Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Semantic Retrieval for Long-Form Video Question Answering

Krishil Parikh

krishil.prkh75@gmail.com

Dwarkadas J. Sanghvi College of Engineering https://orcid.org/0009-0003-3837-8121

Krisha Maisheri

Dwarkadas J. Sanghvi College of Engineering https://orcid.org/0009-0005-7569-5745

Keyush Nisar

Dwarkadas J. Sanghvi College of Engineering

Aruna Gawade

Dwarkadas J. Sanghvi College of Engineering

Nilesh Rathod

Dwarkadas J. Sanghvi College of Engineering

Angelin Florence

Dwarkadas J. Sanghvi College of Engineering

Ragini Mishra

Dwarkadas J. Sanghvi College of Engineering

Research Article

Keywords: Long-form video understanding, video question answering, hierarchical retrieval-augmented generation, temporal localization, multimodal reasoning

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11823375/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Hierarchical Semantic Retrieval for Long-Form Video Question Answering

Krishil Parikh[1,*], Krisha Maisheri[1], Keyush Nisar[1], Dr. Aruna Gawade[1], Dr. Nilesh Rathod[1], Prof. Angelin Florence[1], Prof. Ragini Mishra[1]

Abstract

Long-form videos pose a fundamental challenge for retrieval-augmented question answering because the visual evidence required to answer a query is often sparsely distributed across lengthy and semantically diverse temporal sequences. Existing retrieval frameworks typically perform coarse retrieval over entire videos or broad temporal segments, making it difficult to accurately localize the fine-grained evidence required for complex, multi-step reasoning.

To address this limitation, we propose **VRAG**, a hierarchical coarse-to-fine retrieval framework that progressively refines the search space through three levels of semantic granularity. Starting from a user query, VRAG decomposes complex questions into semantically meaningful sub-questions, retrieves the most relevant topics across candidate videos, refines retrieval to scene-level segments, and finally identifies the keyframes that provide the strongest visual evidence for answer generation. This progressive retrieval strategy reduces semantic ambiguity, suppresses unnecessary temporal search, and improves evidence localization in long-form videos.

We evaluate VRAG on a large-scale benchmark comprising 463 instructional videos and 1,103 temporally grounded queries. Experimental results demonstrate consistent improvements over state-of-the-art VideoRAG baselines across video-, topic-, and scene-level retrieval, achieving 42.16% Video Recall@1, 55.21% Topic Recall@1, and 31.82% Scene Recall@1 while substantially improving fine-grained temporal localization. These results demonstrate that hierarchical retrieval provides an effective and scalable solution for retrieval-augmented long-form video understanding.

[1]Department of Artificial Intelligence and Machine Learning, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

[*]Correspondence: Krishil Parikh, krishil.prkh75@gmail.com

Keywords: Long-form video understanding, video question answering, hierarchical retrieval-augmented generation, temporal localization, multimodal reasoning

1 Introduction

The rapid growth of long-form videos across domains such as education, healthcare, cooking, and technical training has created an increasing demand for intelligent systems capable of retrieving precise information from lengthy multimodal content. Unlike short videos, long-form videos often span tens of minutes and contain multiple semantic events, topic transitions, and temporally dispersed visual evidence. Consequently, answering a seemingly simple question frequently requires identifying a small set of relevant moments

hidden within thousands of video frames, making accurate evidence localization a fundamental challenge for long-form video understanding.

Recent advances in Retrieval-Augmented Generation (RAG) have significantly improved question answering by grounding language models with external knowledge (Guu et al. 2020; Lewis et al. 2020; Izacard and Grave 2021; Izacard et al. 2023). Motivated by this success, several video retrieval frameworks have extended RAG to long-form video understanding by retrieving temporally relevant video segments before answer generation (Jeong et al. 2025; Ren et al. 2025; Zeng et al. 2025). While these approaches improve retrieval efficiency, they generally rely on flat retrieval strategies that independently rank videos or temporal segments using global semantic similarity (Karpukhin et al. 2020; Khattab and Zaharia 2020; Zhao et al. 2024). Such approaches often overlook the hierarchical structure naturally present in instructional and procedural videos, where high-level topics consist of multiple scenes, each containing fine-grained visual evidence necessary to answer specific questions. As a result, coarse retrieval frequently retrieves semantically related yet temporally imprecise content, limiting localization accuracy for complex multi-step queries.

Consider the question, *"When should heavy cream be added while preparing the dish?"* Answering this question requires more than retrieving the correct video. The system must first identify the relevant cooking stage, then localize the corresponding scene within that stage, and finally retrieve the precise visual evidence showing the addition of heavy cream. Performing retrieval over the entire temporal space at once introduces unnecessary search complexity and increases the likelihood of retrieving semantically similar but incorrect moments.

To address these challenges, we propose **VRAG**, a hierarchical coarse-to-fine retrieval framework for long-form video question answering. Instead of treating retrieval as a single-stage ranking problem, VRAG progressively reduces the temporal search space through three semantic levels: topic-level retrieval, scene-level localization, and evidence-level keyframe selection. To further improve retrieval precision, we introduce sub-question decomposition that decomposes complex user queries into semantically coherent retrieval units, allowing each retrieval stage to focus on localized semantic intent. Finally, Gaussian-smoothed keyframe scoring suppresses noisy frame-level similarity fluctuations and produces temporally stable evidence localization for answer generation.

The proposed hierarchical retrieval strategy offers two key advantages. First, progressively constraining the search space substantially reduces semantic ambiguity by eliminating irrelevant temporal regions before fine-grained localization. Second, evidence retrieval becomes increasingly precise at each stage, enabling the system to identify temporally grounded visual evidence rather than relying solely on coarse semantic similarity. Together, these properties improve retrieval quality for long-form instructional videos while maintaining practical computational efficiency.

We evaluate VRAG on a benchmark consisting of 463 long-form instructional videos and 1,103 temporally grounded queries. Experimental results demonstrate consistent

improvements over existing VideoRAG baselines across video-, topic-, and scene-level retrieval benchmarks. Extensive ablation studies further validate the contribution of hierarchical retrieval, sub-question decomposition, and Gaussian-smoothed keyframe localization toward improving fine-grained temporal evidence retrieval.

Our contributions are summarized as follows:

- We propose **VRAG**, a hierarchical coarse-to-fine retrieval framework that progressively localizes evidence through topic-level retrieval, scene-level refinement, and keyframe-level evidence selection for long-form video question answering.
- We introduce a **sub-question decomposition** strategy that transforms complex user queries into semantically meaningful retrieval units, improving hierarchical evidence localization for multi-step instructional queries.
- We propose a **Gaussian-smoothed keyframe localization** method that reduces frame-level retrieval noise and produces temporally consistent evidence selection for answer generation.
- We conduct extensive experiments on a benchmark containing 463 instructional videos and 1,103 temporally grounded queries, demonstrating consistent improvements over VideoRAG baselines across video-, topic-, and scene-level retrieval tasks, together with comprehensive ablation studies validating each component of the proposed framework.

2 Related Work

Research in long-form video understanding has advanced rapidly through developments in retrieval-augmented generation (RAG), video-language models, and temporal localization. Existing approaches can broadly be categorized into retrieval-based video understanding, long-form video question answering, temporal grounding, and vision-language representation learning. We briefly review these directions and discuss the limitations that motivate our work.

2.1 Retrieval-Augmented Video Understanding

Retrieval-Augmented Generation (RAG) was originally introduced to improve language generation by retrieving relevant external knowledge prior to reasoning (Lewis et al. 2020). More recently, this paradigm has been extended to long-form video understanding, where retrieved video content provides grounding for question answering.

Early video RAG systems typically perform *flat retrieval*, where a query is independently matched against a collection of transcripts, clips, or fixed-length temporal segments (Krishna et al. 2017; Lei et al. 2021). While effective for relatively short videos,

such strategies become increasingly challenging as video duration grows and relevant evidence becomes sparsely distributed across multiple temporal regions.

VideoRAG (Jeong et al. 2025) addresses the modality gap between textual transcripts and visual content by jointly leveraging transcript information and vision-language embeddings. SceneRAG (Zeng et al. 2025) further improves retrieval by segmenting transcripts into semantically coherent scenes and constructing knowledge graphs to support multi-hop retrieval. Although these methods significantly improve retrieval quality, retrieval remains largely performed over a single semantic level, limiting fine-grained evidence localization within lengthy instructional videos.

