

VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos

Xubin Ren^{1*} Lingrui Xu^{1*} Long Xia² Shuaiqiang Wang² Dawei Yin² Chao Huang^{1†}

¹The University of Hong Kong ²Baidu Inc.

{xubinrengs, lingruixu.db, long.phil.xia, chaochuang75}@gmail.com
wangshuaiqiang@baidu.com yindawei@acm.org

Abstract

Retrieval-Augmented Generation (RAG) has demonstrated remarkable success in enhancing Large Language Models (LLMs) through external knowledge integration, yet its application has primarily focused on textual content, leaving the rich domain of multi-modal video knowledge predominantly unexplored. This paper introduces VideoRAG, the first retrieval-augmented generation framework specifically designed for processing and understanding extremely long-context videos. Our core innovation lies in its dual-channel architecture that seamlessly integrates (i) graph-based textual knowledge grounding for capturing cross-video semantic relationships, and (ii) multi-modal context encoding for efficiently preserving visual features. This novel design empowers VideoRAG to process unlimited-length videos by constructing precise knowledge graphs that span multiple videos while maintaining semantic dependencies through specialized multi-modal retrieval paradigms. Through comprehensive empirical evaluation on our proposed LongerVideos benchmark-comprising over 160 videos totaling 134+ hours across lecture, documentary, and entertainment categories-VideoRAG demonstrates substantial performance compared to existing RAG alternatives and long video understanding methods. The source code of VideoRAG implementation and the benchmark dataset are openly available at: <https://github.com/HKUDS/VideoRAG>.

1 Introduction

Recent advances in Large Language Models (LLMs) have revolutionized NLP, yet their performance is inherently limited by the knowledge captured during pre-training. To address this limitation, Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm that enhances LLMs by dynamically retrieving and incorporating external knowledge during inference [1, 2]. While RAG has demonstrated success across various text-based applications, such as question answering, and factual reasoning, its potential remains largely untapped in the rich domain of multi-modal content, particularly video understanding. The extension of RAG to video content presents unique challenges and opportunities, as videos contain complex multi-modal features, temporal dynamics, and intricate semantic relationships that go beyond traditional text-based knowledge integration approaches.

Although large vision models have achieved impressive progress in video understanding tasks, they face limitations when processing long-context videos. These models (*e.g.*, VideoLLaMA3 [3] and LLaVA-Video [4]), primarily designed for short video clips, struggle to effectively capture and reason about temporal dependencies spanning multiple hours. The challenge becomes particularly acute in scenarios requiring cross-video understanding and knowledge integration, such as lecture series comprehension, documentary analysis, or sequential entertainment content interpretation. Current

*Equal contribution.

†Chao Huang is the Corresponding Author.

approaches often fragment long videos into isolated clips, leading to loss of contextual information and inability to establish meaningful connections across different videos. This limitation severely hampers applications in educational content analysis, media archiving, and video-based knowledge extraction, where understanding the broader context across multiple videos is essential.

The key challenges in realizing Retrieval-Augmented Generation for extreme long-context videos are multifaceted. (i) **Capturing Heterogeneous Video Knowledge**. Videos contain rich information across multiple modalities, including visual frames, audio streams, and textual descriptions. Effectively capturing and organizing this diverse knowledge presents a unique challenge that cannot be addressed by existing text-based RAG approaches. Existing methods are ill-equipped to handle the complexity of cross-modal information and their relationships. (ii) **Preserving Semantic Coherence for Cross-Video Understanding**. Maintaining the semantic connections across numerous videos, which may span hours or days, is more complex than a single video. Preserving these intricate relationships and comprehensive knowledge interdependencies is crucial for holistic video understanding. (iii) **Efficient Video Knowledge Retrieval**. When the video knowledge base consists of an unrestricted number of lengthy videos, quickly identifying the most pertinent clips in response to user queries becomes significantly more challenging. The retrieval system must provide users with the most relevant information to answer queries accurately.

By addressing these key challenges, the VideoRAG aims to unlock the full potential of RAG in the domain of extreme long-context videos, enabling powerful and comprehensive video understanding capabilities. At the heart of VideoRAG are two interlocking components - the **Multi-Modal Video Knowledge Indexing** framework and the **Knowledge-Grounded Multi-Modal Retrieval** paradigm. The indexing framework transforms video content into structured textual and visual representations, with graph-based textual knowledge grounding to preserve semantic relationships across videos, complemented by multi-modal context encoding to capture fine-grained cross-modal interactions. This dual-channel architecture enables VideoRAG to effectively organize and index long-context videos, preserving the rich semantics of the multimedia content. The knowledge-grounded retrieval paradigm then integrates textual semantic and visual content matching, leveraging the indexed knowledge graph and embeddings to identify the most relevant video content. Finally, VideoRAG employs a two-stage content extraction process that combines LLM-powered keyword extraction and vision-language model-based text grounding to enrich the visual analysis with text-based retrieval, generating comprehensive outputs for the final response.

The comprehensive evaluation on the benchmark datasets demonstrates the advantages and effectiveness of the VideoRAG framework in understanding extremely long-context videos, going beyond the limitations of existing RAG alternatives and large vision models. The results showcase VideoRAG’s superior performance in effectively organizing and indexing long-form video content, allowing for precise retrieval of relevant segments across different video sources in response to user queries. Our ablation studies provide deeper insights into the individual contributions of VideoRAG’s key components, highlighting the importance of the graph-based knowledge grounding and multi-modal retrieval mechanisms in elevating its performance. Furthermore, case studies demonstrate VideoRAG’s practical applications in real-world scenarios, such as video-based knowledge extraction and educational content analysis, unlocking new possibilities for cross-video comprehension.

Moreover, the proposed LongerVideos curates a diverse collection of over 160 long-form videos spanning 134+ hours across lecture, documentary, and entertainment categories - a substantial advancement over existing datasets that are limited to inference on single [5, 6] or relatively short-video content [7, 8]. LongerVideos enables the assessment of models’ capabilities in reasoning across multiple long-context videos, a crucial requirement for real-world cross-video understanding scenarios, like video-based knowledge extraction and educational content analysis. By providing a robust testbed for evaluating long video understanding methods, this benchmark will advance the development of systems that can comprehend and reason about long-form video content at scale.

2 Preliminary

Retrieval-Augmented Generation (RAG) represents a significant advancement in addressing the inherent limitations of LLMs. By intelligently incorporating external knowledge bases, RAG effectively reduces model hallucinations and enables access to domain-specific information without requiring costly model retraining. At its core, the RAG architecture consists of two fundamental components:

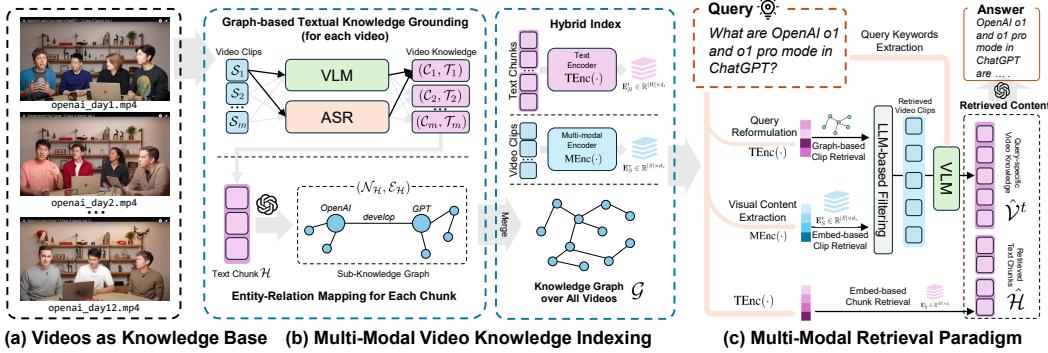


Figure 1: The overall framework of our proposed RAG framework VideoRAG for videos.

