# VideoRAG: Retrieval-Augmented Generation over Video Corpus

Anonymous ACL submission


## Abstract

Retrieval-Augmented Generation (RAG) is a powerful strategy for improving the factual accuracy of models by retrieving external knowledge relevant to queries and incorporating it into the generation process. However, existing approaches primarily focus on text, with some recent advancements considering images, and they largely overlook videos, a rich source of multimodal knowledge capable of representing contextual details more effectively than any other modality. While very recent studies explore the use of videos in response generation, they either predefine query-associated videos without retrieval or convert videos into textual descriptions losing multimodal richness. To tackle these, we introduce VideoRAG, a framework that not only dynamically retrieves videos based on their relevance with queries but also utilizes both visual and textual information. The operation of VideoRAG is powered by recent Large Video Language Models (LVLMs), which enable the direct processing of video content to represent it for retrieval and the seamless integration of retrieved videos jointly with queries for response generation. Also, inspired by that the context size of LVLMs may not be sufficient to process all frames in extremely long videos and not all frames are equally important, we introduce a video frame selection mechanism to extract the most informative subset of frames, along with a strategy to extract textual information from videos (as it can aid the understanding of video content) when their subtitles are not available. We experimentally validate the effectiveness of VideoRAG, showcasing that it is superior to relevant baselines.


## 1 Introduction

Recently, large foundation models, such as large language models and their extension to the vision modality called large vision-language models, have become the standard for addressing diverse tasks due to their remarkable capabilities (OpenAI, 2023; Li et al., 2024; Yang et al., 2024; Dai et al., 2024).

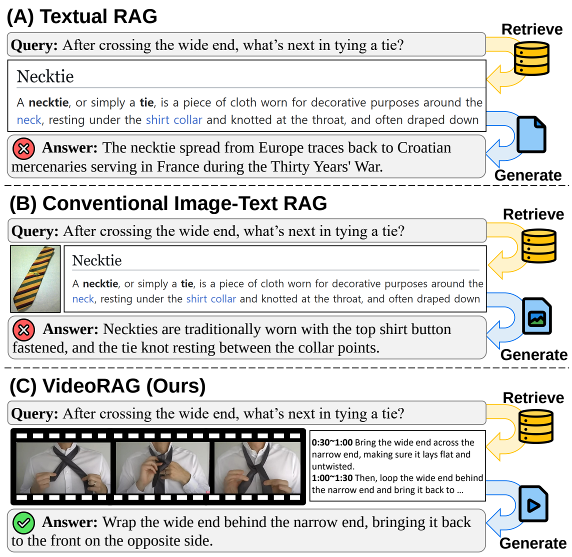



Figure 1: Illustration of existing and the proposed RAG scenarios. (A) Textual RAG retrieves documents (relevant to queries) from a text corpus and incorporates them when generating answers. (B) Conventional image-text multimodal RAG extends retrieval to include static images. (C) VideoRAG (ours) further extends the external knowledge source to videos.

In particular, these models, trained on extensive textual and multimodal corpora, encode vast amounts of knowledge within their large-scale parameters. However, they are still prone to generating factually incorrect outputs, as their parametric knowledge can be inaccurate or outdated (Lewis et al., 2020; Ram et al., 2023). This limitation highlights the need for incorporating knowledge from external knowledge sources, with Retrieval-Augmented Generation (RAG) emerging as an essential mitigator for it. Specifically, RAG typically operates by retrieving query-relevant information and then generating answers grounded in the retrieved content (Niu et al., 2024; Ayala and Béchard, 2024).

However, while existing RAG approaches have been widely adopted for various real-world applications, they have primarily focused on retrieving and incorporating textual content (Ram et al., 2023; Jeong et al., 2024a), with only recent attempts beginning to explore images (or text-image pairs) as the additional source of external knowledge (Yu et al., 2024; Riedler and Langer, 2024). On the other hand, we argue that there remains a rapidly expanding yet underutilized medium, called videos, which provides unparalleled multimodal richness and might be a compelling resource for augmenting the knowledge landscape of current RAG systems. Specifically, videos combine temporal dynamics, spatial details, and multimodal cues, which collectively enable them to capture complex processes, context-dependent interactions, and non-verbal signals that static modalities (e.g., text and images) often fail to convey. Moreover, given the increasing popularity of video-sharing platforms (such as YouTube), the availability of diverse, high-quality video data has grown, ranging from educational tutorials and scientific demonstrations to personal experiences and real-time events, all of which may be useful when formulating responses to user queries.

A few recent studies have started considering video content to handle user queries; however, they have limitations. For instance, some assume that videos relevant to queries are already known and instead focus on identifying query-relevant frames within that specified video (Luo et al., 2024; Ma et al., 2024). While effective in scenarios where the relevant video is explicitly provided, it is suboptimal for more general-use cases, where users expect systems to dynamically identify and retrieve videos to provide answers. On the other hand, other studies handle videos by converting them into textual formats, such as subtitles, and utilizing these textual representations under off-the-shelf text-based RAG pipelines (Arefeen et al., 2024; Zhang et al., 2024b). However, while this text-only strategy may offer a convenient workaround, it inherently sacrifices the multimodal richness of video data by discarding critical information, such as temporal dynamics captured in the visual context, during the conversion process. For example, consider a query: "How does the expression of the dog change when it is angry?". While textual transcriptions might describe the dog's barking or growling, they fail to capture visual cues (baring teeth, raised hackles, or narrowed eyes), which are needed for accurately interpreting the emotional state of the dog and subsequently formulating the answer to the query.

To address the aforementioned limitations, we introduce a novel framework, called VideoRAG, which aims to offer another fruitful angle to existing RAG frameworks by enabling a more comprehensive utilization of video content for its holistic retrieval and incorporation (See Figure 1). Specifically, in response to queries, the proposed VideoRAG retrieves relevant videos from a large video corpus but also integrates both visual and textual elements into the answer-generation process. Also, we operationalize this by harnessing the advanced capabilities of recent Large Video Language Models (LVLMs), which are capable of directly processing video content, consisting of visual and textual information, within the unified framework, thereby more effectively capturing its multimodal richness.

However, there exist a couple of remaining challenges in integrating videos into RAG frameworks. First, videos are inherently long and redundant, oftentimes making it infeasible for LVLMs to process all frames due to their limited context capacity as well as unnecessary since not all frames contribute meaningfully for retrieval and generation. To address this, we introduce a frame selection model that is trained to extract the most informative subset of frames to maximize retrieval and generation performance. Also, we observe that, while the joint utilization of visual and textual features is needed for the effective representation of videos and subsequently their retrieval, the textual descriptions of videos (e.g., subtitles) are oftentimes not available. To tackle this, we further present a simple yet effective mitigation strategy that utilizes automatic speech recognition techniques to generate textual transcripts from videos, allowing us to leverage both visual and textual modalities for every video.

To validate the effectiveness of VideoRAG, we conduct experiments by using overlapping queries from the WikiHowQA dataset (Bolotova-Baranova et al., 2023) (consisting of query-answer pairs) and the HowTo100M dataset (Miech et al., 2019) (including query-video pairs without answers). Also, based on this, we automatically collect the dataset for RAG over videos and then evaluate models on it. Then, the experimental results show the significant performance improvement of the proposed VideoRAG framework over relevant baselines, demonstrating the efficacy of leveraging videos for RAG.

## 2 Method

We present VideoRAG that retrieves query-relevant videos and generates responses grounded in them.

### 2.1 Preliminaries

We begin with describing RAG and LVLMs.

