ViG-RAG: Video-aware Graph Retrieval-Augmented Generation via Temporal and Semantic Hybrid Reasoning

Zongsheng Cao[1,4*], Anran Liu[2*], Yangfan He[2], Jing Li[3], Bo Zhang[1†], Zigan Wang[3,4†]

[1]Shanghai Artificial Intelligence Laboratory
[2]Independent Researcher
[3]School of Economics and Management, Tsinghua University
[4]Shenzhen International Graduate School, Tsinghua University
agiczsr@gmail.com, anniegogo1008@gmail.com, he00577@umn.edu, lijing3@sem.tsinghua.edu.cn
bo.zhangzx@gmail.com, wangzigan@sz.tsinghua.edu.cn

Abstract

Retrieval-augmented generation (RAG) has greatly improved large language models (LLMs) by adding external knowledge. However, existing RAG-based methods face two major challenges in long-context video understanding. First, they struggle to jointly encode multimodal and long-range temporal information, leading to fragmented and context-insensitive knowledge representations. Second, their retrieval mechanisms typically rely on static text matching, which fails to dynamically align user queries with the most relevant video segments, ultimately degrading downstream performance. To overcome these issues, we introduce **ViG-RAG**, a new framework to enhance long-context video understanding through structured textual knowledge grounding and multi-modal retrieval. Specifically, we treat video transcripts into structured units, extract key entities, form temporal connections and confidence for evidence, enabling coherent long-range reasoning. In this way, it utilizes a knowledge-aware grounding mechanism and a context-aware retrieval process that dynamically builds a probabilistic temporal knowledge graph to organize multi-video content. To improve retrieval accuracy, we propose a hybrid retrieval strategy for semantic and temporal features, with an adaptive distribution modeling the relevance. In this way, it achieves the optimal retrieval distribution for each query, enhancing generation efficiency by reducing unnecessary computations. On top of this, ViG-RAG uses a vision-language model to integrate semantic anchors, expanded contextual fields, and selected video frames, generating an accurate response. We evaluate ViG-RAG on several benchmarks, demonstrating that it significantly surpasses current RAG-based methods.

Introduction

Large language models (LLMs) (Yu et al. 2024a; Cao et al. 2025c; Faysse et al. 2024; Cao et al. 2025b) have shown strong performance in a variety of NLP tasks(Cao et al. 2024), such as question answering, summarization, and dialogue generation. However, their reasoning ability remains limited by the static knowledge encoded during pre-training. To overcome this, retrieval-augmented generation (RAG) (Gao et al. 2023a; Cao et al. 2025a) has emerged as a promising framework that enables LLMs to retrieve and incorporate external information during inference, leading to more accurate and context-aware responses (Allahverdiyev et al. 2024; Gutiérrez et al. 2024).

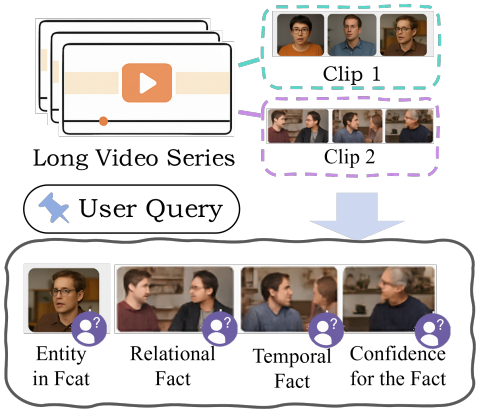

Figure 1: The illustration of the complexitity in the video, including entities, relations, temporal information and confidence for the fact.

While RAG (Edge et al. 2024; Guo et al. 2024) has been widely applied to text-based tasks, its extension to the multimodal domain, particularly video understanding, remains underexplored. In fact, many real-world video applications (e.g., documentary indexing, or instructional content understanding) demand up-to-date reasoning grounded in external or distributed knowledge. RAG offers a natural pathway to enhance these tasks by retrieving supporting evidence across different video segments or even across videos.

At the same time, large vision-language models have achieved notable progress in short video understanding. However, they still struggle in long-context scenarios. These models usually process isolated clips and fail to reason across extended temporal spans. When long visual context reasoning is required, splitting videos into short, unlinked segments causes a substantial loss of temporal and semantic continuity.

*These authors contributed equally.
†Corresponding author

Under these circumstances, this paper addresses this limitation by asking:

Can we design a RAG framework for long video understanding that preserves semantic coherence, enables temporal reasoning, and supports context-aware retrieval?

To explore this, we investigate the design of RAG models specifically for long and complex video content. Recent studies (Edge et al. 2024; Zhang et al. 2025) have made progress in this direction, as shown in Figure 1, but two key challenges remain unresolved:

(C1) **Structurally modeling for multi-modal and temporal information**: Existing methods struggle to build unified representations that combine visual, textual, and temporal features and the confidence for the factual evidence over long durations. This leads to fragmented and incomplete knowledge integration.

(C2) **Query-sensitive retrieval under ambiguity**: Current retrieval mechanisms often fail to align ambiguous or underspecified user queries with the most relevant video segments, especially when cross-video information is required.

To this end, we propose **Vi**deo-aware **G**raph **R**etrieval-**A**ugmented **G**eneration (**ViG-RAG**) via temporal and semantic hybrid reasoning, which leverages temporal-semantic hybrid reasoning to unify knowledge-grounded generation and context-aware retrieval for efficient organization and access of multi-video content. Specifically, to tackle (C1), the framework begins with constructing a probabilistic temporal knowledge graph, which systematically segments video transcripts into structured units, extracts key entities, and establishes temporal connections and confidence for fact evidence. Unlike traditional RAG methods that rely on static textual retrieval, this dynamic knowledge grounding mechanism preserves contextual dependencies across videos, enabling effective cross-video indexing. In this way, textual retrieval identifies relevant knowledge chunks based on entity relationships, while visual retrieval aligns query embeddings with video representations to extract semantically relevant clips.

