CFVBench: A Comprehensive Video Benchmark for Fine-grained Multimodal Retrieval-Augmented Generation

Kaiwen Wei*
weikaiwen@cqu.edu.cn
Chongqing University
Chongqing, China

Xiao Liu*
Liu-xiao-@outlook.com
Chongqing University
Chongqing, China

Jie Zhang*
zhang15827347943@163.com
Independent Researcher
China

Zijian Wang*
wangzijian@stu.cqu.edu.cn
Chongqing University
Chongqing, China

Ruida Liu
dang9ab3e5_1@163.com
Chongqing University
Chongqing, China

Yuming Yang
ymyang@cqu.edu.cn
Chongqing University
Chongqing, China

Xin Xiao
20241401023@stu.cqu.edu.cn
Chongqing University
Chongqing, China

Xiao Sun
sunx@stu.cqu.edu.cn
Chongqing University
Chongqing, China

Haoyang Zeng
202514156683@stu.cqu.edu.cn
Chongqing University
Chongqing, China

Changzai Pan
panpanaqm@126.com
Independent Researcher
China

Yidan Zhang
zhangyidan19@mails.ucas.ac.cn
University of the Chinese Academy of Sciences
China

Jiang Zhong†
zjstud@cqu.edu.cn
Chongqing University
Chongqing, China

Peijin Wang†
wangpeijin17@mails.ucas.ac.cn
Aerospace Information Research Institute, Chinese Academy of Sciences
China

Yingchao Feng†
fengyc@aircas.ac.cn
Aerospace Information Research Institute, Chinese Academy of Sciences
China

Abstract

Multimodal Retrieval-Augmented Generation (MRAG) enables Multimodal Large Language Models (MLLMs) to generate responses with external multimodal evidence, and numerous video-based MRAG benchmarks have been proposed to evaluate model capabilities across retrieval and generation stages. However, existing benchmarks remain limited in modality coverage and format diversity, often focusing on single- or limited-modality tasks, or coarse-grained scene understanding. To address these gaps, we introduce CFVBench, a large-scale, manually verified benchmark constructed from 599 publicly available videos, yielding 5,360 open-ended QA pairs. CFVBench spans high-density formats and domains such as chart-heavy reports, news broadcasts, and software tutorials, requiring models to retrieve and reason over long temporal video spans while maintaining fine-grained multimodal information. Using CFVBench, we systematically evaluate 7 retrieval methods and 14 widely-used MLLMs, revealing a critical bottleneck: current models (even GPT5 or Gemini) struggle to capture transient yet essential fine-grained multimodal details. To mitigate this, we propose Adaptive Visual Refinement (AVR), a simple yet effective framework that adaptively increases frame sampling density and selectively invokes external tools when necessary. Experiments show that AVR consistently enhances fine-grained multimodal comprehension and improves performance across all evaluated MLLMs.

CCS Concepts

• **Information systems** → **Retrieval Augmented Generation**.

Keywords

Multimodal Retrieval Augmented Generation, Video Understanding, Benchmark Evaluation, Multimodal Large Language Models

*These authors contributed equally to this work.
†Corresponding author.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Kaiwen Wei*, Xiao Liu*, Jie Zhang*, Zijian Wang*, Ruida Liu, Yuming Yang, Xin Xiao, Xiao Sun, Haoyang Zeng, Changzai Pan, Yidan Zhang, Jiang Zhong†, Peijin Wang†, and Yingchao Feng†. 2018. CFVBench: A Comprehensive Video Benchmark for Fine-grained Multimodal Retrieval-Augmented Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, the Retrieval-Augmented Generation (RAG) paradigm has become indispensable for enhancing the factual accuracy, knowledge grounding [26], and explainability [82] of Large Language Models (LLMs). This paradigm has been further extended to the Multimodal RAG (MRAG) framework, which leverages external multimodal knowledge (such as images, audio, and video) to address increasingly complex, real-world information-seeking tasks [41, 81]. Among these modalities, video-based MRAG [23, 37, 48, 57, 79] poses unique challenges due to its temporal dynamics, the tight coupling of visual and auditory streams, and the presence of dense, transient details [39].

To evaluate the increasingly sophisticated capabilities of video-based MRAG systems, numerous benchmarks [5, 7, 30, 75] have been proposed, typically focusing on tasks such as basic factual retrieval in visual questions, general video event understanding, or cross-modal reasoning. However, as illustrated in Fig. 1 (see full-size case in Appendix A) and Table 1, existing benchmarks face two key limitations when confronted with real-world complexity. (1) **limited modality and format coverage**: current benchmarks often involve restricted modality pairings (e.g., text-video or audio-video), overlooking datasets that simultaneously encompass audio-text-video streams and densely embedded domain-specific formats, such as data charts, complex tables, or dynamic on-screen text, which are critical for real-life video comprehension. (2) **insufficient fine-grained reasoning**: most benchmarks rely on coarse-grained analysis, such as scene classification or major entity recognition, and do not adequately evaluate a model's ability to perform precise, detailed multimodal reasoning. For example, extracting exact numerical values from rapidly changing charts or following sequential, understanding the complex operations and timing relationships in a tutorial, both of which are essential for grounded, factual multimodal question answering task.

To address these limitations, we introduce the Comprehensive Fine-grained Video-based retrieval-augmented generation Benchmark (CFVBench), a large-scale, manually verified, multimodal dataset specifically designed to evaluate video-based MRAG systems. CFVBench is constructed from real-world videos on YouTube. It contains 599 videos and 5,360 open-ended question-answering (QA) pairs, spanning high-density content such as chart-heavy reports, news broadcasts, and software tutorials. The benchmark requires models to retrieve information across long temporal video spans and multiple modalities while maintaining a high level of fine-grained visual fidelity. Using CFVBench, we conduct a comprehensive evaluation of 7 popular retrieval methods and 14 publicly available Multimodal Large Language Models (MLLMs), revealing a critical bottleneck: *existing methods (even GPT5 or Gemini) struggle significantly with fine-grained multimodal comprehension.* Further human evaluation shows that models frequently miss fine-grained visual cues and transient events, and exhibit internal issues such as hallucination, resulting in incomplete or inaccurate understanding.

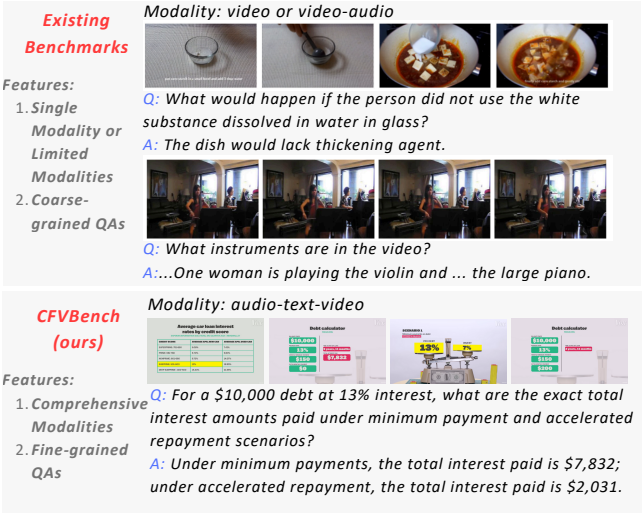


Figure 1: Comparison of video-based MRAG benchmarks with CFVBench, where clues are embedded in tables/on-screen text of video frames, requiring fine-grained reasoning.

Table 1: Comparison of video-based MRAG benchmarks. MR: Multimodal Fine-grained Reasoning; RQA: Retrieval-based Question-Answering; IM: Intra-Modal Multi-hop Questions; OE: Open-Ended Questions; CMF: Comprehensive Modalities and Format (e.g., video, text, audio, chart, table, etc.). Notation / indicates that benchmark is not presented in QA format.

