

CadenceRAG: Context-Aware and Dependency-Enhanced Retrieval Augmented Generation for Holistic Video Understanding

Heng Liu, Siru Jiang, Fangyun Duan, Yongzhe Lyu, Xiusong Wang, Hanlin Ge and Chao Liang*

School of Computer Science, Wuhan University

National Engineering Research Center for Multimedia Software (NERCMS)

Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan

{liuheng, sirujiang, fangyunduan, yongzhelv, 2023302111390, gehanlin, cliang}@whu.edu.cn

Abstract

This paper addresses the challenging problem of holistic video understanding, focusing on rich-text-based video retrieval and question answering. Compared to simple video retrieval tasks with concise queries, complex multi-scene queries faces two major challenges: (1) The rich semantics in text queries often depict complex multi-scene narratives with fine-grained details, making it difficult to align them with target video segments both locally and globally; (2) Relevant segments, though spanning multiple scenes, are still a minuscule fraction compared to the vast video corpus, making precise localization extremely challenging. To address these challenges, we propose a novel approach, CadenceRAG, which introduces a unified Retrieval-Augmented Generation (RAG) framework for Known-Item Search (KIS) and Question Answering (QA). By strategically decomposing rich textual queries into temporally ordered sub-queries and employing Hierarchical Sliding Window (HSW), our method precisely aligns and locates relevant video segments. The retrieved segments, along with their associated multimodal information, are integrated into the RAG framework to enhance contextual grounding and generate accurate, context-aware results. Our team WHU-NERCMS evaluated CadenceRAG on the IViSE competition, achieving a score of 6/10 in KIS and 7/10 in QA, demonstrating strong performance in both tasks.

1. Introduction

Holistic Video Understanding (HVV) [7] remains a significant challenge in computer vision. While recent advances have led to impressive progress in understanding short queries where the target content corresponds to brief segments within videos [6, 8, 15] analyzing long, complex

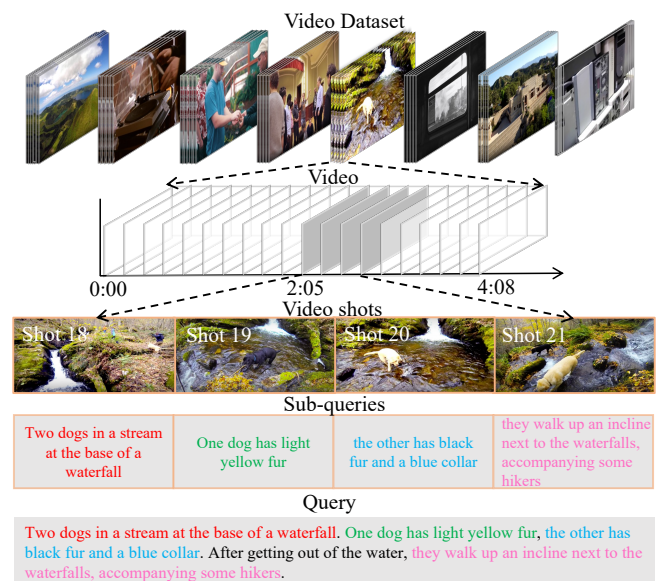


Figure 1. The challenge of “Rich text query vs Multi-Shots” and “Relevant segments vs Vast Corpus”.

queries that span multiple scenes and extended durations remains largely unsolved [46]. The core difficulty arises from the interplay between rich, temporally structured textual queries and fragmented yet semantically dense visual content. Unlike short queries, which often target single scenes or brief moments, long queries frequently describe multi-scene narratives with intricate temporal dependencies, necessitating a nuanced understanding of narrative flow, scene transitions, and contextual relationships (see Figure 1 bottom). Additionally, the vast scale of video collections further exacerbates retrieval challenges, as locating the correct video and pinpointing the relevant segments demand efficient and scalable search strategies (see Figure 1 top).

To evaluate the effectiveness of holistic video under-

*Corresponding author.

standing methods, the 1st International Workshop on Interactive Video Search and Exploration (IViSE) introduces a benchmark challenge covering two key tasks: Known-Item Search (KIS) and Question Answering (QA). KIS requires retrieving the correct video and identifying the precise temporal segment based on a detailed textual description. QA involves retrieving a specific video and generating accurate textual responses to questions about its content.