To address (C2) and refine retrieval precision, ViG-RAG introduces a lightweight, training-free filtering mechanism based on Gaussian Mixture Modeling (GMM). By modeling the distribution of similarity scores among retrieved candidates, our method automatically identifies high-confidence segments without relying on handcrafted thresholds or additional supervision. This enables robust Top-K selection tailored to each query's characteristics. Complementing this, ViG-RAG also employs a query-aware re-ranking strategy guided by LLMs, further filtering noisy segments while preserving essential contextual information. This multi-stage pipeline ensures that generated responses are both aligned with user queries and rich in multimodal semantics. Extensive evaluation on benchmark datasets, including LongerVideos, demonstrates that ViG-RAG significantly outperforms existing RAG-based methods and long-context video understanding frameworks, particularly in retrieving and organizing long-form video content with high precision and coherence. The code can refer to https://github.com/AI-Researcher-Team/ViG-RAG.

In summary, our contributions are threefold:

- We propose ViG-RAG, a novel retrieval-augmented framework for long-context video understanding that jointly models multimodal and temporal information via a fuzzy temporal knowledge graph, enabling unified and context-aware knowledge integration across videos.
- We design a semantic-temporal dual-level retrieval module with a GMM filtering mechanism, which adaptively identifies high-confidence evidence segments and robustly aligns ambiguous user queries with relevant video content.
- We demonstrate that ViG-RAG achieves state-of-the-art performance on multiple established benchmarks, significantly surpassing existing RAG-based and long-context video understanding models in both retrieval accuracy and downstream reasoning.

Related Work

Retrieval-Augmented Generation. The paradigm of RAG has rapidly emerged as a key innovation for empowering large language models to deliver knowledge-rich and contextually accurate outputs. Central to the RAG pipeline are the processes of constructing a structured information base, efficiently retrieving contextually pertinent segments, and integrating them into generative reasoning. Unlike traditional models constrained by static pretraining, RAG frameworks dynamically ground responses in up-to-date and domain-specific external knowledge, thereby substantially expanding their factual coverage and adaptability (Guo et al. 2024; Qian et al. 2024; Gao et al. 2023a). This transformation has enabled LLMs to effectively address real-world information needs that demand both precision and depth.

Recent research in RAG has pursued a spectrum of methodological advances, reflecting the growing complexity and diversity of knowledge resources. On one front, chunk-oriented approaches (Gao et al. 2023b; Allahverdiyev et al. 2024; Chan et al. 2024) refine information granularity and retrieval relevance by leveraging powerful embedding techniques and optimized segmentation protocols. Meanwhile, graph-based RAG systems (Edge et al. 2024; Guo et al. 2024; Li, Miao, and Li 2024) harness explicit structural representations to enhance both retrieval efficiency and semantic precision. In parallel, multi-modal RAG research RAG (Yu et al. 2024b; Faysse et al. 2024; Lin et al. 2023) has focused on integrating heterogeneous content (e.g., images, audio, video) to enable richer evidence synthesis and more robust reasoning across diverse applications.

Afterward, leveraging video data as a source of structured knowledge remains a largely untapped frontier. Early explorations, such as MM-VID (Lin et al. 2023) and iRAG (Arefeen et al. 2024), demonstrate that extracting and utilizing meaningful information from video content poses unique technical challenges. These include not only the inherent temporal and semantic complexity of videos but also the lack of established methodologies for integrating video-derived knowledge into large language model workflows.

Long Video Understanding. Deriving meaningful insights from long-context videos remains a complex challenge in video understanding. Conventional techniques, including

large video language models (LVLMs), have made considerable progress by converting video frames into vision tokens, enabling their interpretation by large language models (Wu et al. 2024a; Wang et al. 2025; Chandrasegaran et al. 2024; Li, Wang, and Jia 2025; Shu et al. 2025; Shang et al. 2024). However, as video lengths and dataset sizes grow, computational costs scale sharply, making it critical to develop more resource-efficient and scalable methods for long-video processing. VideoRAG (Ren et al. 2025) proposes to utilize a knowledge graph to conduct video RAG; however, they neglect the complex association of components in the video, and cannot conduct adaptive selection for the retrieval.

To address this, we introduce a novel framework ViG-RAG. By constructing a probabilistic knowledge graph that fuses multi-video information with visual embeddings, our method enhances the depth and accuracy of query responses while accommodating videos of arbitrary length and scale.

Methodology

In this section, we propsoe a new model termed ViG-RAG, which stands for Video-based RAG. The model is designed to address the challenges of long-context video understanding by leveraging a structured probabilistic temporal knowledge graph (PTKG) to organize and retrieve multi-modal information from videos. The overall framework is illustrated in Figure 2. The problem setting are as follow:

Definition 1 (RAG for Video)**.** *Given a user query q, a collection of videos $\mathcal{V} = \{v_1, \ldots, v_N\}$, and their associated multi-modal streams (visual, textual, temporal), the goal is to generate a response r by retrieving a relevant subset $\mathcal{C}_q \subset \mathcal{V}$ and grounding the answer in semantically and temporally aligned evidence. Formally,*

$$r = \mathcal{G}(q, \mathcal{R}(q, \mathcal{I}(\mathcal{V}))),$$

where $\mathcal{I}$ denotes the multi-modal indexing function, $\mathcal{R}$ the context-aware retriever, and $\mathcal{G}$ the generator conditioned on the retrieved context.