**Retrieval-Augmented Generation** RAG aims to enhance the capabilities of foundation models by

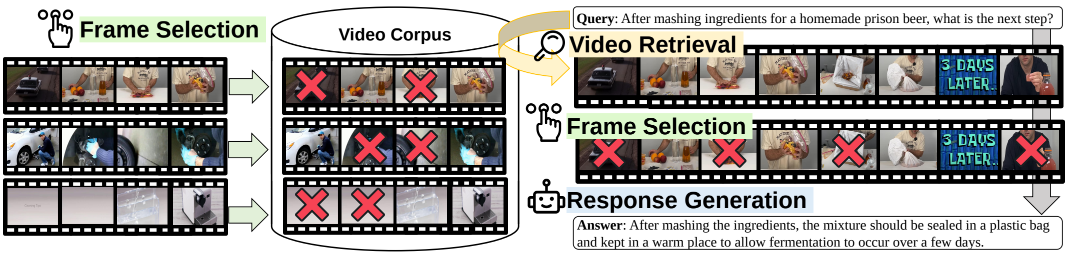



Figure 2: Illustration of the overall pipeline of our VideoRAG, which selects informative frames for retrieval and generation.

grounding their outputs in external knowledge retrieved from the external knowledge source, such as Wikipedia, which consists of two main components: retrieval and generation modules. Formally, given a query $\boldsymbol{q}$, RAG retrieves a set of documents (or knowledge elements) $\mathcal{K} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \dots, \boldsymbol{k}_k\}$ from an external corpus $\mathcal{C}$ ($\mathcal{K} \subseteq \mathcal{C}$) based on their relevance with $\boldsymbol{q}$ using a retrieval module, which can be formalized as follows: $\mathcal{K} = \texttt{Retriever}(\boldsymbol{q}, \mathcal{C})$. Here, the query $\boldsymbol{q}$ and knowledge $\boldsymbol{k}$ are represented as a sequence of tokens $\boldsymbol{q} = [q_1, q_2, \dots, q_i]$ and $\boldsymbol{k} = [k_1, k_2, \dots, k_j]$. Also, during retrieval, the relevance between the query and each knowledge element within the corpus is determined by the scoring function, defined as follows: $\texttt{Sim}(\boldsymbol{q}, \boldsymbol{k})$, which typically measures their representational similarity over the embedding space. In the subsequent generation step, the retrieved knowledge elements are then used as additional input to the generation module, to augment the query to produce an answer $\boldsymbol{y}$, as follows: $\boldsymbol{y} = \texttt{Model}(\boldsymbol{q}, \mathcal{K})$, where $\texttt{Model}$ is typically implemented as the foundation model, such as LLMs. We note that, unlike existing RAG that focuses mainly on retrieving and incorporating textual content (or, in some recent cases, extra static images), we explore the extension toward videos.

**Large Video Language Models** On top of the extensive language understanding capabilities of LLMs, LVLMs are designed to handle and incorporate the features from video content, including temporal, spatial, and multimodal information, within the unified token processing framework. Formally, let us denote a video $\boldsymbol{V}$ as a sequence of visual frames: $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n]$ and its associated textual data (such as subtitles, or any other textual information such as the video-specific query) $\boldsymbol{t}$ as a sequence of tokens: $\boldsymbol{t} = [t_1, t_2, \dots, t_m]$. Then, the typical LVLM, denoted as $\texttt{LVLM}$, enables the joint processing of these multimodal inputs by employing two specialized components: a vision encoder and a text encoder. Specifically, the vision encoder processes the sequence of video frames $\boldsymbol{V}$ (which can span multiple videos), resulting in a sequence of visual feature embeddings (or visual tokens): $\boldsymbol{F}_{\texttt{visual}} = \texttt{VisionEncoder}(\boldsymbol{V})$. Concurrently, the text encoder processes the given textual information $\boldsymbol{t}$ to generate corresponding feature embeddings: $\boldsymbol{F}_{\texttt{text}} = \texttt{TextEncoder}(\boldsymbol{t})$. Then, the overall process to obtain the video representation (with the goal of capturing both visual and textual features) can be denoted as follows: $\boldsymbol{f}_{\texttt{video}} = \texttt{LVLM}(\boldsymbol{V}, \boldsymbol{t})$. Traditionally, $\boldsymbol{f}_{\texttt{video}}$ is obtained by the simple interpolation of the visual and textual representations: $\boldsymbol{f}_{\texttt{video}} = \alpha \cdot \boldsymbol{F}_{\texttt{text}} + (1 - \alpha) \cdot \boldsymbol{F}_{\texttt{visual}}$ (Xu et al., 2021), and, more recently, it can be done by further jointly processing the visual and textual embeddings through several LVLM layers (that sit on top of existing LLMs) (Zhang et al., 2024c), which allows the model to learn a more effective representation and continue generating the next sequence of tokens (for example, an answer to a query).

### 2.2 VideoRAG

We now turn to introduce our VideoRAG, which extends the existing RAG paradigm by leveraging the video corpus as the external knowledge source.

**Video Retrieval** The initial step to operationalize RAG over the video corpus is to implement video retrieval, whose goal is to identify query-relevant videos $\mathcal{V} = \{\boldsymbol{V}_1, \boldsymbol{V}_2, \dots, \boldsymbol{V}_k\}$ from the corpus $\mathcal{C}$, consisting of a large number of videos, as follows: $\mathcal{V} = \texttt{Retriever}(\boldsymbol{q}, \mathcal{C})$. Recall that this retrieval process involves calculating the similarity between the query $\boldsymbol{q}$ and each knowledge element (which is video $\boldsymbol{V}$ in our case) to determine their relevance. To achieve this, we first forward the video $\boldsymbol{V}$ (composed of image frames and, if available, subtitles) as well as the query $\boldsymbol{q}$ (without visual information) into $\texttt{LVLM}$, to obtain their representations $\boldsymbol{f}_{\texttt{query}}$ and $\boldsymbol{f}_{\texttt{video}}$. After that, the relevance is computed based on their representation-level similarity, for example, using a cosine similarity, and the top-$k$ videos with the highest similarity scores are retrieved.

**Video-Augmented Response Generation** After the retrieval of query-relevant videos is done, the next step is to incorporate the retrieved videos into the answer generation process, to formulate the answer grounded in them. To operationalize this, we first concatenate frames of each retrieved video with its associated textual data (e.g., subtitles), then concatenate these multimodal pairs across all videos retrieved, and lastly append the user query, as follows: $[\boldsymbol{V}_1, \boldsymbol{t}_1, \ldots, \boldsymbol{V}_k, \boldsymbol{t}_k, \boldsymbol{q}]$. Then, this input is forwarded into LVLM, which enables the joint processing of the combined visual, textual, and query-specific information, to generate the response while capturing their multimodal richness and dynamics.

### 2.3 Frame Selection for VideoRAG

Unlike conventional RAG with text or images, incorporating videos into RAG presents an additional challenge: some videos contain a large number of visual frames, making it inefficient to process them all (and sometimes impractical due to the limited context size of LVLMs). As a simple workaround, a common approach is to uniformly sample frames; however, this method risks discarding key information while retaining redundant or irrelevant frames, leading to suboptimal retrieval and response generation when augmented with suboptimal frames.

**Adaptive Frame Selection** To overcome these limitations, we introduce an adaptive frame selection strategy, whose objective is to extract the most informative and computationally feasible subset of frames. Let $\texttt{Comb}(\cdot)$ represent a selection function that randomly samples a subset of $m$ frames from total $n$ frames within the video based on the combination, and let $f(\cdot)$ be a scoring function that evaluates and assigns a relevance score to these selected frames. Then, during retrieval, the frame selection operation for the given video $\boldsymbol{V}$ is denoted as follows: $\tilde{\boldsymbol{V}} = \arg\max_{\boldsymbol{V}' \in \texttt{Comb}(\boldsymbol{V}, m)} f(\boldsymbol{V}')$, which is extended to $\tilde{\boldsymbol{V}} = \arg\max_{\boldsymbol{V}' \in \texttt{Comb}(\boldsymbol{V}, m)} f(\boldsymbol{V}', \boldsymbol{q})$ for generation, where $\tilde{\boldsymbol{V}}$ is the optimal subset. The distinction between retrieval and generation arises because retrieval operates over a large video corpus $\mathcal{C}$, making exhaustive query-based processing infeasible, whereas in generation, the top-$k$ retrieved videos allow for query-guided frame selection (i.e., enabling the use of different frames for different queries even if the retrieved video is the same).

**Frame Space Reduction with Clustering** While the adaptive frame selection strategy enables the use of the most effective subset of frames for RAG, the combinatorial space of possible frame subsets (obtained from Comb) remains prohibitively large. For instance, selecting 32 frames from a video of 1000 frames results in more than $10^{60}$ possible combinations, making exhaustive search impossible. To address this, we reduce the frame selection space by extracting representative samples via $k$-means++ clustering. Specifically, we cluster all frames into $k$ groups and, from each of the $k$ clusters, we select the frame closest to its centroid. After that, we constrain the frame selection process to operate within this reduced set; for example, with $k = 64$, the search space is drastically reduced to ${}_{64}\mathrm{C}_{32}$ from ${}_{1000}\mathrm{C}_{32}$, making it computationally feasible while preserving the diversity of selected frames[1].