Our work differs by introducing a hierarchical coarse-to-fine retrieval strategy that progressively reduces the temporal search space from semantic topics to scenes and finally to key visual evidence, enabling more precise temporal localization while preserving contextual information.

2.2 Long-Form Video Question Answering

Long-form video question answering requires models to reason over temporally distributed visual and textual evidence. Recent approaches primarily improve this capability through enhanced representations or stronger temporal reasoning.

LLoVi (Zhang et al. 2024) generates dense captions for video segments before aggregating them with large language models for reasoning across extended videos. Momentor (Qian et al. 2024) introduces large-scale temporal supervision to train video-language models capable of fine-grained temporal localization. UMaT (Bi and Xu 2025) transforms multimodal information into textual memory representations, enabling retrieval over unified text descriptions, while RAVU (Malik et al. 2025) employs graph-based retrieval to model complex spatio-temporal relationships between entities.

These approaches demonstrate that effective reasoning requires identifying temporally relevant evidence rather than processing entire videos. However, they either rely on expensive graph reasoning, localization-specific supervision, or text-centric abstractions that may discard important visual information. In contrast, our framework maintains visual embeddings throughout retrieval and progressively localizes evidence using hierarchical semantic refinement without requiring specialized temporal localization training.

2.3 Temporal Localization in Long Videos

Temporal localization aims to identify the precise temporal interval corresponding to a natural-language query. Classical approaches formulate this problem through proposal generation or direct boundary regression, while recent transformer-based methods model fine-grained interactions between textual queries and temporal video features (Lin et al. 2023b; Zala et al. 2023).

Although these methods achieve impressive localization accuracy, they generally assume that retrieval is performed within a single video and often require large amounts of

localization-specific supervision. Our work addresses a complementary problem: retrieval-based temporal localization across large collections of long-form videos. Rather than directly predicting temporal boundaries, we progressively narrow the retrieval space using hierarchical semantic retrieval followed by keyframe-level evidence localization.

2.4 Vision-Language Models for Video Retrieval

Contrastive vision-language models have substantially improved cross-modal retrieval by learning aligned representations of visual and textual data. CLIP (Radford et al. 2021) and SigLIP (Zhai et al. 2023) demonstrate that jointly embedding images and text enables effective zero-shot retrieval across diverse downstream tasks. OpenCLIP (Cherti et al. 2023), ALIGN (Jia et al. 2021), and BLIP-2 (Li et al. 2023) further extend this paradigm through reproducible open-source scaling, noisy web-scale supervision, and frozen-encoder bootstrapping with large language models, respectively.

Recent video-language models extend these representations to video understanding. Frozen-in-Time (Bain et al. 2021) learns joint video-text embeddings for retrieval, Video-LLaVA (Lin et al. 2023a) unifies image and video representations for multimodal reasoning, while InternVideo2 (Wang et al. 2024) scales video representation learning through large-scale multimodal pretraining. These models provide powerful visual representations but primarily focus on representation learning rather than hierarchical retrieval.

Our framework leverages pretrained vision-language embeddings as the foundation for fine-grained keyframe retrieval while integrating them within a hierarchical retrieval pipeline designed specifically for long-form instructional video understanding.

2.5 Research Gap

Despite significant advances in video retrieval and long-form video understanding, existing approaches generally follow one of three paradigms: (i) flat retrieval over transcripts or clips, (ii) text-centric abstractions that reduce multimodal information to textual memory, or (iii) computationally intensive graph-based reasoning. None simultaneously exploit hierarchical retrieval, progressive temporal refinement, and vision-language grounded evidence localization for scalable retrieval across large collections of long-form videos.

Our work addresses this gap through a hierarchical coarse-to-fine retrieval framework that combines semantic query decomposition, topic-level retrieval, scene-level refinement, and keyframe-level visual grounding within a unified retrieval pipeline.

3 Method

3.1 Problem Formulation

We consider the task of *long-form video question answering* over a corpus of instructional videos. Let $\mathcal{V} = \{V_1, \dots, V_N\}$ denote a collection of videos and let q represent a natural-

language query. The objective is to retrieve a compact set of temporally grounded *visual* and *textual* evidence that enables a language model to generate an accurate answer.

Each video V_i is a continuous multimodal stream comprising of (i) a visual stream of frames $\mathcal{F}_i = \{f_1, \dots, f_T\}$ and (ii) an aligned transcript $\mathcal{T}_i = \{t_1, \dots, t_M\}$ obtained via automatic speech recognition (ASR). Given q, it should return a minimal sufficient evidence set, $\mathcal{E} = \{e_1, \dots, e_K\}$ with which a language model can generate a correct and temporally grounded answer a.

Challenges. Long-form instructional videos present several challenges that make accurate retrieval and temporal grounding considerably more difficult than in short-form videos:

- **Long temporal duration:** videos may be many minutes long, making an exhaustive frame-level evaluation computationally intensive.
- **Multi-step procedural semantics:** often, queries may require several actions or steps (for instance, preparation, mixing, and cooking), which can occur at various time segments (Zhou et al. 2018; Zhukov et al. 2019).
- **Cross-modal semantic misalignment:** critical information may be presented visually (e.g., cutting technique) or verbally (e.g., ingredient quantity) but only when combining the modalities.
- **Limitations of flat retrieval:** retrieving fixed-length chunks or arbitrary frames can break semantic continuity, return irrelevant context, or miss the precise moment containing the answer (Sarthi et al. 2024; Jeong et al. 2025; Zeng et al. 2025).

Objective. To address these challenges, we propose a hierarchical semantic retrieval framework that (i) organizes each video into a topic→scene→frame hierarchy, (ii) performs multi-stage retrieval (topic selection, scene localization, keyframe extraction), and (iii) fuses multimodal evidence for answer generation.

Formally, the system learns a retrieval function

$$R(q, \mathcal{V}) \to \mathcal{E},$$

which returns a compact, high-relevance evidence set $\mathcal{E}$. A generative model then produces the final answer via

$$a = G(q, \mathcal{E}).$$

At a high level, our approach consists of an *offline stage* for hierarchical indexing of video content (illustrated in Fig. 1) and an *online stage* for multi-stage semantic retrieval and evidence fusion (illustrated in Fig. 2).

3.2 Offline Video Processing and Hierarchical Indexing

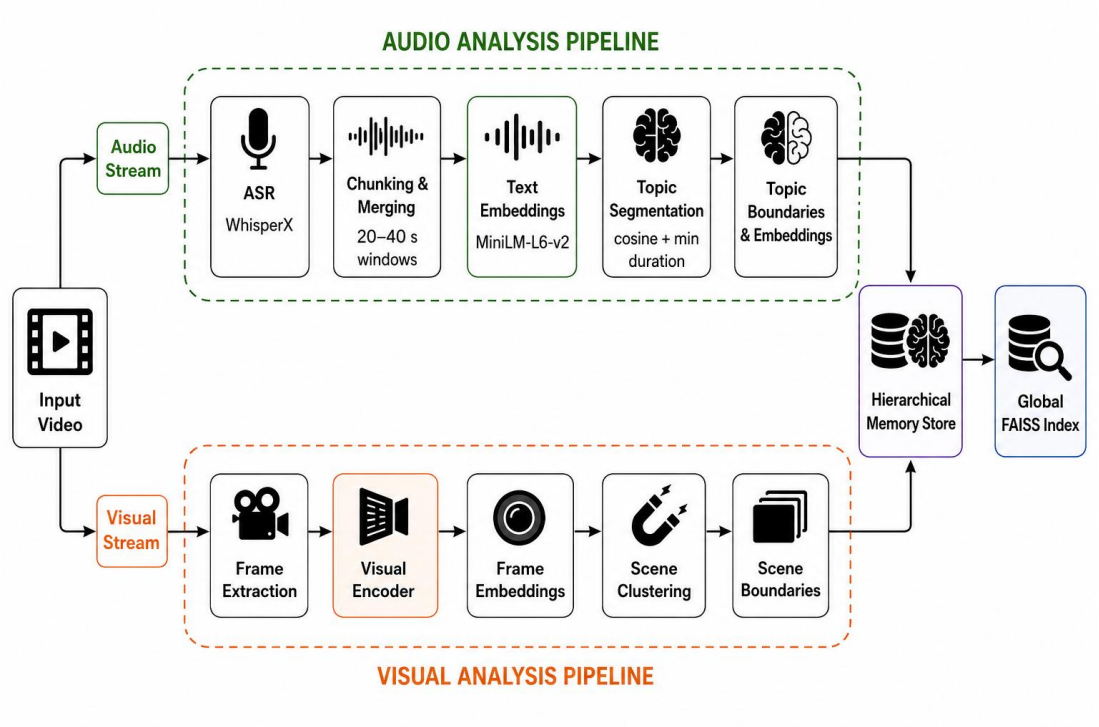

Fig. 1. Offline video processing and hierarchical indexing pipeline.