Preliminary for Knowledge Structure. The knowledge graph (Zhang et al. 2019; Cao et al. 2021) is an useful tool to model complex structures. Let $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities and relations, respectively. A knowledge graph fact is represented as a triple (h, r, t), where $h \in \mathcal{E}$ is the head entity, $r \in \mathcal{R}$ is the relation, and $t \in \mathcal{E}$ is the tail entity. Let $\tau \in \mathbb{T}$ denote a timestamp or time interval, and $p \in [0, 1]$ denote a plausibility score indicating confidence or truth degree. Then we have the following definitions:

Definition 2 (Probabilistic Temporal Knowledge Graph)**.** *A probabilistic Temporal Knowledge Graph (PTKG) is a set of quintuples (h, r, t, τ, p), where each fact is annotated with both a temporal marker τ and a plausibility score p. PTKG enables joint modeling of temporal evolution and uncertainty.*

Knowledge Representation for Videos

Our framework converts multi-modal video content into structured textual representations by PTKG to enhance both indexing and retrieval efficiency. This transformation process encompasses two core modalities: for visual data, we utilize advanced VLMs to produce detailed textual descriptions that encapsulate scene interactions and contextual elements; for audio streams, we apply high-accuracy Automatic Speech Recognition (ASR) to extract spoken content while maintaining temporal alignment. This dual-processing strategy ensures that both visual and auditory semantics are preserved within our textual knowledge base.

Multimodal Content Extraction. To capture both spoken and visual information from arbitrarily long videos, we first divide each video $\mathcal{V}$ into segments $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ following (Ren et al. 2025; Yin Song and Chen Wu and Eden Duthie 2024; Lin et al. 2024). For each segment $\mathcal{S}_j$, we apply ASR to transcribe the audio, yielding $\mathcal{A}_j = \mathrm{ASR}(\mathcal{S}_j)$, which preserves dialogue and narration in a synchronized transcript. Concurrently, we uniformly sample up to $k \leq 10$ frames $\{\mathbf{F}_1, \ldots, \mathbf{F}_k\}$ in chronological order to capture key visual moments. We then feed both the transcript $\mathcal{A}_j$ and the sampled frames into a vision–language model, producing semantically rich captions $\mathcal{B}_j = \mathrm{VLM}\big(\mathcal{A}_j, \{\mathbf{F}_1, \ldots, \mathbf{F}_k\}\big)$, which integrate a series of entities and events.

PTKG Construction for Videos. Existing RAG methods (such as GraphRAG, VideoRAG) typically represent knowledge as static entity-relation pairs, overlooking the crucial dimensions of temporal sequencing and assertion confidence that underpin coherent cross-segment reasoning. Real-world video content is inherently dynamic, including entities, events, and their relationships often change or unfold over time, and the reliability of detected facts can vary due to modality noise or ambiguous signals.

To fill this gap, we introduce the PTKG, which jointly encodes entities, relations, timestamps, and confidence scores in a unified framework. We begin by segmenting each video transcript into coherent text chunks $\{\mathcal{V}_1^t, \ldots, \mathcal{V}_n^t\}$, then extract from each chunk $\mathcal{H}$ its set of entities $\mathcal{N}_\mathcal{H}$, relations $\mathcal{E}_\mathcal{H}$, associated timestamps $\mathcal{T}_\mathcal{H}$, and confidence values $\mathcal{P}_\mathcal{H} \in [0, 1]$ by LLM. Formally, the global PTKG is constructed as

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{T}, \mathcal{P}) = \bigcup_{\mathcal{H} \in \{\mathcal{V}_1^t, \ldots, \mathcal{V}_n^t\}} \big(\mathcal{N}_\mathcal{H},\, \mathcal{E}_\mathcal{H},\, \mathcal{T}_\mathcal{H},\, \mathcal{P}_\mathcal{H}\big). \tag{1}$$

By integrating temporal links and confidence alongside traditional triples, it overcomes the rigid querying limitations of prior static graphs and enables robust, context-aware knowledge retrieval across extensive video collections.

For each video clip $\mathcal{S}$, we generate a unified textual representation by integrating visual captions, ASR transcripts, and knowledge graph information $(\mathcal{B}, \mathcal{A}, \mathcal{G})$. Given a video $\mathcal{V}$ composed of n sequential clips, the complete knowledge extraction process is defined as:

$$\mathcal{V}_\mathcal{S}^t = \{(\mathcal{B}_l, \mathcal{A}_l, \mathcal{G}) \mid l \in [1, n]\}. \tag{2}$$

This representation captures both the visual and auditory semantics of each clip, along with the structured knowledge graph, enabling effective indexing and retrieval.

Temporal-Semantic Dual-Level Retrieval

Multi-Modal Video Knowledge Indexing. Our approach enhances video knowledge retrieval by integrating textual semantics into a unified and structured indexing system. Specif-

