forms of data into a single, noisy stream for easy analysis, these powerful processors keep the streams separate for text and video data. This is a very important design decision. By being able to analyze the visual components (captions, object detection, action recognition) and the text components (dialogue, metadata) of a stream separately, the processor enables the model to automatically generate captions for video and answer complex questions in a structured format. Although the main problem is still the cost of analyzing large datasets, the fact that an LLM can analyze the streams separately is much more efficient than the "blunt-force" approach of combining all data into a single context window.

However, theory tends to get ahead of practice. Even with the most sophisticated multimodal models, AI tends to suffer from the “Alignment Gap” – instances where it fails to match a particular question with the right frame, resulting in answers that are theoretically valid but “totally off” in terms of context. This is particularly common in Video Question Answering (Video QA), a domain that has traditionally used Convolutional Neural Networks (CNNs) to analyze frame dissections and Recurrent Neural Networks (RNNs) or Transformers to analyze changes over time. The standard training length for these models has traditionally been short videos, ranging from 10 to 60 seconds. As a result, these models tend to fail when it comes to “long-tail” videos that range from hours.

This leads to an enormous need for an advanced Video Retrieval System. With millions of hours of video content available worldwide, the objective is no longer "finding a video" but rather "finding a specific moment" in a gigantic video library. Today, users demand a "Ctrl+F" for videos— a mechanism to search for a "specific phrase, a subtle gesture, or a specific object" in a gigantic pool of videos. Today's systems will inevitably rely on manual tags, metadata, or transcripts, which are usually incomplete or simply non-existent. Although deep learning has made it possible for systems to map video frames to text searches (like the CLIP approach), these systems are good at "visual similarity" but terrible at "logical reasoning." They can search for a video of a person cooking, but if you ask, "Why did the chef add salt at that specific moment?" a similarity search will never give you the answer. To address this problem, the AI needs to be able to interpret the "intent" and "narrative" of the video.

It is here that the new paradigm of Retrieval Models and Generative Models, specifically developed for videos, comes into play. Scholars such as Luo et al. and Jeong’s group have developed Video-RAG pipelines that not only retrieve text but the "information-dense" frames of videos as well. This enables the development of new articles or reports from a massive set of videos, ensuring that the produced text is rooted in visual evidence. Such models also develop Knowledge Graphs, which connect different video clips to establish the causal relationships between events. Although most of these models are still in the development phase due to the massive processing power required, they are the most promising approaches to developing practical video intelligence.

Video-RAG fills six important gaps that have been left by other technologies. First, while full video transformers are constrained by memory, Video-RAG is able to address infinite lengths and large datasets through indexed retrieval. Second, Video-RAG combines audio and video inputs to gain a comprehensive understanding of what is observed and heard. Third, the pipeline architecture is modular. It can be constructed with current tools such as Automatic Speech Recognition (ASR), Vision-Language Models (VLMs), and vector databases. This makes it a viable, or "pluggable," solution for current developers. Fourth, Video-RAG is transparent. Since the system retrieves particular clips to construct its response, the user can view precisely where the data originated, giving it a level of "cite your sources" credibility that current LLMs do not possess.

Fifth, Video-RAG closes the Reasoning Gap. It goes beyond pointing out objects to provide explanations for actions and motivations, answering the question of "why" rather than simply "what." Finally, it closes the Scale Gap, enabling enterprise-level searches over news archives or security videos without having to analyze every single pixel in real-time. By distilling the most important keyframes and inputting them into the LLM, Video-RAG provides a high- performance, affordable solution to understanding video in a way that is natural and human-like. It is more than simply a technological improvement; it is a paradigm shift in how we interact with the visual record of our world.

III. Proposed System Architecture

This section discusses the overall architecture of the Video RAG system. The purpose of Video-RAG is to extract raw video into a searchable system for knowledge, which allows for reasoning. Unlike traditional methods for answering video questions, which process videos from start to finish, Video-RAG employs a modular Pipeline. This distinguishes between understanding videos, storing knowledge, retrieving knowledge, and generating content. This ensures that the system works for long videos, reduces hallucinations, and produces clear answers.