Dataset	MR	RQA	IM	OE	CMF
Video Benchmarks					
ShareGPT4Video [5]	×	×	×	/	×
NExT-QA [71]	×	×	×	✓	×
Video-MME [15]	×	×	✓	×	×
LongVideoBench [70]	×	×	✓	×	×
MovieChat [56]	×	×	✓	✓	×
EgoSchema [38]	×	×	✓	×	×
MVBench [27]	×	×	×	✓	×
MMBench-Video [13]	×	×	×	✓	×
SOK-Bench [64]	×	×	×	×	×
Multimodal Benchmarks					
UnAV-100 [18]	×	×	/	/	×
VAST-27M [6]	×	×	✓	/	×
OmniBench [30]	✓	×	×	×	×
Cinepile [50]	×	×	×	×	×
Video-Bench [44]	×	×	×	✓	×
AVQA [72]	×	×	×	×	×
AVInstruct [75]	×	×	×	✓	×
Music-AVQA2.0 [34]	×	×	×	×	×
VGGSound [4]	×	×	/	/	×
AVHaystacks [10]	×	✓	✓	✓	×
AVHBench [58]	×	×	×	×	×
CFVBench (ours)	✓	✓	✓	✓	✓

Motivated by this critical finding, we propose the Adaptive Visual Refinement (AVR) framework. AVR is a simple yet effective, two-stage framework that enables MLLMs to assess their current information state and execute a resource-efficient refinement strategy. In the first stage, adaptive frame interpolation mechanism employs a scoring system to evaluate information sufficiency and automatically increase frame sampling density only when an incomplete evidence chain is detected. In the second stage, on-demand tool invocation strategy intelligently activates specialized external tools, such as Optical Character Recognition (OCR) [63] and object detectors [8], when necessary to extract embedded visual text or entity relevant to the query. Experimental results show that AVR effectively enhances the fine-grained multimodal comprehension of existing MLLMs, enabling them to capture subtle visual details and achieve state-of-the-art (SOTA) performance on the proposed CFVBench. In summary, the main contributions of this work are:

1) We introduce CFVBench, a large-scale MRAG benchmark (599 videos, 5,360 QA pairs) with high-density content (e.g., charts, tutorials), specifically designed to assess complex fine-grained multimodal comprehension across long temporal video spans.

2) We evaluate 7 retrieval methods and 14 MLLMs on CFVBench, uncovering a key bottleneck of existing MLLMs in fine-grained multimodal comprehension, likely caused by internal hallucinations and insufficient attention to transient details.

3) We propose AVR, a simple yet effective framework that leverages adaptive frame interpolation and on-demand tool invocation to enhance fine-grained visual understanding and improve performance across all evaluated MLLMs. The dataset, code, and prompts will be released upon acceptance of the paper.

2 Related Works

2.1 Multimodal RAG Benchmarks

Existing Multimodal Retrieval-Augmented Generation (MRAG) benchmarks have progressively expanded the integration of diverse modalities [41, 81], including text, images, audio, and video. Early knowledge-based visual question answering datasets, such as KB-VQA [66] and KVQA [55], rely on closed-domain knowledge, while OK-VQA [40] and A-OKVQA [54] incorporate external knowledge but mainly focus on single-step retrieval rather than complex reasoning. More recent benchmarks have broadened modality coverage and task complexity. MRAG-Bench [22] and MRAMG-Bench [76] combine text and images, Chart-MRAG Bench [73] integrates textual and chart data, ManyModalQA considers text, images, and tables, WavCaps [42], and MusicCaps [1] emphasize audio-language understanding, KnowIT VQA [17] and SOK-Bench [64] focus on video reasoning, and InfoSeek [7] and Encyclopedic VQA [43] address knowledge-intensive, information-seeking tasks. Benchmarks with broader modality coverage, such as MultiModalQA [59] and AVHaystacks [10], still tend to simplify question formats through templates or multiple-choice settings (further comparison of the benchmarks are illustrated in Table 1.) Despite these advancements, most existing benchmarks remain limited in modality and format diversity and lack systematic evaluation of fine-grained reasoning. To address this gap, we introduce CFVBench, a benchmark designed to evaluate fine-grained multimodal comprehension ability of different video-based MRAG systems.

2.2 Multimodal RAG Methods

Early MRAG methods typically converted multimodal data into unified textual representations for a text-based RAG pipeline [37, 79], causing substantial information loss, while later approaches preserved original multimodal data and introduced MLLMs for cross-modal retrieval and generation [57, 78]. Recent MRAG systems incorporate advanced modules such as multimodal search planners [35, 65], interleaved text-image output modules [28, 80], LLM-based cross-modal rerankers [33, 48], and cross-modal refiners [23, 69] to achieve end-to-end multimodal interaction. Despite these advances, evaluation still relies heavily on VQA [29, 32] or VidQA datasets [53, 77], limiting assessment of fine-grained reasoning. By conducting experiments on the proposed CFVBench, we find that existing video-based MRAG methods struggle with detailed multimodal comprehension, which motivates the Adaptive Visual Refinement (AVR) framework that evaluates retrieved visual information and selectively enhances frame sampling or invokes specialized tools to improve fine-grained comprehension.

3 CFVBench

3.1 Preliminary

Video-based MRAG is typically formulated as a two-stage process: (1) **Retrieval.** Given a textual question T_q, a retriever R selects a subset of videos (or video segments) $V = \{V_1, V_2, \ldots, V_k\}$ that are most relevant to the question. The goal of this stage is to identify a minimal set of videos containing sufficient information to answer question T_q, thereby enabling efficient downstream processing. (2) **Generation.** A MLLM M then takes the question T_q and the retrieved video subset V as input, i.e., (T_q, V), and synthesizes information across multiple modalities and temporal video segments to generate the final answer A. Formally, this can be expressed as

$$A = M(T_q, V) \tag{1}$$

To evaluate retrieval and generation while assessing fine-grained multimodal comprehension ability, we introduce CFVBench, a large-scale, manually verified benchmark of high-density videos.

3.2 Dataset Construction

As shown in Fig. 2, the construction of CFVBench follows a 4-stage pipeline designed to ensure diversity, difficulty, and quality:

(1) Video Curation. We curated videos from YouTube spanning diverse domains such as finance, climate, and environment. Our focus was on content that inherently requires multimodal and fine-grained reasoning. The videos fall into three main categories: (i) *Structured-Data Videos*: Videos containing dense visual information such as charts and tables, which require accurate extraction of fine-grained details and numerical reasoning, aligned with transcripts. (ii) *Tutorials*: Instructional content (e.g., software or websites utilizing workflows) where UI operations are tightly coupled with audio scripts, demanding cross-modal reasoning. (iii) *News*: News reports that integrate live footage, scrolling captions, and occasional charts or texts, requiring models to capture objects, entities, scenes, dynamics, and transient information.