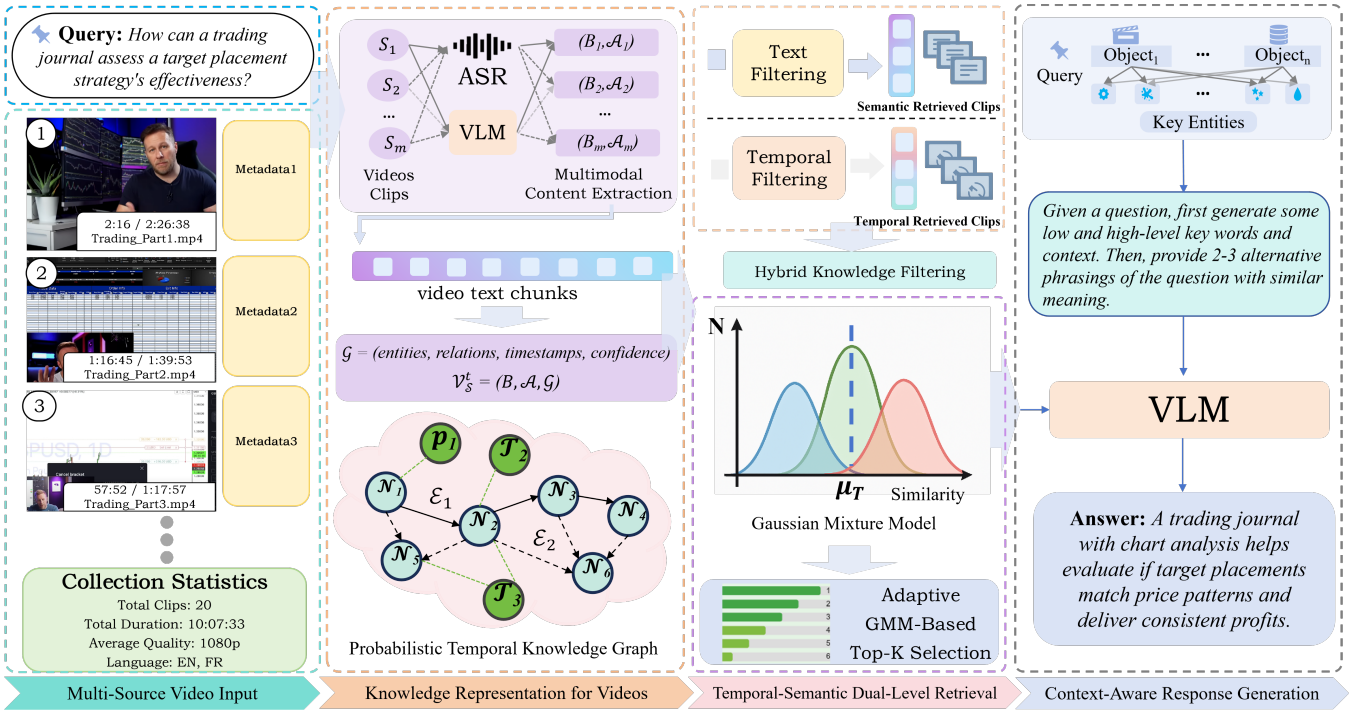

Figure 2: The overall framework of our ViG-RAG for videos. It details how input videos are converted into structured probabilistic temporal knowledge representations via knowledge-aware grounding. Subsequently, temporal-semantic dual-level retrieval and query-aware generation modules leverage multi-modal information to produce semantically precise and contextually enriched responses. The $\mathcal{N}$, $\mathcal{E}$, $\mathcal{T}$, $\mathcal{P}$ represents entity, relation, temporal and probabilistic information.

ically, the textual retrieval leverages a structured PTKG, enabling LLMs to refine queries by accurately identifying entities and their temporal relationships. Given a user query q, the textual retrieval initially generates a candidate set $\mathcal{S}_q^t$ by matching entities and semantic anchors in the TKG.

However, not all segments within $\mathcal{S}_q^t$ are equally relevant. To further improve precision, we introduce a structured filtering mechanism that jointly considers both semantic relevance and temporal coherence. Formally, the final filtered set of relevant video clips is defined as:

$$\{\hat{\mathcal{S}} \mid (\hat{\mathcal{S}} \in \mathcal{S}_q^t) \wedge topK(\alpha\text{Text-F}(\hat{\mathcal{S}}, q)) + (1-\alpha)(\text{Temp-F}(\hat{\mathcal{S}}, q))\}, \tag{3}$$

where the filtering functions are defined as binary classifiers using structured prompting strategies in LLMs. Text-F$(\hat{\mathcal{S}}, q)$: Evaluates the semantic alignment between the textual content of the clip ($\mathcal{V}_{\hat{\mathcal{S}}}^t$) and the refined user query. Temp-F$(\hat{\mathcal{S}}, q)$: Assesses whether the retrieved clip maintains temporal coherence and relevance, capturing meaningful long-range dependencies.

Adaptive GMM-Based Top-K Selection. Fixed thresholds or manually chosen K in Eq.(3) often fail to match the varied score distributions in video retrieval, leading to either excessive false positives or the loss of relevant clips. To adaptively select the top K candidates from similarity scores without manual tuning, we employ a lightweight, training-free strategy based on a Gaussian Mixture Model (GMM). Unlike learning-based ranking methods, our approach analyzes the empirical score distribution directly, requiring no optimization or loss function. Given similarity scores $\{x_1, x_2, \ldots, x_N\}$, we fit a K-component univariate GMM:

$$\log \mathcal{L}_K = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} w_k \cdot \mathcal{N}(x_i \mid \mu_k, \sigma_k^2) \right), \tag{4}$$

where w_k, μ_k, and σ_k^2 are the mixture weight, mean, and variance of each component. Parameters are estimated using the standard EM algorithm, with no gradient-based training.

To prevent overfitting, we determine the optimal number of components K^* using the bayesian information criterion (BIC) (Neath and Cavanaugh 2012). We then identify the component $k^\star$ with the highest mean μ_k, representing the high-confidence region. The posterior probability that a score x belongs to this component is:

$$p(x \in \mathcal{C}_{k^\star}) = \frac{w_{k^\star} \cdot \mathcal{N}(x \mid \mu_{k^\star}, \sigma_{k^\star}^2)}{\sum_{k=1}^{K^*} w_k \cdot \mathcal{N}(x \mid \mu_k, \sigma_k^2)}. \tag{5}$$

In this way, candidates are ranked by posterior confidence or a fixed cutoff, enabling adaptive filtering without handcrafted thresholds. The filtered set ensures semantic and temporal coherence, retaining clips that fully address user queries in both meaning and timing. This dual-filtering process reduces ambiguity, improves retrieval accuracy, and enhances the quality of subsequent response generation.