Current approaches to video understanding are largely inspired by contrastive image-text pretraining, as demonstrated by models like CLIP [33], and extend this paradigm to video-text matching [12, 27, 29, 45]. Some methods introduce video-text interaction mechanisms [17] or temporal modeling [26] to enhance representation learning, while others explore fine-grained video-text alignment [2]. These approaches have been effective for short-text video retrieval and QA tasks [24], forming the backbone of TRECVID’s Ad-hoc Video Search (AVS) and Deep Video Understanding (DVU) tasks [14]. However, when dealing with long-text video retrieval, existing methods face two fundamental limitations as Figure 1: (1) It is difficult to align the rich semantics of the query with the video’s structure at a consistent granularity, as queries often span multiple scenes with varying levels of detail and temporal dependencies. (2) The challenge of pinpointing short target segments within vast video collections. Direct video-text matching often struggles due to the overwhelming volume of irrelevant content.

To address these challenges, we propose a novel approach, CadenceRAG, which creates a unified framework for retrieval and question answering. On the one hand, to achieve granularity alignment between videos and queries, we leverage the capabilities of Large Language Model (LLM), which have demonstrated exceptional proficiency in processing sequential information and capturing intricate textual dependencies [1, 4, 41], to decompose rich queries into temporally ordered sub-queries. On the other hand, to identify candidate target segments within a vast video corpus, we employ a Hierarchical Sliding Window (HSW) algorithm to aggregate the retrieval results of sub-queries, ensuring optimal temporal alignment. The retrieved segments, along with their associated multimodal information, are then processed by a Multimodal Large Language Model (MLLM) within the RAG framework. The MLLM handles both KIS and QA tasks, where the retrieved segments enhance answer generation accuracy, while feedback from the QA task helps refine the retrieval process. By unifying KIS and QA, we enable a seamless reinforcement between retrieval and reasoning.

The contributions of this work are summarized as:

- We introduce CadenceRAG, a novel approach that effectively handles the core IViSE tasks, KIS and QA, under complex textual queries.
- We propose a decomposition and aggregation strat-

egy that efficiently breaks down complex queries into temporally ordered sub-queries, enabling precise video segment retrieval and localization.

- We unify KIS and QA within the RAG framework, achieving comprehensive contextual understanding by integrating multimodal evidence.

2. Related Works

The IViSE challenge includes KIS and QA. KIS focuses on precise retrieval and localization, while QA requires comprehension and reasoning based on retrieval results.

2.1. Known-Item Search

The KIS task focuses on retrieving precise video segments based on complex text queries. Traditional video retrieval methods [11, 20, 27, 33], designed for short queries, struggle with the “bag of concepts” limitation [39] and restricted text context windows [43], making them less effective for complex queries. Recent methods like Long-clip [49] and ViCLIP-XL [43] attempt to improve long-text retrieval but fail to resolve the granularity mismatch between queries and videos, often leading to misaligned retrieval results.

CadenceRAG addresses this by decomposing queries into temporally ordered sub-queries using LLM to maintain narrative coherence. HSW then aligns sub-queries with relevant video segments for temporal consistency and precise mapping, bridging the gap between text and video.

2.2. Video Question Answering

Video QA requires models to answer natural language questions based on video content [3], which can cover a broad spectrum of domains such as daily life scenarios, movie, TV shows, instructional videos, and dramatic narratives [5, 23, 30, 35, 40]. Early methods used spatio-temporal attention [10, 16], while recent LLM-based approaches [21, 36] enable open-ended reasoning. However, existing models [5, 31, 40, 47] struggle with long temporal contexts due to input truncation.

While MLLM-based QA [22, 25, 36, 49] rely on visual encoders, they often ignore crucial audio, titles, and user comments [28]. CadenceRAG addresses this by extending RAG to integrate multimodal context (visual, auditory, textual). An MLLM reasons over retrieved segments and dynamically refines retrieval, ensuring contextually grounded answers for HVU.

