

RAG-Adapter: A Plug-and-Play RAG-enhanced Framework for Long Video Understanding

Xichen Tan

College of Computer Science and Technology,
National University of Defense Technology
Changsha, China

tanxc23@nudt.edu.cn

Yunfan Ye

School of Design,
Hunan University
Changsha, China

Yuanjing Luo

College of Computer and Mathematics,
Central South University of Forestry and Technology
Changsha, China

Qian Wan

Faculty of Artificial Intelligence in Education,
Central China Normal University
Wuhan, China

Fang Liu

School of Design,
Hunan University
Changsha, China

Zhiping Cai

College of Computer Science and Technology,
National University of Defense Technology
Changsha, China

Abstract

*Multi-modal Large Language Models (MLLMs) capable of video understanding are advancing rapidly. To effectively assess their video comprehension capabilities, long video understanding benchmarks, such as Video-MME and MLVU, are proposed. However, these benchmarks directly use uniform frame sampling for testing, which results in significant information loss and affects the accuracy of the evaluations in reflecting the true abilities of MLLMs. To address this, we propose **RAG-Adapter**, a plug-and-play framework that reduces information loss during testing by sampling frames most relevant to the given question. Additionally, we introduce a **Grouped-supervised Contrastive Learning (GCL)** method to further enhance RAG-Adapter’s sampling effectiveness through fine-tuning on our constructed **MMAT** dataset. Finally, we test numerous baseline MLLMs on various video understanding benchmarks, finding that RAG-Adapter sampling consistently outperforms uniform sampling (e.g., GPT-4o’s accu-*

racy increases by 9.3% on Video-MME), providing a more accurate testing method for long video benchmarks.

1. Introduction

In the field of video understanding, research on short videos progresses earlier and more extensively than on long videos, primarily due to the quadratic complexity constraint of transformer-based models in handling long sequences. To mitigate this, many long video models, such as MovieChat [35] and LlamaVid [22], introduce input token compression algorithms to reduce computational costs.

To evaluate the long video understanding capabilities of MLLMs, several specialized long video benchmarks have been proposed, including Video-MME [12] and MLVU [44]. However, these benchmarks do not standardize the number of input frames during testing due to variations in models’ maximum frame capacities. Moreover, not all MLLMs support one-frame-per-second sampling (as-

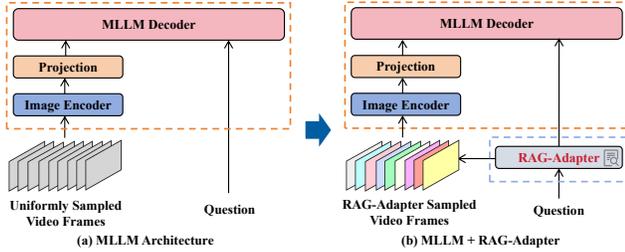


Figure 1. (a) and (b) show a comparison between scenarios with and without the RAG-Adapter framework, respectively.

sumed sufficient to capture content). For these models, testing relies on uniformly sampled frame subsets. In VideoMME, for instance, the longest test video spans one hour, yet only four uniformly sampled frames are used at minimum, often omitting critical information. This leads to responses resembling random guesses and makes it challenging to accurately evaluate true model performance.

To address the testing challenges in existing long video benchmarks, we propose a plug-and-play RAG-enhanced (Retrieval Augmented Generation) optimization framework, **RAG-Adapter**. As illustrated in Figure 1, RAG-Adapter operates without modifying the internal architecture of MLLMs, instead focusing on video frame input. By retrieving the Top K most relevant video frames, it replaces the uniform sampling method, significantly reducing information loss. This straightforward yet effective approach more accurately evaluates the true long video understanding capabilities of MLLMs.

Although the approach is straightforward, research directly integrating RAG with MLLMs is limited. This is mainly because RAG-Adapter’s retrieval performance depends heavily on similarity matching between embeddings generated by its text and image encoders (Figure 2). The embeddings produced by open-source encoders may be sub-optimal for long video understanding tasks. Therefore, we fine-tune these encoders through contrastive learning to better align similar embeddings, thereby enhancing the retrieval effectiveness of RAG-Adapter.

Given the challenge of directly locating relevant frames in long videos, we further construct a fine-tuning dataset, **MMAT**, using short video understanding benchmarks. We extract video frames and pair them with corresponding questions to create positive pairs for fine-tuning.

Additionally, as a single video may correspond to multiple questions, the Self-supervised Contrastive Learning (SCL) assumption that treats other questions as negative samples may mislead the model during training. To address this, we propose **Grouped-supervised Contrastive Learning (GCL)**, where all positive pairs involving the same video’s frame share a common group label. GCL enables clearer differentiation between intra-group and inter-group embeddings, thereby enhancing RAG-Adapter’s retrieval capabilities for video understanding tasks.

Using retrieval results from RAG-Adapter fine-tuned with GCL (RAG-Adapter, unless specified otherwise, is GCL fine-tuned), we introduce two metrics: **Average Similarity Score (ASS)** and **Necessary Information Frame (NIF)**. ASS measures the average similarity between the Top K frames retrieved by RAG-Adapter and the corresponding questions, while NIF represents the average minimum number of frames containing essential information needed to answer each question. The NIF reveals that, even for long video understanding benchmarks, a small subset of frames typically contains the required information, validating our approach of using a fixed number of frames (Top K) across models for fair evaluation.

Notably, the ASS and NIF metrics offered by RAG-Adapter serve as important indicators for evaluating benchmark quality. A lower ASS may indicate insufficient relevance between video content and questions, suggesting potential flaws in question formulation, while a lower NIF implies that fewer frames are needed, indicating lower question complexity.

In summary, the main contributions of this work are:

- 1) We propose **RAG-Adapter**, a plug-and-play enhancement framework for MLLMs. By supplying input-level video frames relevant to test questions, RAG-Adapter enhances the video understanding capabilities of MLLMs without structural modifications.
- 2) We construct the **MMAT** fine-tuning dataset and propose **Grouped-supervised Contrastive Learning (GCL)** for long video understanding scenarios, enhancing RAG-Adapter’s retrieval performance.
- 3) We introduce two metrics through RAG-Adapter: **Average Similarity Score (ASS)** and **Necessary Information Frame (NIF)**, as standards for evaluating benchmark quality and complexity in long video understanding. NIF further confirms that RAG-Adapter provides information that is both sufficient and effective.
- 4) Extensive experiments on open-source long video understanding benchmarks demonstrate the effectiveness of RAG-Adapter in enhancing the video understanding capabilities of existing MLLMs.

2. Related Work

2.1. Multi-model LLMs (MLLMs)

MLLMs extend traditional LLMs by incorporating a visual encoder and projection layer, enabling image and video understanding. Video-based MLLMs [2, 5, 14, 25, 33, 38, 40], process sampled video frames as input, is essentially equivalent to image-based MLLMs [6, 10, 13, 19, 23] that support multiple images, even if not explicitly trained on video data. To handle more frames for long video understanding, many MLLMs reduce computational complexity by compressing the number of visual tokens at the input level.

MovieChat [35] applies ToMe [7] methods to merge similar tokens between adjacent frames. LLaMa-VID [22] reduces image tokens through average pooling, while Chat-UniVi [17] uses the k-nearest-neighbor based density peaks clustering algorithm (DPC-KNN) to segment videos into events, and group tokens of each frame within these events.

Although these models can support inputs of up to thousands of video frames, the NIF metric in Table 1 indicates that the relevant information needed to answer questions resides in only a small subset of frames. Furthermore, ablation experiments in Table 6 show that using more uniformly sampled frames can yield inferior performance compared to using only frames directly relevant to the questions.

2.2. Long Video Understanding Benchmarks

To evaluate MLLMs’s long video understanding capabilities, several benchmarks have been proposed, including Video-MME [12], and MLVU [44]. These benchmarks contain numerous manually annotated Q&A pairs, with average video lengths exceeding 10 minutes. The video content covers a wide range of domains, spanning domains such as daily life, art, sports, and television. They comprehensively assess MLLMs’s abilities in cognition, reasoning, summarization, and other aspects of long video comprehension.