- **Indexing Module $\varphi(\cdot)$:** This component processes a knowledge database \mathcal{D} (such as document collections in text-based RAG) to create an optimized index structure $\hat{\mathcal{D}} = \varphi(\mathcal{D})$. The data structure enables rapid and efficient knowledge retrieval during query processing. The indexing process transforms raw information into an organized, searchable format facilitating retrieval operations.
- **Retrieval Module $\psi(\cdot)$:** When presented with a user query q , this module performs query-specific knowledge retrieval from the indexed data structure, denoted as $\psi(q, \hat{\mathcal{D}})$. The process involves identifying and extracting pertinent information from the indexed knowledge database, yielding informative sources that directly support answering the user’s query.

In essence, the RAG framework operates in two phases. In the preprocessing phase, the indexing module organizes all data into searchable structures. In the query phase, the retrieval module finds relevant knowledge for each input query q . The large language model (LLM) then processes both the query and retrieved knowledge to generate responses, expressed as $\text{LLM}(q, \psi(q, \hat{\mathcal{D}}))$.

Retrieval-Augmented Generation with Videos. While text-based RAG techniques are well-established, their extension to video knowledge remains largely unexplored. Our work advances the capability of Large Language Models (LLMs) to comprehend extremely long videos as a rich knowledge source. We achieve this by: i) Effectively capturing multi-modal characteristics (visual, audio, textual) and their temporal dynamics; ii) Modeling complex cross-modal alignment and inter-dependencies between different information streams.

We formulate this real-world challenge as a novel retrieval scenario with an unconstrained video knowledge base $\mathcal{D} = \mathcal{V}_1, \dots, \mathcal{V}_n$, where each video \mathcal{V}_i can be of arbitrary duration and the total number of videos n is unrestricted. To address this challenge, our VideoRAG framework enables effective video knowledge discovery and semantic understanding while ensuring comprehensive responses through effectively multi-modal context modeling.

3 The VideoRAG Framework

We present our retrieval-augmented generation framework designed for understanding unlimited-length video content. Our approach addresses two fundamental challenges: (1) multi-modal knowledge indexing that effectively captures and organizes visual, audio, and semantic information from videos, and (2) knowledge-grounded information retrieval that enables precise retrieval of relevant video clips for generating accurate responses through large language models. In the following sections, we detail these components and their integration into a unified video understanding system.

3.1 Multi-Modal Video Knowledge Indexing

Unlike traditional text documents, videos encapsulate information through multiple modalities - primarily visual frames - creating unique challenges for knowledge extraction and organization. Standard text-based RAG methods prove insufficient for video content due to several fundamental limitations: (1) text-based systems cannot directly capture visual dynamics; (2) temporal dependencies that traditional RAG approaches fail to preserve across video frames; (3) cross-modal interactions that simple text encoding cannot capture between visual and textual information.

To address these challenges, we introduce a comprehensive indexing framework with two components: **Graph-based Textual Knowledge Grounding** that transforms multi-modal signals into structured textual representations while preserving semantic relationships and temporal dependencies, and **Multi-Modal Context Encoding** that captures fine-grained cross-modal interactions through unified embeddings. This dual-channel architecture enables VideoRAG to effectively organize and index long-context videos while preserving the semantic richness of multi-modal content.

3.1.1 Graph-based Textual Knowledge Grounding

Our framework transforms multi-modal video content into structured textual knowledge through graph-based techniques for enhanced indexing and retrieval. The conversion process operates across two key modalities: for visual content, we employ state-of-the-art Vision Language Models (VLMs) to generate comprehensive textual descriptions capturing scene dynamics and contextual information; for auditory streams, we leverage high-fidelity Automatic Speech Recognition (ASR) to extract spoken content with temporal alignment. This dual-stream processing ensures both visual semantics and audio information are preserved in our textual knowledge representation.

- **Vision-Text Grounding:** To analyze visual content effectively, we segment each video \mathcal{V} into short clips $\mathcal{S}_1, \dots, \mathcal{S}_m$, enabling processing of unlimited-length videos. For each clip \mathcal{S}_j , we transform visual information into text through a two-step process: first, we uniformly sample k frames ($k \leq 10$) chronologically to capture key visual elements; then, we employ Vision-Language Models (VLMs) to generate comprehensive natural language descriptions capturing objects, actions, and scene dynamics. The caption generation process follows:

$$\mathcal{C}_j = \text{VLM}(\mathcal{T}_j, \{\mathbf{F}_1, \dots, \mathbf{F}_k\} \mid \mathbf{F} \in \mathcal{S}_j), \quad (1)$$

where \mathbf{F} denotes the chronologically ordered set of k sampled frames from the clip \mathcal{S}_j . We maintain $k \leq 10$ to optimize efficiency while preserving temporal coherence. The model integrates both visual frames and clip transcript \mathcal{T}_j as input prompts, enabling the VLM to generate contextually rich captions \mathcal{C}_j that capture both visual dynamics and associated speech content.

- **Audio-Text Grounding:** To capture crucial elements like dialogue and narration that enrich video understanding, we employ Automatic Speech Recognition (ASR) technology to transcribe each video clip, where $\mathcal{T}_j = \text{ASR}(\mathcal{S}_j)$ represents the extracted transcript from the clip \mathcal{S}_j .

For each video clip \mathcal{S} , we then create a unified and semantically rich textual representation by systematically merging the generated visual captions and ASR transcriptions $(\mathcal{C}, \mathcal{T})$. For a video \mathcal{V} containing m sequential clips, we formalize the complete knowledge extraction process as:

$$\mathcal{V}^t = \{(\mathcal{C}_l, \mathcal{T}_l) \mid l \in [1, m]\}. \quad (2)$$

At the core of our VideoRAG lies the challenge of organizing and retrieving multi-video knowledge efficiently. To address this, we propose a graph-based indexing framework that systematically links textual knowledge across different videos. This architecture enables incremental construction of a comprehensive knowledge graph from the extracted textual information, while maintaining semantic relationships and contextual dependencies. The entire process is executed through these essential steps, each designed to optimize multi-modal knowledge representation and retrieval:

- **Semantic Entity Recognition and Relationship Mapping:** Our framework leverages Large Language Models (LLMs) to construct a high-quality knowledge graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ that comprehensively captures and connects video knowledge. To optimize LLM performance and manage their context window limitations effectively, we implement a strategic processing pipeline:
 - (i) **Text Segmentation.** The first stage focuses on text segmentation, where we methodically divide video textual descriptions \mathcal{V}^t into manageable chunks $\mathcal{H}_i \in \mathcal{V}^t$. Each chunk is carefully constructed to contain multiple video clip descriptions while adhering to a predefined length threshold, ensuring optimal processing while maintaining semantic coherence.
 - (ii) **Entity-Relation Extraction.** In the entity-relation extraction phase, we process each chunk's caption-transcript pairs through LLMs to identify key entities (represented as nodes \mathcal{N}) and extract meaningful relationships (represented as edges \mathcal{E}). For instance, given the text “*GPT-4 utilizes transformer architecture for advanced natural language understanding, while incorporating visual features through ViT’s patch-based image encoding*”, the system extracts entities “*GPT-4*”, “*transformer architecture*”, and “*Vision Transformer (ViT)*”, along with relationships “*GPT-4 utilizes transformer architecture*” and “*GPT-4 incorporates ViT’s encoding*” in the domain of LLMs.