3. Method

The IViSE challenge presents unique complexities: multi-scene narrative queries targeting brief segments (2-20s) dispersed across a 1,000-hour video corpus. To address this, CadenceRAG, shown as Figure 2, employs: (1) LLM-based

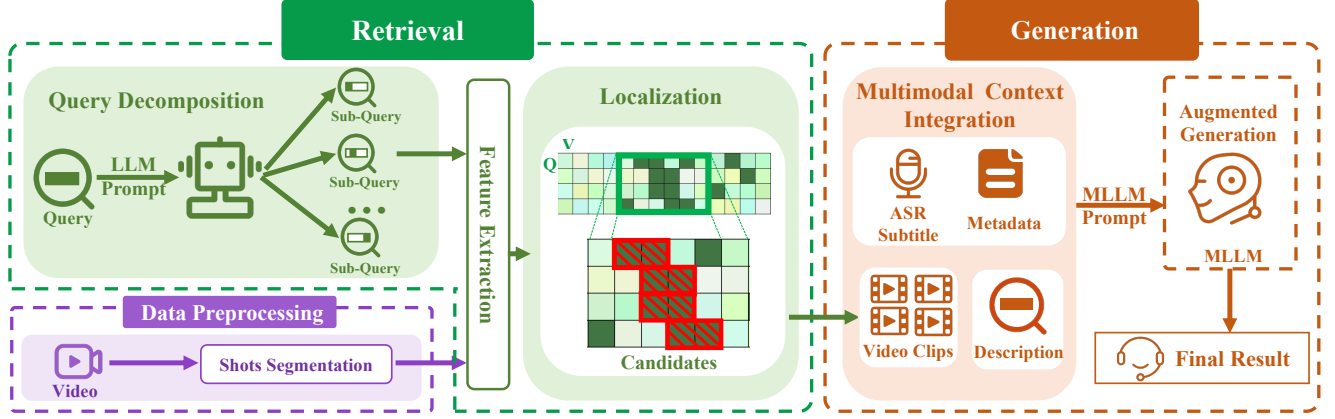


Figure 2. The framework consists of three modules: Data Preprocessing, Retrieval, and Generation. Data Preprocessing structures the data, Retrieval decomposes and aligns queries with video segments, and Generation integrates retrieved information for answer generation.

query decomposition to match single-scene shot granularity, (2) hierarchical sliding window for segments localization, and (3) a progressive RAG framework to handle the extreme search space.

3.1. Data Preprocessing

Videos in the V3C dataset [38] are segmented into 1,082,657 video shots using the segmentation method [37], with each shot accompanied by a representative keyframe. Inspired by the top-performing teams in the TRECVID AVS task [14, 42], we employ a late fusion strategy, aggregating retrieval results from multiple visual-language models (VLM) [9, 18, 19, 32, 44, 48] by averaging their similarity scores, for video retrieval.

3.2. Retrieval

The retrieval stage comprises two key components: Query Decomposition and Segment Localization. Queries are decomposed into temporally ordered sub-queries using LLM. HSW aggregates the decomposed query retrieval results, localizing the most relevant candidate segments.

Query Decomposition Complex video queries span multiple shots, while VLMs handle single keyframe-query pairs. To bridge this gap, we prompt an LLM (Figure 3 top) to generate sub-queries. Figure 3 outlines our decomposition rules, presenting the original query, applied rule, resulting sub-queries (with and without the rule), and rationale. These rules address granularity mismatch and remove non-visual concepts to create more targeted sub-queries. After decomposition, an evaluation prompt assesses quality and further refines the sub-queries.

Segment Localization The location step aims to identify the most relevant candidate segments in the video that align with the decomposed sub-queries. This involves two key considerations: aggregating sub-query results and preserving temporal alignment.

Step 1: Candidate Video Selection Given the similarity matrix $M \in \mathbb{R}^{Q \times V}$ between Q sub-queries and all V keyframes in the corpus, we first perform an efficient candidate video selection to reduce the search space for subsequent fine-grained temporal localization.

For each video k with keyframe indices $\mathcal{V}_k$, we compute a video-level relevance score $S_{\text{video}}(k)$ by multiplying the maximum similarity achieved by each sub-query:

$$S_{\text{video}}(k) = \prod_{i=1}^Q \max_{j \in \mathcal{V}_k} (M_{i,j}) \quad (1)$$

This score ranks videos by their overall relevance to all query components. We sort all videos using $S_{\text{video}}(k)$ and pick the top 10 as the candidate set. The subsequent localization steps will operate exclusively on the similarity sub-matrices corresponding to these candidate videos.

Step 2: Sliding Window for Localization To determine the best-matching video segment, we use a hierarchical sliding window search for temporal alignment, following a bottom-up approach as shown in Figure 4: (1) We initialize the sliding window at the last sub-query layer and progressively move it across the video timeline, computing the maximum similarity score within each window. (2) Once a layer has completed its traversal, the window boundaries are propagated upward, ensuring that higher layers search within the temporal span determined by the previous layer. (3) For each complete window, a final score is computed based on the accumulated similarity across all layers, identifying the best temporal match for the entire query.