Although these benchmarks provide a comprehensive evaluation of different aspects, during the testing phase, a uniform sampling of video frames is used for all questions. Clearly, the information required for each question varies, and there is a high likelihood that the relevant information may not be included in the uniformly sampled frames. Therefore, assessing the long video understanding capabilities of MLLMs in this manner is not entirely reasonable.

2.3. Retrieval Augmented Generation (RAG)

RAG [18] was first introduced in NLP for retrieval augmentation, and rapidly inspired advancements in text retrieval, with optimizations targeting various stages of the RAG framework to enhance retrieval performance. For instance, SPLADE [11] expands query with semantically similar terms, Self-RAG [4] performs self-correction on retrievals, RAT [37] combines RAG with chain-of-thought reasoning, and LoRAG [36] improves text generation quality via iterative looping. Toolformer [32] enables LLMs to call different tool APIs, allowing information gathering from diverse sources.

In the multi-modal domain, integrating RAG with LLMs remains relatively underexplored. FairRAG [34] uses RAG to promote fairness and diversity in image generation, and RAR [24] leverages RAG to assist in image classification and object detection. In the video domain, to our knowledge, only iRAG [3] uses RAG by encoding video information into contextual natural language descriptions, enabling LLMs to interpret video content.

These observations indicate that RAG’s application in the video domain remains very limited. RAG-Adapter is the first to directly integrate RAG with MLLMs, enhancing long video understanding at the input frame level.

3. Method

3.1. RAG-Adapter Pipeline

RAG-Adapter is a simple yet effective plugin to enhance MLLMs’ video understanding, with its main pipeline detailed in Figure 2.

Video Preprocessing. For the test videos, frames are sampled at one frame per second, forming $\{f_i\}_{i=1}^N$. Each frame is then encoded into image embeddings $\{zf_i\}_{i=1}^N$ using the image encoder CLIP-L/14 [30]. As CLIP-L/14 primarily captures global features, which may miss fine-grained details like objects and actions, we also employ the open-source model CogVLM2 [15] to generate captions for each frame, resulting in the set $\{c_i\}_{i=1}^N$. These captions are encoded into text embeddings $\{zc_i\}_{i=1}^N$ using the text encoder BGE-M3 [9], accommodating CLIP’s text length limitations. Here, f_i , zf_i , c_i , and zc_i represent the i_{th} frame, its embedding, the caption, and its embedding, respectively. Finally, $\{zf_i\}_{i=1}^N$ and $\{zc_i\}_{i=1}^N$ are stored in the FramesDB and CaptionsDB databases for retrieval.

Video Frames Retrieval. To address the dimensional discrepancy between the text and image encoder embeddings and avoid the added complexity and potential performance issues of aligning these spaces, we employ a separate retrieval strategy. When a user submits a question, we encode it using both the text and image encoders and independently match it against the FramesDB and CaptionsDB, retrieving the Top M video frames $\{f_i, sf_i\}_{i=1}^M$ and Top N captions $\{c_i, sc_i\}_{i=1}^N$ from each respective databases, where sf_i and sc_i represent the similarity scores of the query with each retrieved frame and caption, respectively.

To effectively integrate the retrieval results from both databases, we introduce the **Dual Reranker** module, comprising two main steps:

1) We sum the similarity scores of the Top M frames and Top N captions (noting that some captions may correspond to frames outside the Top M set), ranking them by these summed scores to obtain the Top X frames, their corresponding captions and scores, where X is determined jointly by M and N . The set $\{f_i^X, c_i^X, s_i^X\}_{i=1}^X$ represents the i_{th} frame, its caption and summed score, respectively.

2) We find that frames ranked closely within the Top X often exhibit high similarity, reducing diversity. To maintain relevance while enhancing diversity, we apply the Maximal Marginal Relevance (MMR) algorithm [8], commonly used in recommendation systems. We begin with an initially selected set $\mathcal{S} = \emptyset$ and an unselected set $\mathcal{U} =$

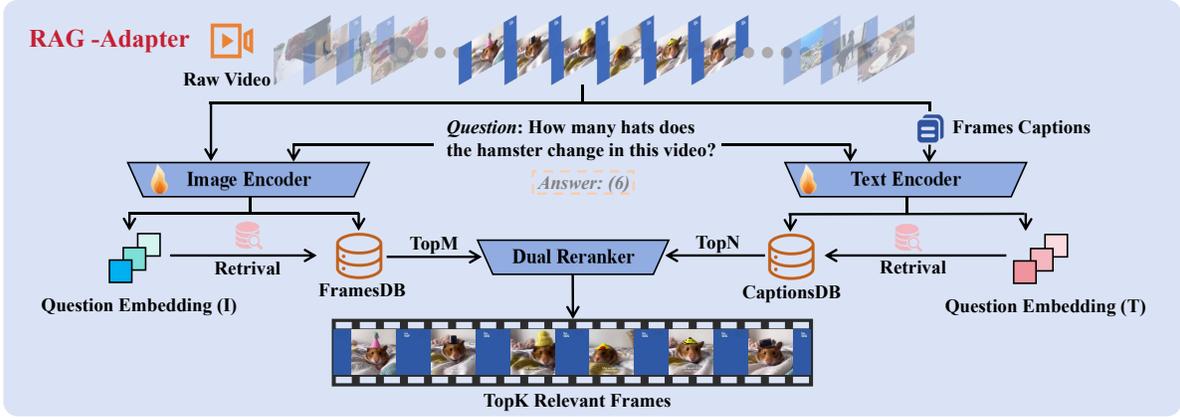


Figure 2. The RAG-Adapter pipeline framework. Given a video and a question, the video frames and corresponding captions are encoded separately using image and text encoders and stored in databases. The question is encoded and retrieved using the same encoders. The Dual Reranker module selects the TopK frames relevant to the question. Details are provided in Section 3.1. To improve retrieval performance, both encoders are fine-tuned using Grouped-supervised Contrastive Learning (GCL), as described in Section 3.2.

$\{f_i^X, c_i^X, s_i^X\}_{i=1}^X$. First, we add the frame with the highest summed score from \mathcal{U} to \mathcal{S} . For each remaining frame in \mathcal{U} , the one with the highest Marginal Relevance (MR) score, $i^* = \arg \max_{i \in \mathcal{U}} MR_i$, is then moved to \mathcal{S} . This step is repeated $K - 1$ times, producing K frames in \mathcal{S} , representing TopK relevant frames selected by RAG-Adapter. The MR_i formula is as follow:

$$MR_i = \theta \cdot s_i^X - (1 - \theta) \cdot \max_{j \in \mathcal{S}} [sim(f_i^X, f_j^X) + sim(c_i^X, c_j^X)] \quad (1)$$

θ is a penalty coefficient to balance the weights of the summed similarity score and diversity score, with $sim()$ computed via cosine similarity.

3.2. RAG-Adapter Fine-tuning

The text and image encoders in RAG-Adapter, BGE-M3 and CLIP-L/14, are trained on large-scale internet-based corpora. However, their embedding spaces may not be fully optimized for video understanding scenarios. To enhance RAG-Adapter’s performance in this domain, we construct a specialized dataset, **MMAT**, consisting of (Q_i, F_i) and (Q_i, C_i) positive pairs for contrastive learning fine-tuning of CLIP-L/14 and BGE-M3, respectively. Here, Q_i , F_i , and C_i denote the question, representative video frame, and corresponding caption for the i_{th} video.