The specific websites corresponding to these categories are listed in Appendix B. Each curated video is paired with its transcript and audio track, ensuring that CFVBench offers comprehensive

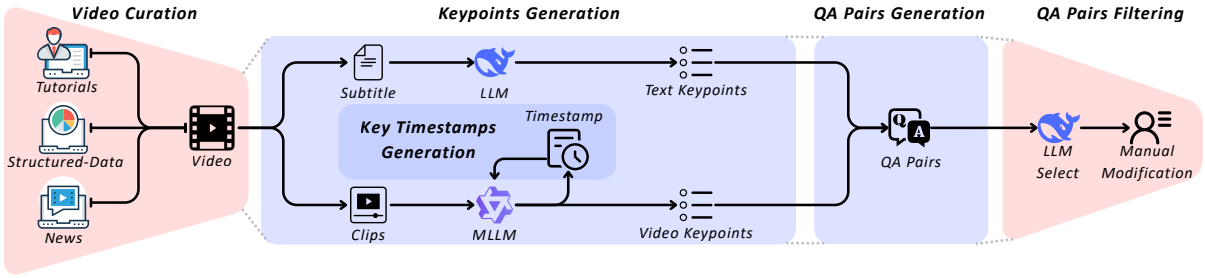


Figure 2: The dataset construction process of CFVBench.

multimodal coverage across video, text, and audio, as well as diverse format (such as charts and tables) in the video frames.

(2) Keypoints Generation. First, we used Qwen2.5-VL-72B-Instruct [61] to automatically identify informative timestamps under a category-specific protocol: For tutorial videos, the model segmented steps and annotated each operation concisely. For structured data videos, it detects comprehensive data formats such as charts, tables, maps, and flowcharts, and outputs start and end times with formatted semantic tags in JSON. For news videos, salient frames were recognized and described at the image level (e.g., polar bears walking on ice). All outputs were stored in a unified JSON schema with temporal boundaries and semantic annotations, providing reliable anchors for keypoint extraction.

Following [73, 84], keypoints are fine-grained, semantically independent facts from videos or subtitles, serving as ground truth for QA construction and evaluation. To handle diverse video domains, we designed tailored extraction protocols. For Structured-Data and News videos, we extracted (i) textual keypoints from subtitles via DeepSeek-R1 [12], and (ii) visual keypoints from frames aligned with timestamps using Qwen2.5-VL-72B-Instruct. For Tutorial videos, where narration and on-screen actions are tightly coupled, we generated hybrid keypoints from subtitles, frames, and timestamps, encoding explicit (action, object, outcome) triples (e.g., "the user clicks the 'File' button → the 'Save As' dialog box opens → a file path is displayed"). This framework ensures QA tasks require cross-modal reasoning rather than a single modality. To improve accuracy, we applied the LLMs three times on each sample and took the intersection of the results.

(3) QA Pair Generation. Building on extracted keypoints, we constructed open-ended QA pairs with DeepSeek-R1. We generate (i) *single-hop* questions, answerable from a single keypoint (e.g., factual or numerical detail) from a single modality, and (ii) *multi-hop* questions, requiring integrating two or more key points (e.g., sequential tutorial steps or combining visual and textual evidence) from multiple modalities. Questions primarily follow factual and procedural forms ("What," "How") while also including causal ("Why"), comparative ("How does … compare"), and conditional ("What happens if …") patterns. Answers are grounded in keypoints without external inference, QA are independent and align with the general cognitive level of human video viewing, supporting reliable evaluation of retrieval and multi-step reasoning.

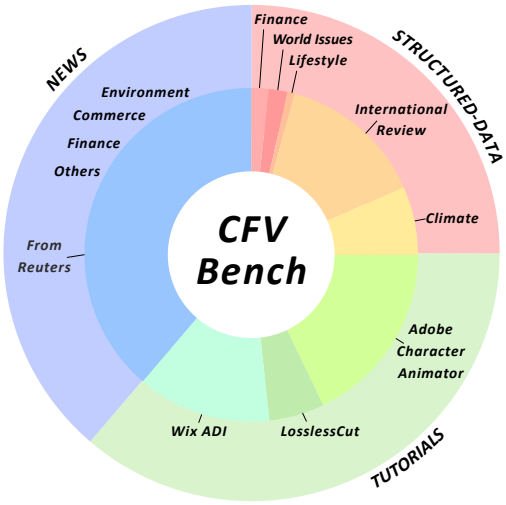


Figure 3: The distribution of videos in CFVBench.

(4) QA Pair Filtering. To ensure the quality of generated QA pairs, we employed DeepSeek-R1 to make each QA pair automatically checked against 4 criteria: (i) for multi-hop questions, all listed keypoints must be necessary to construct the complete answer; redundant or unused keypoints were removed, (ii) duplicates across QA pairs were eliminated, (iii) QA pairs unrelated to the topic were discarded, and (iv) overly mechanical or unnatural questions (e.g., over-detailing or forced associations) were filtered out.

To further enhance the quality of CFVBench and ensure questions are meaningful to humans and align with real-life application scenarios, we conduct human modification. We recruited 13 graduate students from local universities, all with academic backgrounds in computer science, journalism, or English, and IELTS scores above 7.0. Before annotation, they underwent training that introduced the task objectives and detailed guidelines (aligned with QA pair filtering criteria), demonstrated examples of both correct and incorrect annotations, emphasized data privacy and responsible usage, and clarified that they need to modify the QA pairs to make semantics more precise while keeping content unchanged. During annotation, each annotator was assigned a subset of QA pairs, with 30% overlapping between two annotators. Annotators were also instructed to adjust temporal segments when initial video timestamps were

incorrectly split. A QA pair was finalized only when both annotators provided consistent annotations, ensuring dataset reliability. Additionally, for quality control, 2 supervising annotators randomly sampled 20% of every 100 QA pairs. If any issue was found, the process was repeated until all sampled pairs satisfied the standards.

The prompts in this section are shown in Appendix F (1-9).

3.3 Dataset Composition

After the dataset construction process, CFVBench consists of 3 core video categories: structured-data videos, tutorials, and news. as shown in Fig. 3, these categories are further divided into topics such as finance, climate, software operations, and commerce, covering key areas of real-world applications. Statistically, as reported in Table 7 of Appendix C, CFVBench contains a total of 5,360 question-answer pairs. To evaluate different levels of reasoning complexity, the test set includes 3,703 single-hop questions and 1,660 multi-hop questions. On average, multi-hop questions require integrating 2.72 keypoints across different modalities, placing higher demands on the model's reasoning capabilities. Additionally, the average video length is 232.3 seconds, and the average question length is 15.22 words, reflecting rich content and moderate question complexity.

4 Evaluation on CFVBench

MRAG frameworks are typically composed of two core stages: retrieval and generation. To thoroughly assess existing video-based MRAG methods' performance, we separately evaluate both stages of MRAG on the proposed CFVBench. The following sections will first introduce the baselines, evaluation metrics, and implementation details relevant to both stages, followed by a comprehensive analysis of the experimental results.

4.1 Evaluation Settings

Baselines. The retrieval stage employs 3 widely-utilized multimodal retrieval methods, including ImageBind [19], InternVideo [68], and LanguageBind [83] that align video and text representations within a shared embedding space. Building upon this, the nomic-embed-text [45] model is incorporated to specifically capture semantic similarities in the textual modality. After the retrieval stage, the top-k video clips identified by the best-performing retrieval method are passed to the generation stage.

During the generation stage, we combine the retrieval results and conduct a zero-shot evaluation on CFVBench with 14 MLLMs, including: Gemma-3-12b-it [60], Gemma-3-27b-it, Qwen2.5-VL-7B-Instruct [61], InternVL-3.5-8B [67], InternVL3_5-14B, InternVL3_5-30B-A3B, llava-llama-3-8b-v1_1 [11], Intern-S1-mini [3], MiniCPM-V-2_6 [74], Magistral-Small-2509 [49], Mistral-Small-3.2-24B-Instruct-2506, and 3 close-source models: gpt-5-chat-latest [46], gemini-2.5-flash-lite-preview-09-2025 [21], claude-opus-4-20250514 [2].