This hierarchical sliding window method ensures that sub-queries maintain their sequential alignment while avoiding excessive fragmentation, leading to more coherent and accurate video localization. We pass the corresponding video segments to MLLM for further processing to generate the final answer.

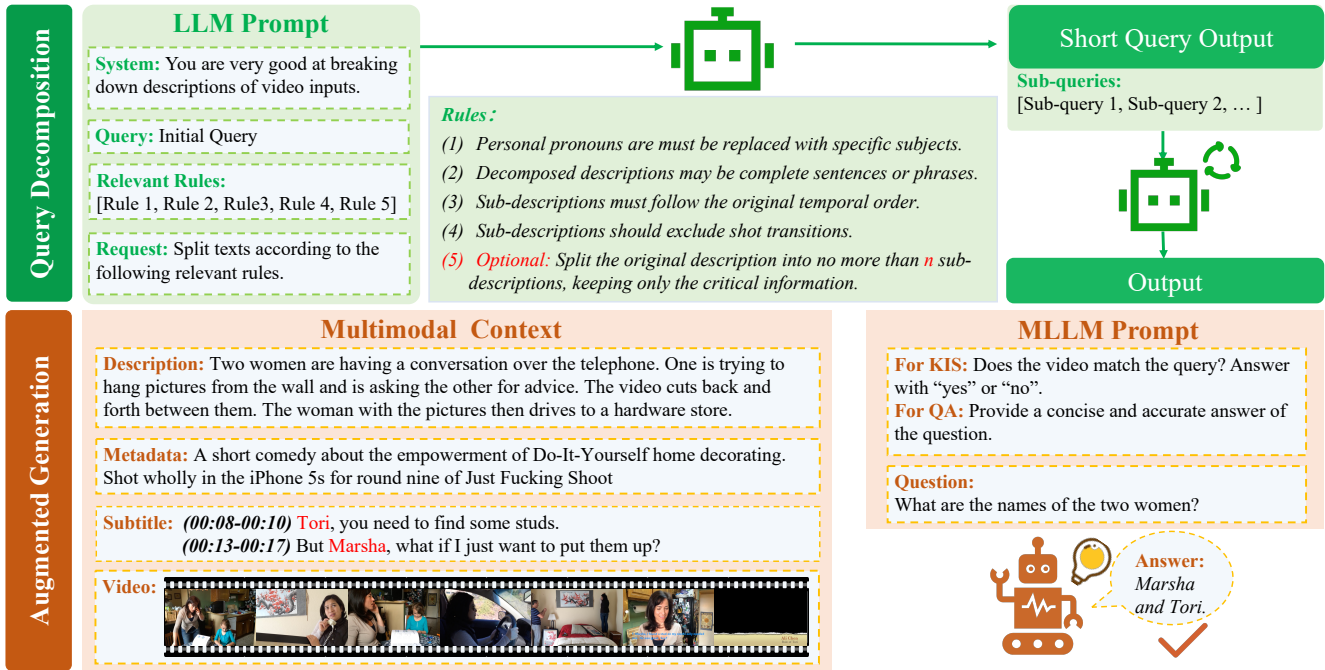


Figure 3. The flowchart of sub-query generation and answer reasoning. The upper part illustrates how LLM prompts are used to decompose a complex query into sub-queries, while the lower part shows how MLLM prompts operate on retrieved video segments to produce final answers for KIS and QA.

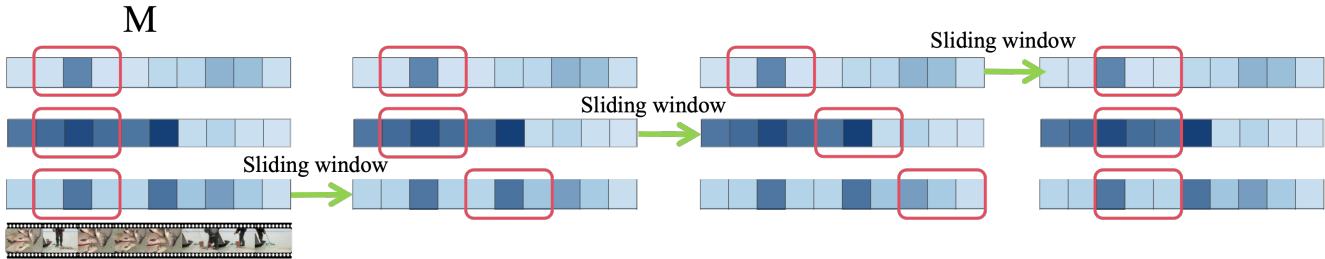


Figure 4. Schematic of the hierarchical sliding window approach, demonstrating layer-wise boundary constraints.