MMAT Construction. We employ a contrastive learning-based fine-tuning method to better fit BGE-M3 and CLIP-L/14’s embedding spaces to the requirements of video understanding benchmarks. Given the challenge of identifying relevant frames in long videos, we start with widely used short video understanding benchmarks, including MSVD-QA [39], MSRVT-QA [39], ActivityNet-QA [43], and TGIF-QA [16], to construct the MMAT. To fully use the available videos, the training and validation sets from these

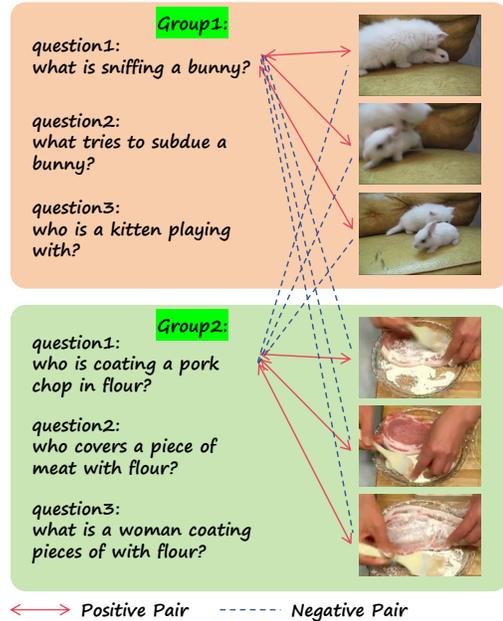


Figure 3. Illustration of Grouped-supervised Contrastive Learning (GCL) constructing positive and negative pairs.

benchmarks are combined to form the MMAT training set, while their test sets create the MMAT test set.

Since the videos in these benchmarks are typically short (usually under 10 seconds) with relatively consistent visual content, we sample frames at one frame per second and select three representative frames from quartile positions within each video. For each question related to the video, one of these frames is randomly chosen to construct the (Q_i, F_i) pairs. For each F_i , we use CogVLM2 to generate captions with detailed descriptions, thereby forming the corresponding (Q_i, C_i) pairs.

To ensure sampled frames align with questions despite potential content inconsistencies, we use a script to auto-

matically exclude videos over 300 seconds and manually filter out those with visibly inconsistent visuals.

We also observe occasional garbled text from CogVLM2 when generating captions for frames with repetitive characters. To address this, we use a script to detect and either regenerate or manually correct such captions, followed by a quick review to ensure semantic consistency with the video frames. These measures ensure the quality of MMAT, resulting in **417, 993** (Q_i, F_i) and (Q_i, C_i) pairs in the training set and **109, 799** pairs in the test set.

Grouped-supervised Contrastive Learning (GCL). In contrastive learning, a self-supervised loss is typically used, where only pairs like (Q_i, F_i) and (Q_i, C_i) are treated as positive samples, while all pairs (Q_i, F_j) and (Q_i, C_j) $\{i \neq j\}$ are treated as negative samples by default.

However, in video understanding scenarios, a single video may correspond to multiple questions. In Self-supervised Contrastive Learning, pairs like (Q_i, F_j) , (Q_i, C_j) $\{i \neq j\}$, which should be positive, may instead be treated as negative samples, disrupting training. As for Fully-supervised Contrastive Learning, label information is incorporated, but it requires the manual construction of negative pairs.

To address this, we propose the **Grouped-supervised Contrastive Learning (GCL)**, designed specifically for video understanding scenarios. In GCL, pairs from the same video are assigned a common group label. Within each group “G”, all possible pairs, such as (Q_i^G, F_j^G) and (Q_i^G, C_k^G) , are treated as positive, while pairs from different groups “G” and “G'”, such as $(Q_i^G, F_j^{G'})$ and $(Q_i^G, C_k^{G'})$, are treated as negative. GCL iterates through all combinations, computes loss values for each, and then averages them. Figure 3 presents an illustration of GCL, with the loss function shown as follows:

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(Q_i, F_p/C_p)/\tau)}{\sum_{a \in A'(i)} \exp(\text{sim}(Q_i, F_a/C_a)/\tau)} \quad (2)$$

$$\mathcal{L}_{GCL} = \frac{1}{|I|} \sum_{i \in I} \mathcal{L}_i \quad (3)$$

Here, I represents the set of all test questions, \mathcal{L}_i denotes the loss function for Q_i , and $P(i)$ is the set of positive samples associated with Q_i within the same group. The notation $A'(i) = A(i) \setminus \{p' \in P(i) \setminus \{p\}\}$ refers to the set of all batch samples $A(i)$, excluding all positive samples except the selected one, F_p or C_p . τ is the temperature coefficient.

This approach offers two main advantages: (1) it eliminates the need for manually constructing numerous negative pairs, and (2) by excludes all positive samples but the selected one F_p or C_p from L_i 's denominator, the model better captures detailed relationships between the question

Table 1. ASS and NIF metrics for evaluation benchmarks.

Benchmarks	ASS (0-2)			NIF
	Top10	Top30	Top50	
Video-MME	0.81	0.69	0.61	2.4
MLVU	0.76	0.61	0.52	2.9
Perception Test	0.74	0.69	0.68	1.9
EgoSchema	0.82	0.67	0.58	2.1

and each specific positive sample, facilitating a more refined understanding of each sample’s unique characteristics.

4. Experiments and Analysis

4.1. Evaluation Benchmarks

We select two commonly used long video understanding benchmarks, Video-MME and MLVU, as well as two relatively shorter benchmarks focusing on human interaction and perception in real-world scenarios: Perception Test [29] and EgoSchema [26], for evaluation. This selection aims to demonstrate RAG-Adapter’s generalization performance across benchmarks with varying temporal spans (ranging from approximately 0.5 minutes to several hours) and contexts. Due to the time-consuming process of generating captions for each video (as discussed in Table 5), we sample 90 videos from each benchmark to manage this. Video-MME videos are categorized by length (short, medium, long) and further divided into six content domains, from which we randomly select five videos per domain. MLVU videos are classified into nine types, from which ten videos are randomly sampled per category. For Perception Test, the 90 longest videos from the Multiple-Choice Video Q&A task are selected. In Egoschema, where all videos are 3 minutes long, 90 videos are randomly sampled.

4.2. Statistics of ASS and NIF

Using the RAG-Adapter, we calculate the ASS and NIF metrics for all evaluation benchmarks.

For ASS, we measure the average summed score, $ASS = \frac{1}{n} \sum_{i=1}^K s^i$ (on a 0-2 scale), between all questions and their Top K relevant frames with captions, where K set to 10, 30, and 50. For NIF, we manually identify the minimum number of frames containing essential visual information needed to answer each question (note: for Video-MME, some information is also found in subtitle files). The NIF value is the average of these frame counts across all questions. Results are summarized in Table 1.

The ASS values for the four benchmarks are similar, as the top K frames retrieved by the RAG-Adapter in each benchmark show little variation in relevance to the questions. Additionally, as K increases, the overall relevance tends to decrease. MLVU has the highest NIF value due to the greater frame requirement for Action Order and Video Summarization tasks. Both MLVU and Video-MME show

Table 2. The test results for various MLLMs on Video-MME include accuracy metrics for 6 domains, along with the overall **Avg. Acc.** (Average Accuracy). The highest accuracy is **bolded**, and the second highest is underlined.