Evaluation Metrics. For retrieval evaluation, we adopt Recall@K [73] (R@K), which measures whether at least one relevant video is included among the top-k retrieved results. We report Recall@1, Recall@3, Recall@5, and Recall@10 to capture performance across different cutoff thresholds. We evaluate two query types: (1) multi-point: queries require retrieving videos containing multiple keypoints across different modalities; (2) single-point: queries target videos with a single keypoint, further divided into text-single-point (speech/transcript-based, where keypoints are derived from audio transcripts) and video-single-point (visual content-based, where keypoints originate from visual information) to assess modality-specific retrieval capabilities.

In the generation stage, following [73, 84], evaluation metrics are divided into two categories. (1) *Automatic evaluation*: To assess a model's ability to leverage multimodal information in CFVBench, we adopt keypoint-based recall, precision, and F1-score [47]. Recall is computed for video ($Recall_v = C_v/N_v$) and textual keypoints ($Recall_t = C_t/N_t$), where C_v, C_t denote correctly matched keypoints and N_v, N_t mean total keypoints, the overall recall is $Recall_{all} = (C_v + C_t)/(N_v + N_t)$. Precision measures the proportion of correctly covered keypoints among all factual claims (M) in the response ($Precision = (C_v + C_t)/M$), and F1 is their harmonic mean. We also introduce ROUGE-L [31] to capture lexical overlap. (2) *LLM-as-Judge evaluation*: We average the results from Qwen3-8B-Instruct [62] and GLM-4-9B [20] to holistically score model outputs. Each response is rated on three dimensions: Factual Coverage (Fact_cov), Visual Detail Usage (Vis_use), Linguistic Precision (Ling_prec), and assigned a final score on a 1–5 Likert scale [24]. Additionally, we compute semantic similarity (St_cos) between generated and reference answers using all-MiniLM-L6-v2 [51]. Relevant prompts are in Appendix F (1-9).

Implementation Details. During the retrieval stage, we follow the procedure in [52], which leverages both audio transcriptions and video captions. Specifically, the audio of each video is transcribed using Faster-Distil-Whisper-Large-v3 [16], while visual frames sampled at 6-second intervals are captioned with MiniCPM-V-4-int4 [74], conditioned on the corresponding audio information. Based on these modalities, we generate multimodal embeddings from the raw videos and extract text embeddings from the generated descriptions. The top-k full videos are first retrieved using cosine similarity, after which each is segmented into 30-second clips, and the results of best-performing retrieval method are then passed to the generation stage.

In the generation stage, we follow [52] and further segment each video into 6-second clips, from which the five most relevant ones are selected. For each clip, its transcript is combined with 5 frames sampled at 6-second intervals, and the resulting multimodal input is fed into the MLLM in the retrieval order to generate the final answer. To ensure deterministic outputs, we set the temperature to 0.1 and top-p to 1. All MLLMs are evaluated using identical prompts Appendix F (11) on NVIDIA A100 GPUs with 80GB of memory.

4.2 Evaluation Results

Analysis of Retrieval Results. As shown in Table 2, we could find: combining textual and multimodal embeddings yields consistent improvements, with LanguageBind + nomic-embed-text achieving the best overall R@10 of 81.37%, underscoring the complementary strengths of text and video representations. In comparison, multimodal embeddings alone generally outperform text-only retrieval on single-point queries, where the best video model (LanguageBind) achieves 76.17% R@10, slightly above the text baseline (76.11%). However, multi-point queries remain particularly challenging. An interesting observation is that multimodal embeddings underperform text-only embeddings under multi-point setting, suggesting that the video modality may introduce noise and hinder multi-hop

Table 2: Retrieval methods performance comparison under different scenarios. The top 3 best scores for each metric are highlighted as first, second, and third.

Model	Overall				Multi-Point				Single-Point							
									Text-Single-Point				Video-Single-Point			
	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10
Nomic-embed-text	51.01	64.53	69.42	76.11	41.74	56.20	60.84	68.19	58.33	71.01	75.74	82.13	49.26	63.12	68.62	75.05
Imagebind	37.16	54.03	61.21	70.63	25.42	38.97	45.18	54.91	42.12	61.23	68.82	79.97	42.98	59.95	67.62	74.67
Internvideo	30.05	45.12	52.32	63.21	20.00	29.27	35.36	43.73	34.16	52.03	59.70	71.97	35.32	52.59	60.34	71.88
Languagebind	46.18	61.88	68.13	76.17	32.16	43.37	48.67	57.28	53.93	71.84	78.48	86.31	49.72	67.07	73.81	81.48
Imagebind + nomic-embed-text	55.24	70.05	75.64	81.33	43.49	58.25	63.79	71.62	63.14	77.81	83.33	88.05	55.61	70.72	76.52	81.25
Internvideo + nomic-embed-text	54.09	67.55	72.83	78.76	42.83	56.74	62.65	69.06	61.69	74.83	79.35	85.15	54.37	67.85	73.74	79.24
Languagebind + nomic-embed-text	57.51	71.47	76.31	81.37	45.78	57.71	62.59	68.13	65.46	79.85	84.78	89.38	57.70	73.50	78.07	83.42

Table 3: Generation performance comparison of different MLLMs on CFVBench. The top 3 best scores for each metric are highlighted as first, second, and third. (red and blue to distinguish open-source and close-source models).

Model	$Recall_t$	$Recall_v$	$Recall_{all}$	Precision	F1	Rouge_L	St_cos	Likert	Fact_cov	Vis_use	Ling_prec
claude-opus-4	0.2031	0.2162	0.2079	0.1667	0.1850	0.0909	0.6545	3.2000	1.8909	2.3818	3.1273
gpt-5-chat	0.2079	0.1552	0.1887	0.1220	0.1482	0.1477	0.6705	3.5455	2.2386	2.3409	3.7045
gemini-2.5-flash	0.2843	0.3051	0.2919	0.1673	0.2127	0.1778	0.8000	3.6889	2.1000	2.5333	3.9444
Gemma-3-12b-it	0.0685	0.0601	0.0656	0.0837	0.0736	0.1287	0.6045	2.8205	1.4817	2.0272	3.0366
Qwen2.5-VL-7B-Instruct	0.1152	0.1220	0.1176	0.1067	0.1119	0.1376	0.6223	3.0996	1.6807	2.0953	3.7785
InternVL-3.5-8B	0.1805	0.2188	0.1936	0.1440	0.1652	0.1601	0.6814	3.4934	2.1734	2.3189	4.1022
llava-llama-3-8b	0.1583	0.2810	0.2018	0.0985	0.1324	0.2209	0.7933	3.5095	2.1308	2.2473	3.8312
MiniCPM-V-2_6	0.2407	0.3793	0.2892	0.1424	0.1908	0.1702	0.7766	3.6383	2.4286	2.3736	3.9560
InternVL3_5-14B	0.1858	0.2354	0.2034	0.1470	0.1707	0.2017	0.7272	3.7070	2.2843	2.3807	4.1494
Mistral-Small-3.2-24B-Instruct	0.1883	0.2041	0.1931	0.1069	0.1376	0.1639	0.6995	3.7104	2.2623	2.3880	4.0273
InternVL3_5-30B	0.2156	0.1818	0.2046	0.1193	0.1507	0.2000	0.7444	3.8296	2.4556	2.4000	4.1148
Intern-S1-mini	0.2941	0.2353	0.2745	0.1609	0.2029	0.2017	0.7647	3.8319	2.2797	2.3814	4.4153
Magistral-Small	0.2303	0.1512	0.2045	0.1057	0.1394	0.1883	0.7324	3.9351	2.5490	2.5621	4.1569
Gemma-3-27b	0.2655	0.1935	0.2400	0.1180	0.1582	0.2079	0.8020	3.9703	2.4851	2.5644	4.2277

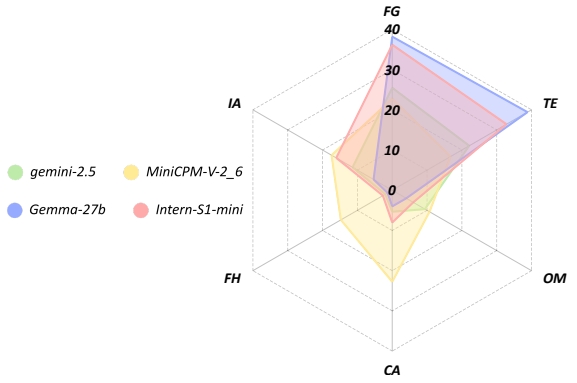


Figure 4: Human evaluation of typical MLLMs on CFVBench.

reasoning. Moreover, across all multimodal methods, performance in multi-point setting lags behind single-point results, with even the best hybrid approach reaching only 71.62% R@10. This substantial drop highlights that current models still struggle with fine-grained temporal integration and multi-hop reasoning.