Context-Aware Response Generation

To generate responses closely aligned with user queries, we propose a hierarchical semantic coordination framework with two key modules: explicit semantic extraction and implicit context construction. Given a refined user query q, we first extract explicit semantic components K_q including low/high-level keywords using our knowledge-aware grounding mechanism. We then construct the implicit context field C_p through token-level local context refinement, capturing subtle semantics and preserving contextual coherence.

Using the semantic anchors K_q, the expanded context field C_p, and the filtered relevant clips $\hat{\mathcal{S}}$, our model synthesizes the response $\mathcal{R}$ as:

$$\mathcal{R} = \text{VLM}(K_q, C_p, \hat{\mathcal{S}}), \tag{6}$$

where K_q denotes the query-specific semantic anchors extracted from the question, guiding the response toward precise and relevant semantics. C_p enriches these anchors with additional implicit context, ensuring fluency and coherence in the output. $\hat{\mathcal{S}}$ is the refined set of video clips retrieved by our multi-modal context-aware retrieval mechanism, providing both visual and textual evidence to support the generated response.

Experiments

Experimental Settings

Evaluation Datasets. Our empirical study utilizes several large-scale video datasets, each designed to evaluate different facets of multimodal understanding and extended temporal reasoning. The LongerVideos benchmark (Ren et al. 2025) provides a diverse selection of more than twenty video sets, spanning educational, documentary, and entertainment genres. Because these videos are all sourced from publicly accessible YouTube content, our experiments are fully reproducible and transparent. In addition, we employ the Video-MME dataset (Fu et al. 2025), which spans real-world clips from 11 seconds to nearly an hour, to assess models ' ability to capture fine-grained daily-life activities. For a further challenge in long-range multimodal reasoning, LongVideoBench (Wu et al. 2024b) offers 6,678 carefully crafted multiple-choice questions across 17 topical areas, specifically targeting retrieval and inference over lengthy and information-rich video content.

Baseline. To benchmark the capabilities of our proposed system on challenging long-form and multi-video tasks, we conduct a comprehensive evaluation against a range of state-of-the-art retrieval-augmented generation frameworks. This comparison spans classical designs, such as NaiveRAG (Gao et al. 2023b), GraphRAG (Edge et al. 2024), and LightRAG (Guo et al. 2024), as well as advanced video-focused methods like VideoRAG (Ren et al. 2025). All baselines are re-implemented under a standardized experimental protocol to ensure the reliability and fairness of the comparison. We also take our model as a plug-in for other VLM, including four widely used 7B-parameter systems (Video-LLaVA (Lin et al. 2024), LLaVA-NeXT-Video (Zhang et al. 2024b), LongVA (Zhang et al. 2024a), and Long-LLaVA (Yin Song and Chen Wu and Eden Duthie 2024)), alongside two prominent 72B-parameter architectures (Qwen2-VL (Wang et al. 2024) and LLaVA-Video (Zhang et al. 2024c)), and other models such as Gemini (Reid et al. 2024).

Evaluation Protocols and Metrics. To ensure a multifaceted evaluation of our method, we adopt two complementary assessment protocols, building upon methodologies recently established in the literature (Ren et al. 2025). The first protocol, known as the Win-Rate metric, leverages GPT-4o-mini to perform head-to-head comparisons between candidate responses from different models, offering both rankings and explanatory rationale for its choices. In addition, we implement a Quantitative Scoring procedure, where responses are rated on a 5-point scale relative to a gold-standard reference, with scores ranging from 1 (markedly inferior) to 5 (significantly superior), thereby allowing for fine-grained, query-specific analysis. To minimize bias and increase reliability, we randomly permute answer positions within each prompt and conduct two separate rating rounds per query, adhering to established evaluation standards (Edge et al. 2024; Guo et al. 2024). Each protocol is repeated across five independent trials to further reduce variance, and final results are presented as aggregate statistics, either in terms of mean scores or win tallies. Full implementation details and the structure of our evaluation prompts are provided in Appendix.

Implementation Details. Following (Ren et al. 2025; Yin Song and Chen Wu and Eden Duthie 2024; Lin et al. 2024), videos are first partitioned into 30-second segments, with each interval yielding five representative frames selected for initial visual analysis. To capture both visual and linguistic cues, we utilize a quantized MiniCPM-V (Yao et al. 2024) as the vision-language interface, while spoken audio tracks are transcribed into text using Distil-Whisper (Radford et al. 2023; Gandhi, von Platen, and Rush 2023). Multi-modal information, spanning both images and transcripts, is jointly encoded via ImageBind (Girdhar et al. 2023), which embeds heterogeneous inputs into a unified latent space. For downstream retrieval, we generate text embeddings using OpenAI's text-embedding-3-small model, enabling semantic search over entity mentions and candidate passages. To strengthen section-level visual summarization, a denser frame sampling scheme ($\hat{k} = 15$ frames per segment) is incorporated during caption extraction. Orchestration of all core processes, including data indexing, evidence retrieval, and response construction, is handled by GPT-4o-mini, which serves as the central large language model driving end-to-end pipeline integration.