Models	Sampling Method	Category						Avg. Acc. (%)
		Knowledge	Film & Television	Sports Competition	Artistic Performance	Life Record	Multilingual	
<i>Image MLLMs</i>								
Otter-I [19]	Uniform	28.9	28.9	33.3	33.3	24.4	33.3	30.4
	RAG-Adapter	37.8 (+8.9)	37.8 (+8.9)	42.2 (+8.9)	40.0 (+6.7)	28.9 (+4.5)	31.1 (-2.2)	36.3 (+5.9)
LLaVA-1.6 [23]	Uniform	33.3	22.2	33.3	44.4	26.7	24.4	30.7
	RAG-Adapter	35.6 (+2.3)	28.9 (+6.7)	37.8 (+4.5)	48.9 (+4.5)	31.1 (+4.4)	31.1 (+6.7)	35.6 (+4.9)
GPT4-Turbo [1]	Uniform	60.0	<u>71.1</u>	48.9	57.8	53.3	55.6	57.8
	RAG-Adapter	<u>71.1 (+11.1)</u>	<u>71.1 (+0.0)</u>	51.1 (+2.2)	60.0 (+2.2)	53.3 (+0.0)	60.0 (+4.4)	61.1 (+3.3)
<i>Video MLLMs</i>								
Otter-V [19]	Uniform	31.1	28.9	33.3	22.2	31.1	26.7	28.9
	RAG-Adapter	37.8 (+6.7)	31.1 (+2.2)	35.6 (+2.3)	28.9 (+6.7)	37.8 (+6.7)	31.1 (+4.4)	33.7 (+4.8)
mPlug-Owl-V [41]	Uniform	22.2	31.1	24.4	28.9	17.8	24.4	24.8
	RAG-Adapter	35.6 (+13.4)	37.8 (+6.7)	28.9 (+4.5)	31.1 (+2.2)	26.7 (+8.9)	28.9 (+4.5)	31.5 (+6.7)
MovieChat [35]	Uniform	24.4	28.9	22.2	31.1	28.9	22.2	26.2
	RAG-Adapter	33.3 (+8.9)	35.6 (+6.7)	33.3 (+11.1)	31.1 (+0.0)	28.9 (+6.7)	35.6 (+6.7)	33.0 (+6.7)
VideoChat [20]	Uniform	26.7	28.9	33.3	26.7	37.8	22.2	29.3
	RAG-Adapter	31.1 (+4.4)	35.6 (+6.7)	33.3 (+0.0)	33.3 (+6.6)	40.0 (+2.2)	33.3 (+11.1)	34.4 (+5.1)
VideoChat2 [21]	Uniform	33.3	17.8	24.4	35.6	44.4	28.9	30.7
	RAG-Adapter	42.2 (+8.9)	22.2 (+4.4)	26.7 (+2.3)	37.8 (+2.2)	42.2 (-2.2)	31.1 (+2.2)	33.7 (+3.0)
LLaMA-VID [22]	Uniform	31.1	17.8	24.4	37.8	22.2	26.7	26.7
	RAG-Adapter	31.1 (+0.0)	26.7 (+8.9)	33.3 (+8.9)	37.8 (+0.0)	28.9 (+6.7)	28.9 (+2.2)	31.1 (+4.4)
TimeChat [31]	Uniform	31.1	33.3	28.9	46.7	31.1	26.7	33.0
	RAG-Adapter	33.3 (+2.2)	42.2 (+8.9)	31.1 (+2.2)	48.9 (+2.2)	33.3 (+2.2)	33.3 (+6.6)	37.0 (+4.0)
Chat-UniVi [17]	Uniform	33.0	26.7	24.4	37.8	31.1	24.4	29.6
	RAG-Adapter	42.2 (+8.9)	35.6 (+8.9)	35.6 (+11.2)	46.7 (+8.9)	42.2 (+11.1)	35.6 (+11.2)	39.7 (+10.1)
GPT-4o [28]	Uniform	66.7	62.2	66.7	64.4	55.6	53.5	61.5
	RAG-Adapter	77.8 (+11.1)	73.3 (+11.1)	68.9 (+2.2)	75.6 (+11.2)	68.9 (+13.3)	60.0 (+6.7)	70.8 (+9.3)

Table 3. Comparison across more benchmarks.

Models	Sampling Method	MLVU	Perception Test	EgoSchema
MovieChat [35]	Uniform	29.6	32.5	23.3
	RAG-Adapter	41.5 (+11.9)	37.8 (+5.3)	28.9 (+5.6)
LLaMA-VID [22]	Uniform	34.8	33.1	24.4
	RAG-Adapter	43.0 (+8.2)	37.2 (+4.1)	31.1 (+6.7)
TimeChat [31]	Uniform	37.8	37.8	27.8
	RAG-Adapter	45.2 (+7.4)	41.1 (+3.3)	32.2 (+4.4)
Chat-UniVi [17]	Uniform	32.6	38.1	32.2
	RAG-Adapter	40.0 (+7.4)	41.6 (+3.5)	41.1 (+8.9)

slightly higher NIF values than the other two benchmarks, given their longer average durations, though the number of frames containing essential information remains limited. Supplementary materials provide each test video’s ID, corresponding question (or ID), minimum frame count, frame timestamps and identified issues in the benchmarks using RAG-Adapter.

4.3. Baselines and Experimental Setups

We classify the baselines into two categories: image-based MLLMs supporting multiple image inputs and video MLLMs. Open-source models are tested locally on an NVIDIA 4090 GPU, while proprietary models are accessed via official APIs. Based on the NIF metrics (Table 1) and the maximum number of key frames—Video-MME (9), MLVU (20), Perception Test (8), and EgoSchema (6)—we set $K = 10$ frames for Video-MME, Perception Test, and EgoSchema, and $K = 20$ for MLVU to ensure sufficient information retrieval by RAG-Adapter and maintain evaluation fairness (M , N , and θ are set to 50, 50, and 0.7, respectively). Frames are input in chronological order to preserve temporal information. We compare each MLLM’s performance under identical frame input conditions, contrasting uniform sampling with RAG-Adapter sampling. The accuracy (ranging from 0 to 100) for multiple-choice questions across the four benchmarks is calculated by comparing the

Table 4. Comparison under different fine-tuning methods.

Models	Fine-Tuning Method	Avg. Acc. (%)
TimeChat [31]	No Fine-Tuning	33.7
	SCL	32.6 (-1.1)
	CB	34.8 (+1.1)
	GCL	37.0 (+3.3)
GPT-4o [28]	No Fine-Tuning	65.9
	SCL	65.2 (-0.7)
	CB	66.7 (+0.8)
	GCL	70.8 (+4.9)

predicted results with the ground truth.

4.4. Results and Analysis

Table 2 compares the performance of uniform sampling and RAG-Adapter sampling across six domains in the Video-MME (without subtitle information). Table 3 compares the performance on the remaining three benchmarks, and the more comprehensive experimental results for MLVU are provided in the supplementary materials. From the experimental results, we draw the following key conclusions:

- **Performance:** RAG-Adapter sampling improves overall performance across all models compared to uniform sampling, indicating that the information loss from uniform sampling adversely affects model performance in testing. Thus, uniform sampling does not fully reflect MLLMs’ true long video understanding capabilities.
- **Unified Improvement:** In Table 2, while commercial MLLMs outperform open-source models, RAG-Adapter consistently enhances performance. For example, GPT-4o shows accuracy gains exceeding 10% in the Knowledge, Film & Television, Artistic Performance, and Life Record domains, with an average accuracy increase of 9.3%. Models like mPLUG-Owl-V show a 13.4% improvement in Knowledge, with an overall accuracy increase of 6.7%, while Chat-UniVi achieves over a 10%

Table 5. Ablation Study of the RAG-Adapter Components, where “T&E” denotes Text Encoder, “I&E” represents Image Encoder, and “D&R” stands for the Dual Ranker.

Component Ablation								
Fine-Tuning Method	T&E	I&E	D&R	Avg. Acc. (%)	Preprocessing	Retrieval	Inference	Recall@10
GCL	✓			33.3	48.8min	4.7s	0.88s	19.7
		✓		34.5	11.4s	8.7s		18.7
	✓	✓		35.2		13.5s		24.8
	✓	✓	✓	39.7				30.4
No Fine-Tuning	✓	✓	✓	33.3	48.8min	15.4s		24.1
SCL	✓	✓	✓	32.2				23.7
CB	✓	✓	✓	35.6				25.1
Other Baselines								
Sampling Method				Avg. Acc. (%)	Preprocessing	Retrieval	Inference	Recall@10
Uniform				29.6	11.4s	N/A	0.88s	5.5
Two-stage Retrieval				<u>37.0</u>	4.3min	(8.5+2.5)s		<u>25.8</u>

Table 6. Comparison of different sampling strategies and input frame counts.