To facilitate efficient retrieval over long-form instructional videos, we construct a *hierarchical multimodal memory* for each video during an offline preprocessing stage. This memory organizes content into three semantic levels:

- **Topics:** coarse semantic segments aligned with major instructional steps.
- **Scenes:** finer-grained visual segments within each topic.
- **Frames:** individual visual observations used for keyframe selection.

The resulting hierarchy enables the retrieval process to progressively narrow the search space from videos to topics, topics to scenes, and finally scenes to individual keyframes.

3.2.1 Video Preprocessing

Each input video is decomposed into visual and textual modalities.

Frame Extraction. Frames are uniformly sampled at 1 FPS following sparse sampling strategies proposed in prior work (Gao and Lu 2022).

$$\mathcal{F} = \{f_1, \ldots, f_T\}.$$

The extracted frames constitute the fundamental visual units used for downstream scene clustering and keyframe retrieval. Each frame is associated with a timestamp, ensuring that visual evidence remains temporally grounded.

Speech Transcription. We process the audio track using an automatic speech recognition (ASR) system (Radford et al. 2022; Bain et al. 2023) to generate a time-aligned transcript

$$\mathcal{T} = \{(t_1, \tau_1), (t_2, \tau_2), \dots\},$$

where t_i denotes the spoken text segment and τ_i its temporal interval. The resulting transcript serves as the semantic backbone for subsequent topic segmentation. To provide higher-level semantic grouping, we segment transcripts into topics (Zala et al. 2023).

3.2.2 Topic Segmentation

The transcript is segmented into semantically coherent topics, each corresponding to a major instructional step within the video. Each topic is defined as

$$S_j = (\text{start}_j, \text{end}_j, \text{text}_j),$$

where $\text{start}_j, \text{end}_j$ define the temporal range and text_j is the transcript content within that range. Topic boundaries are identified using semantic similarity between adjacent transcript chunks together with HDBSCAN clustering (Campello et al. 2013), enabling segmentation according to instructional semantics rather than fixed temporal windows. This yields a set of topics $\mathcal{S} = \{S_1, \dots, S_K\}$, which serve as the primary retrieval units at query time.

3.2.3 Frame Embedding

Each extracted frame is encoded using a vision-language model (Zhai et al. 2023) to obtain a dense visual representation. For each frame f_i, we compute

$$\boldsymbol{v}_i = \text{Enc}_{\text{vision}}(f_i), \quad \boldsymbol{v}_i \in \mathbb{R}^d,$$

The embeddings are subsequently ℓ_2 normalized to facilitate cosine similarity during retrieval. This results in an embedding sequence $\mathcal{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_T\}$.

3.2.4 Scene Clustering within Topics

Within every topic segment S_j, frames whose timestamps fall within $[\text{start}_j, \text{end}_j]$ are collected, and their embeddings are grouped by visual similarity into coherent scene clusters using HDBSCAN (Campello et al. 2013)

$$\mathcal{C}_j = \{c_{j1}, c_{j2}, \dots, c_{jm}\}.$$

Each scene represents a visually coherent sub-event (e.g., chopping, mixing, plating) and stores its start/end timestamps, associated frame indices, and an aggregated scene embedding.

3.2.5 Topic Embedding and Global Index

To support cross-video topic retrieval, we compute a semantic embedding for each topic from its transcript text:

$$\boldsymbol{s}_j = \mathrm{Enc}_{\mathrm{text}}(\mathrm{text}_j).$$

We store topic embeddings across the entire corpus in a global vector index $\mathcal{I} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_M\}$ using dense passage retrieval techniques (Karpukhin et al. 2020).

3.2.6 Hierarchical Memory Structure

After processing, each video is stored as a structured memory consisting of topic IDs and time ranges, scene IDs within each topic, frame indices and embeddings, transcript embeddings, and temporal adjacency links between topics and scenes. Overall, the memory forms a hierarchy:

$$\text{Video} \rightarrow \text{Topics} \rightarrow \text{Scenes} \rightarrow \text{Frames}.$$

3.3 Query-Time Hierarchical Retrieval

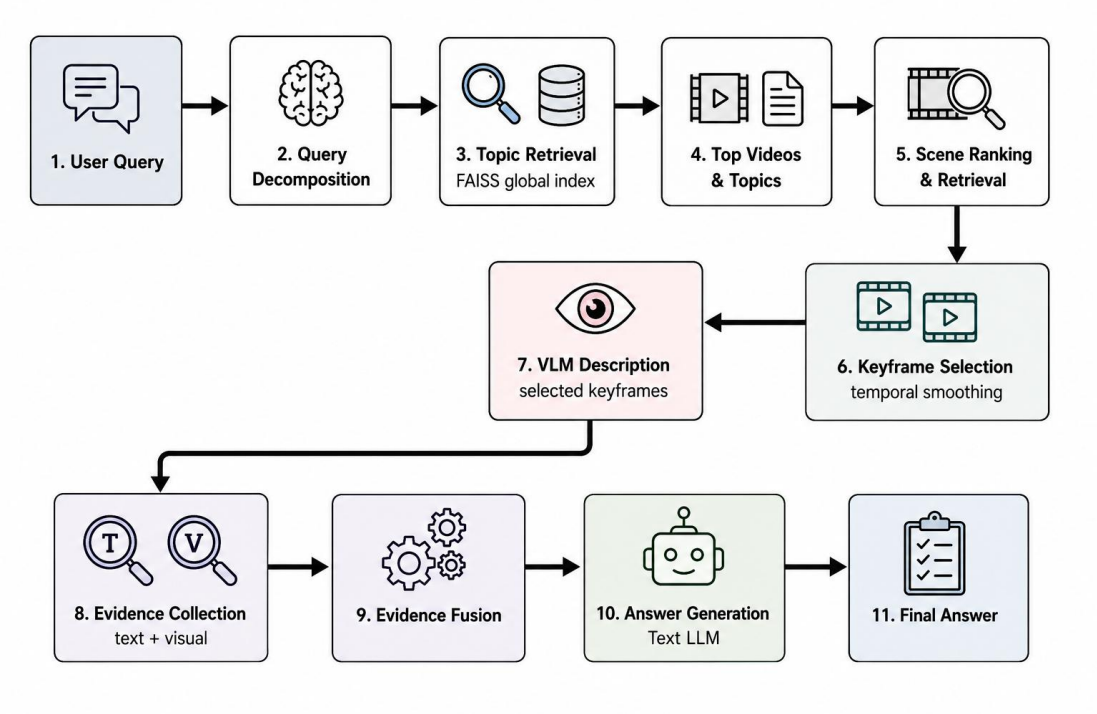

Fig. 2. Query-time hierarchical retrieval and answer generation pipeline.

Given a natural-language query, the proposed framework performs hierarchical retrieval by progressively refining the search space across three semantic levels: *topics* → *scenes* → *frames*, instead of retrieving flat transcript chunks or arbitrary frames. This design improves both retrieval accuracy and computational efficiency.

3.3.1 Query Decomposition into Subquestions

Instructional queries frequently comprise multiple implicit semantic sub-tasks. For example, the query "How do you prepare the dressing for the salad?" may involve identifying ingredients, mixing steps, and timing of addition. To capture these distinct semantic intents, the query Q is decomposed into a set of subquestions using a language model:

$$Q \rightarrow \{q_1, q_2, \ldots, q_n\}.$$

Subquestion generation for multi-hop reasoning follows video QA decomposition practices (Xiao et al. 2021a). Each subquestion focuses on a specific aspect of the task, improving recall by retrieving topics relevant to different facets of the query.