3.3. Generation

The generation step receives candidate video segments. In video-based QA tasks, key details relevant to answering questions may appear in audio and textual modalities. Our generation stage combines two critical components: multimodal information integration and augmented generation, which has demonstrated that LLM achieve stronger performance when provided with relevant knowledge [51].

Multimodal Context Integration Accurate video question answering requires understanding beyond visual signals, because key evidence is often embedded in spoken dialogue, contextual descriptions, and external metadata such as titles, subtitles, and uploader-provided annotations. To address this, we aggregate multimodal context (see Figure 3 bottom) to improve both retrieval grounding and generative reasoning:

- **Visual Information:** Precisely localized video segments that align with the query semantics, extracted via our hierarchical sliding window mechanism;
- **Textual Information:** Derived video metadata (e.g., titles, descriptions), and original query text, which provide semantic scaffolding beyond visual scenes;
- **Audio Information:** Captured from the Whisper-generated speech-to-text transcription, synchronized with video timestamps, which were created using the methodology in combination with OpenAI’s Whisper [34] speech recognition model¹.

Augmented Generation The primary role of the unified generation stage is to leverage the MLLM to identify target video segments for the KIS task and generate answers

¹Vimeo Creative Commons Collection (V3C) Whisper Transcripts, available at: <https://zenodo.org/records/7383303>.

Table 1. Results of the Known-Item Search Tasks. Each task is scored on a scale from 0 to 1 and the final score is the sum of the 10 task scores. #V signifies for how many tasks the correct video was part of the result set and #S signifies for how many tasks the correct video segment was part of the result set

Team	Run	Score	#S	#V
nii-uit	1	9.9	10	10
vifi	4	9.8	10	10
vifi	5	9.8	10	10
vifi	3	9.6	10	10
nii-uit	2	8.9	10	10
vifi	1	8.9	9	9
Exquisitor	1	8.3	9	9
Exquisitor	2	7.9	8	8
Exquisitor	4	7.8	8	8
vifi	2	7.6	8	8
Exquisitor	3	7.3	8	8
WHU-NERCMS	3	6.0	6	9
WHU-NERCMS	2	5.8	6	9
WHU-NERCMS	1	4.0	4	9
nii-uit	3	3.5	5	8
certh_iti	1	3.5	4	8
certh_iti	2	1.6	3	4

for the QA task. Reformulating KIS as a QA-like task ensures task consistency and simplifies the overall pipeline. This unified framework design leverages the core strengths of RAG, progressively narrowing down the search scope to achieve accurate retrieval.

As shown in Figure 3 *bottom*, we design task-specific prompts for KIS and QA, framing KIS as a QA-like problem. For KIS, the model outputs “yes” or “no” to indicate segment relevance, along with a confidence score, leveraging its reasoning capabilities within a unified framework. The retrieved segments are then ranked accordingly with detailed temporal positioning from the MLLM for the best-matching segment. Similarly, for QA tasks, the question is processed by MLLM alongside retrieved video-related textual information, speech transcripts, and the original query.

4. Experiments

To validate the effectiveness of the framework, our team WHU-NERCMS present a comprehensive experimental evaluation in this section.

4.1. Experiment Settings

Dataset and Tasks We use V3C1 [38], the first partition of the Vimeo Creative Commons Collection, containing 7,475 videos (approximately 1,000 hours) ranging from 3 to 30 minutes. The dataset includes rich visual content and structured metadata, spanning diverse real-world categories. The fully automated track evaluates two tasks:

Table 2. Results of the Question Answering Tasks. Each task is scored on a scale from 0 to 1 and the final score is the sum of the 10 task scores. #V signifies for how many tasks the correct video was part of the result set and #A signifies for how many tasks the correct answer was part of the result set.