Models	Sampling Method	Frames Count	Avg. Acc. (%)
TimeChat [31]	Uniform	5	30.7
		10	33.0
		20	33.0
		64	32.2
	RAG-Adapter	5	36.3
		10	37.0
20		37.0	
Chat-UniVi [17]	Uniform	5	27.8
		10	29.6
		20	32.6
		256	31.9
	RAG-Adapter	5	35.6
		10	39.7
20		40.0	

Table 7. Comparison between no subtitles (w/o subs), subtitles corresponding to RAG-Adapter sampled frames (w/ subs (Corresp.)), and subtitles sampled by RAG-Adapter (w/ subs (RAG-Adapter)).

Models	Subtitles	Avg. Acc. (%)
TimeChat [31]	w/o subs	37.0
	w/ subs (Corresp.)	38.2 (+1.2)
	w/ subs (RAG-Adapter)	39.6 (+2.6)
Chat-UniVi [17]	w/o subs	39.7
	w/ subs (Corresp.)	40.0 (+0.3)
	w/ subs (RAG-Adapter)	41.5 (+1.8)

improvement in Sports Competition, Life Record, Multilingual, and overall accuracy increase. This demonstrates the versatility of RAG-Adapter, as its effectiveness is not directly linked to the intrinsic capabilities of the models.

- **Generalization:** In Table 3, Perception Test and Egochema have shorter average durations (35s&3min) compared to MLVU (12min), leading to a less pronounced improvement of RAG-Adapter over uniform sampling. Nonetheless, its performance across benchmarks of varying lengths demonstrates the method’s effectiveness and generalization.
- **Constraint:** In Table 2, RAG-Adapter does not consistently improve accuracy across all domains. GPT-4 Turbo shows no improvement in Film & Television, while MovieChat and LLaMA-VID remain unchanged in Artistic Performance. Otter-I and VideoChat2 experience a 2.2% accuracy drop in Multilingual and Life Record.

This stability or slight decline is primarily due to RAG-Adapter occasionally failing to retrieve all relevant information, resulting in the omission of key frames. In such cases, RAG-Adapter may mislead the model, affecting its accuracy. We aim to further refine the retrieval process of RAG-Adapter to minimize these issues.

4.5. Ablation Study

The following ablation experiments are conducted on the Video-MME benchmark, with additional results provided in the supplementary materials.

Effect of RAG-Adapter Fine-Tuning. In Table 4, we evaluate RAG-Adapter’s performance across different fine-tuning approaches, including No Fine-Tuning, Self-supervised Contrastive Learning (SCL) fine-tuning, Customizing Batch (CB) fine-tuning (where each question in a batch belongs to a different video), and Grouped-supervised Contrastive Learning (GCL) fine-tuning. The results demonstrate that GCL fine-tuning achieves superior performance for all models. This enhancement is primarily due to GCL’s ability to train RAG-Adapter’s text and image encoders to effectively learn rich positive sample features within each group while avoiding the adverse effects of treating intra-group samples as negatives, as seen in SCL. Moreover, GCL retains all inter-group negative samples from SCL and CB, ensuring robust learning.

Discussion on the effectiveness and efficiency of RAG-Adapter components. In Table 5, we utilize Chat-UniVi to conduct an ablation study on RAG-Adapter’s components, evaluating pipeline efficiency (scaled to 10-minute videos per query) and average recall. The preprocessing phase involves frame sampling and captioning. Our method achieves optimal accuracy and recall only with all components and GCL fine-tuning. To address the impracticality of captioning every frame in long videos, we propose a **Two-stage Retrieval** method: initially retrieving the top50 frames, then using captions to refine the selection to the final topK frames, achieving a favorable balance between accuracy and efficiency. Also, retrieval accuracy using only Image Encoder outperforms uniform sampling, offering an-



Figure 4. Comparison of RAG-Adapter and uniform sampling results: RAG-Adapter accurately identifies two consecutive key frames relevant to the question, whereas uniform sampling tends to miss them.

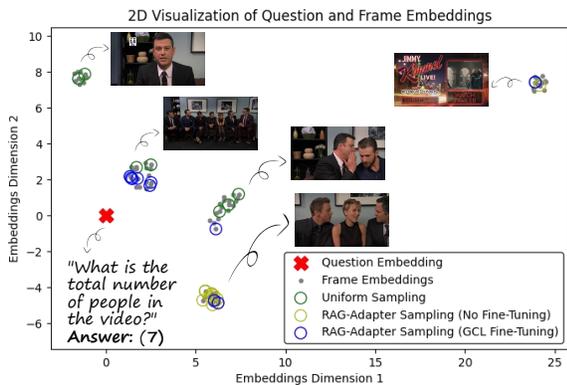


Figure 5. The relationship between the embedding spaces of video frames sampled using different methods and that of the corresponding questions. The frame embeddings are primarily grouped into five clusters, each representing a set of consecutive shots, with each cluster labeled by a representative frame.

other viable alternative in practice. Furthermore, inspired by SeViLA [42], we employ the tested MLLM for frame filtering but find it time-intensive and ineffective.

Comparison of Different Input Frame Counts. Since some long-video MLLMs can support a larger number of input frames, we compare uniform sampling (using 5, 10, 20 frames or the model’s maximum supported frames on a single NVIDIA 4090) with RAG-Adapter sampling (using 5, 10, and 20 frames), as shown in Table 6. Results indicate that, despite utilizing more frames, uniform sampling does not outperform RAG-Adapter and even exhibits slight performance degradation compared to using fewer frames. For RAG-Adapter sampling, performance improves from $K = 5$ to $K = 10$, suggesting information loss at $K = 5$, and stabilizes at $K = 20$. This aligns with the NIF metric for Video-MME, which is below 10, implying most questions require fewer than 10 frames to capture essential information. Additionally, increasing uniformly sampled frames does not guarantee inclusion of critical details and may in-

roduce greater redundancy.

Impact of Subtitle Information. In the Video-MME benchmark, subtitle files are available and contain some information relevant to certain questions. In Table 7, we examine the impact of subtitles on MLLMs using 10 frames sampled by RAG-Adapter. We evaluate two subtitle inclusion methods: providing subtitles directly correspond to the sampled frames and using RAG-Adapter to select the 10 most relevant subtitles for each question.

Our experiments reveal two main insights. First, model accuracy consistently improves with subtitle inclusion, as subtitles often provide question-relevant information. Second, subtitles filtered by RAG-Adapter outperform those directly tied to sampled frames, as critical subtitle information may not align with key video content, and complex questions often rely more heavily on subtitle data.

5. Visualization

5.1. Visualization of Frame Sampling Methods

In Figure 4, we compare the results of uniform sampling and RAG-Adapter sampling for the same question in the Video-MME benchmark. The specific scene referenced in the question - “How many people are shown having lunch with the woman in the video?”, occurs only between 73-74 seconds in the original video. As a result, uniform sampling fails to capture any relevant frames, whereas RAG-Adapter successfully identifies the two pertinent frames (sampled at one frame per second). Additional visualizations of the video frames are provided in the supplementary materials.

5.2. Differences of Embedding spaces

In Figure 5, we reduce the embedding space of the question and all video frames to two dimensions using UMAP [27] (Uniform Manifold Approximation and Projection) to preserve the global structure of the data. This visualization illustrates the spatial relationship between the question embedding and the embeddings of frames sampled by uniform

sampling, the non-fine-tuned RAG-Adapter, and the GCL fine-tuned RAG-Adapter. It can be observed that the embeddings of uniformly sampled frames are highly scattered, while the embeddings of frames sampled by the non-fine-tuned RAG-Adapter cluster around a few similar frames. In contrast, the embeddings from the GCL fine-tuned RAG-Adapter exhibit greater diversity and are closer to the question embedding.