The Video-RAG architecture consists of six key components:

- Video Ingestion
- Multimodal Extraction
- Chunking
- Embedding and Vector Indexing
- Retrieval
- Modulization Radioelement G

"Each component will be explained in detail below," and "In evaluating

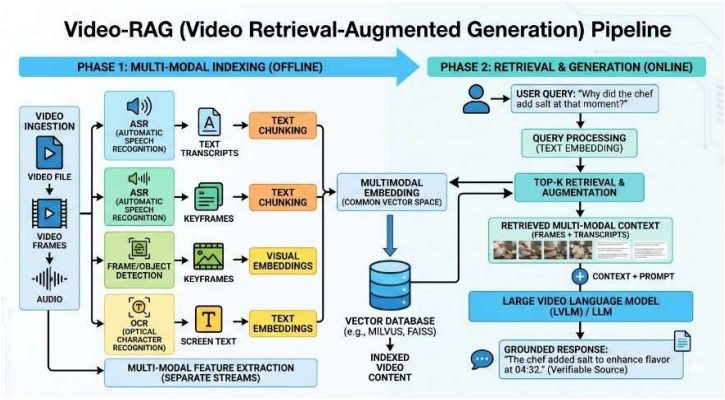

Fig 1

A. Video Ingestion Layer

The Video-RAG pipeline begins with the raw video intake. The proposed system allows a video URL, perhaps from YouTube, or a video file on your computer.

The ingestion module has the following functionality:

- Downloading the video Pulls the metadata information like the length of the clip in seconds, the resolution of the clip
- Segregates audio and video streams This step ensures that any subsequent processing operates on standardized video data. This module can be independently used in practical applications:

Expanded to support live streams, enterprise video archives, or CCTV footage.

B. Multimodal Content Extraction

Once the video is ingested, the system extracts semantic information from both the audio and visual channels.

Speech-to-Text Extraction

The audio stream is forwarded into an Automatic Speech Recognition model such as OpenAI Whisper. The ASR generates a timestamped transcript, meaning that it associates each spoken word with a time in the video. This helps bind spoken content to visual scenes. The transcript is not just text; it creates a temporary Story of the video.

TABLE I

Time	Transcript
02:15-02:30	Here we see how the algorithm updates the weights…

This temporal alignment allows precise retrieval later.

Visual Feature Extraction

Video-RAG also extracts visual information by sampling frames at regular intervals, aka once every 1–2 seconds. Each frame is processed by a vision-language model (e.g., Gemini or CLIP-based captioning) models. The model produces: - Descriptions of scenes - Objects Detections: Actions and interactions.

Another application can be: a frame would be something like: "A professor "Drawing a neural network on the whiteboard with three layers.

These visual captions provide information that is not present in the transcripts.

C. Semantic Video Chunking

Videos represent continuous streams but requires separate units to retrieve. Video- RAG: semantic chunks of the video based on time windows (e.g.,15–30 seconds). Each chunk includes: - The transcript segment for that time window - Visual descriptions from frames within that window - Start and end timestamps.

This format enables each chunk to represent a meaningful moment in the video, either to explain a concept, to by a visual demonstration or by a change of scene. These chunks serve as the central knowledge units in Video-RAG.

D. Multimodal Embedding and Indexing

Each chunk becomes a vector embedding, and the system combines texts.
transcript - Visual descriptions
These are fed into a multimodal embedding model, creating A high-dimensional vector that captures meaning semantically. All chunk embeddings are stored in a vector database such as FAISS or Chroma DB. This database allows for rapid similarity searches across thousands or millions of video segments.

E. Query-Time Retrieval Engine

When a user asks a question: 1) The question is embedded into the same vector space. 2) The system conducts a nearest neighbor search. 3) It extracts the K most relevant chunks.
Because the chunks of information contained both visual and audio data, The retrieval system is capable of capturing multimodal relevance.