Overall Comparison

Comparision with Graph-based RAG. Table 1 reports the win rate evaluation outcomes, comparing ViG-RAG against the baseline approaches. For the GraphRAG series of benchmarks, we consider three variants: a local search baseline (GraphRAG-l), a global search configuration (GraphRAG-g), and a fully hybrid retrieval scheme as implemented by LightRAG. In Table 1, we present comparative win rate results, where our method is systematically benchmarked against all competing approaches. Across a range of evaluation criteria, our model demonstrates a clear and consistent performance

	GraphRAG-l	VideoRAG	**Ours**	LightRAG-h	VideoRAG	**Ours**	NaiveRAG	VideoRAG	**Ours**	GraphRAG-g	VideoRAG	**Ours**
Lecture												
Comprehensiveness	21.73%	32.15%	**46.12%**	22.05%	31.91%	**46.04%**	21.86%	32.95%	**45.19%**	21.91%	32.12%	**46.97%**
Clarity	22.44%	33.33%	**44.23%**	22.61%	33.33%	**44.06%**	22.44%	33.42%	**44.14%**	22.45%	33.58%	**43.97%**
Depth	20.99%	34.57%	**44.45%**	21.44%	34.31%	**44.25%**	21.41%	34.27%	**44.32%**	21.18%	34.51%	**44.31%**
Relevance	21.96%	34.91%	**43.13%**	21.99%	35.01%	**43.01%**	22.17%	34.84%	**43.99%**	21.88%	35.06%	**43.06%**
Practical Value	20.24%	35.05%	**44.71%**	20.47%	34.89%	**44.63%**	20.73%	34.75%	**44.52%**	20.27%	35.17%	**44.56%**
Overall	21.86%	35.05%	**43.10%**	22.09%	34.88%	**43.03%**	22.17%	34.76%	**43.07%**	22.10%	34.89%	**43.01%**
Documentary												
Comprehensiveness	22.82%	33.15%	**44.03%**	23.05%	32.91%	**44.04%**	22.86%	33.95%	**43.19%**	22.91%	33.12%	**43.97%**
Clarity	23.44%	34.33%	**42.23%**	23.61%	34.33%	**42.06%**	23.44%	34.42%	**42.14%**	23.45%	34.58%	**41.97%**
Depth	21.99%	35.57%	**42.45%**	22.44%	35.31%	**42.25%**	22.41%	35.27%	**42.32%**	22.18%	35.51%	**42.31%**
Relevance	22.96%	35.91%	**41.13%**	22.99%	36.01%	**41.01%**	23.17%	35.84%	**41.99%**	22.88%	36.06%	**41.06%**
Practical Value	21.24%	36.05%	**42.71%**	21.47%	35.89%	**42.63%**	21.73%	35.75%	**42.52%**	21.27%	36.17%	**42.56%**
Overall	22.86%	36.05%	**41.10%**	23.09%	35.88%	**41.03%**	23.17%	35.76%	**41.07%**	23.10%	35.89%	**41.01%**
Entertainment												
Comprehensiveness	23.82%	34.15%	**42.03%**	24.05%	33.91%	**42.04%**	23.86%	34.95%	**41.19%**	23.91%	34.12%	**42.97%**
Clarity	24.44%	35.33%	**40.23%**	24.61%	35.33%	**40.06%**	24.44%	35.42%	**40.14%**	24.45%	35.58%	**39.97%**
Depth	22.99%	36.57%	**40.45%**	23.44%	36.31%	**40.25%**	23.41%	36.27%	**40.32%**	23.18%	36.51%	**40.31%**
Relevance	23.96%	36.91%	**39.13%**	23.99%	37.01%	**39.01%**	24.17%	36.84%	**39.99%**	23.88%	37.06%	**39.06%**
Practical Value	22.24%	37.05%	**40.71%**	22.47%	36.89%	**40.63%**	22.73%	36.75%	**40.52%**	22.27%	37.17%	**40.56%**
Overall	22.86%	37.05%	**40.10%**	23.09%	36.88%	**40.03%**	23.17%	36.76%	**40.07%**	23.10%	36.89%	**40.01%**
Overall												
Comprehensiveness	22.59%	33.41%	**44.00%**	22.81%	33.23%	**43.96%**	22.68%	33.62%	**43.70%**	22.74%	33.38%	**43.88%**
Clarity	23.33%	34.43%	**42.24%**	23.54%	34.39%	**42.07%**	23.41%	34.51%	**42.08%**	23.38%	34.58%	**42.04%**
Depth	21.99%	35.31%	**42.70%**	22.43%	35.10%	**42.47%**	22.38%	35.14%	**42.48%**	22.18%	35.32%	**42.50%**
Relevance	22.96%	35.62%	**41.42%**	22.99%	35.71%	**41.30%**	23.16%	35.58%	**41.26%**	22.88%	35.76%	**41.36%**
Practical Value	21.24%	36.26%	**42.50%**	21.47%	36.12%	**42.41%**	21.73%	36.00%	**42.27%**	21.27%	36.36%	**42.37%**
Overall	22.42%	35.81%	**41.77%**	22.64%	35.67%	**41.69%**	22.70%	35.57%	**41.73%**	22.49%	35.82%	**41.69%**

Table 1: Performance Comparison of ViG-RAG against Different RAG Baselines. The best results are in **bolded**.