6. Conclusion

In this paper, we integrate the RAG framework with MLLMs, introducing RAG-Adapter, a plugin that enhances the long video understanding capabilities of MLLMs without modifying their internal structure. By providing question-relevant video frames during testing, RAG-Adapter ensures that the evaluation of long video understanding benchmarks accurately reflects the model’s true video comprehension capabilities. To better adapt RAG-Adapter to the video question-answering context, we construct a fine-tuning dataset, MMAT, and introduce Grouped-supervised Contrastive Learning (GCL) to help RAG-Adapter learn rich and relevant embedding between questions and video frames. Additionally, we proposed two metrics, ASS and NIF, to assess the benchmarks quality and complexity, using NIF as a basis for determining the number of frames sampled by RAG-Adapter. Tests on Video-MME, MLVU, Perception Test and EgoSchema demonstrate that RAG-Adapter consistently improves accuracy across all baseline MLLMs, demonstrating our approach’s simplicity and effectiveness.

Limitations. RAG-Adapter does not always retrieve all relevant frames, especially when key information is dispersed across multiple segments, often returning only a subset. Additionally, complex tasks like sentiment analysis or video summarization, which lack explicit visual cues, may further constrain its effectiveness. Moreover, the substantial preprocessing time required to encode video data into the database makes RAG-Adapter unsuitable for real-time video processing. While the proposed Two-Stage Retrieval and purely visual retrieval strategies mitigate this issue, future work will focus on further optimizing retrieval efficiency.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [3] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. irag: An incremental retrieval augmented generation system for videos. *arXiv preprint arXiv:2404.12309*, 2024. 3
- [4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. 3
- [5] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 2
- [6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2
- [7] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3
- [8] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998. 3
- [9] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024. 3
- [10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2
- [11] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021. 3
- [12] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1, 3
- [13] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 2
- [14] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-llm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 2
- [15] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 3
- [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. 4
- [17] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 3, 6, 7
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 3
- [19] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: a multi-modal model with in-context instruction tuning. corr abs/2305.03726 (2023), 2023. 2, 6
- [20] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhao Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 6
- [21] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6
- [22] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 1, 3, 6
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 6
- [24] Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv preprint arXiv:2403.13805*, 2024. 3
- [25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2
- [26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 5
- [27] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8
- [28] OpenAI. Gpt-4o, 2024. 6
- [29] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 5
- [30] 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. 3
- [31] 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. 6, 7
- [32] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [33] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2
- [34] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. Fairrag: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12005, 2024. 3
- [35] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 1, 3, 6
- [36] Ayush Thakur and Rashmi Vashisth. Loops on retrieval augmented generation (lorag). *arXiv preprint arXiv:2403.15450*, 2024. 3
- [37] Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024. 3
- [38] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*, 2024. 2
- [39] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 4

- [40] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. [2](#)
- [41] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. [6](#)
- [42] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. [8](#)
- [43] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. [4](#)
- [44] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. [1](#), [3](#)

RAG-Adapter: A Plug-and-Play RAG-enhanced Framework for Long Video Understanding

Supplementary Material

Appendix Contents

A. Training Hyperparameters	1
B. Additional Experiments	1
B.1. Comparison of more MLLMs under different fine-tuning methods	1
B.2. Comparison of more MLLMs across different input frame numbers and sampling strategies	1
B.3. Comparison of more MLLMs under different subtitle input conditions	1
B.4. More Comparison Results on MLVU	2
C. Unreasonable Questions in the Benchmark	3
D. Additional Visualization Results	6
E. Detailed NIF Statistics	8

A. Training Hyperparameters

Table 8 provides the hyperparameters used for fine-tuning the image encoder (BGE-M3) and text encoder (CLIP-L/14) in RAG-Adapter. Both Self-supervised Contrastive Learning and Grouped-supervised Contrastive Learning use the same hyperparameter configurations.

Table 8. Training Hyperparameters for BGE-M3 and CLIP-L/14.

Hyperparameter	Encoders	
	BGE-M3	CLIP-L/14
Batch size	32	32
Fine-tuning epochs	2	2
Fine-tuning iterations	26126	26126
Temperature	20	20
Weight decay	0.01	0.01
Learning rate	2e-5	1e-5
Warm-up iterations	2612	2612
Optimizer	AdamW	AdamW
Schedule	linear decay	cosine decay
AdamW β_1	0.9	0.9
AdamW β_2	0.999	0.98
AdamW ϵ	1e-6	1e-6

B. Additional Experiments

B.1. Comparison of more MLLMs under different fine-tuning methods

Table 9. Comparison under different fine-tuning methods.

Models	Fine-Tuning Method	Avg. Acc. (%)
MovieChat [35]	No Fine-Tuning	29.2
	SCL	28.5 (-0.7)
	CB	29.3 (+0.1)
	GCL	33.0 (+3.8)
LLaMA-VID [22]	No Fine-Tuning	28.9
	SCL	28.9 (+0.0)
	CB	29.6 (+0.7)
	GCL	31.1 (+2.2)

B.2. Comparison of more MLLMs across different input frame numbers and sampling strategies

Table 10. Comparison of different sampling strategies and input frame counts.

Models	Sampling Method	Frames Count	Avg. Acc. (%)
MovieChat [35]	Uniform	5	25.6
		10	26.3
		20	27.0
		512	28.9
	RAG-Adapter	5	30.0
		10	33.0
LLaMA-VID [22]	Uniform	20	33.3
		5	26.7
		10	26.7
		20	27.1
	RAG-Adapter	512	27.8
		5	28.9
		10	31.1
		20	30.8

B.3. Comparison of more MLLMs under different subtitle input conditions

Table 11. Comparison between no subtitles (w/o subs), subtitles corresponding to RAG-Adapter sampled frames (w/ subs (Corresp.)), and subtitles sampled by RAG-Adapter (w/ subs (RAG-Adapter)).

Models	Subtitles	Avg. Acc. (%)
MovieChat [35]	w/o subs	33.0
	w/ subs (Corresp.)	34.1 (+1.1)
	w/ subs (RAG-Adapter)	34.8 (+1.8)
LLaMA-VID [22]	w/o subs	31.1
	w/ subs (Corresp.)	31.9 (+0.8)
	w/ subs (RAG-Adapter)	32.2 (+1.1)

B.4. More Comparison Results on MLVU

Due to page limitations in the main body of the paper, we have included the comprehensive experiments on MLVU mentioned in Section 4.4 in the supplementary materials. Table 12 presents the performance of various MLLMs on the MLVU benchmark using uniform sampling versus RAG-Adapter sampling, with all models evaluated using 20 frames. The results are consistent with those on the Video-MME benchmark: RAG-Adapter sampling improves the performance of all MLLMs compared to uniform sampling. For M-Avg, GPT-4o achieves the highest improvement of 12.6%, while VideoChat2 shows the lowest gain of 6.6%. However, for G-Avg, the improvements are generally modest, with even a slight decline (e.g., VideoChat decreases by 0.03). This is because generative tasks require a more comprehensive understanding of the video content, meaning that the sampled frames must adequately represent the entire video. In such scenarios, RAG-Adapter sampling offers no significant advantage over uniform sampling.

Table 12. The test results for various MLLMs on MLVU. The evaluation includes nine types of tasks: PQA (Plot QA), NQA (Needle QA), ER (Ego Reasoning), AC (Action Count), AO (Action Order), AR (Anomaly Recognition), TR (Topic Reasoning), SSC (Sub-Scene Captioning), and VS (Video Summary). “M-Avg” (0-100) represents the average performance across multiple-choice tasks, while “G-Avg” (0-10, marked by *) indicates the average performance for generative tasks.