**Operationalizing Frame Selection** Notably, the design of $f$ to score the selected frame is flexible, allowing us to use any models capable of processing visual features (and textual features particularly for generation), such as CLIP (Radford et al., 2021). Also, we collect examples for training $f$, by performing retrieval and generation with randomly selected frames (from possible combinations), and then labeling them as true or false based on their success, from which we use the conventional loss functions (such as cross-entropy) for optimization. We provide more details on it in Appendix A.3.

### 2.4 Auxiliary Text Generation

In both the retrieval and generation steps, the inclusion of video-associated textual data, such as subtitles, can play a crucial role in enhancing video representation since it provides additional context and semantic cues that complement the visual content. However, not every video in the corpus comes with subtitles since they require additional annotations. Therefore, for such videos, we propose generating auxiliary textual data by extracting audio from the video and converting it into text using off-the-shelf automatic speech recognition techniques. Formally, given a video $\boldsymbol{V}$, this process can be formalized as follows: $\boldsymbol{t}_{\mathsf{aux}} = \texttt{AudioToText}(\texttt{Audio}(\boldsymbol{v}))$, where $\texttt{Audio}(\boldsymbol{V})$ extracts the audio track from the video, and AudioToText converts the extracted audio signal into textual content. Therefore, for those videos without subtitles, auxiliary text $\boldsymbol{t}_{\mathsf{aux}}$ can be used in place of $\boldsymbol{t}$ in both the retrieval and generation steps.

## 3 Experiment

We now describe experimental setup and results.

### 3.1 Experimental Setup

**Datasets** We evaluate VideoRAG in question answering tasks, following the convention for validating RAG approaches (Asai et al., 2024; Jeong et al.,

[1]In inference, evaluating all possible combinations from this reduced set might still be computationally expensive; thus, we further perform random sampling over them.

Table 1: Overall RAG results across four metrics. The best results are highlighted in **bold**, and the second-best results are highlighted with underline. Note that the ORACLE setting (that uses ideal retrieval results) is not comparable to others.

| | | WikiHowQA with HowTo100M | | | | Synthetic QA with HowTo100M | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Methods | ROUGE-L | BLEU-4 | BERTScore | G-Eval | ROUGE-L | BLEU-4 | BERTScore | G-Eval |
| LLaVA-Video (7B) | NAÏVE | 14.08 | 1.352 | 83.43 | 1.579 | 10.68 | 1.574 | 84.51 | 1.634 |
| | TEXTRAG (BM25) | 17.22 | 2.327 | 84.66 | 1.633 | 14.70 | 2.382 | 86.03 | 1.681 |
| | TEXTRAG (DPR) | 16.65 | 2.173 | 84.61 | 1.591 | 14.58 | 2.397 | 85.85 | 1.686 |
| | TEXTIMAGERAG | 22.43 | 4.222 | 86.88 | 2.022 | 25.19 | 6.149 | 88.56 | 2.175 |
| | TEXTVIDEORAG | 22.81 | 4.388 | 86.97 | 1.979 | 23.41 | 5.435 | 88.40 | 2.278 |
| | VIDEORAG-V | **24.95** | 5.080 | 87.85 | 2.140 | 29.38 | 7.530 | **89.77** | **2.479** |
| | VIDEORAG-VT | 24.93 | **5.276** | **87.92** | **2.142** | **29.74** | **8.043** | 89.72 | 2.476 |
| | ORACLE-V | 26.19 | 5.480 | 88.41 | 2.225 | 32.16 | 8.769 | 90.34 | 2.884 |
| | ORACLE-VT | 25.37 | 5.237 | 87.95 | 2.166 | 32.31 | 8.885 | 90.46 | 2.938 |
| InternVL2.5 (8B) | NAÏVE | 16.54 | 1.859 | 84.30 | 1.720 | 12.60 | 2.381 | 85.12 | 1.725 |
| | TEXTRAG (BM25) | 17.41 | 2.275 | 84.89 | 1.552 | 26.66 | 6.760 | 88.48 | 1.938 |
| | TEXTRAG (DPR) | 17.21 | 2.077 | 84.84 | 1.563 | 26.72 | 6.579 | 88.56 | 1.917 |
| | TEXTIMAGERAG | 22.39 | 3.917 | 86.91 | 1.904 | 27.65 | 7.187 | 88.99 | 2.176 |
| | TEXTVIDEORAG | 19.88 | 3.199 | 85.81 | 1.686 | 26.36 | 6.542 | 88.68 | 1.983 |
| | VIDEORAG-V | **25.11** | 4.243 | **88.15** | 1.863 | **33.68** | 9.454 | **90.29** | **2.452** |
| | VIDEORAG-VT | 23.75 | **4.271** | 87.42 | **1.906** | 32.90 | **9.572** | 90.14 | 2.427 |
| | ORACLE-V | 25.59 | 4.318 | 88.29 | 1.958 | 35.21 | 10.57 | 90.70 | 2.813 |
| | ORACLE-VT | 24.60 | 4.421 | 8.770 | 2.002 | 34.99 | 10.69 | 90.68 | 2.820 |
| Qwen2.5-VL (3B) | NAÏVE | 17.96 | 2.077 | 84.97 | 1.765 | 15.05 | 2.729 | 86.13 | 1.843 |
| | TEXTRAG (BM25) | 19.65 | 2.989 | 85.41 | 1.721 | 19.70 | 3.911 | 86.88 | 1.877 |
| | TEXTRAG (DPR) | 19.45 | 2.863 | 85.38 | 1.708 | 19.04 | 3.903 | 86.77 | 1.831 |
| | TEXTIMAGERAG | 20.66 | 3.327 | 85.80 | 1.838 | 20.36 | 4.298 | 87.11 | 1.931 |
| | TEXTVIDEORAG | 22.18 | 4.180 | 86.56 | 1.821 | 24.29 | 5.722 | 88.37 | 2.156 |
| | VIDEORAG-V | **23.24** | 3.963 | **87.13** | **1.899** | 26.28 | 5.998 | 88.97 | 2.258 |
| | VIDEORAG-VT | 23.22 | **4.531** | 87.00 | 1.876 | **27.54** | **7.279** | **89.11** | **2.274** |
| | ORACLE-V | 21.53 | 3.156 | 86.05 | 1.912 | 26.82 | 6.683 | 88.96 | 2.515 |
| | ORACLE-VT | 24.37 | 4.811 | 87.43 | 1.994 | 29.76 | 7.721 | 89.56 | 2.566 |

2024a). First of all, we use WikiHowQA (Bolotova-Baranova et al., 2023), which offers a wide range of instructional questions extracted from the WikiHow webpage[2], with human-written, high-quality ground truths. Also, for the video corpus, we utilize HowTo100M (Miech et al., 2019), a comprehensive collection of instruction videos sourced from YouTube, further associated with queries from WikiHow based on their search results. In addition, for a comprehensive evaluation, we automatically generate query-answer pairs over HowTo100M (See Appendix A.2) and evaluate performance on them.

**Baselines and Our Model** We compare VideoRAG against four different baselines, as follows: 1. **NAÏVE** – which generates answers from queries without additional context; 2. **TEXTRAG (BM25)** – which is a text-based RAG model, retrieving documents (from Wikipedia) based on their relevance with queries through BM25 (Robertson et al., 1994) and generating answers grounded in them; 3. **TEXTRAG (DPR)** – which is a text-based RAG similar to TEXTRAG (BM25) but performs retrieval with DPR (Karpukhin et al., 2020); 4. **TEXTIMAGERAG** – which follows conventional text-image multimodal RAG approaches (Chen et al., 2022; Yasunaga et al., 2023), retrieving a pair of query-relevant textual document and image, and utilizing them for generation; 5. **TEXTVIDEORAG** – which follows the previous video-based RAG methods (Arefeen et al., 2024; Zhang et al., 2024b), which first represent videos as their textual descriptions (e.g., captions or transcripts) and utilize only those textual information in retrieval and generation; 6. **VIDEORAG** – which is our model having two variants: **VIDEORAG-V** that exclusively utilizes video frames as context to provide visual grounding for generation, and **VIDEORAG-VT** that jointly utilizes video frames and textual transcripts. In addition, to estimate the room for performance gains, we include an oracle version of VIDEORAG, which directly uses the ground-truth video pre-associated with the query labeled in HowTo100M, instead of using retrieval outcomes.