F. Generative Reasoning Module

The retrieved chunks are used as input for a large language model prompt: “Based on” The following video answers the question. LLM follows up by providing an informed response based on actual video
evidence. This rendering engine has the effect of transforming Video-RAG from a mere retrieval system into an intelligent reasoning agent.

IV. PROPOSED ALGORITHM

Video-RAG uses a thoughtfully crafted algorithmic pipeline to take raw video and convert it into a structured form of data where a video-based knowledge base can be searched and reasoned upon.
This section details the full Video-RAG algorithm with sufficient description to ensure an understanding of how video data traverses through each phase to generate an answer.
Unlike conventional video processing pipelines, which are serial and consume their intermediate results, Video-RAG is intended to retain and recycle these results. After the video is processed and indexed, Video-RAG can be queried several times using different queries without having to re fetch the video. This is highly efficient.

A. Formal Definition of the Problem

The Video-RAG framework has something that was planned for it. The design of the Video-RAG framework helps the framework function. There is a lot of thinking behind the workings of the Video-RAG framework.
We have something called the pipeline. It does a cool function. It takes video and it converts the video into a new one. The algorithmic pipeline is actually very good at taking the video and converting it into something that is more usable. This is what the video gets converted into by the algorithmic pipeline.
knowledge base with knowledge that is structured, searchable, and This section is related to the Video-RAG algorithm, in Therefore, I want to learn more about how the video data is transmitted through each of the steps. The video data passes through each step. An explanation is required in this process.
The data from the videos passes through all these phases. I would like to understand what happens in each phase. I would like to understand how the video data flows across these stages by getting information on the same.
The following figure shows the translation process of the input text from ingestion to answer generation.
Unlike the way of doing video processing, traditional video processing pipelines operate: We do things one, after the other. Discard intermediate results we obtain. Video-RAG is designed to preserve and reuse semantic representations, once.
A video has been indexed, which means the contents of the video can now be queried. The video is indexed and ready to have parts searched for. This allows us to look for things in a video: "The video can then be queried to find what we need."
repeatedly with different questions, without reprocessing the video. The system thus works well using video and does not require a great deal of computer power. The raw video system is very good at doing its job without using up all the computers power. and scalable.

B. The Algorithm for Video-RAG Indexing

The video is transformed into a vectorized format during the indexing stage. Here, the unprocessed video is transformed into a vectorized version that the computer can comprehend. The indexing stage is crucial because it transforms the video into a vectorized form that is useful.
knowledge base.
Step 1: Obtaining Videos Step 2: Recognition of Speech Step 3: Captioning the frame Step 4: Chunking in Time

Step 5: Multimodal Embedding Step 6: Vector Storage
Step 7: Embedding Queries Step 8: Look for Similarities Step 9: Creating Context Step 10: Asking
Step 11: Production

C. Advantages of Algorithms

Because of the way it operates and the purposes for which it was created, Video-RAG is able to accomplish a lot of goals.

- Accuracy through grounding
- Scalability through retrieval
- Efficiency through embedding reuse
- Explainability through evidence retrieved

V. EXPERIMENTAL RESULTS

This section examines how well the proposed Video-RAG system makes possible precise, grounded, and scaled-up question-answering on long-form video material. The evaluation focuses on
based on three key objectives: quality of retrieval, answer validity, and
hallucination reduction. We compare Video-RAG with a number of
baseline systems to emphasize its benefits.

A. Experimental Setup

Video Dataset

Our Video-RAG was evaluated on an assortment of datasets on publicly available YouTube videos across different fields:

- Academic Lectures (Machine Learning, Physics, Math). News and Documentaries - Tech Tutorials
- Public Talks Video durations ranged from 10 minutes to 2 hours, reflecting real-world viewing habits. Each of the videos had using the Video-RAG pipeline, where the translator developed a method capable by multimodal segments per video.