- **Incremental Graph Construction and Cross-Video Knowledge Integration:** The construction of our comprehensive knowledge graph follows an iterative and systematic approach across multiple video sources. Our framework implements a sophisticated incremental construction process that ensures coherent knowledge integration through several key mechanisms:
 - (i) **Entity Unification and Merging.** Our cross-video entity unification process systematically identifies and merges semantically equivalent entities across various videos into unified nodes within the knowledge graph \mathcal{G} . This unification approach not only maintains a consistent representation of entities but also preserves the rich contextual information derived from diverse video sources. As a result, it effectively creates a cohesive and interconnected knowledge network that enhances the overall understanding and usability of the information contained within the graph.
 - (ii) **Dynamic Knowledge Graph Evolution.** As new video content is processed, our knowledge graph undergoes systematic expansion through dual-track evolution: the integration of newly discovered entities and the establishment of previously unobserved relationships. When processing textual chunks from incoming videos, the system not only identifies and incorporates novel entities (e.g., emerging AI architectures or methodologies) but also discovers new semantic connections between existing nodes. This bidirectional growth process simultaneously reinforces established knowledge patterns while accommodating emerging concepts, ensuring the graph maintains both comprehensiveness and adaptability as it scales.
 - (iii) **LLM-Powered Semantic Synthesis.** To maintain semantic coherence, we strategically leverage Large Language Models (LLMs) to generate unified entity descriptions by synthesizing information from multiple video clips. This synthesis process ensures each entity maintains a comprehensive yet consistent representation, effectively consolidating knowledge across different video contexts while preserving semantic accuracy throughout the knowledge structure.

Formally, we define the construction of our complete knowledge graph as follows:

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}) = \bigcup_{\mathcal{H} \in \{\mathcal{V}_1^t, \dots, \mathcal{V}_n^t\}} (\mathcal{N}_{\mathcal{H}}, \mathcal{E}_{\mathcal{H}}), \quad (3)$$

Let $(\mathcal{N}_{\mathcal{H}}, \mathcal{E}_{\mathcal{H}})$ denote the extracted entities and their relations from each text chunk \mathcal{H} , split from the video description \mathcal{V}^t . Through processing of all videos, we construct the complete graph \mathcal{G} .

- **Text Chunk Embedding.** For each text chunk \mathcal{H} , we encode a text embedding $\mathbf{e}_{\mathcal{H}}^t = \text{TEnc}(\mathcal{H})$ using a text encoder $\text{TEnc}(\cdot)$, enabling efficient retrieval of raw chunks. We denote the complete set of chunks as H and represent their collective text embeddings as $\mathbf{E}_H^t \in \mathbb{R}^{|H| \times d_t}$, where $|H|$ represents the total chunk count and d_t denotes the embedding dimension. The knowledge graph \mathcal{G} and chunk embeddings \mathbf{E}_H^t together form the core components of our graph indexing module.

3.1.2 Multi-Modal Context Encoding

In vision-to-text grounding, certain visual nuances are inherently lost, such as lighting dynamics and intricate object details that resist accurate textual representation. To preserve these visual elements, we employ a multi-modal encoder $\text{MEnc}(\cdot)$ that transforms video content into retrieval-optimized embeddings. This encoder is capable of mapping both visual content and textual queries into a shared feature space, enabling efficient semantic retrieval. Building upon powerful multi-modal encoding frameworks like CLIP [9] and ImageBind [10], we formalize our video encoding as:

$$\mathbf{E}_S^v \in \mathbb{R}^{|S| \times d_v} \quad w.r.t. \quad \mathbf{e}_S^v = \text{MEnc}(\mathcal{S}). \quad (4)$$

In this formulation, each video clip \mathcal{S} is encoded into an embedding, collectively forming \mathbf{E}_S^v . Here, we utilize the capital S to represent the complete clip set, with $|S|$ denoting the total clip count and d_v representing the visual embedding dimensionality. Our VideoRAG framework's indexing module $\varphi(\cdot)$ processes the video knowledge base $\mathcal{D} = \mathcal{V}_1, \dots, \mathcal{V}_n$ to create a hybrid index combining both knowledge graph and multi-modal context embeddings:

$$\hat{\mathcal{D}} = \varphi(\mathcal{D}) = (\mathcal{G}, \mathbf{E}_H^t, \mathbf{E}_S^v). \quad (5)$$

3.2 Multi-Modal Retrieval Paradigm

The Multi-Modal Retrieval Paradigm aims to efficiently retrieve relevant knowledge from videos in response to queries by integrating both textual semantic and visual content matching. Leveraging a

hybrid indexing framework $\hat{\mathcal{D}}$, this approach identifies informative video clips and generates query-specific descriptions using VLMs, ensuring a comprehensive retrieval process that captures both semantic understanding and visual context for more accurate responses.

- **Textual Semantic Matching.** The textual retrieval process leverages our constructed knowledge graph \mathcal{G} , where each entity contains a text description derived from relevant text chunks. The retrieval process consists of four sequential steps: (i) **Query Reformulation:** In the initial step, we employ LLMs to reformulate the input query into a declarative sentence, optimizing it for subsequent entity matching operations. (ii) **Entity Matching:** The system then calculates similarity scores between this reformulated query and entity descriptions within the knowledge graph, identifying the most relevant entities along with their associated text chunks. (iii) **Chunk Selection:** Following entity matching, we apply a GraphRAG [11]-based methodology to sort and identify the most pertinent chunks \mathcal{H}_q from the retrieved collection. (iv) **Video Clip Retrieval:** Finally, we extract video clips from the selected chunks, as each chunk contains descriptions of multiple video clips, resulting in our final textual retrieval set \mathcal{S}_q^t .
- **Visual Retrieval via Content Embeddings.** Our framework complements textual matching with direct visual retrieval, enabling semantic alignment between queries and video clips. Building upon our established visual indexing framework (Section 3.1.2), each video clip is encoded through a multi-modal encoder $\text{MEnc}(\cdot)$ to generate content-based embeddings. The visual retrieval process operates in two stages: (i) **Scene Information Extraction from Query:** We leverage LLMs to distill the query q into its core visual scene components, creating a focused scene description. For instance: Original question: “*In the movie, what color is the car that chases the main character through the city?*”; Scene-focused reformulation: “*An intense urban chase sequence featuring a car pursuing someone through city streets, with buildings and traffic in the background*” (ii) **Cross-Modal Feature Alignment:** This scene-centric query reformulation is projected into the same feature space as our visual embeddings using the multi-modal encoder, leveraging its cross-modal capabilities [9, 10] to align the context from different modalities. We compute similarity scores between the query embedding and video clip embeddings \mathbf{E}_S^v through cosine similarity, denoted as $\text{Sim}(\text{MEnc}(q), \mathbf{E}_S^v)$. The top-K matching clips form the visual retrieval result \mathcal{S}_q^v .
- **LLMs-based Video Clip Filtering:** To filter out noisy clips from the retrieved results, we employ LLMs to evaluate each clip $\mathcal{S} \in \mathcal{S}_q^t \cap \mathcal{S}_q^v$ for its relevance to query q using textual and visual information $\mathcal{V}_{\mathcal{S}}^t$ (Section 3.1.1). The filtered clips are formally defined as:

$$\{\hat{\mathcal{S}} \mid (\hat{\mathcal{S}} \in \{\mathcal{S}\}_q^t \cap \{\mathcal{S}\}_q^v) \wedge \text{LLMs-Judge}(\mathcal{V}_{\hat{\mathcal{S}}}^t) = 1\}, \quad (6)$$

where the function $\text{LLMs-Judge}(\cdot)$ serves as a binary decision maker that evaluates clip relevance via carefully-designed prompting instructions, returning 1 if the clip contains information vital to answering query q . This approach leverages LLMs’ advanced semantic reasoning capabilities to effectively filter out irrelevant clips while preserving key information.