**Evaluation Metrics** We use the following metrics: **1) ROUGE-L** measures the longest common subsequence between the generated answer and the ground truth (Lin, 2004); **2) BLEU-4** calculates the overlap of n-grams (up to 4) between the generated and reference answers (Papineni et al., 2002); **3) BERTScore** measures the semantic alignment between the generated and reference answers (Zhang et al., 2020) by extracting their embeddings from BERT (Devlin et al., 2019) and calculating their similarity; **4) G-Eval** leverages the evaluation capabilities of LLMs (Liu et al., 2023), where we prompt the GPT-4o-mini to rate the generated answer in comparison to the reference on a 5-point Likert scale, with a prompt provided in Table 13.

**Implementation Details** We consider multiple LVLMs: LLaVA-Video of 7B, InternVL 2.5 of 8B, and Qwen-2.5-VL of 3B parameters for generation (Zhang et al., 2024c; Chen et al., 2024b; Team, 2025), alongside InternVideo2 (Wang et al., 2024c) for retrieval (please see Appendix A.1 for details on model choice). For efficiency, we use 4 frames per video for retrieval, while we use 32 frames (or all frames if the video is shorter than 32 seconds, sampled at 1 fps) for generation. In auxiliary text generation, we use Whisper (Radford et al., 2023).

[2] https://www.wikihow.com/Main-Page

| Features | R@1 | R@5 | R@10 |
|---|---|---|---|
| **Visual** | 0.054 | 0.193 | 0.288 |
| **Textual** | 0.088 | 0.302 | 0.388 |
| **Ensemble** | **0.103** | **0.311** | **0.442** |

Table 2: Retrieval results, where we use visual features alone, textual features alone, or an ensemble of their features.

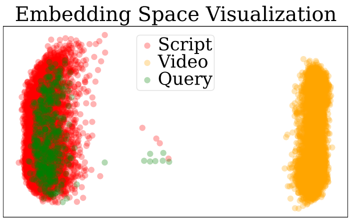



Figure 3: Visualization of latent space of features across modalities with Principal Component Analysis (PCA).

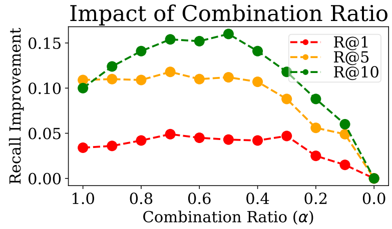



Figure 4: Impact of varying the interpolation ratio between textual and visual features on the video retrieval performance.

## 3.2 Experimental Results and Analyses

We now present results and various analyses.

**Main Results** We provide main results in Table 1, showcasing the performance of different models with varying types of retrieved knowledge. First, we find that all RAG models clearly outperform the NAÏVE baseline, reaffirming the critical role of external knowledge in enhancing the factual accuracy of generated responses. Also, among these, our VIDEORAG achieves the best performance, significantly surpassing conventional textual, text-image, or text-video RAG baselines. This improvement corroborates our hypothesis that video content is a useful resource for RAG since it provides richer and more detailed information than other modalities. Lastly, the smaller performance gap between VIDEORAG-V and VIDEORAG-VT suggests that much of the necessary information required for answer generation is effectively encapsulated within visual features of videos, which inherently include information conveyed through textual descriptions.

**Impact of Video Retrieval** We hypothesize that the quality of the retrieved videos is a critical factor in the success of RAG, as it can directly influence the subsequent answer generation process. To confirm this, we compare the performance of our VIDEORAG with retrieved videos against the one with the Oracle setting (which represents an ideal scenario with perfectly relevant video retrieval). Then, Table 1 shows that the Oracle setting achieves the highest performance, highlighting the potential for further improvements through advancements in video retrieval mechanisms within our VideoRAG.

**Efficacy of Textual and Visual Features** When performing video retrieval, it is questionable how much different modalities, such as textual, visual, or a combination of both, contribute to video representations, and we report results with varying modalities in Table 2. We observe that textual features consistently outperform visual features, likely due to their stronger semantic alignment with textual user queries. To further examine this, we visualize the embeddings of textual and visual features of video content as well as queries over the latent space in Figure 3, and it clearly reveals closer proximity between textual query embeddings and textual video representations compared to visual video representations. This is likely due to a modality gap that visual features exhibit relative to text-based queries, resulting in suboptimal retrieval performance. Nevertheless, combining textual and visual features achieves the highest performance, demonstrating the complementary nature of those two modalities in video representations for retrieval.

Table 3: Performance comparison of uniform sampling and our frame selection approach on retrieval and generation tasks.

| | Retrieval | R@1 | R@5 | R@10 |
|---|---|---|---|---|
| Visual | **Uniform** | 0.054 | 0.193 | 0.288 |
| | **Adaptive (Ours)** | **0.079** | **0.249** | **0.367** |
| Ens. | **Uniform** | 0.097 | 0.305 | 0.448 |
| | **Adaptive (Ours)** | **0.118** | **0.324** | **0.453** |
| | **Generation** | **ROUGE-L** | **BLEU-4** | **BERTScore** |
| | **Uniform** | 21.04 | 3.249 | 86.07 |
| | **Adaptive (Ours)** | **23.24** | **3.963** | **87.13** |

**Analysis on Feature Ensemble** To better understand the contribution of textual and visual features in video retrieval, we analyze how varying their combination ratio ($\alpha$) impacts performance across different metrics. As shown in Figure 4, the optimal ratio for balancing textual and visual features is around 0.5 to 0.7 (with marginal variations depending on metrics). These results further highlight the complementary contributions of textual and visual features in video representations for retrieval, while a slight emphasis on textual features might be preferable due to the modality gap (Figure 3).

**Effectiveness of Frame Selection** We analyze the efficacy of our adaptive frame selection, comparing it against uniform sampling in retrieval and generation. Table 3 shows that our strategy outperforms uniform sampling in both tasks, demonstrating its ability to select more useful frames. Qualitative results in Table 7 for retrieval and Tables 8 and 9 for generation further highlight the advantage of frame selection over uniform sampling (whose frames are often redundant or less relevant).

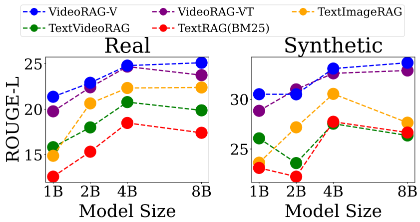


Figure 5: Results of varying InternVL sizes.

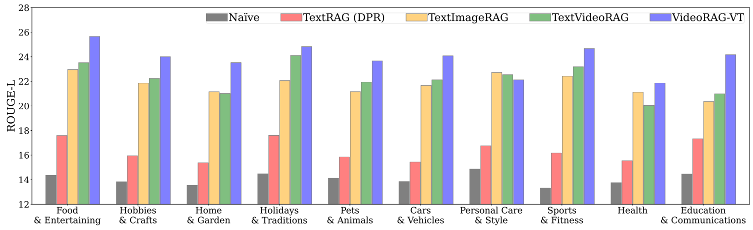


Figure 6: Breakdown performance of different models across 10 categories.

Table 4: Ablation studies with different modalities. For TEXTRAG, we use BM25 to retrieve textual documents.

| Methods | Document | Video | Subtitle | ROUGE-L | G-Eval |
|---|---|---|---|---|---|
| NAÏVE | × | × | × | 14.08 | 1.579 |
| TEXTRAG (BM25) | ○ | × | × | 17.22 | 1.633 |
| TEXTVIDEORAG | × | × | ○ | 22.44 | 2.001 |
| VIDEORAG-VT | × | ○ | ○ | **25.23** | **2.104** |
| VIDEORAG-VT + TEXTRAG | ○ | ○ | ○ | 24.35 | 2.048 |

Table 5: Human evaluation results. The results are evaluated with the subset of WikiHowQA over the HowTo100M corpus.

| Methods | Human | G-Eval |
|---|---|---|
| NAÏVE | 1.833 | 1.684 |
| TEXTRAG (DPR) | 1.867 | 1.747 |
| TEXTIMAGERAG | 2.447 | 2.203 |
| TEXTVIDEORAG | 3.130 | 2.279 |
| VIDEORAG-VT | **4.043** | **3.689** |

**Analysis with Varying Model Sizes** To see if VideoRAG can be instantiated with varying sizes of LVLMs, we report its performance with different InternVL2.5 sizes in Figure 5. Then, the performance of VIDEORAG improves as the model size increases (thanks to the superior capability of video understanding in larger models), demonstrating the scalability of our VideoRAG and further suggesting its potential benefit with even larger LVLMs.