Analysis of Generation Results. As shown in Table 3, the experimental results on CFVBench reveal several insights: (1) there are substantial performance differences among different MLLMs. For example, considering $Recall_t$, the highest-performing model achieves 0.2941 while the lowest is only 0.0685, indicating a large gap in the ability to extract key fine-grained information. (2) Although closed-source models that are not specifically designed for video tasks, such as gemini, do not show significant advantages over newer open-source video-oriented models, they demonstrate relatively stable performance across metrics, highlighting their effectiveness and reliability in video-based MRAG task. (3) The overall performance of all evaluated models remains modest across automatic and LLM-as-Judge evaluation, demonstrating the challenge posed by CFVBench and underscoring the necessity and importance of studying fine-grained video-based MRAG.

To further investigate model behavior, we conducted a manual error analysis. We randomly sampled 100 generated answers from Gemma-3-27B, MiniCPM-V-2_6, Gemma-27B, and Intern-S1-mini, and asked 3 annotators to label each response according to 6 predefined error types. Majority voting was applied, and Fleiss's kappa [14] of 0.81 indicated strong inter-annotator agreement. The evaluated error categories include Fine-Grained Detail Omission

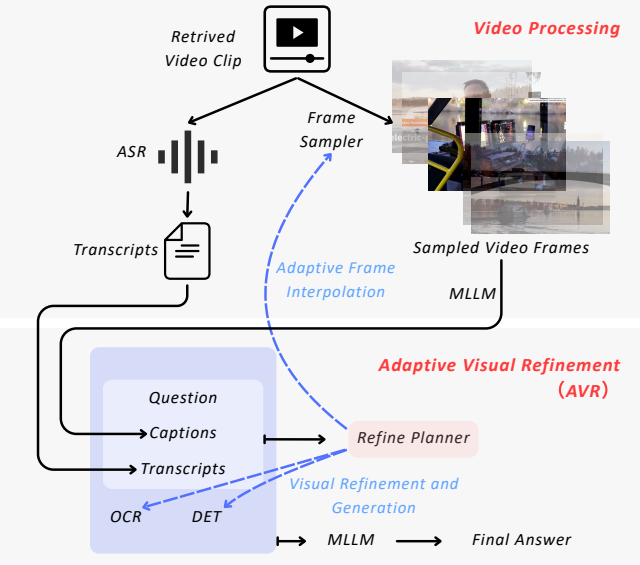


Figure 5: The workflow of the AVR framework.

(FG), Transient Event Neglect (TE), Object/Entity Misidentification (OM), Contextual Association Error (CA), Factual Hallucination (FH), and Incomplete Answer (IA). As shown in Fig. 4, we observe: (1) all models frequently suffer from FG and TE errors, with Gemma-27B and Intern-S1-mini being most affected, highlighting the challenge of fine-grained information capture; (2) MiniCPM-V-2_6 produces fewer FG and TE errors but more CA, FH, and IA cases, suggesting that although it excels at detail perception, it is more vulnerable to hallucination and reasoning instability; (3) Gemini-2.5 demonstrates the most balanced performance, showing stronger robustness in the fine-grained video-based MRAG task.

5 Methodology

To address the challenge of fine-grained comprehension, we propose Adaptive Visual Refinement (AVR) framework. First, adaptive frame interpolation mechanism assesses evidence from sparse frame sampling and adaptively increases frame density for critical segments. Second, visual refinement and generation strategy performs on-demand tool invocation, such as OCR [63] and object detection (DET) [8], and aggregates multimodal evidence to generate the final answer. The workflow of AVR is shown in Fig. 5.

5.1 Adaptive Frame Interpolation

Adaptive frame interpolation evaluates the sufficiency of initial evidence and determines whether denser sampling is needed. For each retrieved video segment V_i, we first conduct video processing: (1) *Sparse Frame Sampling*: extract N_{sparse} frames (e.g., $N_{sparse} = 5$) to form a sparse frame set F_{sparse}. (2) *Audio Transcription*: generate subtitles T_{asr} using Automatic Speech Recognition (ASR) model Whisper [16]. (3) *Preliminary Summary*: obtain an initial summary C_{init} from F_{sparse} using the MLLM to be evaluated.

We then feed the query T_q, subtitles T_{asr}, and summary C_{init} into the MLLM-based refinement planner with a refinement prompt (Appendix F (10)), which explicitly instructs the model to assess whether the current evidence suffices to answer the query and to identify potentially missing information. The planner outputs an *Information Sufficiency Score* $S \in [0, 10]$, which integrates an overall *answerability score* measuring how well the query can be addressed given the current evidence and an *information density score* evaluating the richness and granularity of visual content in F_{sparse} and T_{asr}. A higher score S indicates sufficient evidence, while a lower score signals missing fine-grained details.

Based on the information sufficiency score S, AVR determines the number of frames N_{target} to sample according to:

$$N_{target} = f(S) = \begin{cases} N_{sparse} & \text{if } S > \theta \\ N_{min} + (N_{max} - N_{min}) \cdot \frac{\theta - S}{\theta} & \text{if } S \le \theta \end{cases} \quad (2)$$

where θ is the threshold (set to 5.0) and $[N_{min}, N_{max}]$ is the refinement interval (set to $[20, 40]$; following Kandhare and Gisselbrecht [25] and Luo et al. [36], they both identify that the sampling rate achieves optimal results at around 1 fps). If $S > \theta$, the original sparse frames are retained; otherwise, denser sampling is applied proportionally to the severity of the information gap. Interpolated frames are filtered by visual similarity to reduce redundancy, and only diverse frames are retained for caption generation. Captions from these refined frames are merged with existing subtitles to construct a richer, fine-grained representation, which is subsequently fed into the visual refinement and generation stage of AVR.

5.2 Visual Refinement and Generation

Guided by the MLLM-based refinement planner, we perform on-demand tool invocation and multimodal evidence aggregation to construct a rich feature representation for final answer generation.

(1) On-demand Tool Invocation. The refinement planner determines whether specialized tools are required (Appendix F (10)). Specifically, OCR and DET tools are invoked only when explicitly required: OCR is used when the query depends on reading symbolic or structured visual data (e.g., numbers, UI elements, tables, charts), while DET is applied when answering requires recognizing entity categories, attributes, or spatial relations (e.g., object state, interactions, fine-grained classification).

If OCR is triggered, we apply EasyOCR [63] to the selected target frames F_{target} to extract symbolic or numeric information, such as textual overlays, tables, or chart values, and aggregate the results into T_{ocr}. If object detection (DET) is requested, we derive a task-specific detection vocabulary from the query T_q using a prompt-based keyword generator (Appendix F (13)) and run YOLO-World [9] to detect corresponding entities in F_{target}. The detection results (object class and spatial attributes) are converted into textual descriptions T_{det}.