3.3 Query-Aware Content Integration and Response Generation

With the retrieved video clips, we implement a two-stage content extraction process. First, we utilize LLMs to extract keywords \mathcal{K}_q from query q , which are then integrated into VLM prompts alongside sampled frames to generate detailed visual captions $\hat{\mathcal{C}}$:

$$\hat{\mathcal{C}} = \text{VLM}(\mathcal{K}_q, \hat{\mathcal{T}}, \{\mathbf{F}_1, \dots, \mathbf{F}_{\hat{k}}\} \mid \mathbf{F} \in \hat{\mathcal{S}}), \quad (7)$$

where $\hat{\mathcal{T}}$ represents the audio transcription for clip $\hat{\mathcal{S}}$, with a larger $\hat{k} > k$ sampled frames. For each clip $\hat{\mathcal{S}}_j$, we create a comprehensive description $\hat{\mathcal{V}}_j^t = (\hat{\mathcal{C}}_j, \hat{\mathcal{T}}_j)$ by combining its visual caption and transcription. These descriptions are collected into set $\hat{\mathcal{V}}^t$ for enhanced generation. We then enrich this visual analysis with traditional text-based retrieval, employing semantic similarity matching between query q and text chunks \mathcal{H} to obtain relevant textual information $\hat{\mathcal{H}}$. The complete output of our retrieval module $\psi(\cdot)$ thus comprises both query-specific video descriptions and relevant text chunks: $\psi(q, \hat{\mathcal{D}}) = (\hat{\mathcal{V}}^t, \hat{\mathcal{H}})$. Finally, VideoRAG leverages a general-purpose LLM (*e.g.*, GPT4 or DeepSeek) to generate responses based on the query q and retrieved content, as detailed in Section 2.

Table 1: Statistics of the experimental dataset *LongerVideos*.

Video Type	#video list	#video	#query	#avg. queries per list	#overall duration
Lecture	12	135	376	31.3	~ 64.3 hours
Documentary	5	12	114	22.8	~ 28.5 hours
Entertainment	5	17	112	22.4	~ 41.9 hours
All	22	164	602	27.4	~ 134.6 hours

4 Evaluation

We conduct comprehensive empirical evaluations of our VideoRAG framework on established benchmark datasets to address the following key research questions (RQs): **RQ1**: How effectively does VideoRAG perform in handling long-form video content compared to existing RAG alternative approaches? **RQ2**: What advantages does VideoRAG demonstrate over large vision models (LVMs) in understanding extremely long-context videos? **RQ3**: How do ablation studies reveal the effectiveness of individual components (textual and visual retrieval) in VideoRAG? **RQ4**: What insights can be derived from qualitative case studies of VideoRAG across diverse application scenarios?

4.1 Experimental Settings

Evaluation Datasets. Current benchmarks for video-based question answering are limited by relatively short durations (average <1 hour per video) [7] or single-video understanding scenarios (e.g., MLVU [5] and LVBench[6]). These constraints make it challenging to evaluate models’ capabilities in processing and reasoning across multiple extremely long-context videos for question-answering. To address this limitation in existing evaluation frameworks, we introduce **LongerVideos**, a comprehensive benchmark comprising over twenty video collections across three distinct categories:

- **Lecture Video:** Open-access educational content featuring contemporary technical topics, including AI Agents and RAG Techniques, delivered through comprehensive tutorials.
- **Documentary Video:** High-quality documentaries spanning wildlife exploration, natural landscapes, and expert interviews, each produced with professional cinematography.
- **Entertainment Video:** Diverse content including award ceremonies, gaming commentary with strategic analysis, and travel experiences documenting global cultural explorations.

All content is sourced from open-access YouTube videos, ensuring broad accessibility and reproducibility. Using NotebookLM³, we systematically generate an average of 25+ high-quality queries per collection by processing video transcripts. Each collection averages over 4 hours in total duration, containing between 1 to 20+ individual videos, ultimately yielding a robust evaluation set of 600+ diverse queries. Detailed statistical analysis of the benchmark is presented in Table 1.

Evaluation Protocols and Metrics. We implement two distinct protocols to evaluate model performance across different scenarios. The first protocol, **Win-Rate Comparison**, follows established Retrieval-Augmented Generation (RAG) evaluation methodologies [11, 12] using LLM-based judgment. This approach employs GPT-4o-mini to comparatively rank responses generated by two models, providing explanatory justification for each ranking decision. The second protocol, **Quantitative Comparison**, extends the LLM-based judgment by incorporating score assignment. It establishes a standard baseline answer for each query, against which other responses are evaluated on a 5-point scale, ranging from 1 (strongly worse) to 5 (strongly better).

We strategically apply these protocols for different evaluation purposes. The Win-Rate Comparison protocol is utilized to assess our methods against various RAG techniques and their ablation variants, enabling competitive analysis of our VideoRAG. Conversely, the Quantitative Comparison protocol facilitates fine-grained analysis when comparing VideoRAG with long video understanding methods. Following the framework established in [11], our evaluation encompasses multiple dimensions for comprehensive analysis, focusing on five distinct aspects detailed as follows:

³<https://notebooklm.google/>

Table 2: We analyze the performance of VideoRAG against RAG baselines on the LongerVideos dataset, presenting results both by individual video categories and across the complete dataset.

	Lecture		Documentary		Entertainment		All	
	NaiveRAG	VideoRAG	NaiveRAG	VideoRAG	NaiveRAG	VideoRAG	NaiveRAG	VideoRAG
Comprehensiveness	47.63%	52.37%	44.04%	55.96%	46.43%	53.57%	46.73%	53.27%
Empowerment	45.85%	54.15%	40.00%	60.00%	45.36%	54.64%	44.65%	55.35%
Trustworthiness	46.73%	53.27%	42.54%	57.46%	44.46%	55.54%	45.51%	54.49%
Depth	46.70%	53.30%	43.25%	56.75%	46.07%	53.93%	45.93%	54.07%
Density	46.73%	53.27%	44.21%	55.79%	44.29%	55.71%	45.80%	54.20%
Overall Winner	47.66%	52.34%	44.04%	55.96%	46.43%	53.57%	46.74%	53.26%
	GraphRAG- <i>l</i>	VideoRAG						
	44.60%	55.40%	48.68%	51.32%	49.29%	50.71%	46.25%	53.75%
Comprehensiveness	42.34%	57.66%	47.54%	52.46%	49.02%	50.98%	44.57%	55.43%
Empowerment	42.79%	57.21%	47.11%	52.89%	46.07%	53.93%	44.22%	55.78%
Trustworthiness	42.34%	57.66%	48.33%	51.67%	49.55%	50.45%	44.82%	55.18%
Depth	39.26%	60.74%	45.26%	54.74%	46.52%	53.48%	41.74%	58.26%
Density	44.44%	55.56%	48.68%	51.32%	49.20%	50.80%	46.13%	53.87%
	GraphRAG- <i>g</i>	VideoRAG						
	42.66%	57.34%	46.23%	53.77%	48.48%	51.52%	44.42%	55.58%
Comprehensiveness	39.55%	60.45%	44.04%	55.96%	48.30%	51.70%	42.03%	57.97%
Empowerment	38.54%	61.46%	41.49%	58.51%	43.48%	56.52%	40.02%	59.98%
Trustworthiness	40.61%	59.39%	45.26%	54.74%	47.23%	52.77%	42.72%	57.28%
Depth	37.55%	62.45%	46.93%	53.07%	48.04%	51.96%	41.28%	58.72%
Density	42.23%	57.77%	46.32%	53.68%	48.75%	51.25%	44.22%	55.78%
	LightRAG	VideoRAG	LightRAG	VideoRAG	LightRAG	VideoRAG	LightRAG	VideoRAG
	42.42%	57.58%	45.09%	54.91%	43.84%	56.16%	43.19%	56.81%
Comprehensiveness	39.55%	60.45%	38.95%	61.05%	42.05%	57.95%	39.90%	60.10%
Empowerment	39.52%	60.48%	42.11%	57.89%	40.00%	60.00%	40.10%	59.90%
Trustworthiness	40.13%	59.87%	41.93%	58.07%	41.96%	58.04%	40.81%	59.19%
Depth	39.57%	60.43%	42.37%	57.63%	41.61%	58.39%	40.48%	59.52%
Density	42.15%	57.85%	44.30%	55.70%	43.75%	56.25%	42.86%	57.14%