Model	**Text**	**LLM Params**	**Frames**	**Short**	**Medium**	**Long**	**Overall**	**Gain**
Proprietary LVLMs								
GPT-4o	-	-	384	80.0	70.3	65.3	71.9	-
Gemini-1.5-Pro	-	-	0.5 fps	81.7	74.3	67.4	75.0	-
Open-Source LVLMs								
Video-LLaVA	-	7B	8	44.6	38.3	35.8	39.6	-
Video-LLaVA + ViG-RAG	2.0K	7B	8	48.6	42.3	40.1	43.5	+3.9
LLaVA-NeXT-Video	-	7B	16	49.4	43.0	36.7	43.0	-
LLaVA-NeXT-Video + ViG-RAG	2.0K	7B	16	59.8	49.2	53.7	54.2	+11.2
LongVA	-	7B	32	60.9	49.3	44.0	51.4	-
LongVA + ViG-RAG	1.8K	7B	32	64.5	56.7	55.2	58.8	+7.4
Long-LLaVA	-	7B	32	60.3	51.4	44.1	52.0	-
Long-LLaVA + ViG-RAG	1.9K	7B	32	63.2	53.1	56.7	57.7	+5.7
Qwen2-VL	-	72B	32	75.0	63.3	56.3	64.9	-
Qwen2-VL + ViG-RAG	2.1K	72B	32	76.3	68.6	72.2	72.4	+7.5
LLaVA-Video	-	72B	32	78.0	63.7	59.6	67.1	-
LLaVA-Video + ViG-RAG	2.1K	72B	32	**79.3**	**71.1**	**72.8**	**74.4**	**+7.3**

Table 2: Performance evaluation on the Video-MME (Fu et al. 2025) benchmark. The Text field reflects the mean number of additional tokens introduced per example by ViG-RAG. All open-source baselines and their ViG-RAG variants were re-benchmarked.

lead. This improvement is attributable to our comprehensive multi-modal indexing strategy, which seamlessly integrates probabilistic temporal knowledge graph construction with deep cross-modal representation learning. By encoding not only entity relationships and temporal structure but also associating each knowledge fact with a calibrated confidence measure, our framework enables robust semantic organization and reliable evidence synthesis over long video sequences. Additionally, our adaptive retrieval mechanism, powered by distribution-based candidate selection, balances both text-based semantic matching and visual content relevance, thus dynamically identifying the most trustworthy video segments for each query without reliance on fixed heuristics. This precise alignment between linguistic and visual signals results in

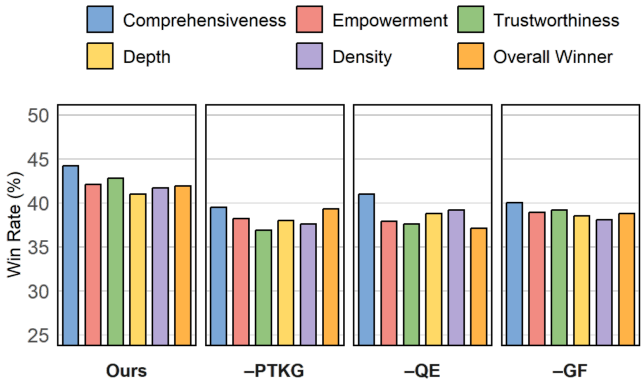

Figure 3: Ablation on graph-based knowledge grounding and cross-modal retrieval components.

more contextually relevant and semantically precise retrieval, enabling our system to generate outputs with enhanced contextual depth and factual accuracy compared to traditional retrieval-augmented pipelines.

The Performance on Other Datasets. In our evaluation of video-language models, as shown in Table 2, we integrate ViG-RAG as an auxiliary retrieval-augmented module, standardizing the input to a fixed 32-frame window across all LVLM baselines. This constraint ensures fair comparisons, particularly for resource-intensive 72B-parameter systems, while also providing a unified benchmark for smaller 7B models. Distinct from prior methods, our framework leverages a probabilistic temporal knowledge graph to construct multi-modal indices and facilitate advanced retrieval. Integrating ViG-RAG as a retrieval-augmented module consistently enhances the performance of a wide range of open-source LVLM backbones. Regardless of model scale or architecture, the addition of ViG-RAG leads to marked improvements across short, medium, and especially long-duration video scenarios. This demonstrates that ViG-RAG functions as a robust, plug-and-play component, systematically strengthening multi-modal retrieval and temporal reasoning capabilities for diverse base models. More results can refer to Appendix.

Ablation Study

To systematically evaluate the contribution of each core module within our multi-modal retrieval framework, we conduct an ablation analysis comprising three distinct model variants: (1) w/o PTKG: Disables indexing and retrieval based on the probabilistic temporal knowledge graph, thereby removing the system's capacity for building meaningful temporal and cross-video relationships; (2) w/o Query enhancement (QE): Omits explicit semantic augmentation and implicit contextual enrichment, reducing the model's effectiveness in leveraging nuanced query information; (3) w/o GMM filtering (GF): Eliminates the GMM-based candidate filtering, so segment selection relies solely on raw similarity scores without adaptive thresholding.

Results, visualized in Figure 3, demonstrate that the complete ViG-RAG pipeline depends critically on the synergy of all three modules. Disabling any single component results in marked decreases in retrieval accuracy and overall performance on all benchmarks, revealing the necessity of each design element for coherent and context-aware evidence aggregation. Models lacking these mechanisms exhibit notable deficiencies in integrating information across disparate video segments, often generating outputs that are disjointed and lacking depth. These findings reinforce the value of our graph-augmented, probabilistic, and adaptive retrieval architecture for robust multi-modal video understanding.

Case Study Analysis

To further assess the effectiveness of ViG-RAG, we examine its performance on a complex query requiring long-context video reasoning as shown in Appendix. The query *"Do chimpanzees have a designated leader who dictates the strategy, or is it a more fluid process based on individual initiative and cues from the environment?"* demands a nuanced understanding of chimpanzee social structures, requiring ViG-RAG to retrieve, integrate, and reason over multimodal evidence that captures hierarchical leadership, environmental adaptability, and individual behavioral agency. The input video, *fights-in-animal-kingdom*, contains rich social interactions illustrating these dynamics. It shows ViG-RAG's response alongside the retrieved video segments. ViG-RAG successfully identifies key scenes illustrating both structured leadership and fluid social hierarchies. The retrieved clips comprehensively capture three core aspects: (1) Leadership roles, dominant males often guide foraging and territorial navigation. (2) Environmental adaptability, leadership shifts dynamically with social context and environmental pressures. (3) Individual agency and social learning, younger members align with dominant figures to strengthen their social standing.