(2) Evidence Aggregation and Final Generation. For the selected target frames F_{target}, we employ an MLLM-based captioner to generate a detailed video summary C_{target}. All available evidence sources are then concatenated to form the enriched feature set $F_{rich} = [T_q, T_{asr}, C_{target}, T_{ocr}, T_{det}]$, where T_{ocr} and T_{det} are optional. Finally, the MLLM integrates F_{rich} to generate the final answer $A = M(T_q, F_{rich})$.

All MLLM-based components in AVR (including the refinement planner, captioner, and answer generator) employ the same MLLM

Table 4: Generation performance comparison on CFVBench with the proposed AVR framework (marked with *).

Model	$Recall_t$	$Recall_v$	$Recall_{all}$	Precision	F1	Rouge_L	St_cos	Likert	Fact_cov	Vis_use	Ling_prec
claude-opus-4*	0.3125	0.3077	0.3103	0.2000	0.2432	0.1765	0.7059	3.6471	2.4706	2.8235	3.8824
gpt-5-chat*	0.2946	0.2727	0.2874	0.1644	0.2092	0.2211	0.7263	4.1158	2.8526	2.6632	4.0947
gemini-2.5-flash*	0.2210	0.4625	0.2950	0.1778	0.2219	0.1986	0.8582	3.7801	2.3901	2.5887	3.9929
Gemma-3-12b-it*	0.0621	0.0730	0.0660	0.1013	0.0799	0.1354	0.6247	3.0143	1.5879	2.1817	3.1508
Qwen2.5-VL-7B-Instruct*	0.1069	0.1449	0.1204	0.1049	0.1121	0.1513	0.6454	3.2943	1.6980	2.1510	3.9584
InternVL-3.5-8B*	0.1638	0.2941	0.2051	0.1457	0.1704	0.1667	0.7583	3.5542	2.2000	2.3958	4.1208
llava-llama-3-8b*	0.1523	0.3321	0.2160	0.1116	0.1472	0.2328	0.8017	3.5701	2.1490	2.3552	3.9380
MiniCPM-V-2_6*	0.2178	0.4561	0.3038	0.1846	0.2297	0.2111	0.8111	3.7111	2.4778	2.4333	4.0222
InternVL3_5-14B*	0.1692	0.2839	0.2099	0.1600	0.1816	0.2127	0.7485	3.8613	2.3691	2.4914	4.2580
Mistral-Small*	0.1857	0.2451	0.2035	0.1090	0.1420	0.1799	0.7143	3.7937	2.2910	2.4762	4.0952
InternVL3_5-30B*	0.2239	0.2941	0.2475	0.1217	0.1632	0.2712	0.8136	3.9831	2.5085	2.5339	4.1525
Intern-S1-mini*	0.2946	0.3667	0.3198	0.1993	0.2456	0.2626	0.8182	3.9495	2.3232	2.4343	4.4646
Magistral-Small*	0.2143	0.2958	0.2400	0.1176	0.1579	0.2598	0.7480	4.0000	2.6220	2.6220	4.2126
Gemma-3-27b*	0.2768	0.2581	0.2701	0.1247	0.1706	0.2500	0.8500	4.1000	2.5100	2.7800	4.2600

Table 5: Ablation study on AVR, with green values showing the change in indicator scores relative to results w/o AVR.

Model	$Recall_v$	Vis_use
Gemma-3-27b	0.1935	2.5644
Gemma-3-27b (+ frames)	0.2222 +14.83%	2.7059 +5.52%
Gemma-3-27b (+ frames & DET)	0.2407 +24.39%	2.7326 +6.56%
Gemma-3-27b (+ frames & OCR)	0.2545 +31.52%	2.7529 +7.35%
InternVL3_5-30B	0.1818	2.4000
InternVL3_5-30B (+ frames)	0.2593 +42.63%	2.4824 +3.43%
InternVL3_5-30B (+ frames & DET)	0.2778 +52.81%	2.5059 +4.41%
InternVL3_5-30B (+ frames & OCR)	0.2647 +45.60%	2.5085 +4.52%

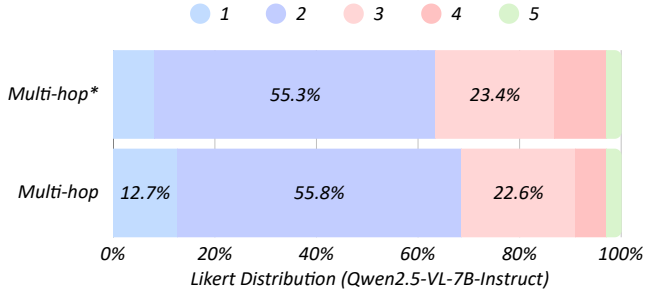


Figure 6: Performance on multi-hop questions, with (*) indicating use of AVR.

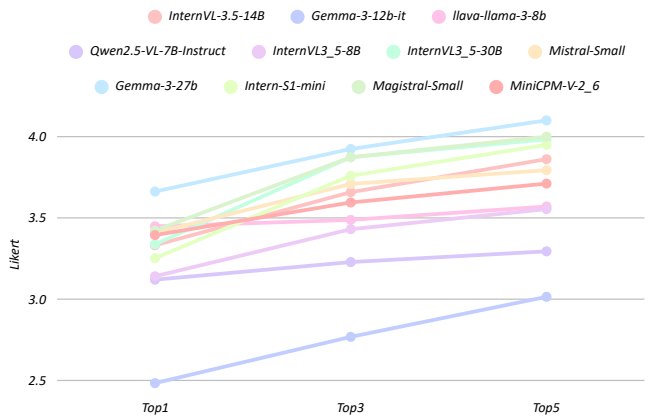


Figure 7: Retrieval effect of different top-k settings.

that is being evaluated, ensuring consistency across stages and eliminating potential confounding effects from external models.

6 Experiment with AVR

Results with AVR. Table 4 demonstrates the effectiveness of AVR. Incorporating AVR leads to consistent metric improvements across most MLLMs, with particularly notable gains in $Recall_v$; for instance, MiniCPM-V-2_6 rises from 0.3793 to 0.4561, confirming AVR's ability to enhance fine-grained multimodal comprehension through dynamic frame sampling and selective OCR/object detection. These results suggest that AVR enables models to better capture critical visual details and reason over dense video content. A slight drop in $Recall_t$ for some models reflects the trade-off between richer visual coverage and potential retrieval noise. We also present performance and robustness analyses, including method comparisons and retrieval reordering experiments. Detailed results are available in Appendix **??** and Appendix D.

Ablation Study. To assess the contribution of each AVR component, we perform ablation studies on two representative MLLMs: Gemma-3-27B and InternVL3.5-30B. As shown in Table 5, all modules improve fine-grained multimodal comprehension. The adaptive frame interpolation strategy consistently boosts both $Recall_v$ and Vis_{use} by mitigating information loss through dynamic sampling. Further integrating OCR and DET tools yields additional gains, with OCR particularly effective for structured scenes involving textual or numeric content.