(i) **Comprehensiveness** evaluates answer coverage of question aspects. (ii) **Empowerment** measures how effectively the answer enables reader understanding and informed judgment. (iii) **Trustworthiness** assesses the answer’s credibility through detail sufficiency and alignment with common knowledge. (iv) **Depth** examines the presence of thorough analysis versus superficial information. (v) **Density** evaluates the concentration of relevant information while minimizing redundancy.

We implement two key strategies to ensure reliable results. First, to mitigate position-related bias in LLM inference, we alternate the answer positions within each prompt and collect two judgments per query during win-rate comparisons. Second, to minimize statistical variance, we perform five evaluation repetitions for both win-rate and quantitative assessments, then aggregate wins or calculate mean scores to determine final results. Complete evaluation prompts are provided in Appendix C.

Implementation Details of VideoRAG. For vision-text grounding (Section 3.1.1), we segment videos into 30-second clips and use $k = 5$ frames for initial visual captioning. We employ the quantized MiniCPM-V [13] as the VLM model and Distil-Whisper [14, 15] as the VSR model. For multi-modal encoding (Section 3.1.2), we utilize ImageBind [10] as MEnc(\cdot) for both visual and textual encoding. Entity and textual chunk retrieval leverage OpenAI’s text-embedding-3-small model, while section-run visual captioning uses an increased frame count of $\hat{k} = 15$. Throughout the implementation, GPT-4o-mini serves as our core LLM for indexing, retrieval, and answer generation. Complete implementation details are available in our open-source codebase.

4.2 Overall Comparison (RQ1 & RQ2)

We assess VideoRAG’s capabilities in comprehending long-form, multi-video content by comparing its retrieval-augmented generation performance against state-of-the-art RAG baselines.

- **NaiveRAG [16]:** A standard RAG implementation that segments documents into uniform-sized chunks and retrieves contextually relevant content through text embedding similarity matching, serving as a widely-adopted baseline for retrieval-augmented generation systems.
- **GraphRAG [11]:** An enhanced RAG system that leverages LLMs to construct entity knowledge graphs from input documents. It improves answer generation by performing community-based graph summarization to capture global context and relationships between entities.

Table 3: We conduct quantitative comparisons between VideoRAG and existing long-context video understanding models on the benchmark dataset. Each model’s performance is rated against NaiveRAG (our baseline) on a 5-point scale, where 1 indicates ‘strongly worse than baseline’ and 5 represents ‘strongly better than baseline’. We evaluate across three video categories: lectures (‘lec’), documentaries (‘doc’), and entertainment content (‘ent’), with ‘all’ representing the aggregate performance.

	LLaMA-VID				VideoAgent				NotebookLM				VideoRAG			
	lec	doc	ent	all	lec	doc	ent	all	lec	doc	ent	all	lec	doc	ent	all
Comprehensiveness	2.36	2.62	2.54	2.44	2.02	1.99	1.80	1.98	3.53	3.20	2.96	3.36	4.48	4.51	4.44	4.48
Empowerment	2.79	3.03	2.86	2.85	2.42	2.37	2.10	2.35	3.88	3.62	3.29	3.72	4.51	4.55	4.45	4.51
Trustworthiness	3.15	3.30	3.35	3.22	2.83	2.73	2.65	2.78	3.95	3.80	3.71	3.88	4.50	4.54	4.48	4.50
Depth	2.01	2.06	2.00	2.02	1.79	1.75	1.62	1.75	3.14	2.89	2.55	2.98	4.34	4.42	4.31	4.35
Density	3.15	3.28	3.21	3.18	2.82	2.73	2.52	2.75	4.07	3.82	3.61	3.94	4.59	4.63	4.56	4.59
Overall Score	2.36	2.61	2.54	2.44	2.03	2.01	1.80	1.98	3.54	3.21	2.97	3.37	4.45	4.49	4.41	4.45

- **LightRAG [12]:** A lightweight graph-based RAG framework that implements dual-level retrieval architecture, integrating both low-level and high-level semantic knowledge discovery. The system enables efficient and contextually-aware document retrieval to process complex queries.

Details of Baseline Implementation. To ensure fair comparison, we implement all baseline methods with the following specifications: • **Input Data:** We utilize grounded textual knowledge (*e.g.*, visual captions and transcripts) from all videos, employing identical chunk-splitting protocols as our method. • **Visual Processing:** For frame-level analysis, we maintain 15 frames per video clip for visual caption generation, matching the fine-grained section-run captions produced by VideoRAG’s retrieval process. • **Baseline Variants:** GraphRAG: Implemented with both local (GraphRAG-*l*) and global (GraphRAG-*g*) search capabilities; LightRAG: Deployed with full hybrid search functionality.

Comparison Results and Analysis (RQ1). Table 2 presents the win rate evaluation results comparing VideoRAG with baseline methods. Our analysis reveals several significant findings:

- **Superior Video-based RAG Performance.** Our evaluation demonstrates that VideoRAG consistently outperforms all baseline methods across performance metrics. The superior performance stems from our innovative multi-modal video knowledge indexing framework, which combines graph-based knowledge grounding with multi-modal context encoding, enabling effective capture and organization of visual dynamics and semantic information across videos.

Furthermore, VideoRAG’s multi-modal retrieval paradigm significantly enhances performance through its hybrid approach to knowledge discovery. By integrating textual semantic matching with visual content embedding-based retrieval, the system achieves precise and contextually relevant video clip retrieval. This comprehensive retrieval strategy enables more accurate and nuanced responses compared to traditional single-modality approaches, while ensuring the extracted information remains semantically coherent and contextually appropriate.

- **Performance Analysis Across Baseline Methods.** In comparison with NaiveRAG, VideoRAG demonstrates exceptional performance across evaluation dimensions, with particular strengths in Comprehensiveness and Empowerment. This superiority stems from our effective knowledge indexing framework, which interlinks information across multiple videos, enabling sophisticated synthesis of diverse information during retrieval and yielding more comprehensive responses.

When benchmarked against GraphRAG and LightRAG, VideoRAG achieves superior performance through multi-modal context integration capabilities. Our approach excels in two aspects: (1) sophisticated knowledge indexing that effectively fuses visual-textual information, and (2) query-aware retrieval that leverages unified multi-modal representations for precise content selection. This architecture enables more nuanced understanding and contextually coherent response generation, significantly outperforming existing methods in knowledge-grounded video question answering.