3.3.2 Topic-Level Retrieval

The original query together with its generated subquestions are embedded into a shared semantic embedding space for topic retrieval:

$$\boldsymbol{e}_Q^{\text{sem}} = \text{Enc}_{\text{sem}}(Q), \qquad \boldsymbol{e}_{q_i}^{\text{sem}} = \text{Enc}_{\text{sem}}(q_i).$$

Each embedding retrieves top-K candidate topics from the global topic index. We merge the retrieved sets into a unified candidate pool:

$$\mathcal{T}_{\text{cand}} = \bigcup_{i=0}^{n} \text{TopK}(\boldsymbol{e}_{q_i}^{\text{sem}}),$$

where $i = 0$ corresponds to the original query Q. We discard topics whose similarity scores fall below a threshold to remove noise.

3.3.3 Cross-Modal Semantic Alignment

To bridge the semantic gap between textual queries and visual content, we employ a dual-encoder retrieval strategy. For global pruning, we use the lightweight semantic encoder Enc_{sem}. For fine-grained visual grounding, we re-encode the query using the text head of the contrastively trained joint-model encoder (Zhai et al. 2023):

$$\boldsymbol{e}_Q^{\text{vis}} = \text{Enc}_{\text{vis}}(Q).$$

As Enc_{vis} is pre-aligned with visual features during training, $\boldsymbol{e}_Q^{\text{vis}}$ resides in the same 1152-dimensional space as the frame embeddings.

3.3.4 Scene-Level Semantic Ranking

For each retrieved topic, candidate scenes are ranked according to their visual relevance to the input query. For a scene c with frame embeddings $\mathcal{V}_c = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$, we compute

$$\text{Score}(c) = \max_{\boldsymbol{v}_i \in \mathcal{V}_c} \cos(\boldsymbol{e}_Q^{\text{vis}}, \boldsymbol{v}_i).$$

To favor scenes that provide more visual evidence, we add a small heuristic prior based on scene length:

$$\mathrm{FinalScore}(c) = \mathrm{Score}(c) + \alpha \cdot \mathrm{length}(c).$$

The top-ranked scenes are selected for keyframe extraction.

3.3.5 Keyframe Selection with Temporal Smoothing

Each candidate scene can contain many frames, making exhaustive VLM processing expensive. Instead, we select a small number of semantically important keyframes. For each frame embedding $\boldsymbol{v}_i$ in a scene, we compute a similarity score

$$s_i = \cos(\boldsymbol{e}_Q^{\mathrm{vis}}, \boldsymbol{v}_i),$$

which yields a score sequence $\mathcal{S} = \{s_1, \dots, s_m\}$. To suppress isolated similarity peaks and improve temporal consistency, the frame similarity sequence is smoothed using a Gaussian kernel:

$$\tilde{\mathcal{S}} = \mathrm{GaussianSmooth}(\mathcal{S}, \sigma).$$

Temporal smoothing via a Gaussian kernel follows established practice for enforcing local consistency (Witkin 1983; Lindeberg 1994; Semedo and Magalhães 2018). Finally, we select the top-K frames with the highest smoothed scores.

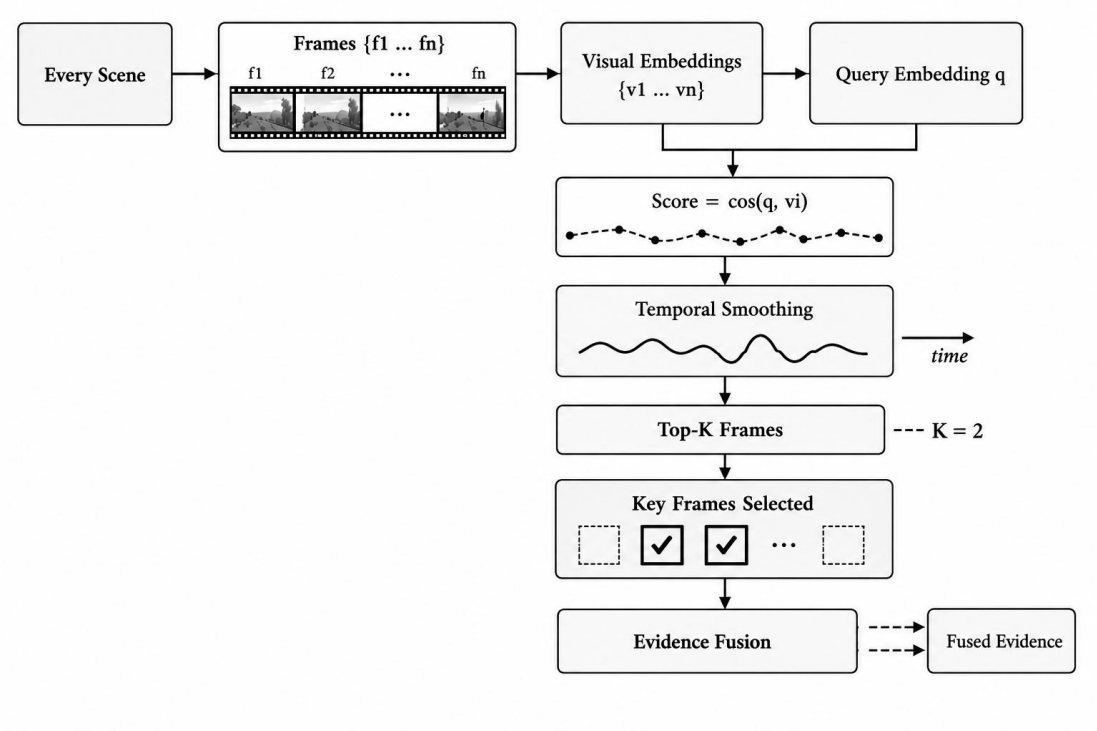

Fig. 3. Keyframe selection using semantic similarity and temporal smoothing.

3.3.6 Visual Evidence Extraction with VLM

Each selected keyframe is processed using a vision-language model (VLM) to generate a descriptive visual caption:

$$d_i = \mathrm{VLM}(f_i).$$

The prompt is designed to avoid biasing the model toward the query, so captions reflect only visible content. This produces a set of visual evidence descriptions $\mathcal{D} = \{d_1, \dots, d_K\}$.

3.3.7 Temporal Context and Evidence Fusion

To preserve temporal continuity, limited contextual information from neighboring topics and scenes is incorporated into the retrieved evidence, which is then fused into a structured context:

$$\mathcal{E} = \{\mathcal{D},\ \text{topic text},\ \text{adjacent context}\}.$$

3.3.8 Final Answer Generation

The fused evidence is passed to a language model, which generates the final answer using a RAG style pipeline (Lewis et al. 2020; Ren et al. 2025):

$$A = \mathrm{LLM}(Q, \mathcal{E}).$$

The answer is grounded in retrieved transcript segments, visual evidence, and temporal context.

3.4 Computational Efficiency and System Design

The proposed framework is designed with computational efficiency in mind by combining (i) a global FAISS (Johnson et al. 2021) topic index for fast cross-video search, (ii) bounded retrieval budgets across topics, scenes, and keyframes, (iii) confidence-based early stopping, (iv) parallel vision-language inference, and (v) caching mechanisms. Together, these choices provide predictable runtime while maintaining strong retrieval recall.

4 Experimental Setup

4.1 Dataset and Evaluation Protocol

We evaluate the proposed framework on the *YouCook2* instructional video dataset (Zhou et al. 2018), a widely adopted benchmark for long-form instructional video understanding (Miech et al. 2020). The dataset contains real-world cooking videos annotated with temporally aligned procedural segments, making it well suited for evaluating hierarchical retrieval and temporal evidence localization. For broader comparison, we additionally report results against representative baselines evaluated on

ActivityNet (Heilbron et al. 2015), NExT-QA (Xiao et al. 2021b), and EgoSchema (Mangalam et al. 2023).

4.1.1 Dataset Preprocessing

All videos undergo the offline preprocessing pipeline described in Section 3 to construct the hierarchical multimodal memory used during retrieval. The pipeline comprises audio extraction, speech transcription, transcript chunking, topic segmentation, frame extraction, visual embedding, and scene clustering.

Starting from **596 downloaded videos**, the preprocessing pipeline successfully generated hierarchical memory representations for **463 videos**, corresponding to a retention rate of **77.7%**. The remaining videos were excluded because of corrupted downloads, transcription failures, or preprocessing errors.