Team	Run	Score	#A	#V
WHU-NERCMS	1	7.0	7	10
WHU-NERCMS	2	7.0	7	10
vifi	1	6.3	7	10
vifi	3	6.3	7	10
nii-uit	3	6.5	7	10
nii-uit	1	5.8	6	10
nii-uit	2	5.8	6	10
nii-uit	4	5.8	6	10
vifi	2	5.2	6	9
Exquisitor	2	4.2	5	7
Exquisitor	3	3.6	4	6
Exquisitor	4	3.6	4	6
Exquisitor	1	3.4	4	6
certh_iti	1	2.4	3	6
certh_iti	2	2.4	3	5
certh_iti	3	1.8	3	6

- **KIS:** A textual description is provided which uniquely describes a segment of one particular video from the challenge dataset of 7,475 videos.
- **QA:** Each query consists of two parts. The first provides salient information sufficient to uniquely identify a video, while the second poses a question about the video, typically unrelated to the first part.

Setup and Metric The final score is computed as:

$$\text{score} = 1 - \frac{(r - 1)}{n}, \quad (2)$$

where r represents the rank of the correct answer, and $n = 10$, indicating that only the top ten candidates will be considered.

Implementation Details In the data preprocessing stage, we extract features using a set of vision-language models, including CLIP, BLIP, SLIP, LaCLIP, BLIP2, and beit3 [9, 18, 19, 32, 44, 48], and perform late fusion by averaging the similarity scores. Additionally, we utilize Deepseek-R1 [13] to decompose complex queries and employ Videollama3 [50] as the generator with its default settings.

4.2. Overall performance

Tables 1 and 2 show the final competition rankings for the KIS and QA tasks, respectively. Figures 5 and 6 visualize the retrieval performance.

In the KIS task, our team WHU-NERCMS successfully retrieved 9 out of 10 relevant videos, achieving a perfect match in 8 out of 10 queries. However, due to a submission error on our part, our final recorded score was 6, lower

KIS-01 : Close-up of the hand of an elderly Asian man putting down a green beer bottle on the table. There are a full beer glass, some bowls and a box containing a cognac bottle on the table. The camera moves to the man's face, before going back to his hands, as he uses chopsticks to put food from one bowl into sauce in another bowl.

video : 00726 Time : 5:05-5:07



KIS-03 : A sequence of shots taken from a moving motorbike. In the first shot we see a view under the rider's left arm, the handlebar, both mirrors and the rider's hands are visible. The motorbike moves along a country road, and a red triangular street sign indicates a railway crossing coming up. The following shot is a view down to the road on the right side of the motorbike, first showing the right part of the handlebar and then moving forward.

video : 05739 Time : 3:23-3:26



KIS-05 : Two shots from the window of a moving train: In the first one we see some of the cars in front on the right, and a steep wall, covered with plants, close to the track on the left. In the second shot the view is out of a window on the other side, with the train on the left and a young man standing at the door, and bushes on the right. A black and white mile marker with 101 is visible.

video : 06844 Time : 1:05-1:07



KIS-07 : Several shots of old electronics. The first one shows a wall of stacked desktop PCs, the next a close-up of a processor socketed in a mainboard. The next shots respectively show a closeup of a hand holding a processor pins-up, a box full of old processors, and a box full of old mobile phones.

video : 01425 Time : 1:36-1:38



KIS-09 : A man is sitting in a camping chair outside, drinking from a brown glass bottle. He is holding a green bottle in his other hand. Next to him is another chair with a large white teddy bear in it. The man turns to the bear and places the green bottle in the bear's arm.

video : 07309 Time : 3:22-3:24



KIS-02 : A shot of mostly wooden outdoor fitness equipment, installed between trees. The camera pans to the left, before the next shot shows a closeup of a wooden sculpture, with the words "Peace, Hope & Love" engraved.

video : 04873 Time : 2:07-2:10



KIS-04 : A woman in a wedding dress is standing with outstretched arms on an embankment made of rocks. She is holding a white translucent shawl or veil behind her. In the background, there is a cliff face with a waterfall feeding into a little lake. The next shot shows her more zoomed in as she is wrapping her shawl/veil around her.

video : 04755 Time : 0:56-0:58



KIS-06 : A man wearing red gloves is standing on ice, only his lower half is in frame. At his feet are several fish, still alive. He bends down and throws some of the fish back into the water through a hole in the ice. He puts the remaining fish into a black sled next to him. He picks up a blue rope and pulls away the sled.

video : 06884 Time : 1:06-1:08



KIS-08 : Two dogs in a stream at the base of a waterfall. One dog has light yellow fur, the other has black fur and a blue collar. After getting out of the water, they walk up an incline next to the waterfalls, accompanying some hikers.