To establish comprehensive performance benchmarks, we evaluate VideoRAG against state-of-the-art large vision models specifically designed for long-context video understanding, encompassing both advanced vision-language models and intelligent agent systems.

- **LLaMA-VID [17]:** A vision-language framework that leverages context and content tokens for efficient long video processing, addressing token complexity in video understanding tasks.

- **VideoAgent [18]**: A multi-modal agent that integrates diverse foundation models through a unified memory architecture. It enables powerful video understanding through fine-grained object detection, tracking, and modeling of temporal dependencies within short video clips.
- **NotebookLM**: An assistant system by Google designed for video content analysis. It enables efficient multi-video comprehension and information retrieval through advanced transcript analysis, allowing users to extract contextually coherent insights across multiple video sources.

Details of Baseline Implementation. We implement all baselines using their official codebases or available platforms for fair evaluation. For vision-language model baselines like LLaMA-VID, we standardize the implementation through three crucial modifications: (i) Replacing their original ASR model with the one used in VideoRAG for consistent transcript extraction; (ii) Uniformly sampling 3,600 frames per video due to GPU memory constraints (48GB per GPU); and (iii) Employing prompting instructions consistent with the compared RAG methods.

Comparison Results and Analysis (RQ2). We present comprehensive performance comparisons between VideoRAG and existing video understanding methods in Table 3. The results demonstrate that our model consistently outperforms all compared long-context video understanding methods across various metrics. We attribute these improvements to the following aspects:

- **Enhanced Long-Context Modeling.** VideoRAG proposes a graph-enhanced multi-modal indexing and retrieval mechanism that significantly extends video processing capabilities beyond existing vision models. Unlike conventional approaches that face length constraints, our model can effectively handle unlimited long-context videos by establishing and leveraging cross-video knowledge connections and inter-dependent relations. This architectural advantage enables comprehensive knowledge extraction and integration across extended video sequences, surpassing models like LLaMA-VID that are limited by computational constraints when processing video frames directly.
- **Superior Multi-Modal Fusion.** VideoRAG excels in capturing, reasoning, and aligning diverse multi-modal contexts through advanced cross-modal knowledge integration. Our approach effectively fuses information across different modalities (visual, audio, and textual), enabling superior cross-modal alignment and comprehensive understanding. This multi-modal synthesis surpasses single-modality focused approaches like VideoAgent (visual-only) and NotebookLM (transcript-only), leading to more nuanced, coherent, and expressive video understanding.

4.3 Ablation Study (RQ3)

To evaluate the effectiveness of our multi-modal indexing and retrieval design, we conduct comprehensive ablation studies using two model variants:

- **Variant 1 (-Graph):** Removes the graph-based index-retrieval pipeline, limiting the model’s ability to establish multi-video relationships.
- **Variant 2 (-Vision):** Eliminates the visual indexing and retrieval component from the multi-modal encoder.

The ablation study results in Figure 2 reveal the crucial contribution of each component to the model performance of VideoRAG, as evidenced by the following analyses:

- The *-Graph* variant exhibits significant performance degradation across all evaluation metrics, demonstrating that our graph-based index-retrieval mechanism is essential for two key capabilities: (1) capturing complex inter-video relationships and (2) establishing cross-video knowledge dependencies. This validates the effectiveness of our graph-enhanced architecture in connecting and synthesizing information across multiple videos.
- The *-Vision* variant shows substantially decreased win rates, underscoring the critical role of visual information processing in our framework. This performance drop validates our model’s effective multi-modal context fusion mechanism, which successfully integrates and aligns visual features with other modalities for comprehensive video understanding.

These comprehensive findings underscore that the synergistic integration of graph-based architecture and visual modality processing serves as a cornerstone for achieving superior performance in multi-modal indexing and retrieval tasks, validating our architectural design choices.

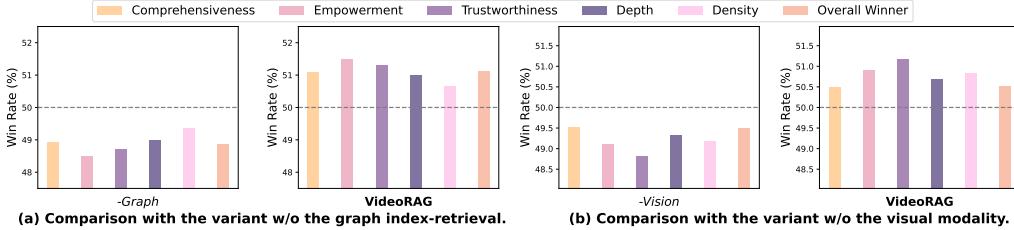


Figure 2: Ablation on graph-based knowledge grounding and cross-modal retrieval components.

4.4 Case Study Analysis (RQ4)

To comprehensively evaluate VideoRAG’s capabilities, we conduct a detailed case study examining its response to a specific query: “*the role of graders in reinforcement fine-tuning*”, drawn from OpenAI’s landmark 12-day video series released in late 2024 [as educational materials]. The query’s target information is primarily located in Day 2’s content, which details OpenAI’s systematic and innovative approach to model enhancement through reinforcement fine-tuning techniques.

Retrieval Accuracy and Response Quality. Table 6 presents VideoRAG’s response alongside its retrieved video clips. Our analysis reveals that VideoRAG successfully identified and extracted relevant content from Day 2, specifically focusing on reinforcement fine-tuning discussions within the broader context of the 12-video series. The retrieved two-minute clips comprehensively cover: (1) Fundamental concepts of graders; (2) Operational mechanisms of the grading system; (3) Practical examples of partial credit assignment. Within the table, we highlight portions of VideoRAG’s response that directly correspond to the retrieved video clips. This visualization demonstrates how VideoRAG leverages retrieved information to construct detailed and well-supported answers.

Comparative Performance. A comparative analysis with LightRAG (detailed in Appendix B) reveals performance distinctions in handling technical content. While both models successfully convey the core concepts of “*the grading system*” in reinforcement learning, LightRAG’s response demonstrates notable limitations in granularity and scope. Specifically, LightRAG’s output lacks crucial technical elements in explaining “*grader scoring mechanisms*”. Although LightRAG’s response maintains fundamental accuracy, it falls short of the comprehensive depth and technical precision exhibited by VideoRAG, which provides a nuanced explanation of the grading system’s intricacies.

Key Findings. This case study provides compelling evidence of VideoRAG’s effectiveness in three critical areas: its ability to construct precise knowledge graph structures that capture complex relationships, its successful leverage of multi-modal information for highly accurate content retrieval, and its enhanced capability to process and synthesize information from multiple long-context videos. These capabilities collectively demonstrate VideoRAG’s advanced proficiency in handling sophisticated multi-modal tasks while maintaining high standards of accuracy and relevance.

5 Related Work

Retrieval-Augmented Generation. RAG has emerged as a pivotal paradigm in elevating performance of LLMs. By seamlessly integrating relevant information retrieved from external databases, these systems are able to ground their responses in rich, factual, and domain-specific knowledge [12, 19, 20]. At the core of the RAG process lie three essential components: indexing, retrieval, and generation. First, raw data is meticulously processed and structured into a comprehensive database. Next, this database is intelligently queried to retrieve most pertinent information based on user inputs. Finally, this retrieved knowledge is leveraged to generate informed and insightful responses.

Recent advancements in RAG have followed two distinct methodological trajectories. Chunk-based approaches [16, 1, 21] have focused on optimizing text segmentation and retrieval through advanced vector space embeddings. In parallel, graph-based methods [11, 12, 22] have explored the use of structured knowledge representations to enhance the efficiency and precision of the retrieval process. Concurrent to these text-centric innovations, the research community has also made significant strides in developing multi-modal RAG systems [23–25], leveraging databases as rich, multi-faceted documents to enrich the knowledge retrieval and generation capabilities.