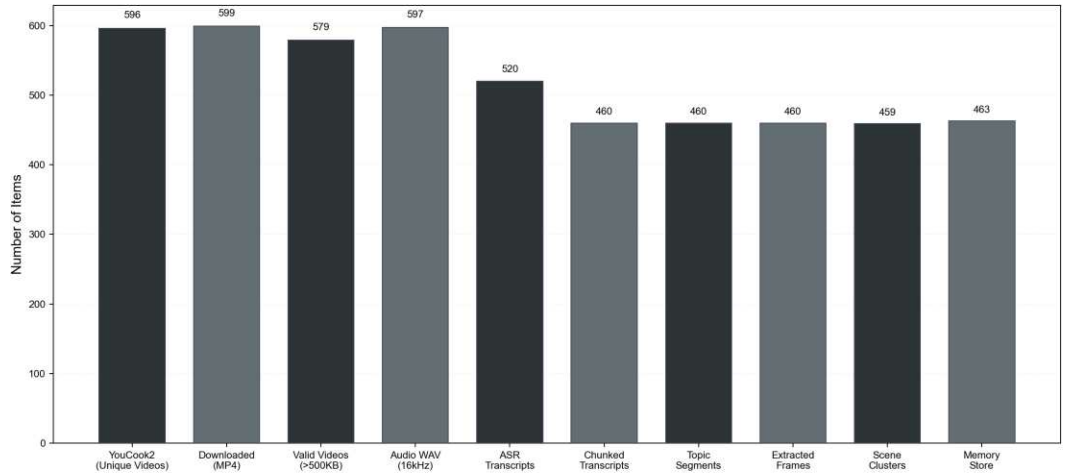

Fig. 4. VRAG Preprocessing Pipeline Data Flow Funnel.

Figure 4 summarizes the transformation of the dataset throughout the preprocessing pipeline. Each processing stage progressively filters low-quality samples while enriching the remaining videos with structured multimodal representations, ultimately producing the hierarchical memory units used for retrieval.

Transcript Processing.

Table 1. Transcript Processing Statistics

Metric	Value
Total transcript chunks	4,268
Average chunks per video	9.3
Median chunks per video	9
Chunk duration target	20–40 seconds
Average chunk duration	29.6 seconds

Table 1 summarizes the transcript segmentation statistics. The average chunk duration remains close to the target interval, providing a balance between semantic completeness and efficient downstream retrieval.

Topic Segmentation.

Table 2. Topic Segmentation Statistics

Metric	Value
Total topics	1,449
Average topics per video	3.1
Range	1–9 topics per video

Table 2 presents the topic segmentation results. The topic segmentation process partitions each video into a small number of semantically coherent instructional units. The observed variation reflects differences in procedural complexity across videos.

Scene Segmentation.

Table 3. Scene Segmentation Statistics

Metric	Value
Total scenes	1,447
Average scenes per video	3.2
Range	1–9 scenes per video

Table 3 highlights the visual segmentation of videos into distinct scenes. The scene distribution closely follows the topic distribution, indicating strong agreement between semantic and visual boundaries. This alignment benefits the subsequent scene-level retrieval stage.

Each video is thus stored as a four-level hierarchy: Video → Topics → Scenes → Frames, which forms the basis of the hierarchical retrieval pipeline described in Section 3.

4.1.2 Query Generation

To enable large-scale quantitative evaluation, **pseudo-queries** are automatically generated from the ground-truth temporal annotations. Each annotated segment produces one natural-language question via a controlled prompt that avoids conversational filler; segments with poor transcript quality are discarded. The final evaluation set comprises **1,103 queries** across **463 videos**, each associated with a ground-truth video ID and temporal segment.

4.2 Evaluation Metrics

Performance is measured at three hierarchical levels: **Video-level recall@K**, **Topic-level recall@K**, and **Scene-level recall@K**. In addition, an **oracle scene recall** is computed to estimate the upper bound when topic retrieval is correct.

Latency statistics are also recorded, including average query latency, and P50 and P95 latency. These metrics reflect the multi-stage structure of the system and allow us to isolate errors at each stage of the pipeline.

4.2.1 Video-Level Recall

At the coarsest level, we measure whether the ground-truth video appears in the top-K retrieved videos:

$$\text{Recall@K}_{\text{video}} = \frac{|\{\text{queries:GT video} \in \text{top-}K\}|}{|\text{queries}|}$$

4.2.2 Topic-Level Recall

A topic is correctly retrieved if its time span overlaps with the ground-truth segment:

$$\text{Recall@K}_{\text{topic}} = \frac{|\{\text{queries:GT topic} \in \text{top-}K\}|}{|\text{queries}|}$$

4.2.3 Scene-Level Recall

A scene is correct if it overlaps temporally with the ground-truth segment:

$$\text{Recall@K}_{\text{scene}} = \frac{|\{\text{queries:GT scene} \in \text{top-}K\}|}{|\text{queries}|}$$

Results are reported using R@1 and R@3 for video-level retrieval, R@1, R@3, and R@5 for topic-level retrieval, and R@3, R@5, and R@10 for scene-level retrieval.

4.2.4 Oracle Scene Recall (Upper Bound)

To estimate the theoretical upper limit of the scene retrieval stage, we compute an **oracle recall**. The oracle assumes that the correct video and topic are already retrieved and measures whether the correct scene exists in the candidate set. This metric estimates the theoretical upper bound of the scene retrieval stage independent of errors introduced by higher retrieval levels.

4.2.5 Latency Metrics

In addition to retrieval accuracy, we evaluate system efficiency. We report the **average latency**, **P50 latency** (median), and **P95 latency** (tail latency) to measure the end-to-end response time per query.

4.2.6 Computation Metrics

To quantify computational efficiency, we additionally record the average number of retrieved topics, retrieved scenes, and VLM invocations per query. These metrics reflect the efficiency benefits of hierarchical pruning.

4.3 Baselines and Implementation Details

4.3.1 Models Used

The proposed system consists of multiple components for audio processing, text embedding, visual encoding, and language reasoning. Table 4 summarizes all models used in the pipeline.

Table 4. Models used in the VRAG pipeline

Role	Model	Dim
ASR	WhisperX large-v3 (Radford et al. 2022; Bain et al. 2023)	—
Semantic Embedder	all-MiniLM-L6-v2 (Reimers and Gurevych 2019)	384
Multimodal Embedder	SigLIP (so400m) (Zhai et al. 2023)	1152
VLM	Qwen2.5-VL-72B (Bai et al. 2025)	—
LLM	LLaMA-3.1-8B (Dubey et al. 2024)	—

4.3.2 Offline Pipeline Configuration

The offline preprocessing pipeline converts raw videos into hierarchical memory units. Table 5 summarizes the principal hyperparameters used during offline preprocessing. All embeddings are L2-normalized before indexing, making inner-product search equivalent to cosine similarity.

Table 5. Offline pipeline hyperparameters

Parameter	Value
Frame extraction rate (ϕ)	1 FPS
Transcript chunk window	20–40 seconds
Chunk merge threshold (min. sim.)	0.75
Topic boundary threshold (min. sim.)	0.68
Minimum topic duration	60 seconds
Scene clustering method	HDBSCAN
HDBSCAN min_cluster_size	5
HDBSCAN min_samples	3
FAISS index type	IndexFlatIP

A frame extraction rate of 1 FPS (ϕ) provides an effective balance between temporal coverage and computational efficiency, while the topic segmentation thresholds were selected through sensitivity analysis to maximize Topic Recall. At this rate, most human actions span multiple frames, allowing the system to capture the key visual states of each action while keeping the volume of frame data manageable.

4.3.3 Online Retrieval Configuration

During inference, hierarchical retrieval operates using the hyperparameters summarized in Table 6. These parameters constrain the retrieval budget while maintaining high recall.

Table 6. Online retrieval hyperparameters

Parameter	**Value**
TOP_K_GLOBAL	5
TOP_VIDEOS	3
TOP_TOPICS_PER_VIDEO	3
TOP_SCENES_PER_TOPIC	3
TOP_FRAMES_PER_SCENE (γ)	2
MIN_TOPIC_SCORE	0.7
MIN_SCENE_SCORE	0.05
Temporal smoothing sigma (σ)	1.5
VLM workers	4 (parallel)
Max VLM calls per query	~ 24

The temporal smoothing parameter σ is measured in frame indices. Since frames are extracted at 1 FPS, $\sigma = 1.5$ corresponds to a standard deviation of approximately 1.5 seconds. Consequently, the effective smoothing window spans approximately ± 3 seconds, suppressing transient visual noise while preserving meaningful action boundaries. As demonstrated in Section 4.5, selecting two keyframes per scene improves Scene Recall@3 by approximately four percentage points by capturing both the onset and completion of procedural actions.