video : 05506 Time : 2:34-2:36



KIS-10 : A timelapse of several people mounting large posters with several color images and black text onto wooden structures. The structures are curved so they stand upright to present the posters, but they are laid flat on the floor for mounting them. The whole sequence is filmed from a fixed camera placed high in a corner of the room.

video : 04962 Time : 0:32-0:34



Figure 5. Our KIS results, showing the query, retrieved video segments (rank 1) and related interval (format: minutes:seconds). Green indicates perfect retrieval, while red indicates incorrect retrieval.

than the expected value. In the QA task, our approach successfully retrieved all relevant videos and achieved a 70% accuracy rate, securing first place among all teams. Even when providing incorrect answers, the correct video segments were consistently retrieved.

5. Conclusions

This work addresses the granularity mismatch between long-text queries and video content via a decomposition-based approach. Our findings highlight the importance of granularity alignment and multimodal integration for

QA-01: Grainy black/white shot of a radio antenna mast, panning down until the roof of a car becomes visible, followed by a shot showing a hand turning both dials of a radio receiver.

Question: What was the call sign of the first radio station founded by the man portrayed in this documentary?

video: 05052 **Ans:** W2XJY



QA-03: A group of people is holding a workshop in a room with pastel green walls with white trim, and parquet flooring. Throughout the video, the people are shown in various constellations in the room, intercut with several short sequences where individual people are talking directly into the camera. There are some shots with different arrangements of chairs and tables. On one side of the room, there is a raised stage with a short red curtain at the top.

Question: What institution hosted that workshop?

video : 04222 **Ans:** The University of Edinburgh



QA-05: Two shots showing a wooden hut with a chimney made of stone on the left end. It is sunny, but the ground and the roof of the hut are mostly covered with snow. A woman (wearing jeans and greenish/petrol jacket, gloves, bonnet) and a child (jeans, blue jacket, gloves, bonnet) walk towards the hut. In the second shot, the woman opens the door of the hut, which makes a squeaking sound.

Question: Which drink does the woman serve when they are back in their car?

video : 02431 **Ans:** Coffee



QA-07: Two shots of a wooden pavilion on a public square, with benches below and cars parked in front. A long red boat is stored below the ceiling. The first shot shows a closeup of the boat, the second a larger view of the square. This clip was made on a sailing trip.

Question: Which is the lowest flag on the rope down from the mast to the right side of the bow of the sailboat?

video : 00063 **Ans:** The Yellow one



QA-09: A series of aerial shots of buildings in different locations. The shots include a round white building with a sign saying "CAPITOL RECORDS". Another shot shows a view of the "GRAND LAKE THEATRE". There is a watermark in the corner of the entire video.

Question: How many different aerial video sequences are in the video in total?

video : 04093 **Ans:** 12



QA-02: The video is filmed from the perspective of a paraglider flying over a lush green landscape. There is a deep blue lake surrounded by grass-covered hills, along which the paraglider is flying. In the distance, the ocean is visible. The paraglider is often flying very close to the steep hills on their right.

Question: At the very beginning of that video, there is a title card indicating the year and the location the video was recorded. Where and when was that?

video: 06208 **Ans:** Azoren, 2016



QA-04: An urban setting filmed in black and white. Early in the video, there is a shot where the camera is looking out through the window of a moving train. The video is then composed of several outdoor shots with falling snow. During the whole video, there is music playing.

Question: According to the end card, what is the name of the song?

video : 03343 **Ans:** The song is called 'Talking To Ghost' by Foxes



QA-06: A woman is sitting in a chair, playing the accordion and singing a song. She has long light-brown hair and wears a red sweater. She is alternatingly shown singing from several perspectives for several minutes.

Question: What language is the song in?

video : 01155 **Ans:** French



QA-08: Two women are having a conversation over the telephone. One is trying to hang pictures from the wall and is asking the other for advice. The video cuts back and forth between them. The woman with the pictures then drives to a hardware store. What are the names of the two women?

Question: What are the names of the two women?

video : 03288 **Ans:** Marsha and Tori



QA-10: A man wearing dark trousers and jacket, a scarf and a horse mask walks down a path. There is a wooden bench on the right, cars and houses in the background, and litter on the path and grass. He shakes his head and puts his hand to the head. The camera follows him as he walks to another wooden bench and sits down.