Table 4: A case study examining VideoRAG’s response to a query about “*graders*” in OpenAI’s 12-day educational series demonstrates its video retrieval capabilities and knowledge synthesis.

Query: Explain the purpose and functionality of ‘graders’ in the context of reinforcement fine-tuning.
Input Videos: 12 Days of OpenAI (A total of 12 videos)
VideoRAG Response:
<pre>### Understanding Gradation in Reinforcement Fine-Tuning **Purpose of Graders** Graders are integral components in the process of reinforcement fine-tuning, mainly utilized to evaluate and improve the performance of machine learning models... (Omitted for brevity) **Functionality of Graders** The grading system operates by comparing a model's outputs against those of verified correct responses. The mechanics of this assessment involve assigning scores to the model's outputs based on their correctness. These scores range from zero to one, where: - A score of **0** indicates that the model's answer is completely incorrect. - A score of **1** signals that the answer is fully correct. - Scores can also be fractional, allowing for partial credit. For instance, a score of **0.7** may denote that the model's suggestion was correct but not the highest in a ranked list of possible answers.</pre>
<p>The process includes:</p> <ol style="list-style-type: none"> 1. Input Comparison: The grader takes both the model’s output and the correct answer, analyzing them to derive a score. 2. Graded Outputs: This interaction allows for a dynamic reflection of the model’s performance, where the scores directly influence how the model is fine-tuned in subsequent training iterations. 3. Feedback Loop: The scores obtained from graders provide feedback that can be used to adjust learning parameters and improve the model’s future responses. This highlights the feedback mechanism that is crucial for reinforcement learning.
(The remaining content are omitted for brevity.)

Retrieved Sources by VideoRAG:
 VideoRAG accurately retrieved four continuous video segments from the OpenAI’s 12-day show, ranging from 10:00 to 12:00 in Day 2 of the video “Reinforcement Fine-Tuning.” Here, we highlight key moments relevant to the detailed content in the answer. From left to right, these are retrieved moments at timestamps **10:35**, **10:39**, and **11:10**, which provide informative insights that help VideoRAG give a comprehensive answer to the query.



However, one crucial medium of knowledge remains relatively underexplored in the context of RAG – videos. Preliminary efforts, such as MM-VID [25] and iRAG [26], have taken initial steps to bridge this gap, but substantial challenges remain in effectively organizing and extracting video-based knowledge. This is where the VideoRAG framework stands as a groundbreaking innovation. By synthesizing state-of-the-art techniques from both text-based and multi-modal RAG approaches, VideoRAG constructs a comprehensive knowledge graph that seamlessly integrates knowledge from multiple long-form video sources. Coupled with its multi-modal retrieval matching capabilities,

VideoRAG empowers LLMs to tap into the wealth of information inherent in video content, elevating their ability to provide accurate, informed, and contextually relevant responses to user queries.

Long Video Understanding Extracting meaningful knowledge from long-context videos poses a challenge in the domain of video understanding. Traditional approaches, such as large video language models (LVLMs), have made significant strides by converting video frames into vision tokens for comprehension by large language models [27, 6, 28, 17, 29–31]. However, as video length and quantity increase, the computational demands also escalate, necessitating a balance between video length and available resources. This has motivated the exploration of more efficient and scalable solutions to address the growing need for effective long-video understanding.

To this end, our innovative approach presents a novel framework that leverages an efficient video language model for video segment-based general knowledge extraction and query-specific information retrieval. By constructing a graph that integrates information from multiple videos with visual features, we enhance the precision and comprehensiveness of query responses, while accommodating input videos of arbitrary lengths and quantities. This contrasts with existing agent- or retrieval-augmented generation-based methods [18, 32–34], which rely heavily on external tools for frame-level information extraction, limiting their capacity to respond to diverse queries due to the inherent constraints of these tools. Our holistic approach, which seamlessly combines efficient video understanding with advanced knowledge organization and retrieval, represents a significant advancement in the field, poised to unlock new possibilities in long-video comprehension and query-answering.

6 Conclusion

This paper introduces VideoRAG, a novel retrieval-augmented generation framework designed for understanding extremely long-context videos. Through a dual-channel architecture that seamlessly integrates graph-based textual knowledge grounding with multi-modal context encoding, VideoRAG effectively processes, indexes, and retrieves information from unlimited-length videos for large language model enhancement. Our comprehensive empirical evaluation on the established LongerVideos benchmark demonstrates VideoRAG’s superior performance compared to existing RAG alternatives and long video understanding methods across multiple dimensions. The framework’s demonstrated capabilities—constructing precise video knowledge structures, leveraging multi-modal information for accurate content retrieval, and processing information from multiple long-context videos—showcase its significant potential for advancing video-based knowledge retrieval and generation tasks.

References

- [1] Ritvik Aggarwal Ishneet Sukhvinder Singh Ibrahim Allahverdiyev, Muhammad Taha, Aslihan Akalin, and Kevin Zhu. Chunkrag: Novel llm-chunk filtering method for rag systems. *arXiv preprint arXiv:2410.19572*, 2024.
- [2] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *NeurIPS*, 2024.
- [3] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.
- [4] Yuanhan Zhang, Jinning Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.
- [5] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.
- [6] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.
- [7] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [8] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36:46212–46244, 2023.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.
- [10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023.
- [11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- [12] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.
- [13] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR, 2023.
- [15] Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.
- [16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [17] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2025.

- [18] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, pages 75–92. Springer, 2025.
- [19] Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*, 2024.
- [20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [21] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- [22] Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. *arXiv preprint arXiv:2410.20724*, 2024.
- [23] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.
- [24] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.
- [25] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023.
- [26] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. irag: Advancing rag for videos with an incremental approach. In *CIKM*, pages 4341–4348, 2024.
- [27] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.
- [28] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*, 2024.
- [29] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024.
- [30] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
- [31] Yuzhang Shang, Bingxin Xu, Weitai Kang, Mu Cai, Yuheng Li, Zehao Wen, Zhen Dong, Kurt Keutzer, Yong Jae Lee, and Yan Yan. Interpolating video-lmms: Toward longer-sequence lmms in a training-free manner. *arXiv preprint arXiv:2409.12963*, 2024.
- [32] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, pages 58–76. Springer, 2025.
- [33] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024.
- [34] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. *arXiv preprint arXiv:2406.12846*, 2024.

A Details of LongerVideos

LongerVideos is a comprehensive benchmark dataset designed to evaluate a model’s ability to comprehend and extract knowledge from long-form videos. By leveraging semantic connections across multiple sources, the dataset facilitates the development of efficient, knowledge-based question-answering. The core methodology presents the model with diverse video lists of varying lengths, then assesses the model’s output in terms of completeness, accuracy, and diversity. This holistic approach ensures the evaluated models demonstrate a robust understanding of the content, the ability to synthesize information, and the aptitude to generate well-rounded responses.

- **Input Data:** A diverse collection of long videos, with durations ranging from minutes to hours.
- **Question:** A series of open-ended questions carefully tailored to the provided video list.
- **Expected Output:** Individual responses generated based on the information extracted from videos.