Performance on Multi-hop Questions. To further investigate model performance in complex scenarios that require integrating fine-grained visual cues for multi-step reasoning, we take Qwen2.5-VL-7B-Instruct as an base model and compare the 1–5 Likert [24]

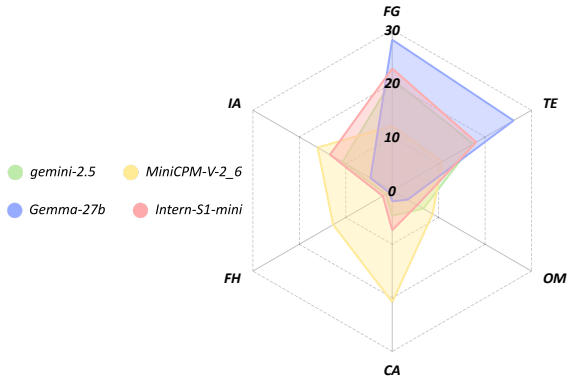


Figure 8: Human evaluation on CFVBench with AVR.

score distribution for multi-hop questions with (Multi-hop*) or without AVR (Multi-hop). Higher Likert scores indicate better balanced use of multimodal information and more complete coverage of keypoints. As shown in Fig. 6, incorporating AVR leads to a noticeable increase in average Likert scores, particularly for score 4, demonstrating that AVR enables the model to better comprehend fine-grained content, which in turn supports more accurate multimodal multi-hop reasoning.

Impact of Retrieval Top-k. As shown in Fig. 7, we also conducted experiments comparing different retrieval top-k settings, using the Likert score as the evaluation metric. The Likert [24] score provides a holistic assessment of answer clarity, balanced use of multimodal information, and coverage of keypoints. We observe that as k increases from 1 to 5, the performance of nearly all models improves, indicating that retrieving more relevant segments is generally beneficial for the fine-grained video-based MRAG task. However, the magnitude of improvement varies across models, suggesting that different MLLMs have varying abilities to capture and utilize fine-grained multimodal information.

Human Evaluation. Following the same protocol as Section 4.2, we further evaluated 100 responses from 4 representative MLLMs after integrating the AVR framework. Each response was independently annotated by 3 human raters, achieving a Fleiss's kappa [14] of 0.83, indicating strong agreement. Majority voting results are shown in Fig. 8, we find: (1) compared to methods without AVR in Fig. 4, both FG and TE errors of all MLLMs notably decreased, validating AVR's effectiveness in capturing fine-grained visual cues via adaptive frame selection and OCR/DET tools; while (2) hallucination and contextual association errors persist, particularly for MiniCPM-V-2_6, suggesting that intrinsic reasoning instability remains even when visual details are accurately captured. A case study experiment was also conducted, please refer to Appendix E.

7 Conclusions

In this work, we introduce CFVBench, a large-scale, human-checked benchmark of 599 real-world videos and 5,360 open-ended QA pairs spanning tables, charts, and other formats. The benchmark covers audio-text-video modalities and is designed to evaluate fine-grained multimodal comprehension in existing video-based MRAG systems. Comprehensive evaluation of 7 retrieval methods and 14 MLLMs reveals that current approaches struggle with transient but critical multimodal details. To address this, we propose Adaptive Visual Refinement (AVR), a simple yet effective framework that combines adaptive frame sampling with on-demand tool invocation, consistently enhancing fine-grained understanding and overall performance across all the evaluated MLLMs. CFVBench and AVR could provide a practical testbed and methodology for advancing video-based multimodal reasoning in real-world scenarios.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[2] Anthropic. 2025. Introducing Claude 4. https://www.anthropic.com/news/claude-4

[3] Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. 2025. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763* (2025).

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[5] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems* 37 (2024), 19472–19495.

[6] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* 36 (2023), 72842–72866.

[7] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713* (2023).

[8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16901–16911.

[9] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.

[10] Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. 2025. MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks. *arXiv preprint arXiv:2506.07016* (2025).

[11] XTuner Contributors. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.

[12] DeepSeek-AI, Daya Guo, and etc. Dejian Yang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[13] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems* 37 (2024), 89098–89124.

[14] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24108–24118.

[16] Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling.

[17] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10826–10834.

[18] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22942–22951.

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[20] Team GLM, :, Aohan Zeng, and et.al. Bin Xu. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [cs.CL] https://arxiv.org/abs/2406.12793

[21] Google. 2025. Continuing to bring you our latest models, with an improved Gemini 2.5 Flash and Flash-Lite release. https://developers.googleblog.com/en/continuing-to-bring-you-our-latest-models-with-an-improved-gemini-2-5-flash-and-flash-lite-release/

[22] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182* (2024).

[23] Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 4075–4083.

[24] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4 (2015), 396.

[25] Mahesh Kandhare and Thibault Gisselbrecht. 2024. An empirical comparison of video frame sampling methods for multi-modal rag retrieval. *arXiv preprint arXiv:2408.03340* (2024).

[26] Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446* (2024).

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[28] Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. 2025. Revolutionizing Text-to-Image Retrieval as Autoregressive Token-to-Voken Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 813–822.

[29] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, et al. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937* (2024).

[30] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. 2024. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272* (2024).

[31] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[32] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems* 36 (2023), 22820–22840.

[33] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271* (2024).

[34] Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4478–4487.

[35] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. 2022. Audio-text retrieval in context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4793–4797.

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

[37] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. Prop: Pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 14th ACM international conference on web search and data mining*. 283–291.

[38] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.

[39] Mingyang Mao, Mariela M Perez-Cabarcas, Utteja Kallakuri, Nicholas R Waytowich, Xiaomin Lin, and Tinoosh Mohsenin. 2025. Multi-RAG: A Multimodal Retrieval-Augmented Generation System for Adaptive Video Understanding. *arXiv preprint arXiv:2505.23990* (2025).

[40] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[41] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748* (2025).

[42] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 3339–3354.

[43] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3113–3124.

[44] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103* (2023).

[45] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic Embed: Training a Reproducible Long Context Text Embedder. arXiv:2402.01613 [cs.CL]

[46] OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/

[47] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[48] Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2024. Tiger: Unifying text-to-image generation and retrieval with large multimodal models. *arXiv preprint arXiv:2406.05814* (2024).

[49] Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. 2025. Magistral. *arXiv preprint arXiv:2506.10910* (2025).

[50] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813* (2024).

[51] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[52] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos. *arXiv preprint arXiv:2502.01549* (2025).

[53] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347* (2018).

[54] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*. Springer, 146–162.

[55] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8876–8884.

[56] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18221–18232.

[57] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 46345–46361.

[58] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2024. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325* (2024).

[59] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039* (2021).

[60] Gemma Team. 2025. Gemma 3. https://goo.gle/Gemma3Report

[61] Qwen Team. 2025. Qwen2.5-VL. https://qwenlm.github.io/blog/qwen2.5-vl/

[62] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[63] DR Vedhaviyassh, R Sudhan, G Saranya, Mozhgan Safa, and D Arun. 2022. Comparative analysis of easyocr and tesseractocr for automatic license plate recognition using deep learning algorithm. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE, 966–971.

[64] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13384–13394.

[65] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition. 16551–16560.
[66] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).
[67] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265* (2025).
[68] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2023. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. *arXiv preprint arXiv:2307.06942* (2023).
[69] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. Promptmm: Multi-modal knowledge distillation for recommendation with prompt-tuning. In *Proceedings of the ACM Web Conference 2024*. 3217–3228.
[70] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems* 37 (2024), 28828–28857.
[71] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9777–9786.
[72] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*. 3480–3491.
[73] Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking Multimodal RAG through a Chart-based Document Question-Answering Generation Framework. *arXiv preprint arXiv:2502.14864* (2025).
[74] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2025. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *Nat Commun 16, 5509 (2025)* (2025).
[75] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*. Springer, 146–164.
[76] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025. MRAMG-Bench: A BeyondText Benchmark for Multimodal Retrieval-Augmented Multimodal Generation. *arXiv e-prints* (2025), arXiv–2502.
[77] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9127–9134.
[78] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*. 1441–1452.
[79] Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. *arXiv preprint arXiv:2110.03611* (2021).
[80] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, et al. 2023. IRGen: Generative Modeling for Image Retrieval. CoRR abs/2303.10126 (2023).
[81] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868* (2023).
[82] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924* (2024).
[83] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2023. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. arXiv:2310.01852 [cs.CV]
[84] Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 8520–8544. doi:10.18653/v1/2025.acl-long.418

A The Full-sized Example in CFVBench

The full-sized example in Fig. 1 from the Introduction section is presented in Fig. 9.