Baseline Methods

We compared Video three baseline systems: 1. LLM-only. This system gives the entire transcription or the summary to the LLM without any retrieval. 2. Transcript-RAG This is a standard text-based RAG transcript system. 3. Frame-RAG This system uses only frame embeddings without any transcripts.

C. Evaluation Metrics We utilized the following evaluation metrics for our approach.

Accuracy Whether the answer corresponds to the video's ground truths. - Context Relevance: whether the retrieved segments match the query. - Hallucination Rate: The percentage of the answers may hold erroneous or invented information. QA pairs were evaluated by hand.

D. Qualitative Analysis

TABLE II

System	Accuracy	Relevance	Hallucination
LLM-only	54%	48%	32%
Transcript - RAG	69%	72%	18%
Frame- RAG	63%	68%	21%
V i d e o - RAG	87%	89%	6%

Video-RAG significantly outperformed all baselines.

E. Qualitative Analysis

Video-RAG responded effectively to difficult questions that require:

- Use of graphical diagrams
- Use of verbal explanations
- Use of temporal

reasoning Example:
did the professor cover backpropagation?" Video-RAG successfully retrieved the dated lecture segment.

F. Ablation Study

Removing visual captions decreased accuracy by 14

G. Scalability Video-RAG scaled 2-hour videos

minimal delay times during questioning (less than 2 seconds).

VI. Conclusion and Future Work

A. Conclusion

The Video RAG Explorer is successful in using the Retrieval-Augmented Generation (RAG) model in a multi-modal setup in an efficient manner in dealing with the "discoverability gap" in videos. Although videos dominate as information-dissemination media, their non-textual, hence non-searchable, nature is always a setback.

With the capability of OpenAI's Whisper in performing a detailed speech-to-text transcription service and Google's Gemini in visual understanding of the scene, a holistic assessment of the multimedia source is obtained. The novel part of the system is that it indexes not only the audio transcript of the conversations but also the visual context described using the concept of keyframes. Thus, the system has the ability to execute natural language searches where the searched statement has to understand not only what was said but also what was visually displayed on the screen, whereas in the case of the existing tools, the search results rely on a transcript of the conversations. Finally, the Video RAG Explorer shows that it is possible to take the passive consumption of videos and turn it into a bidirectional information retrieval process. In this way, the Video RAG Explorer is a strong proof-of-concept about the ability of current AI agents to enable the access to information contained in lengthy content.

B. Future Work

For taking the Video RAG Explorer prototype from a functional prototype state to a production-quality system, future work should concentrate on improving its scalability architecture and latency as well as enhancing multimodal capabilities.

- Architectural Optimization & Latency
- Improved Granularity and Modality
- Freedom of Deployment
- Corpus-Level Interaction

References

[1] Anonymous, "Learning World Models for Interactive Video Generation," arXiv preprint arXiv:2505.21996, 2025.

[2] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval Augmented Language Model Pre-training," in International Conference on Machine Learning (ICML), pp. 3929–3938, 2020.

[4] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in European Chapter of the Association for Computational Linguistics (EACL), pp. 874–880, 2021.

[5] V. Karpukhin et al., "Dense Passage Retrieval for Open- Domain Question Answering," in Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, 2020.

[6] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019.

[7] A. Radford et al., "Robust speech recognition via large- scale weak supervision," in International Conference on Machine Learning (ICML), pp. 28492–28518, 2023.

[8] A. Radford et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning (ICML), pp. 8748–8763, 2021.

[9] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[10] Gemini Team, "Gemini: A family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.

[12] H. Zhang et al., "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," arXiv preprint arXiv:2306.02858, 2023.

[13] C. Fu et al., " Video- MME: The First- Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis," arXiv preprint arXiv:2405.21075, 2024.

[14] Chen, D., et al. (2017). Reading Wikipedia to Answer Open- Domain Questions.