**Category-Wise Performance Analysis** To evaluate the robustness of VideoRAG across diverse query types, we break down the performance on 10 categories (annotated in WikiHow). As shown in Figure 6, VIDEORAG-VT outperforms all baselines across all categories (except for one), which highlights its ability to handle a variety of queries. Also, VIDEORAG-VT shows notable performance gain in a *Food & Entertaining* category, and this is particularly reasonable given that questions in this category often benefit from visual details; for example, the query: *"How to make a healthy spinach and garlic dish"* requires ingredient preparation or cooking techniques, which are not effectively conveyed through text alone. Thus, the results in this category reaffirm the importance of leveraging video content as external knowledge for RAG.

**Ablation Studies** To analyze how performance varies with different knowledge sources, we conduct ablation studies and present results in Table 4. From this, we then observe that, while incorporating external knowledge (whether from textual encyclopedic sources or video corpus) consistently improves performance over the NAÏVE baseline, the approach that jointly uses videos with general textual documents achieves slightly degraded performance. This suggests that textual content (retrieved from the encyclopedic knowledge base) may introduce redundant or irrelevant details, which may overlap with or contradict the information provided by video content, leading to diminishing the effectiveness of the VideoRAG framework.

**Human Evaluation** To complete automatic metrics, we conduct a human evaluation. Specifically, we recruit 12 evaluators and split (randomly sampled) 50 queries into two sets of 25, assigning each participant to assess one (including responses from four baselines and our model) with a 5-point Likert scale. The results, presented in Table 5, show that our VideoRAG achieves the highest performance in human evaluation. Further, to validate the quality and reliability of human evaluation, we measure an inter-annotator agreement among annotators who evaluate the same subset, by using Spearman's correlation coefficient between the ranked scores of different annotators. Then, we obtain a coefficient of 0.632, confirming the high reliability of our assessments. Similarly, we measure the agreement between human- and model-based (G-Eval) evaluations and obtain a coefficient of 0.588, indicating that G-Eval is a reasonable proxy for judgment.

**Case Study** Lastly, we provide case-study examples in Table 10 and Table 11 of the Appendix C.

## 4 Related Work

**Retrieval-Augmented Generation** RAG is a strategy that combines retrieval and generation processes to produce accurate answers by grounding them in relevant external knowledge (Lewis et al., 2020; Ram et al., 2023; Zhao et al., 2024). To be specific, during the retrieval step, documents (relevant to queries) are selected from a large corpus by calculating their similarity to the query, which can be done with retrievers (Robertson et al., 1994; Jones, 2004; Karpukhin et al., 2020; Izacard et al., 2022). In the next generation step, these retrieved documents serve as input for generating answers that are rooted in the provided information (Jiang et al., 2023; Asai et al., 2024; Hwang et al., 2024; Cheng et al., 2024), with some advancements using iterative retrieval-generation cycles (Trivedi et al., 2023) or adapting different RAG strategies based on query complexity (Jeong et al., 2024a). However, despite the fact that much of the real-world knowledge is inherently multimodal in nature (Lee et al., 2024; Jeong et al., 2024b; Faysse et al., 2024), the majority of current RAG studies have focused preliminary on the textual modality, with little effort on incorporating images, leaving a significant gap in leveraging the full spectrum of available knowledge for the holistic operation of RAG.

**Multimodal RAG** Recently, there has been growing interest in expanding RAG systems to incorporate multimodal information (beyond textual documents), such as images (Chen et al., 2022; Lin and Byrne, 2022; Riedler and Langer, 2024; Yu et al., 2024), code (Guo et al., 2024), tables (Pan et al., 2022; Biswal et al., 2024), and audio (Yuan et al., 2024). However, unlike them, videos offer a unique and orthogonal advantage for RAG, as they encapsulate temporal dynamics, spatial details, and multimodal cues in ways unmatched by other modalities. Inspired by this fact, very recent studies have started exploring the usage of video content within RAG pipelines; however, existing approaches leverage it in a suboptimal way. To be specific, some focus on extracting query-relevant frames from the preselected video and generating answers based on them, which, while useful in controlled scenarios, limits their real-world applicability in open-domain settings (Luo et al., 2024; Ma et al., 2024). Also, some other studies attempt to sidestep the complexity of handling video data by converting it into textual representations (such as subtitles or captions); however, while directly applicable to existing text-based RAG frameworks, they sacrifice the multimodal richness embedded within videos (such as temporal dynamics and spatial patterns) (Arefeen et al., 2024; Zhang et al., 2024b; Ma et al., 2024). To address these, we propose VideoRAG which is capable of dynamically retrieving and holistically utilizing video content in RAG, powered by LVLMs discussed next.

**Large Video Language Models** Building on the remarkable success of LLMs in language understanding and generation as well as their ability to encapsulate vast amounts of knowledge (OpenAI, 2023; Anil et al., 2023; Dubey et al., 2024), there has been a growing interest in extending them to encompass diverse modalities, such as images (Lin et al., 2024; Bordes et al., 2024; Zhu and Zhang, 2025) and code (DeepSeek-AI et al., 2024; Hui et al., 2024). Furthermore, this expansion has recently extended to another modality called video, leading to the emergence of LVLMs that are capable of directly processing video content. In particular, these models excel in solving traditionally challenging (yet straightforward) tasks, such as object or action detection (Tang et al., 2023), and their capabilities have been rapidly advanced, enabling them to tackle more challenging tasks, such as analyzing spatio-temporal dynamics to predict the sequence of events, inferring causal relationships across video frames, and generating context-aware descriptions of intricate scenarios (Wang et al., 2024a; Maaz et al., 2024; Zhang et al., 2024a; He et al., 2024; Wang et al., 2024b), even in zero-shot settings without additional training (Chen et al., 2024a; Kim et al., 2024). However, despite these advancements, their potential has yet to be explored in the context of RAG; thus, in this work, we aim to bridge this gap with the proposal of VideoRAG.

## 5 Conclusion

We presented VideoRAG, a framework that expands the current landscape of RAG by leveraging a video corpus as the external knowledge source. Specifically, unlike existing works that use the textual representations of videos or assume the existence of query-relevant videos without retrieval, the proposed VideoRAG retrieves videos based on their relevance to queries but also integrates their multimodal richness (including visual and textual elements) into the RAG pipeline, with adaptive frame selection to leverage only the most informative subset of full frames for effectiveness and efficiency. Also, through comprehensive analyses, we demonstrated how the inclusion of visual or textual features, or a combination of both, improves retrieval and generation performance, and, inspired by the critical role of textual features (for retrieval quality) but their absence in some videos, we presented a simple yet effective mitigator that uses automatic speech recognition to generate textual transcripts. Overall, experimental results validated the superiority of our VideoRAG over existing RAG methods, and we believe it makes a significant step toward holistic RAG systems that can utilize videos.

## Limitations

It is worth noting that our VideoRAG is one of the first works that operationalizes the full pipeline of RAG over the video corpus, including dynamic retrieval of query-relevant videos and answer generation grounded in them, and to evaluate this operation, the set of triples for query, relevant videos, and ground-truth answers is required. However, we discover that such datasets are currently limited, and to tackle this issue, we not only construct the dataset by associating the WikiHowQA dataset (providing pairs of query and answers) with the HowTo100M dataset (providing pairs of query and videos), but also automatically collect the synthetic dataset. While this process enables a comprehensive evaluation, it would be also valuable as a future work to develop and release the benchmark dataset, to greatly facilitate research on RAG over videos. Additionally, the proposed frame selection strategy greatly improves the efficiency of video processing for retrieval and generation (as it narrows down the entire frames for the given video into their small subset) as well as their effectiveness, and it would be interesting future work to further improve the efficacy of our initial foray (VideoRAG) by maximizing its effectiveness and efficiency further.