Table 6: Data Source of CFVBench.

Topic	Website Link
Structured-Data Videos Tutorials	Vox
	Adobe Character Animator
	LosslessCut
	Wix ADI
News	Reuters

Table 7: CFVBench properties and question statistics.

Metric	Value
Average Video Length	232.3s
Average Question Length	15.22 words
Total Questions	5,360
Single-hop Questions	3,703
Multi-hop Questions	1,660
Average Keypoints per Multi-hop Question	2.72
What	2,581 (48.13%)
How	1,374 (25.62%)
Mixed	719 (13.41%)
Which	251 (4.68%)
Why	239 (4.46%)
Where	68 (1.27%)
When	59 (1.10%)
Who	30 (0.56%)
Other	42 (0.78%)

B CFVBench Data Source

The data source of different topics of CFVBench is in Table 6.

C Statistics of CFVBench

CFVBench is a large-scale, manually curated multimodal benchmark for evaluating fine-grained reasoning in video-based MRAG systems. It contains 5,360 open-ended QA pairs from 599 real-world YouTube videos across diverse domains such as news, reports, and tutorials, with an average duration of 232.3 seconds. Questions are concise (15.22 words on average) and include 3,703 single-hop and 1,660 multi-hop items, the latter averaging 2.72 reasoning keypoints. "What" and "How" questions dominate (48.13% and 25.62%), while other types such as "Which," "Why," and "Where" contribute additional diversity.

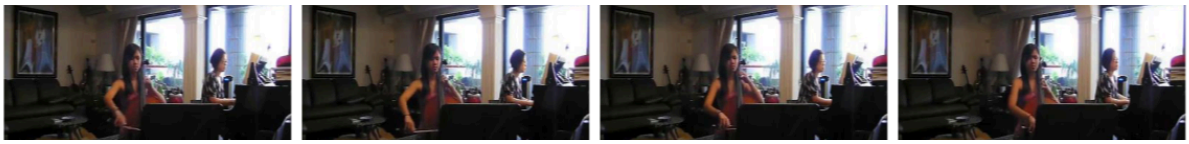
D Robustness Experiment

We further evaluate AVR's robustness via a reorder experiment, in which the five retrieved video clips are randomly shuffled. As shown in Table 8, models with AVR consistently outperform their base counterparts across F1, St_cos, and Likert scores. For example, Gemma-3-27B shows a 6.26% F1 and 2.68% St_cos improvement, indicating that AVR reliably enhances fine-grained reasoning even when the input clip order is disrupted.

E Case Study

To qualitatively illustrate AVR's capabilities, we present a case study with Qwen2.5-VL-7B-Instruct (Figure 10), where key clues appear in

Q: What would happen if the person did not use the white substance dissolved in water in glass?
A: The dish would lack thickening agent.



Q: What instruments are in the video?
A: ...One woman is playing the violin and the other is playing the large piano.

Q: For a $10,000 debt at 13% interest, what are the exact total interest amounts paid under minimum payment and accelerated repayment scenarios?
A: Under minimum payments, the total interest paid is $7,832; under accelerated repayment, the total interest paid is $2,031.

Figure 9: A full-sized cases comparison between existing video-based MRAG benchmarks and CFVBench. In CFVBench, crucial clues are embedded in tables and images of video frames, requiring MLLMs to perform fine-grained reasoning.

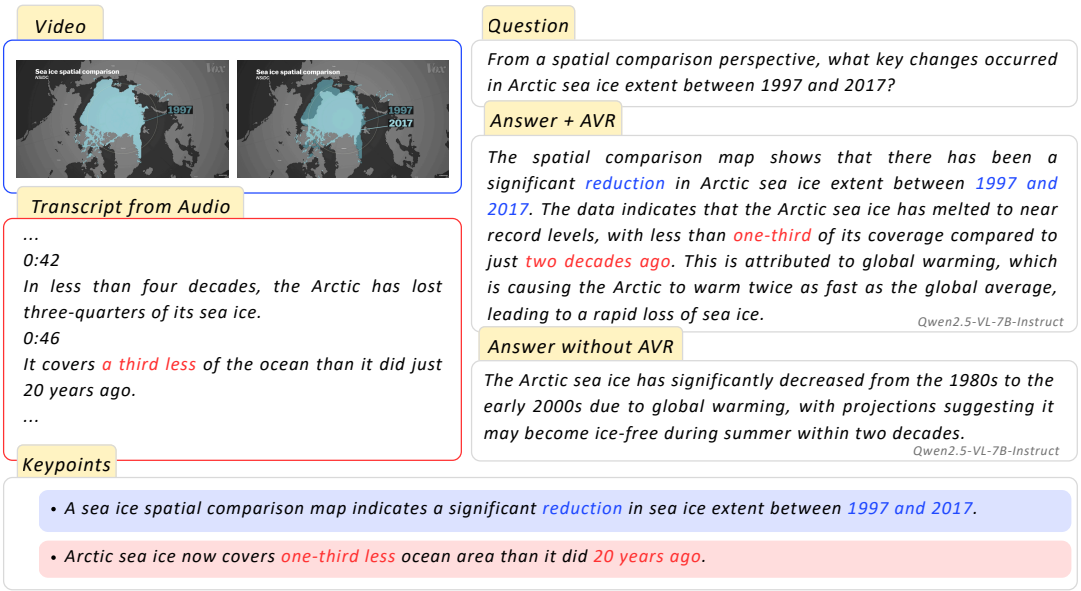


Figure 10: The case study experiment, where key clues are embedded in the on-screen text and audio transcript of the video.

on-screen text and the audio transcript. Without AVR, the baseline fails to link multimodal evidence and produces an incorrect answer, demonstrating hallucination. With AVR, the model successfully integrates multimodal cues to generate a complete, accurate, and coherent answer, showing that AVR effectively resolves incomplete evidence and enhances fine-grained visual comprehension.

Table 8: Effect of reordering on generation quality. Green values indicate relative changes compared to the base models.

Model	F1	St_cos	Likert
Gemma-3-12b-it	0.0748 +1.63%	0.6128 +1.37%	2.9580 +4.88%
Qwen2.5-VL-7B-Instruct	0.1181 +5.54%	0.6295 +1.16%	3.2078 +3.49%
InternVL3_5-14B	0.1688 +2.18%	0.6824 +0.15%	3.5294 +1.03%
llava-llama-3-8b	0.1368 +3.32%	0.7969 +0.45%	3.5630 +1.52%
Mistral-Small	0.1393 +1.24%	0.7059 +0.91%	3.7529 +1.15%
InternVL-3.5-8B	0.1756 +2.87%	0.7307 +0.48%	3.7850 +2.10%
Intern-S1-mini	0.2108 +3.89%	0.7882 +3.07%	3.9059 +1.93%
InternVL3_5-30B	0.1515 +0.53%	0.7882 +5.88%	3.9412 +2.91%
Gemma-3-27b	0.1681 +6.26%	0.8235 +2.68%	4.0235 +1.34%
MiniCPM-V-2_6	0.2165 +13.47%	0.7882 +1.49%	3.6824 +1.21%

F Prompts

The dataset, code, and prompts will be released upon acceptance of the paper.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009