Question: Which writer is quoted at the start of this video?

video : 03087 **Ans:** Balzac.



Figure 6. Our QA results, including the query, question, answer, and retrieved video. Green indicates correct retrieval, while green and red answers denote correct and incorrect responses, respectively.

holistic video understanding, especially when visual data is limited or ambiguous. CadenceRAG significantly enhances performance, especially in QA, where our system ranked first. Despite strong results, challenges remain. The system’s reliance on LLM introduces sensitivity to query

phrasing, and multimodal reasoning incurs non-trivial computational cost. Future work should explore more efficient retrieval-generation pipelines and adaptive mechanisms to enhance scalability and robustness while keeping inference efficient.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China (No. 62372339). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [2] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10638–10647, 2020. 2
- [3] Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. Dam: Dynamic adapter merging for continual video qa learning. *arXiv preprint arXiv:2403.08755*, 2024. 2
- [4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna>, 3(5), 2023. 2
- [5] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1166–1174, 2021. 2
- [6] Giorgio De Magistris, Riccardo Caprari, Giulia Castro, Samuele Russo, Luca Iocchi, Daniele Nardi, and Christian Napoli. Vision-based holistic scene understanding for context-aware human-robot interaction. In *International Conference of the Italian Association for Artificial Intelligence*, pages 310–325. Springer, 2021. 1
- [7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhofen, and Luc Van Gool. Holistic large scale video understanding. *CoRR*, 2019. 1
- [8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhofen, and Luc Van Gool. Large scale holistic video understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 593–610. Springer, 2020. 1
- [9] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023. 3, 5
- [10] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6576–6585, 2018. 2
- [11] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 1(2):6, 2021. 2
- [12] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022. 2
- [13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5
- [14] Jiangshan He, Ruizhe Li, Jiahao Guo, Hong Zhang, Mingxi Li, Zhengqian Wu, Zhongyuan Wang, Bo Du, and Chao Liang. Whu-nercms at trecvid2023: Adhoc video search (avs) and deep video understanding (dву) tasks. In *TRECVID*, 2023. 2, 3
- [15] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 1
- [16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2
- [17] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482, 2023. 2
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3, 5
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3, 5
- [20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [21] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2
- [22] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

- [23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, 2020. Association for Computational Linguistics. 2
- [24] Ruizhe Li, Jiahao Guo, Mingxi Li, Zhengqian Wu, and Chao Liang. A hierarchical deep video understanding method with shot-based instance search and large language model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9425–9429, 2023. 2
- [25] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2
- [26] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer, 2022. 2
- [27] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2
- [28] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023. 2
- [29] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022. 2
- [30] Kartikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, pages 46212–46244. Curran Associates, Inc., 2023. 2
- [31] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 2
- [32] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 3, 5
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2
- [34] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4
- [35] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 2
- [36] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2
- [37] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. Cineast: a multi-feature sketch-based video retrieval engine. In *2014 IEEE International Symposium on Multimedia*, pages 18–23. IEEE, 2014. 3
- [38] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3c—a research video collection. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pages 349–360. Springer, 2019. 3, 5
- [39] Moming Tang, Chengyu Wang, Jianing Wang, Chuanqi Tan, Songfang Huang, Cen Chen, and Weining Qian. Xtremeclip: Extremely parameter-efficient tuning for low-resource vision language understanding. In *Findings of the association for computational linguistics: ACL 2023*, pages 6368–6376, 2023. 2
- [40] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhausen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2
- [41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [42] Kazuya Ueki, Yuma Suzuki, Hiroki Takushima, Hideaki Okamoto, Hayato Tanoue, and Takayuki Hori. Waseda meisei softbank at trecvid 2022. In *Proceedings of the TRECVID 2022 Workshop*, pages 1–5, 2022. 3
- [43] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advancing long description understanding for video clip models. *arXiv preprint arXiv:2410.00741*, 2024. 2
- [44] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 5
- [45] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, pages 36978–36989. PMLR, 2023. 2
- [46] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Con-*

- ference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. [1](#)
- [47] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 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)*, pages 9777–9786, 2021. [2](#)
 - [48] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. [3](#), [5](#)
 - [49] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024. [2](#)
 - [50] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. [5](#)
 - [51] Zihang Zhou, Xin Wang, Jinyu Zhang, Han Shi, Jinhui Tang, and Hongyang Chao. Context optimization for retrieval-augmented visual question answering. In *ICCV*, 2023. [4](#)