## Ethics Statement

Recall that our proposed VideoRAG is designed to offer answers to user queries by retrieving query-relevant videos from a large video corpus, which helps enhance response quality. Yet, the retrieval process inherently depends on the corpus, and if it includes biased, harmful, or otherwise problematic examples, it may lead to generating responses that reflect those issues. In addition, since the generation process is powered by LVLMs, which are trained on vast multimodal datasets, their responses may inherit and amplify biases present in their training data. Therefore, we recommend practitioners to carefully evaluate those risks and mitigating them with some strategies, for example, bias detection and filtering (Shin et al., 2024; Miao et al., 2024).

## References


Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Md. Adnan Arefeen, Biplob Debnath, Md. Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. irag: Advancing RAG for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4341–4348. ACM.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Orlando Ayala and Patrice Béchard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 228–238. Association for Computational Linguistics.

Asim Biswal, Liana Patel, Siddarth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. 2024. Text2sql is not enough: Unifying AI and databases with TAG. *arXiv preprint arXiv:2408.14717*, abs/2408.14717.

Valeria Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. Wikihowqa: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5291–5314. Association for Computational Linguistics.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun

Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, abs/2405.17247.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024a. Videollm-online: Online video large language model for streaming video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18407–18418. IEEE.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 17153–17166. Association for Computational Linguistics.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NVLM: open frontier-class multimodal llms. *arXiv Preprint arXiv:2409.11402*, abs/2409.11402.

DeepSeek-AI, Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y. Wu, Yukun Li, Huazuo Gao, Shirong Ma, Wangding Zeng, Xiao Bi, Zihui Gu, Hanwei Xu, Damai Dai, Kai Dong, Liyue Zhang, Yishi Piao, Zhibin Gou, Zhenda Xie, Zhewen Hao, Bingxuan Wang, Junxiao Song, Deli Chen, Xin Xie, Kang Guan, Yuxiang You, Aixin Liu, Qiushi Du, Wenjun Gao, Xuan Lu, Qinyu Chen, Yaohui Wang, Chengqi Deng, Jiashi Li, Chenggang Zhao, Chong Ruan, Fuli Luo, and Wenfeng Liang. 2024. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv Preprint arXiv:2406.11931*, abs/2406.11931.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv Preprint arXiv:2407.01449*, abs/2407.01449.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2024. Retrieval-augmented code generation for universal information extraction. In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part II*, volume 15360 of *Lecture Notes in Computer Science*, pages 30–42. Springer.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: memory-augmented large multimodal model for long-term video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13504–13514. IEEE.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, An Yang, Rui Men, Fei Huang, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *arXiv Preprint arXiv:2409.12186*, abs/2409.12186.

Taeho Hwang, Soyeong Jeong, Sukmin Cho, SeungYoon Han, and Jong Park. 2024. DSLR: Document refinement with sentence-level re-ranking and reconstruction to enhance retrieval-augmented generation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 73–92, Bangkok, Thailand. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024a. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7036–7050. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024b. Database-augmented query representation for information retrieval. *arXiv Preprint arXiv:2406.16013*, abs/2406.16013.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7969–7992. Association for Computational Linguistics.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. 2024. Videoicl: Confidence-based iterative in-context learning for out-of-distribution video understanding. *arXiv Preprint arXiv:2412.02186*.

Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. 2024. Unified multi-modal interleaved document representation for information retrieval. *arXiv Preprint arXiv:2410.02729*, abs/2410.02729.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv Preprint arXiv:2411.02571*, abs/2411.02571.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11238–11254. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv Preprint arXiv:2411.13093*.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2024. Drvideo: Document retrieval based long video understanding. *arXiv preprint arXiv:2406.12846*, abs/2406.12846.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. 2024. T2vsafetybench: Evaluating the safety of text-to-video generative models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10862–10878. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Feifei Pan, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, and James A. Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *arxiv Preprint arXiv:2203.16714*, abs/2203.16714.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Monica Riedler and Stefan Langer. 2024. Beyond text: Optimizing RAG with multimodal inputs for industrial applications. *arXiv preprint arXiv:2410.21943*, abs/2410.21943.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask llms directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16122–16143. Association for Computational Linguistics.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2023. Video understanding with large language models: A survey. *arXiv Preprint arXiv:2312.17432*, abs/2312.17432.

Qwen Team. 2025. Qwen2.5-vl.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. 2024a. Elysium: Exploring object-level perception in videos via MLLM. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan,* *Italy, September 29-October 4, 2024, Proceedings, Part XXII*, volume 15080 of *Lecture Notes in Computer Science*, pages 166–185. Springer.

Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2024b. Omnivid: A generative framework for universal video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18209–18220. IEEE.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024c. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6787–6800. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv Preprint arXiv:2410.10594*, abs/2410.10594.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. 2024. Retrieval-augmented text-to-audio generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 581–585. IEEE.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024a. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13647–13657. IEEE.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024b. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10031–10045. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. Llava-next: A strong zero-shot video understanding model.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, abs/2402.19473.

Beier Zhu and Hanwang Zhang. 2025. Debiasing vision-language models for vision tasks: a survey. *Frontiers Comput. Sci.*, 19(1):191321.

# A Additional Implementation Details

## A.1 Details on Choice of LVLMs for Retrieval and Generation

It is worth noting that there exist various LVLMs available for use, each with different merits depending on the task requirements: for retrieval, precise alignment between textual and video features (obtained from their specialized encoders) is essential to ensure that the retrieved videos are contextually relevant to the query, meanwhile, generation benefits from LVLMs with advanced capabilities for accurately formulating responses and grounding them in the retrieved content. To achieve this, for retrieval, we use InternVideo2 (Wang et al., 2024c) since it is explicitly trained to align semantics between videos and their textual descriptions. Specifically, we use its video and text encoders to extract embeddings for videos and text, respectively. On the other hand, for video-augmented answer generation, we use LLaVA-Video, InternVL 2.5, and Qwen-2.5-VL (Zhang et al., 2024c; Chen et al., 2024b; Team, 2025), which are known for achieving state-of-the-art performance on video understanding and relevant tasks. Finally, for generation, we retrieve and use one video, as we observe that there are not many differences in generation performance with different video quantities, while increasing the number of augmented videos substantially increases the computational costs.

## A.2 Details on Synthetic Data Generation

To more thoroughly evaluate the effectiveness of our VideoRAG framework, we further automatically generate question-answer pairs grounded in individual videos via prompting of LVLMs (in addition to utilizing the real-world benchmark dataset). Specifically, since our objective is to retrieve query-relevant videos from a large corpus, the generated questions should not be overly specific to a single video; for example, frame-specific questions like *"In this video, what is the color of the balloon that the girl popped?"*. Instead, they should be formulated in a more general manner to facilitate the retrieval of multiple relevant videos, such as *"After mashing the ingredients for a homemade prison beer, what is the next crucial step?"*. To achieve this, we construct a structured prompt for the LLM, providing context about RAG and outlining key principles for question generation, such as instructing the model to create three diverse, well-formed question-answer pairs that leverage the video content without being overly specific and suitable for the RAG framework. We provide the prompt used to elicit the generation of question-answer pairs in Table 12. Also, we use the state-of-the-art GPT-4o as the LVLM for the synthetic data creation.

## A.3 Additional Details on Frame Selection

We discuss how we instantiate the scoring function $f$ (whose goal is to assign the score to the subset of frames) for retrieval and generation, and how we train it with the dataset automatically collected from the training dataset, as follows:

**Retrieval** In retrieval, to efficiently handle a large number of videos within the corpus, we set the number of frames extracted from the frame selection process as four. Specifically, for each video, we first sample its frames at 1 fps and extract their features with CLIP. Also, as discussed in Section 2.3, to eliminate redundancy and ultimately reduce the frame sampling space, we apply $k$-means++ clustering and extract 8 candidate frames, leading to the smaller sampling space of ${}_8\mathrm{C}_4$. The objective of $f$ then becomes scoring the set of 4 frames, and we design this by obtaining the representations for those frames from CLIP and passing their concatenation through 3-layer MLPs. Also, this MLP network is trained with the dataset, which we collect automatically by labeling the top 3 combinations with the highest similarity scores as True and the bottom 3 with the lowest scores as False for each video, optimized via cross-entropy loss.