Models	Sampling Method	Category									M-Avg	G-Avg
		PQA	NQA	ER	AC	AO	AR	TR	SSC*	VS*		
<i>Image MLLMs</i>												
Otter-I	Uniform	31.4	20.8	34.4	20.0	40.0	40.0	40.0	2.15	1.10	31.8	1.63
	RAG-Adapter	34.3 (+2.9)	54.2 (+33.4)	46.9 (+12.5)	10.0 (-10.0)	50.0 (+10.0)	40.0 (+0.0)	53.3 (+13.3)	2.05 (-0.1)	1.35 (+0.25)	43.0 (+11.2)	1.70 (+0.07)
LLaVA-1.6	Uniform	34.3	29.2	34.4	20.0	20.0	50.0	73.3	1.30	1.05	37.0	1.18
	RAG-Adapter	57.1 (+22.8)	50.0 (+20.8)	34.4 (+0.0)	20.0 (+0.0)	30.0 (+10.0)	70.0 (+20.0)	66.7 (-6.6)	1.95 (+0.65)	1.30 (+0.25)	48.1 (+11.1)	1.63 (+0.45)
<i>Video MLLMs</i>												
Otter-V	Uniform	28.6	21.7	18.8	20.0	40.0	30.0	33.3	2.20	1.10	26.1	1.65
	RAG-Adapter	31.4 (+2.8)	37.5 (+15.8)	28.1 (+9.3)	40.0 (+20.0)	40.0 (+0.0)	40.0 (+10.0)	40.0 (+6.7)	2.25 (+0.05)	1.20 (+0.10)	34.8 (+8.7)	1.73 (+0.08)
mPlug-Owl-V	Uniform	25.7	33.3	37.5	40.0	20.0	40.0	26.7	2.15	1.10	31.8	1.63
	RAG-Adapter	34.3 (+8.6)	50.0 (+16.7)	40.6 (+3.1)	50.0 (+10.0)	40.0 (+20.0)	50.0 (+10.0)	40.0 (+13.3)	2.80 (+0.65)	1.15 (+0.05)	42.2 (+10.4)	1.98 (+0.35)
MovieChat	Uniform	25.7	29.2	25.0	30.0	30.0	20.0	53.3	1.40	1.05	29.6	1.23
	RAG-Adapter	42.9 (+17.2)	37.5 (+8.3)	34.4 (+9.4)	40.0 (+10.0)	40.0 (+10.0)	50.0 (+30.0)	53.3 (+0.0)	1.45 (+0.05)	1.10 (+0.05)	41.5 (+11.9)	1.28 (+0.05)
VideoChat	Uniform	22.9	16.7	25.0	30.0	20.0	40.0	26.7	2.25	1.40	24.5	1.83
	RAG-Adapter	31.4 (+8.5)	33.3 (+16.6)	25.0 (+0.0)	40.0 (+10.0)	40.0 (+20.0)	50.0 (+10.0)	33.3 (+6.6)	2.30 (+0.05)	1.30 (-0.10)	33.3 (+8.8)	1.80 (-0.03)
VideoChat2	Uniform	34.3	29.2	21.9	30.0	20.0	40.0	26.7	2.25	1.15	28.9	1.70
	RAG-Adapter	37.1 (+2.8)	37.5 (+8.3)	31.3 (+9.4)	40.0 (+10.0)	30.0 (+10.0)	40.0 (+0.0)	33.3 (+6.6)	2.65 (+0.40)	1.20 (+0.05)	35.5 (+6.6)	1.93 (+0.23)
LLaMA-VID	Uniform	25.7	33.3	40.6	40.0	20.0	50.0	40.0	2.45	1.25	34.8	1.85
	RAG-Adapter	31.4 (+5.7)	37.5 (+4.2)	43.8 (+3.2)	50.0 (+10.0)	20.0 (+0.0)	70.0 (+20.0)	66.7 (+26.7)	2.65 (+0.20)	1.25 (+0.00)	43.0 (+8.2)	1.95 (+0.10)
TimeChat	Uniform	34.3	41.7	40.6	40.0	40.0	20.0	40.0	1.69	1.10	37.8	1.40
	RAG-Adapter	42.9 (+8.6)	54.2 (+12.5)	40.6 (+0.0)	50.0 (+10.0)	60.0 (+20.0)	20.0 (+0.0)	46.7 (+6.7)	1.85 (+0.16)	1.05 (-0.05)	45.2 (+7.4)	1.45 (+0.05)
Chat-UniVi	Uniform	34.3	37.5	21.9	30.0	20.0	60.0	33.3	2.75	1.20	32.6	1.98
	RAG-Adapter	37.1 (+2.8)	45.8 (+8.3)	28.1 (+6.2)	50.0 (+20.0)	30.0 (+10.0)	60.0 (+0.0)	46.7 (+13.4)	2.95 (+0.20)	1.20 (+0.00)	40.0 (+7.4)	2.08 (+0.10)
GPT-4o	Uniform	54.3	54.2	37.5	40.0	50.0	50.0	53.3	1.40	1.55	48.9	1.48
	RAG-Adapter	62.9 (+8.6)	70.8 (+16.6)	50.0 (+12.5)	50.0 (+10.0)	60.0 (+10.0)	80.0 (+30.0)	60.0 (+6.7)	2.35 (+0.95)	1.85 (+0.30)	61.5 (+12.6)	2.10 (+0.62)

C. Unreasonable Questions in the Benchmark

While using RAG-Adapter to assist in calculating the NIF values for benchmarks, we identified a few unreasonable questions, detailed in Figures 6 to 11. Despite these issues, both benchmarks maintain high overall quality. This suggests that RAG-Adapter is an effective tool for evaluating and refining long video benchmarks, significantly reducing manual verification effort and enhancing benchmark quality.

Video-MME

Video ID: dH8l--46j6s

Question ID: 214-3

Question and Options:

What does the male performer wear in this video?

A. Black pants and white shorts. B. Black pants with a naked upper body.
C. Black pants with a naked upper body. D. Black pants with a naked upper body.

Answer: (B)

Issue: Options B, C, and D are identical.

Video ID: zNxi2s36tSO

Question ID: 163-3

Question and Options:

What is the shortest time to reach the finish line in the video?

A. 9.5 seconds. B. 10.06 seconds. C. 8.7 seconds. D. 8.5 seconds.

Answer: (B)

Issue: Options C and D are unreasonable; the world record for the 100-meter sprint is 9.58 seconds.

Figure 6. Issues identified in Video-MME during NIF calculations using RAG-Adapter.

MLVU

Video ID: subPlot_new_all_97

Question:

Please describe the situation after a woman in white riding a horse is shot by an arrow and falls off the horse.

Answer:

The man in front of the city gate looks back at the man in yellow clothes on the city tower, then turns back and runs towards the woman in white riding the horse.

Issue: The scene where "a woman in white riding a horse is shot by an arrow" does not appear in the video.

Video ID: needle_51

Question and Options:

What nationality are the kids having fun in the paddy field?

A. American B. Malays C. Chinese D. Indian

Answer: (B)

Issue: It is not possible to determine the children's nationality solely from the frames.

Figure 7. Issues identified in MLVU during NIF calculations using RAG-Adapter.

EgoSchema

Video ID: 51688142-10e7-48ab-adeb-2caa5448b456

Question and Options:

How does the introduction of the palm frond contribute to the development of the final product, and why might c have chosen to include it?

"option 0": "The palm frond contributes to the development of the final product by providing a structural element.",

"option 1": "The versatile palm frond significantly contributes to the development of the final product by providing a highly functional and essential element.",

"option 2": "The palm frond contributes to the development of the final product by providing a decorative element.",

"option 3": "The palm frond significantly contributes to the ultimate development of the final product by reliably providing a crucial protective element.",

"option 4": "The palm frond significantly contributes to the development of the final product by providing an essential nutritional element, enriching its value."

Answer: 2

Issue: Palm does not appear in the video.

Figure 8. Issue 1 identified in EgoSchema during NIF calculations using RAG-Adapter.

EgoSchema

Video ID: dfb6c468-e124-40f6-9c4e-c13ee45a2ad9

Question and Options:

What is the overarching goal of the actions taken by both c and the man throughout the video, and how do their techniques differ?