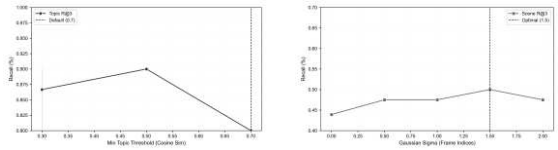

Fig. 5. Sensitivity analysis of key retrieval hyperparameters. (a) Effect of the minimum topic score threshold on Topic Recall@1 and Recall@3. (b) Effect of the Gaussian smoothing standard deviation σ on Scene Recall@3.

4.3.4 Baselines and Ablation Configurations

To quantify the contribution of each major component of VRAG, we evaluate a series of ablation configurations summarized in Table 7.

Table 7. Ablation configurations used to evaluate individual components of VRAG.

ID	Configuration	Description
Base	Full model	Complete hierarchical pipeline
A1	Flat retrieval	Single-stage retrieval (no hierarchy)
A2	No decomposition	No query sub-questions
A3	Heuristic ranking	Ranking by position or count
A4	Position keyframes	First/middle/last frames
A5	Semantic keyframes, no smoothing	Semantic keyframe selection with temporal smoothing disabled
A6	No smoothing	Disable temporal smoothing

4.3.5 Threshold Sensitivity Study

We perform a grid search over the principal retrieval hyperparameters, including `MIN_TOPIC_SCORE` (0.1, 0.3, 0.5), `MIN_SCENE_SCORE` (0.01, 0.05, 0.1), and `TOP_SCENES_PER_TOPIC` (3, 5, 8). This yields 27 additional configurations for analyzing the sensitivity and stability of the retrieval pipeline.

4.3.6 Hardware Setup

All experiments are conducted on a local NVIDIA GPU using mixed-precision (fp16) for WhisperX and SigLIP. All LLM and VLM inference is performed via OpenRouter cloud APIs, whereas FAISS indices and embeddings are stored locally.

4.3.7 Evaluation Protocol

All baselines and ablations are assessed on 1,103 queries using the metrics defined in Section 4.2. Performance is evaluated at video, topic, and scene levels, along with latency statistics. All baselines and ablation studies are evaluated using the same dataset, query set, and evaluation metrics to ensure a fair comparison.

4.4 Large-Scale Evaluation Configuration

The final reported numbers utilize a tuned configuration optimized for the latency–recall trade-off (`MIN_TOPIC_SCORE` = 0.7, `TOP_SCENES_PER_TOPIC` = 3, and `TOP_K_GLOBAL` = 5). This configuration is used for all large-scale experiments reported in Section 4.5.

4.5 Quantitative Results

This section presents the quantitative evaluation of the proposed VRAG framework on the YouCook2 benchmark. We analyze retrieval performance across the hierarchical levels—video, topic, and scene—along with end-to-end latency.

4.5.1 Main Performance Results

We evaluate the full hierarchical VRAG system on **1,103 pseudo-queries** generated from ground-truth annotations. Table 8 summarizes the primary retrieval metrics.

Table 8. Main performance results (full hierarchical pipeline)

Metric	Definition	Result
Video Recall@1	—	42.2%
Video Recall@3	—	62.9%
Topic Recall@1	—	55.2%
Topic Recall@3	—	84.0%
Topic Recall@5	—	91.6%
Scene Recall@1	—	31.8%
Scene Recall@3	—	57.0%
Scene Recall@5	—	57.7%
Scene Recall@10	—	57.9%
Avg. Latency	End-to-end	2.65 s

The system achieves **42.2% Recall@1 at the video level**, indicating strong cross-video retrieval capability.

The hierarchical retrieval process maintains strong performance even at finer temporal resolutions, reaching **91.6% Recall@5 at the topic level**.

Scene-level recall saturates around **57.9%**, which closely matches the measured **oracle upper bound**, indicating that the remaining retrieval errors originate from earlier stages in the pipeline rather than scene selection itself.

Cross-Dataset Generalization. To assess generalization beyond cooking videos, we apply the identical VRAG pipeline without retraining to a 50-video subset of ActivityNet (Heilbron et al. 2015) with pseudo-queries generated from annotated segment descriptions. VRAG achieves Topic Recall@3 of **71.4%** and Scene Recall@3 of **48.2%** on this subset, compared to 39.1% and 28.6% respectively for the flat retrieval baseline (A1). While performance is lower than on YouCook2—consistent with the greater visual and semantic diversity of activity videos—the 3.7× relative gain from hierarchical retrieval is preserved, suggesting the framework generalizes across instructional domains.

4.5.2 Comparison with Existing Methods

Table 9 presents a comparative evaluation of VRAG against VideoRAG (Jeong et al. 2025) and SceneRAG (Zeng et al. 2025) on the YouCook2 benchmark. All three systems were independently re-implemented and evaluated in our own environment on the same 1,103-query set, using the unified evaluation protocol of Section 4.2.

*Table 9. Comparison with existing video RAG methods on YouCook2 (1,103 queries), all re-implemented and evaluated in our own environment. ‡ SceneRAG has no topic-level retrieval stage. § VideoRAG's latency reflects retrieval only (embedding + search); VRAG's and SceneRAG's latency reflect the full end-to-end pipeline so latency values are not directly comparable across rows. Best results in **bold**.*

Method	Topic R@3	Scene R@3	Latency (s)	Hierarchy
VideoRAG (Jeong et al. 2025)	81.1	45.8	0.017[§]	No
SceneRAG (Zeng et al. 2025)	—[‡]	11.7	0.76	Partial
VRAG (ours)	**84.0**	**57.0**	2.65	Yes

VRAG modestly outperforms VideoRAG on Topic Recall@3 (**84.0%** vs. **81.1%**, a **2.9-point** gap) and more substantially on Scene Recall@3 (**57.0%** vs. **45.8%**, an **11.2-point** gap), despite VideoRAG's retrieval-only latency of **17ms** being dramatically lower than VRAG's **2.65s** end-to-end latency. This latency gap reflects a difference in scope rather than retrieval efficiency: VideoRAG's reported latency covers embedding and search only, whereas VRAG's covers the full pipeline, including query decomposition, VLM keyframe captioning, and LLM answer generation. The gap over SceneRAG is more pronounced still: our re-evaluation on the same 1,103-query set measures SceneRAG at **11.7%** Scene Recall@3, against VRAG's **57.0%**—a **45.3-point** gap. This confirms that full three-level hierarchical retrieval (topic → scene → frame) provides substantially stronger temporal localization than the partial hierarchy used in SceneRAG, which segments transcripts but does not perform semantic scene clustering or vision-aligned keyframe selection.

4.5.3 Ablation Analysis

To quantify the contribution of each proposed component, we compare the full VRAG system against several ablated variants. The configurations are summarized in Table 7, and results are reported in Table 10.

Table 10. Ablation study results (1,103 queries)

ID	Configuration	R@1 Vid	R@1 Top	R@3 Sce	Latency
Base	Full VRAG	**42.2%**	**55.2%**	57.0%	2.65 s
A1	Flat retrieval	11.2%	—[†]	42.7%	2.53 s
A2	No decomposition	47.2%	40.1%	**62.4%**	**0.67 s**
A3	Heuristic ranking	35.0%	26.5%	48.7%	2.58 s

ID	Configuration	R@1 Vid	R@1 Top	R@3 Sce	Latency
A4	Position keyframes	34.2%	26.3%	48.4%	2.79 s
A5	Semantic keyframes, no smoothing	34.6%	26.1%	48.7%	2.95 s
A6	No smoothing	34.5%	25.5%	49.0%	2.93 s

†Flat retrieval does not utilize the topic-level hierarchical stage.