By synthesizing these elements, ViG-RAG demonstrates its ability to integrate explicit and implicit video knowledge. Its responses are contextually precise and temporally coherent, accurately linking behavioral patterns with ecological constraints, resource distribution, and intra-group competition. This highlights ViG-RAG's strength in multi-modal reasoning and its capacity to provide evidence-grounded insights into complex behavioral phenomena.

Conclusion

In this work, we propose ViG-RAG, a novel framework designed to advance long-context video understanding by addressing the limitations of existing RAG-based methods. By introducing a probabilistic temporal knowledge graph and leveraging a hybrid retrieval strategy that dynamically models semantic and temporal relevance, ViG-RAG effectively bridges multimodal content and complex temporal dependencies across videos. Our approach enables coherent long-range reasoning, context-aware retrieval, and accurate evidence aggregation, all while improving computational efficiency. Extensive experiments across multiple benchmarks confirm that ViG-RAG delivers significant performance gains over prior methods, highlighting its promise as a robust and generalizable solution for multi-modal, long-context video understanding.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory. This work is also supported by Intern Discovery. This work is also supported by the Hong Kong General Research Fund (Grant No. 17503722), NSFC HY Working Fund (Grant No. 03070100001), Tsinghua SIGS Basic Support Fund (Grant No. 07010100003), and Tsinghua SIGS Research Support Fund (Grant No. 01030100049). This work is also supported in part by the Institute for Industrial Innovation and Finance (IIIF), Tsinghua University.

References

Allahverdiyev, R. A. I. S. S. I.; Taha, M.; Akalin, A.; and Zhu, K. 2024. ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems. *arXiv preprint arXiv:2410.19572*.

Arefeen, M. A.; Debnath, B.; Uddin, M. Y. S.; and Chakradhar, S. 2024. irag: Advancing rag for videos with an incremental approach. In *CIKM*, 4341–4348.

Cao, Z.; He, Y.; Liu, A.; Xie, J.; Chen, F.; and Wang, Z. 2025a. TV-RAG: A Temporal-aware and Semantic Entropy-Weighted Framework for Long Video Retrieval and Understanding. In *ACM MM*, 9071–9079.

Cao, Z.; He, Y.; Liu, A.; Xie, J.; Wang, Z.; and Chen, F. 2025b. CoFi-Dec: Hallucination-Resistant Decoding via Coarse-to-Fine Generative Feedback in Large Vision-Language Models. In *ACM MM*, 10709–10718.

Cao, Z.; He, Y.; Liu, A.; Xie, J.; Wang, Z.; and Chen, F. 2025c. PurifyGen: A Risk-Discrimination and Semantic-Purification Model for Safe Text-to-Image Generation. In *ACM MM*, 816–825.

Cao, Z.; Li, J.; Wang, Z.; and Li, J. 2024. Diffusione: Reasoning on knowledge graphs via diffusion-based graph neural networks. In *ACM SIGKDD*, 222–230.

Cao, Z.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2021. Dual Quaternion Knowledge Graph Embeddings. In *AAAI conference on artificial intelligence*, 6894–6902.

Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Chandrasegaran, K.; Gupta, A.; Hadzic, L. M.; Kota, T.; He, J.; Eyzaguirre, C.; Durante, Z.; Li, M.; Wu, J.; and Fei-Fei, L. 2024. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. In *CVPR*, 24108–24118.

Gandhi, S.; von Platen, P.; and Rush, A. M. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023a. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, 15180–15190.

Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779*.

Gutiérrez, B. J.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. *NeurIPS*.

Li, M.; Miao, S.; and Li, P. 2024. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. *arXiv preprint arXiv:2410.20724*.

Li, Y.; Wang, C.; and Jia, J. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 323–340.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 5971–5984.

Lin, K.; Ahmed, F.; Li, L.; Lin, C.-C.; Azarnasab, E.; Yang, Z.; Wang, J.; Liang, L.; Liu, Z.; Lu, Y.; et al. 2023. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.

Neath, A. A.; and Cavanaugh, J. E. 2012. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2): 199–203.

Qian, H.; Zhang, P.; Liu, Z.; Mao, K.; and Dou, Z. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, 28492–28518. PMLR.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ren, X.; Xu, L.; Xia, L.; Wang, S.; Yin, D.; and Huang, C. 2025. VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos. *arXiv preprint arXiv:2502.01549*.

Shang, Y.; Xu, B.; Kang, W.; Cai, M.; Li, Y.; Wen, Z.; Dong, Z.; Keutzer, K.; Lee, Y. J.; and Yan, Y. 2024. Interpolating

Video-LLMs: Toward Longer-sequence LMMs in a Training-free Manner. *arXiv preprint arXiv:2409.12963*.

Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *CVPR*, 26160–26169.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Ding, M.; Gu, X.; Huang, S.; Xu, B.; et al. 2025. Lvbench: An extreme long video understanding benchmark. In *ICCV*, 22958–22967.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024b. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37: 28828–28857.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yin Song and Chen Wu and Eden Duthie. 2024. aws-prototyping/long-llava-qwen2-7b.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; et al. 2024a. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; et al. 2024b. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; Jin, P.; Zhang, W.; Wang, F.; Bing, L.; and Zhao, D. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhang, S.; Tay, Y.; Yao, L.; and Liu, Q. 2019. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024b. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024c. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.