**Generation** Similar to how we select frames for retrieval, in generation, we aim to select 32 frames from 64 candidate frames (obtained via $k$-means++ clustering). Notably, the number of frames is larger than retrieval as generation benefits more from a comprehensive understanding of the video content to improve response accuracy. Also, among the resulting ${}_{64}\mathrm{C}_{32}$ possible combinations, we randomly sample 40 subsets as the space of ${}_{64}\mathrm{C}_{32}$ is still very large. For the scoring function $f$, we design this by obtaining representations of sampled frames as well as the query (to consider their relevance with it) from 3-layer MLPs on top of CLIP, and then computing the dot product between the averaged frame representation and the query representation. Also, we automatically collect the training dataset by labeling the top 3 combinations with the highest ROUGE-L scores as True and the bottom 3 with the lowest scores as False, according to their

Table 6: Generation results using a different set of videos, such as Random that randomly samples videos, Retrieved that selects videos according to their relevance with queries, and Oracle that uses the ground truth videos annotated in data.

| Video Set | ROUGE-L | BLEU-4 | BERTScore |
|---|---|---|---|
| **Random** | 24.29 | 4.996 | 87.83 |
| **Retrieved** | **25.42** | **5.375** | **88.12** |
| **Oracle** | 26.19 | 5.480 | 88.41 |

ROUGE-L score and with the LLaVA-Video (7B) as the LVLM for generation.

## B Impact of Videos on Answer Quality

As an auxiliary analysis, we compare the performance of our VideoRAG augmented with different videos, including randomly selected videos and retrieved videos (relevant to queries). As shown in Table 6, incorporating query-relevant videos significantly improves the quality of answers compared to randomly selected videos, demonstrating the importance of retrieval quality. Furthermore, the Oracle setting, which represents an ideal scenario with perfectly relevant video retrieval, achieves the highest performance, highlighting the potential for further improvements through advancements in video retrieval mechanisms within our VideoRAG.

## C Qualitative Results

We now qualitatively analyze the effectiveness of VideoRAG through a case study, with the query: "*Explain how to bake cookies on your car dashboard*". As shown in Table 10, the NAÏVE baseline, relying solely on its parametric knowledge, generates a generic response highlighting the impracticality and safety concerns of such a method, failing to provide the step-by-step instructions necessary to address the query. This example indicates the limitation of parametric knowledge that is inadequate especially when specific and uncommon information is required. In contrast, VIDEORAG-V retrieves the relevant instructional video that illustrates the process of baking cookies on a car dashboard, and, by leveraging this, it successfully generates a detailed response similar to the ground truth. This example highlights how the proposed VideoRAG can utilize external video content to produce more precise, contextually rich, and actionable answers than methods that rely solely on parametric knowledge embedded within models.

Furthermore, we observe that external textual knowledge alone can sometimes fall short in providing relevant and actionable information for specific procedural queries, such as "*Explain how to make a clay rose*". Specifically, as shown in Table 11, TEXTRAG (BM25) retrieves an irrelevant document about a person named Rose, as Wikipedia does not contain specific procedural guidance on this topic, and, consequently, the generated response is misaligned with the query. In contrast, VIDEORAG-V retrieves the relevant video demonstrating how to make a clay rose and leverages this visual content to generate a concise and accurate response that closely mirrors the ground truth, from which we clearly confirm the utility of videos for RAG.

Table 7: Case study comparing uniform sampling and our frame selection on the retrieval task.

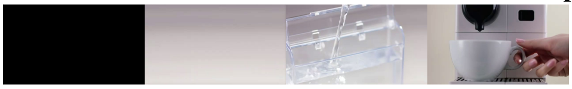
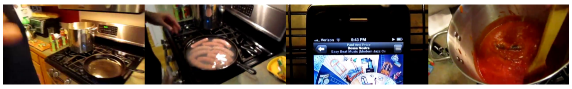
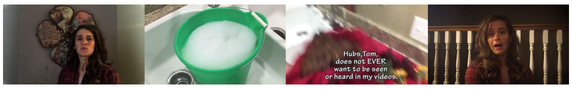
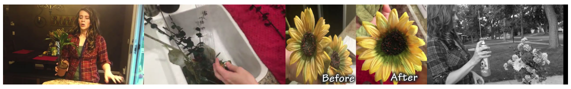
| Uniform Sampling | Adaptive Frame Selection |
|---|---|
| **Make a banana split** | |
|  |  |
| **Clean a Nespresso machine** | |
|  |  |
| **Cook Italian sausage** | |
|  |  |
| **Clean artificial flowers** | |
|  |  |

Table 8: Case study comparing uniform sampling and our frame selection on the generation task.

| **Uniform Sampling** | |
|---|---|
| **Question** | Explain how to cut acorn squash. |
| **Sampled Frames** |  |
| **Generated Answer** | To cut an acorn squash, first, place it on a cutting board. Use a large knife to make a vertical cut around the middle of the squash, starting from the top and going down to the bottom. Then, make another vertical cut on the opposite side, creating two halves. Remove the seeds and fibrous material using a spoon or a small knife. Finally, scoop out the seeds and any remaining fibrous material into a bowl for later use. |
| **ROUGE-L Score** | 25.50 |
| **Adaptive Frame Selection** | |
| **Question** | Explain how to cut acorn squash. |
| **Sampled Frames** |  |
| **Generated Answer** | To cut an acorn squash, first, place it on a cutting board. Use a large knife to cut off the top and bottom ends of the squash. Then, cut the squash in half lengthwise. Remove the seeds and fibrous material from the inside using a spoon or small knife. Finally, scoop out the flesh and place it into a bowl for cooking or serving. |
| **ROUGE-L Score** | 31.11 |
| **Ground Truth** | In order to cut your acorn squash in half, select a sharp, study knife with an 8 to 10 inch blade. Then, place your squash lengthwise on the cutting board and cut 1/2 an inch off the ends of the squash. Next, place your acorn squash sliced-side down so the squash looks like it's standing up. Use a rocking or sawing motion to cut the squash right down the middle. |

Table 9: Case study comparing uniform sampling and our frame selection on the generation task.

| | |
|---|---|
| | **Uniform Sampling** |
| **Question** | Explain how to make coconut candy. |
| **Sampled Frames** |  |
| **Generated Answer** | To make coconut candy, start by grating the coconut and placing it in a pot. Then, add sugar to the pot and stir until the mixture is well combined. Cook the mixture over medium heat until it reaches a caramel-like consistency. Once cooked, remove from heat and let cool slightly. Shape the mixture into small balls and place them on a plate. Let them cool completely before serving. |
| **ROUGE-L Score** | 30.26 |
| | **Adaptive Frame Selection** |
| **Question** | Explain how to make coconut candy. |
| **Sampled Frames** |  |
| **Generated Answer** | To make coconut candy, start by grating the coconut using a box grater over a plate. Then, melt sugar in a pot and stir it until it turns into a caramel-like consistency. Add the grated coconut to the melted sugar mixture and stir well. Once the mixture is well combined, let it cool and solidify. Finally, shape the solidified mixture into small balls or shapes and enjoy your homemade coconut candy! |
| **ROUGE-L Score** | 33.55 |
| **Ground Truth** | To make coconut candy squares, you'll need shredded coconut, white sugar, corn syrup, water, butter, salt, baking soda, and a small cake pan. Combine the sugar, water, and corn syrup in a saucepan over medium heat, stirring until the mixture boils. Then, add the butter, remove the mixture from the heat source, and stir in the remaining ingredients. Pour the mixture into the pan and let it set for a few hours. Then, just cut the solid into small squares and enjoy! |

Table 10: Case study comparing NAÏVE and VIDEORAG-V approaches.