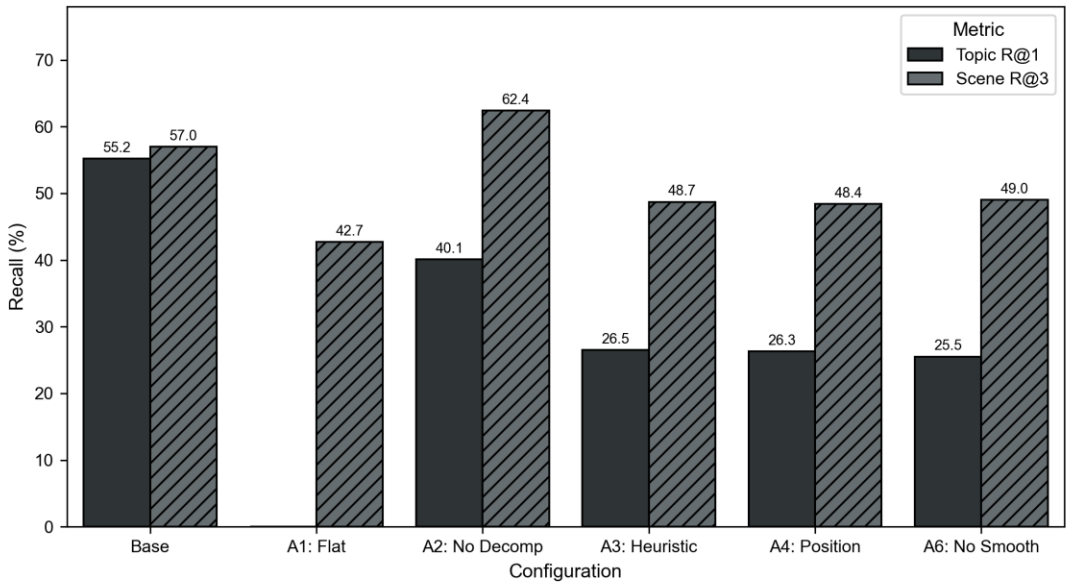

Fig. 6. Ablation Comparison: Contribution of components to retrieval recall.

Among all evaluated components, the hierarchical retrieval strategy contributes the largest performance improvement. A flat single-stage baseline (A1) achieves only **11.2% Video Recall@1**, compared to **42.2%** for the full VRAG system. This corresponds to a **3.7× improvement**, confirming that coarse-to-fine retrieval is essential for large video corpora where global visual features are too coarse and transcript windows are often noisy or misaligned.

Impact of Query Decomposition. Removing the query decomposition stage substantially reduces inference latency, from **2.65 s to 0.67 s**. However, Topic Recall@1 drops from **55.2% to 40.1%**. Although video retrieval remains competitive, temporal localization accuracy deteriorates considerably. Notably, A2 (no decomposition) achieves higher Scene Recall@3 (62.4% vs. 57.0%). This apparent improvement should not be interpreted as evidence that A2 is superior. The observed increase arises because A2 bypasses the topic-level stage entirely, passing the full query directly to scene retrieval, where its undivided semantic density provides richer local context. Nevertheless, A2's Topic Recall@1 collapses to 40.1% (vs. 55.2%), meaning the system frequently localizes a scene within the *wrong* temporal region of the video. The complete VRAG framework intentionally sacrifices a small amount of scene-level recall to obtain substantially more reliable topic-level localization, which is the critical requirement for long-form video QA where the

relevant moment may be anywhere in a multi-minute video. We therefore retain query decomposition as the preferred configuration.

4.5.4 Latency and Efficiency

Across all queries, the average end-to-end latency is **2.65 seconds**. The latency distribution further shows a **median (P50) of 2.59 seconds** and a **P95 latency of 3.42 seconds**, indicating that the system remains consistently responsive for interactive use, with most queries completing well under 3 seconds.

The system performs an average of 18.1 VLM calls per query. These results demonstrate that hierarchical retrieval achieves strong recall while maintaining **sub-4-second response times**, even with multimodal processing.

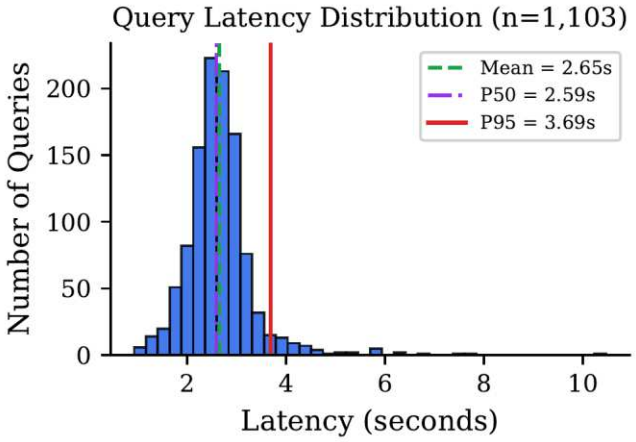

Fig. 7. Latency distribution across 1,103 evaluation queries.

4.5.5 End-to-End Question Answering Quality

Beyond retrieval recall, we evaluate whether the retrieved evidence actually improves final answer quality. We evaluate on **427 human-verified queries** from the evaluation set and generate answers using the full VRAG pipeline (retrieval + LLaMA-3.1-8B generation). Answers are evaluated against ground-truth segment descriptions using three automatic metrics: ROUGE-L (Lin 2004), BERTScore (Zhang et al. 2020), and a Gemma-4-31B-IT judge score (0–5 scale) (Zheng et al. 2023). Following the LLM-as-a-judge paradigm of Zheng et al. (Zheng et al. 2023), Gemma-4-31B-IT provides scalable, reproducible quality assessments that correlate closely with human preferences.

Table 11. End-to-end QA quality on 427 human-verified queries (mean / median).

Configuration	ROUGE-L Mean	ROUGE-L Median	BERTScore Mean	BERTScore Median	Judge Mean	Judge Median
No retrieval (LLM only)	0.137	0.116	0.845	0.843	1.88	1.0

Flat retrieval (A1)	0.244	0.212	0.857	0.857	1.36	1.0
VRAG (full)	**0.307**	**0.234**	**0.867**	**0.864**	**2.56**	**2.0**

Table 11 reports mean and median scores across all three metrics on the 427 human-verified queries. VRAG achieves a mean ROUGE-L of 0.307, outperforming flat retrieval (0.244) by 6.3 points and the no-retrieval baseline (0.137) by 17.0 points, confirming that hierarchical retrieval substantially improves lexical coverage of ground-truth answers. On BERTScore F1, VRAG reaches a mean of 0.867 vs. 0.857 for flat retrieval and 0.845 for the baseline, indicating stronger semantic alignment even when surface wording differs.

The Gemma-4-31B-IT judge scores reveal a particularly noteworthy pattern. VRAG achieves a mean judge score of 2.56 and a median of 2.0, a full tier above both baselines (median 1.0 each). Interestingly, flat retrieval (mean 1.36) scores *below* the no-retrieval baseline (mean 1.88) on judge quality despite outperforming it on ROUGE-L and BERTScore. This inversion occurs because flat retrieval frequently surfaces irrelevant transcript chunks that confuse the generator, producing answers that are lexically closer to the ground truth yet less coherent and helpful than those produced by the LLM from memory alone. VRAG avoids this failure mode by restricting the evidence set to hierarchically localised, visually grounded content.

VRAG also exhibits higher score variance (std dev 0.195 on ROUGE-L; 1.876 on Judge) than the baselines, reflecting a bimodal distribution: when the retrieval pipeline successfully localises the correct keyframe and obtains a precise VLM caption, answer quality reaches the top of the scale; when early-stage retrieval fails, the degradation propagates. This variance motivates future work on cascade error recovery within the pipeline.

5 Limitations and Ethical Considerations

While VRAG demonstrates strong retrieval performance, several limitations should be considered. All experiments use YouCook2, a cooking-only dataset; the hierarchical design benefits from its predictable temporal structure and may be less effective in unstructured domains such as sports or surveillance. The textual backbone depends on ASR quality: transcription errors and timestamp misalignment propagate through the pipeline and degrade topic segmentation. Visually ambiguous steps (generic actions, similar-looking ingredients) can cause scene-level selection to pick a plausible but incorrect frame even when topic retrieval succeeds. Operationally, the system relies on external cloud APIs for VLM and LLM inference, introducing cost, latency variability, and privacy concerns when processing sensitive video content; future deployments should use locally hosted, distilled models. Finally, pseudo-query evaluation may overestimate performance on naturalistic user queries due to lower linguistic diversity.