The LongerVideos dataset was constructed by systematically curating diverse video lists from YouTube, leveraging the platform’s structure where creators often compile thematic content. A major data source comprised online course videos, typically segmented into multiple recordings corresponding to distinct course chapters. For each video, the team employed the yt-dlp tool to download the content in 720P resolution, after which they prepared open-ended questions for each list with the assistance of NotebookLM, a robust multi-video understanding model from Google that can process various videos as input to generate relevant answers. The final LongerVideos dataset consists of 22 carefully curated video lists, with detailed statistics provided in Table 5.

Table 5: Detailed statistics of the *LongerVideos* dataset.

Video Type	video list name	#video	#query	#overall duration
Lecture	climate-week-at-columbia-engineering	4	26	5.91 hours
	rag-lecture	19	38	5.34 hours
	ai-agent-lecture	39	45	9.35 hours
	daubechies-wavelet-lecture	4	25	8.97 hours
	daubechies-art-and-mathematics-lecture	4	21	4.87 hours
	tech-ceo-lecture	4	31	4.83 hours
	dspy-lecture	9	38	4.22 hours
	trading-for-beginners	2	23	4.11 hours
	ahp-superdecision	11	24	2.40 hours
	decision-making-science	4	26	2.20 hours
Documentary	12-days-of-openai	12	35	3.43 hours
	autogen	23	44	8.70 hours
	fights-in-animal-kingdom	1	11	3.00 hours
	nature-scenes	1	17	3.98 hours
	education-united-nations	6	39	8.41 hours
Entertainment	elon-musk	1	13	8.63 hours
	jeff-bezos	3	34	4.47 hours
	black-myth-wukong	10	23	21.36 hours
	primetime-emmy-awards	3	17	7.31 hours
	journey-through-china	1	27	3.37 hours
Entertainment	fia-awards	1	27	3.02 hours
	game-awards	2	18	6.73 hours

B Details of Case Study

This section provides further details on the case study presented in Section 2, which investigates the purpose and functionality of ‘graders’ in the context of reinforcement fine-tuning. The investigation utilizes input from the “12 Days of OpenAI” video series, comprising 12 videos that showcase OpenAI’s activities in late 2024. To effectively answer the question, the model retrieves relevant content that specifically discusses the role of graders within the reinforcement fine-tuning context. To further illustrate our model’s capabilities in retrieving detailed information from videos for generating nuanced answers, we also present a response from another retrieval-augmented generation model, LightRAG, for comprehensive analysis. A comparison of the generated answer by our model, as

Table 6: Further details on the case study concerning the responses of LightRAG to the same question.

Query: Explain the purpose and functionality of 'graders' in the context of reinforcement fine-tuning.
Input Videos: 12 Days of OpenAI (A total of 12 videos)
LightRAG Response:
Purpose and Functionality of Graders in Reinforcement Fine-Tuning Graders serve a crucial role within the context of reinforcement fine-tuning, particularly focusing on evaluating and improving the performance of machine learning models. These evaluative tools assist in determining how well a model aligns with desired outcomes based on specific criteria.
Evaluation Process In reinforcement fine-tuning, graders provide a quantitative measure for assessing the effectiveness of a model's outputs. They help in evaluating the responses generated by the AI against predefined correct answers or standards. This comparison is essential as it offers insights into areas where the model performs well and where it may require further improvements. The process involves analyzing the outputs of the model after it has undergone reinforcement learning techniques. By examining the consistency and accuracy of the generated responses, graders facilitate the identification of discrepancies and help in recalibrating the model to optimize its performance further.
Role in Feedback Loop Grading is also integral to the feedback mechanism within reinforcement fine-tuning. The insights gained from the grading process inform the adjustments necessary for the model's reinforcement learning parameters. In this way, graders contribute to refining the model's learning algorithms, effectively guiding the subsequent training iterations. Overall, the implementation of graders in reinforcement fine-tuning exemplifies an iterative improvement strategy, where ongoing assessment and feedback drive enhancements in model performance. This systematic approach not only improves the quality of outputs but also ensures that the model continues evolving to meet dynamic user and operational requirements.

shown in Table 1, reveals that it provides greater detail, including the scoring criteria used by graders and specific examples illustrating the partial scores assigned. Moreover, it delineates the processes through which graders operate in reinforcement learning, thereby enhancing the overall quality of the generated response. Consequently, the win-rate comparison with GPT-4o-mini emphasizes the empowerment and trustworthiness of the response generated by our VideoRAG.

- **Empowerment:** Our VideoRAG empowers the reader by explaining how graders work in a practical sense, including details on scoring, input comparison, and their role in the feedback loop. This allows readers to grasp how to utilize graders effectively in machine learning contexts.
- **Trustworthiness:** Our VideoRAG aligns closely with established concepts in machine learning and provides more operational details, enhancing credibility. It clearly explains the scoring method and implications for model performance, reinforcing its trustworthy nature.

C Instructions for Win-Rate & Quantitative Performance Comparison

We present the instructions employed for LLM-based evaluation in Figure 3, which includes both win-rate comparison and quantitative comparison. For the win-rate comparison, we input the query alongside two competing answers, designated as `answer1` and `answer2`, while alternating their positions across multiple iterations to mitigate any positional bias affecting LLM inference.

In the quantitative comparison, we leverage a standard answer from NaiveRAG [16] labeled `baseline_answer`, against which the evaluated answer, referred to as `evaluation_answer`, is assessed. The LLM assigns a score from 1 to 5, indicating whether the evaluated answer is inferior or superior to the baseline. This instruction allows us to compare the outputs of multiple models against the same standard answer, thus eliminating the need to adjust their positions. Since all methods are evaluated consistently against the same baseline, positional bias is inherently mitigated, enabling a direct comparison of scores across different methods.

Instructions for Win-Rate & Quantitative Comparison

Win-Rate Comparison

You will evaluate two answers to the same question based on these criteria: **Comprehensiveness**, **Empowerment**, **Trustworthiness**, **Depth** and **Density**.

- **Comprehensiveness**: How much detail does the answer provide to cover all aspects and details of the question?
- **Empowerment**: How well does the answer help the reader understand and make informed judgments about the topic?
- **Trustworthiness**: Does the answer provide sufficient detail and align with common knowledge, enhancing its credibility?
- **Depth**: Does the answer provide in-depth analysis or details, rather than just superficial information?
- **Density**: Does the answer contain relevant information without less informative or redundant content?

For each criterion, choose the better answer (either Answer 1 or Answer 2) and explain why. Then, select an overall winner based on these criteria.

Here is the question: {query}

Here are the two answers:

Answer 1: {answer1}

Answer 2: {answer2}

Evaluate both answers using the criteria listed above and provide detailed explanations for each criterion. Output your evaluation in the following JSON format:

(The remaining content are omitted for brevity.)

Quantitative Comparison

You are an expert evaluating an answer against a baseline answer based on these criteria: **Comprehensiveness**, **Empowerment**, **Trustworthiness**, **Depth** and **Density**.

(We omit the similar part on win-rate comparison here for brevity.)

For the evaluated answer labeled "Evaluation Answer," assign a score from 1 to 5 for each criterion compared to the baseline answer labeled "Baseline Answer." Then, assign an overall score based on these criteria. The evaluation scores are defined as follows:

- 1: Strongly worse than the baseline answer
- 2: Weakly worse than the baseline answer
- 3: Moderate compared to the baseline answer
- 4: Weakly better than the baseline answer
- 5: Strongly better than the baseline answer

Here is the question: {query}

Here are the answers:

Baseline Answer: {baseline_answer}

Evaluation Answer: {evaluation_answer}

Evaluate the answer using the criteria listed above and provide detailed explanations for the scores. Output your evaluation in the following JSON format:

(The remaining content are omitted for brevity.)

Figure 3: Instructions for LLM-based answer comparison and scoring