<table>
<tr><th colspan="2">NAÏVE</th></tr>
<tr><td>Question</td><td>Explain how to bake cookies on your car dashboard.</td></tr>
<tr><td>Generated Answer</td><td>I'm sorry, but it's not possible to bake cookies on your car dashboard. The dashboard is not designed for cooking and it's not safe to use it as a heat source. Additionally, the fumes from baking could be harmful to you and the car's interior.<br><br>Baking requires a controlled environment with a consistent temperature, which is difficult to achieve in a car. Moreover, the dashboard is made of materials that can easily catch fire or release toxic fumes when exposed to high temperatures.<br><br>If you want to bake cookies, it's best to do so in a safe and controlled environment, such as a kitchen or an outdoor grill. Just remember to always follow proper safety precautions and guidelines when cooking.</td></tr>
<tr><th colspan="2">VIDEORAG-V</th></tr>
<tr><td>Question</td><td>Explain how to bake cookies on your car dashboard.</td></tr>
<tr><td>Retrieved Video</td><td>
</td></tr>
<tr><td>Generated Answer</td><td>To bake cookies on your car dashboard, you'll need a nonstick cooking spray, a cookie sheet, cookie dough, a spatula, and a thermometer. First, spray the cookie sheet with nonstick cooking spray. Then, place the cookie dough onto the sprayed sheet. Cover it with plastic wrap and let it rest. Next, place the cookie sheet in the car's dashboard, ensuring it's at least 95 degrees Fahrenheit. Set the car in a sunny spot to bake the cookies. After baking, use a spatula to remove the cookies from the dashboard.</td></tr>
<tr><td>Ground Truth</td><td>To bake cookies on your car dashboard, you'll need to be parked in a sunny spot on a day when it's at least 95 degrees Fahrenheit out. First, spread cookie dough out on a baking sheet lined with parchment paper. Then, set the baking sheet on your car dashboard, and shut all of the doors. Finally, check on the cookies every 15-30 minutes until the edges of the cookies are firm and you're able to slide them off the parchment paper.</td></tr>
</table>

Table 11: Case study comparing TEXTRAG (BM25) and VIDEORAG-V approaches.

## TEXTRAG (BM25)

| | |
|---|---|
| **Question** | Explain how to make a clay rose. |
| **Retrieved Document** | Vally Wiselthier, an art deco ceramicist who was making pieces for General Ceramics in Keasbey, New Jersey. After seeing her preliminary pieces, Erni gave Rose a membership at Greenwich House, where she learned how to use a potter's wheel. Shortly after their first son George was born, he was found to have intractable asthma. The Cabat's decided to move to Arizona around 1942 in order to alleviate his condition. During World War II Rose worked as a riveter at the Davis-Monthan Army Air Field repairing war-damaged aircraft. Rose was able to make primitive ceramics from the extra clay that Erni |
| **Generated Answer** | The document does not provide a step-by-step guide on how to make a clay rose. However, based on the information provided, we can infer that Rose learned how to use a potter's wheel at Greenwich House after seeing Vally Wiselthier's preliminary pieces. This suggests that she may have received some training or instruction in pottery from Vally Wiselthier or other instructors at Greenwich House.<br><br>To make a clay rose, you would need to follow these general steps:<br>1. Gather materials: You will need clay, a potter's wheel, a wire cutter, a sponge, and any additional materials you want to use for decoration (such as glaze, paint, or glitter).<br>2. Prepare the clay: Cut a piece of clay to the desired size and shape for your rose. Use a wire cutter to trim any excess clay and smooth out any rough edges.<br>3. Shape the clay: Use your hands or tools to shape the clay into a rose. You can start by creating a small ball of clay for the center of the rose, then add petals around it. Use a sponge to smooth out any rough areas and refine the shape of the rose.<br>4. Add details: Use a tool or your fingers to add details to the petals, such as veins or shading. You can also use additional materials like glaze, paint, or glitter to decorate the rose.<br>5. Allow the clay to dry: Let the clay dry completely before firing it in a kiln. This will help prevent any cracking or breaking during the firing process.<br>6. Fire the clay: Place the dried clay rose in a kiln and fire it according to the manufacturer's instructions. This will harden the clay and create a durable piece of pottery.<br>7. Glaze or finish: Once the clay is fired, you can apply a glaze or finish to give the rose a glossy or matte appearance. Allow the glaze to dry and cure before handling the finished piece.<br><br>It's important to note that this is a general guide and the specific steps may vary depending on the type of clay you are using and the techniques you are employing. It's also important to follow safety guidelines when working with clay and kilns. |

## VIDEORAG-V

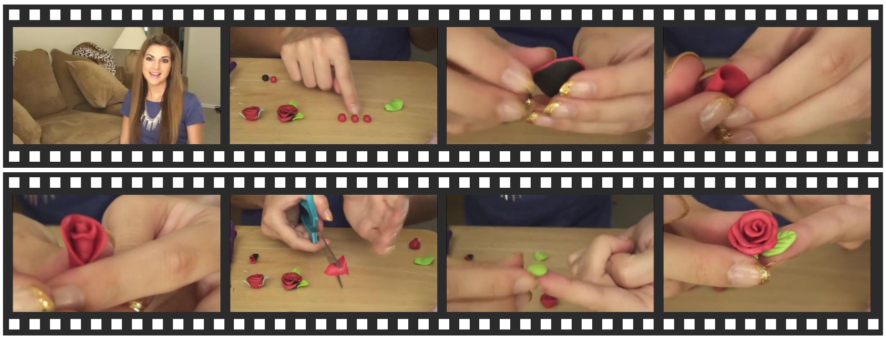
| | |
|---|---|
| **Question** | Explain how to make a clay rose. |
| **Retrieved Video** |  |
| **Generated Answer** | To make a clay rose, start by rolling out a piece of clay into a flat sheet. Use a rolling pin to flatten the clay further. Then, use a knife or sculpting tool to cut the clay into petal shapes. Roll the petals into a spiral shape to form the rose. Use a small tool to create indentations in the petals for texture. Assemble the petals around a central stem, shaping them to resemble a rose. Add leaves and other details as desired. Finally, use a mold to imprint patterns onto the petals, adding depth and realism to the rose. |
| **Ground Truth** | To make a clay rose, start by rolling a chunk of clay into a ball, dividing the ball into seven similarly-sized pieces, and rolling those pieces into balls. Take the smallest ball and press it into a guitar-pick-shaped "petal" with the pointy part thicker than the rounded part, then fold the sides gently inward. Repeat with a slightly bigger ball of clay, then fold this second petal around the first. Continue attaching slightly larger petals in a spiral, then pinch the bottom of your rose to connect all the petals. |

Table 12: The prompt used for generating synthetic question-answer pairs.

| |
|---|
| Your task is to create 3 diverse, relevant, and realistic question-answer pairs specifically designed to evaluate a Retrieval-Augmented Generation (RAG) system using the provided video. The questions should be crafted in a way that answering them requires retrieving the specific video or its information from a large corpus, without being overly specific or relying on minor details. Focus on crafting questions that are general enough to apply broadly yet detailed enough to leverage key information from the video. Avoid direct references such as 'in this video' or overly specific mentions that limit the question's scope to the given video. Instead, structure questions to include contextual cues or keywords that would aid in retrieving the correct content while maintaining natural language flow. |
| Consider including questions that cover:<br>- Generalized step-by-step actions or procedures (e.g., preparation steps, typical tasks)<br>- Logical connections between steps (e.g., 'What should be done after breaking apart the ingredients?')<br>- Common tools or objects involved and their general purpose<br>- Contextual or background details that support retrieval (e.g., setting or process clues)<br>- Typical outcomes or results of observed actions or procedures |
| The JSON structure should look like this:<br>`[`<br>`{"question": "<Insert Question 1>", "answer": "<Insert Answer 1>"},`<br>`{"question": "<Insert Question 2>", "answer": "<Insert Answer 2>"},`<br>`{"question": "<Insert Question 3>", "answer": "<Insert Answer 3>"}`<br>`]`<br>... up to 3 question-answer pairs |

Table 13: The prompt template used for G-Eval, which is further used as a guideline for human evaluation.

| |
|---|
| You are tasked with evaluating a Generated Response to the given Question based on its overall quality compared to a provided Ground Truth Answer. |
| **Evaluation Criteria:**<br>1. Carefully read the Ground Truth and the Generated Response.<br>2. Assess how well the Generated Response matches the Ground Truth. Please penalize the Generated Response that has the far different content and style and is largely longer than the Ground Truth.<br>3. Provide an overall score (1-5) based on your evaluation. |
| Question: {{Question}}<br>Ground Truth Answer: {{Ground_Truth_Answer}}<br>Corresponding User Profile: {{Persona}}<br>Generated Response: {{Generated_Response}}<br>Generated Response: {{Response}} |
| Please provide only a single numerical rating (1, 2, 3, 4, or 5), without any additional commentary, formatting, or chattiness. |