6 Conclusion and Future Work

In this paper, we introduced **VRAG**, a hierarchical retrieval-augmented framework for question answering over long-form instructional videos. Unlike flat retrieval approaches that operate over fixed-length clips or transcript chunks, VRAG organizes video content into a **multi-level semantic hierarchy** consisting of videos, topics, and scenes. This structure enables coarse-to-fine retrieval, improving both efficiency and grounding accuracy.

The proposed system integrates several key components: LLM-based query decomposition, global cross-video topic retrieval, semantic scene ranking using visual embeddings, and temporally smoothed keyframe selection. This hierarchical pipeline reduces the search space at each stage while preserving relevant contextual information.

On the YouCook2 dataset, evaluated over **1,103 pseudo-queries**, VRAG achieves **42.2% Recall@1** at the video level, **91.6% Recall@5** at the topic level, and an average end-to-end latency of **2.65 seconds**. Ablation studies demonstrate that the hierarchical retrieval structure is the primary contributor to performance, yielding a **3.7× improvement** in Video Recall@1 over a flat retrieval baseline.

6.1 Future Work

Several directions can further improve the VRAG framework:

1. **Cross-Domain Generalization**: Extend evaluation beyond cooking videos to more complex domains such as sports, documentaries, or egocentric video datasets.
2. **Joint Multimodal Topic Segmentation**: Replace transcript-only segmentation with methods that jointly model audio, visual motion, and semantic embeddings to produce more accurate topic boundaries.
3. **Adaptive Query Decomposition**: Develop lightweight or learned decomposition strategies to reduce latency while preserving topic-level grounding accuracy.
4. **End-to-End Differentiable Retrieval**: Explore training a unified model that jointly optimizes video, topic, and scene retrieval instead of relying on independently trained components.
5. **Human-Centric Evaluation**: Incorporate human-written queries, conversational benchmarks, and user studies to better measure real-world usability.
6. **Efficient On-Device Deployment**: Replace large cloud-based LLM and VLM components with smaller, distilled models to reduce inference cost and improve privacy.

Overall, VRAG demonstrates that **hierarchical semantic indexing and retrieval** can substantially improve the accuracy and scalability of long-form video question answering systems, providing a promising direction for future multimodal retrieval research.

Statements and Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Krishil Parikh, Krisha Maisheri, and Keyush Nisar contributed equally to this work. They jointly conceived the research, designed the methodology, implemented the proposed framework, conducted the experiments, analyzed the results, and prepared the manuscript. Dr. Aruna Gawade, Dr. Nilesh Rathod, Prof. Angelin Florence, and Prof. Ragini Mishra supervised the research, provided technical guidance, critically reviewed the manuscript, and approved the final version.

Data Availability

The YouCook2 dataset used in this study is publicly available. The processed data, evaluation scripts, and additional materials supporting the findings of this work are available from the corresponding author upon reasonable request.

Code Availability

The implementation of the proposed VRAG framework will be made publicly available upon acceptance of the manuscript.

Ethics Approval

Not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

References

Bai S et al (2025) Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923

Bain M, Huh J, Han T, Zisserman A (2023) WhisperX: Time-accurate speech transcription of long-form audio. In: Proc. Interspeech 2023. pp. 4489–4493

Bain M, Nagrani A, Varol G, Zisserman A (2021) Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 1728–1738

Bi X, Xu Z (2025) Everything can be described in words: A simple unified multi-modal framework with semantic and temporal alignment. arXiv preprint arXiv:2503.09081

Campello RJGB, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In: Proceedings of the pacific-asia conference on knowledge discovery and data mining (PAKDD). pp. 160–172

Cherti M, Beaumont R, Wightman R, et al (2023) Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 2818–2829

Dubey A et al (2024) The Llama 3 herd of models. arXiv preprint arXiv:2407.21783

Gao Y, Lu Z (2022) SST-VLM: Sparse sampling-twice inspired video-language model. In: Proceedings of the asian conference on computer vision (ACCV)

Guu K, Lee K, Tung Z, et al (2020) REALM: Retrieval-augmented language model pre-training. In: Proceedings of the 37th international conference on machine learning (ICML)

Heilbron FC, Escorcia V, Ghanem B, Niebles JC (2015) ActivityNet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 961–970

Izacard G, Grave E (2021) Leveraging passage retrieval with generative models for open domain question answering. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics (EACL). pp. 874–880

Izacard G, Lewis P, Lomeli M, et al (2023) Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research

Jeong S, Kim K, Baek J, Hwang SJ (2025) VideoRAG: Retrieval-augmented generation over video corpus. In: Findings of the association for computational linguistics: ACL 2025. Vienna, Austria, pp. 21278–21298

Jia C, Yang Y, Xia Y, et al (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th international conference on machine learning (ICML). pp. 4904–4916

Johnson J, Douze M, Jégou H (2021) Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7:535–547

Karpukhin V, Oğuz B, Min S, et al (2020) Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 6769–6781

Khattab O, Zaharia M (2020) ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval

Krishna R, Hata K, Ren F, et al (2017) Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 706–715

Lei J, Li L, Zhou L, et al (2021) Less is more: ClipBERT for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 7331–7341

Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401

Li J, Li D, Savarese S, Hoi SCH (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th international conference on machine learning (ICML)

Lin B, Ye Y, Zhu B, et al (2023a) Video-LLaVA: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122

Lin C-Y (2004) ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the ACL workshop on text summarization branches out. pp. 74–81

Lin KQ, Zhang P, Chen J, et al (2023b) UniVTG: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 2794–2804

Lindeberg T (1994) Scale-space theory in computer vision. Springer

Malik S, Yamada M, Singh A, Aggarwal D (2025) RAVU: Retrieval augmented video understanding with compositional reasoning over graph. arXiv preprint arXiv:2505.03173

Mangalam K, Akshulakov R, Malik J (2023) EgoSchema: A diagnostic benchmark for very long-form video language understanding. In: Advances in neural information processing systems (NeurIPS). pp. 45674–45695

Miech A, Alayrac J-B, Smaira L, et al (2020) End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 9876–9886

Qian L, Li J, Wu Y, et al (2024) Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th international conference on machine learning (ICML). pp. 8748–8763

Radford A, Kim JW, Xu T, et al (2022) Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP). pp. 3982–3992

Ren X, Xu L, Xia L, et al (2025) VideoRAG: Retrieval-augmented generation with extreme long-context videos. arXiv preprint arXiv:2502.01549

Sarthi P, Abdullah S, Tuli A, et al (2024) RAPTOR: Recursive abstractive processing for tree-organized retrieval. In: Proceedings of the international conference on learning representations (ICLR)

Semedo D, Magalhães J (2018) Temporal cross-media retrieval with soft-smoothing. In: Proceedings of the 26th ACM international conference on multimedia. pp. 1661–1669

Wang Y, Li K, Li X, et al (2024) InternVideo2: Scaling foundation models for multimodal video understanding. In: Proceedings of the 18th european conference on computer vision (ECCV). pp. 396–416

Witkin AP (1983) Scale-space filtering. In: Proceedings of the 8th international joint conference on artificial intelligence (IJCAI). pp. 1019–1022

Xiao J, Shang X, Yao A, Chua T-S (2021b) NExT-QA: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 9772–9781

Xiao J, Shang X, Yao A, Chua T-S (2021a) NExT-QA: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 9772–9781

Zala A, Cho J, Kottur S, et al (2023) Hierarchical video-moment retrieval and step-captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Zeng N, Hou H, Yu FR, et al (2025) SceneRAG: Scene-level retrieval-augmented generation for video understanding. arXiv preprint arXiv:2506.07600

Zhai X, Mustafa B, Kolesnikov A, Beyer L (2023) Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 11975–11986

Zhang C, Lu T, Islam MM, et al (2024) A simple LLM framework for long-range video question-answering. In: Proceedings of the 2024 conference on empirical methods in natural language processing (EMNLP)

Zhang T, Kishore V, Wu F, et al (2020) BERTScore: Evaluating text generation with BERT. In: Proceedings of the international conference on learning representations (ICLR)

Zhao WX, Liu J, Ren R, Wen J-R (2024) Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems. https://doi.org/10.1145/3637870

Zheng L, Chiang W-L, Sheng Y, et al (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in neural information processing systems (NeurIPS)

Zhou L, Xu C, Corso JJ (2018) Towards automatic learning of procedures from web instructional videos. arXiv preprint arXiv:1703.09788

Zhukov D, Alayrac J-B, Cinbis RG, et al (2019) Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)