"option 0": "The primary, overarching goal of the various actions taken by both c and the man in the video is to efficiently repair a broken light fixture together.",

"option 1": "The primary, overarching goal driving the actions taken by both individual c and the man appearing throughout the entire video is to successfully remove a specific light fixture.",

"option 2": "The overarching goal of the actions taken by both c and the man throughout the video is to clean a light fixture.",

"option 3": "The primary, overarching goal of the various actions taken by both c and the man throughout the entire video is simply to replace a malfunctioning light bulb.",

"option 4": "The overarching goal of the actions taken by both c and the man throughout the video is to install a new light fixture."

Answer: 4

Issue: The goal is to demolish a wall, not to install a new light fixture.

Figure 9. Issue 2 identified in EgoSchema during NIF calculations using RAG-Adapter.

EgoSchema

Video ID: ece69b04-2e67-434a-b923-1329feed590d

Question and Options:

What is the primary objective *c* is trying to achieve in the video, and how does their interaction with various materials (ruler, cutter, craft pieces, glue, etc.) contribute to that objective?

"option 0": "*C* is trying to build a model. they use the ruler to measure the craft pieces, the cutter to cut them out, the glue to put them together, and the paper to decorate them.",

"option 1": "*C* is trying to make a craft. they use the ruler to measure the craft pieces, the cutter to cut them out, the glue to put them together, and the paper to decorate them.",

"option 2": "*C* is diligently attempting to construct a structure. they skillfully utilize the ruler for measuring the various craft elements, employ the cutter to accurately shape them, apply the adhesive glue for securely assembling, and utilize the colorful paper to aesthetically decorate them.",

"option 3": "Creatively, *c* is attempting to make an imaginative toy. diligently, they use the ruler to precisely measure the craft pieces, the cutter to skillfully cut them out, the glue to securely put them together, and the vibrant paper to beautifully decorate them.",

"option 4": "*C* is attempting to create a thoughtful gift. they utilize the ruler to precisely measure the craft pieces, employ the cutter to shape them, the adhesive glue to assemble them securely, and the decorative paper to enhance their appearance."

Answer: 0

Issue: The video does not contain a cutter or decorative paper.

Figure 10. Issue 3 identified in EgoSchema during NIF calculations using RAG-Adapter.

EgoSchema

Video ID: e67de76c-1058-49a7-a47e-12736da4ffc0

Question and Options:

Based on the actions performed by the woman and *c*, determine the critical moments in the video when they accomplish their respective tasks and care for the bearded dragon and the playing cards. how do these moments illustrate their priorities?

"option 0": "The critical moments in the video are when *c* picks up the bottle from the table, touches the cards on the table, and picks up the cards from the table. these moments illustrate his priority of playing with the cards.",

"option 1": "The critical moments in the video are when the woman and *c* are both playing with the cards. these moments illustrate their shared priority of playing with the cards.",

"option 2": "The critical moments in the video are when the woman and *c* are both caring for the bearded dragon. these moments illustrate their shared priority of caring for the bearded dragon.",

"option 3": "The critical moments in the video are when the woman plays with the bearded dragon and *c* plays with the cards. these moments illustrate their different priorities.",

"option 4": "The critical moments in the video are when the woman feeds the bearded dragon water, cleans it, and puts it in a container. these moments illustrate her priority of caring for the bearded dragon."

Answer: 4

Issue: The woman in the video does not place the bearded dragon into a container.

Figure 11. Issue 4 identified in EgoSchema during NIF calculations using RAG-Adapter.

D. Additional Visualization Results

Figures 12 to 16 present additional comparisons between uniform sampling of 10 frames and RAG-Adapter sampling of 10 frames on the Video-MME benchmark. The results demonstrate that RAG-Adapter more accurately identifies video frames relevant to the question. In contrast, uniform sampling often misses these critical frames, resulting in MLLMs lacking essential information when answering questions.

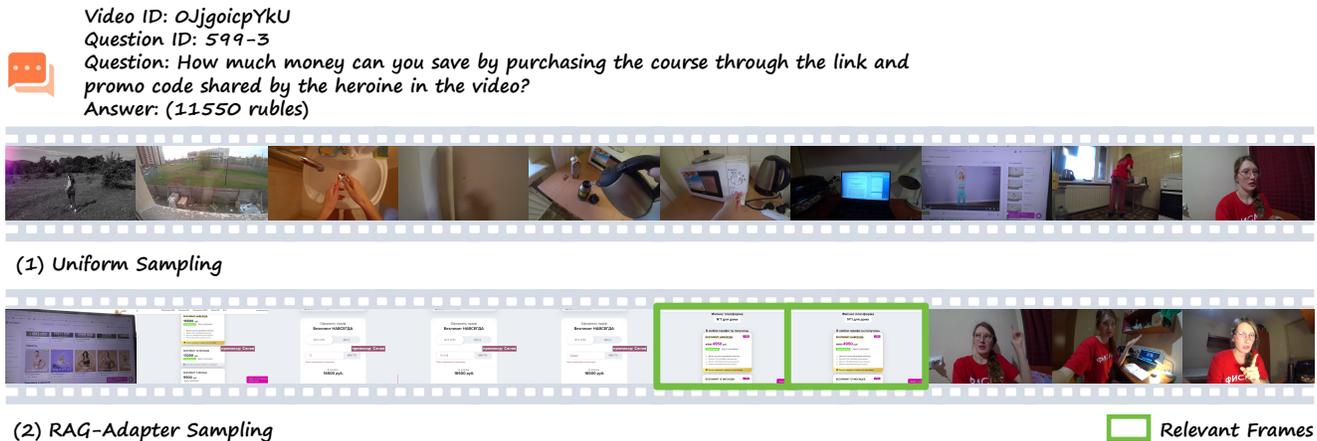


Figure 12. Comparison of RAG-Adapter and uniform sampling results: The answer to the question appears at the 527th and 528th seconds of the video, showing the original course price as 16,500 rubles and the current price as 4,950 rubles, resulting in a total saving of 11,550 rubles. RAG-Adapter accurately identifies these two consecutive key frames, while uniform sampling tends to miss them.



Figure 13. Comparison of RAG-Adapter and uniform sampling results: The answer to the question appears between the 68th and 73rd seconds of the video, where a man picks up a black plastic bag to clean up after his dog. RAG-Adapter accurately identifies one key frame depicting this action (the rest of the frames contain highly similar content), whereas uniform sampling misses it.



Video ID: 1wzgMHRkrys
 Question ID: 779-2
 Question: Who is interviewed both before and after the race based on the video?
 Answer: (Jake Gagne.)



(1) Uniform Sampling



(2) RAG-Adapter Sampling

Relevant Frames

Figure 14. Comparison of RAG-Adapter and uniform sampling results: The answer to the question appears between 270-294 seconds and 2329-2414 seconds of the video, where a reporter interviews Jake Gagne both before and after the race. RAG-Adapter accurately identifies frames at 280s, 284s, and 285s before the race, and 2334s after the race, showing the interview with Jake Gagne. The frames at 284s and 285s explicitly display Jake Gagne's name. In contrast, uniform sampling only captures a frame at 2428s, which shows an interview with another competitor, missing the key moments relevant to Jake Gagne.



Video ID: 81NyR6UvxU
 Question ID: 307-2
 Question: Which of the following features can not describe Spartacus?
 Answer: (Thick beard.)



(1) Uniform Sampling



(2) RAG-Adapter Sampling

Relevant Frames

Figure 15. Comparison of RAG-Adapter and uniform sampling results: The answer to the question appears at the 39s and 40s of the video, clearly displaying the name Spartacus and his appearance (notably without a thick beard). RAG-Adapter accurately identifies these two key frames, whereas uniform sampling fails to capture them, missing essential visual details.



Video ID: 40BIVzjxu-1
 Question ID: 006-1
 Question: What is one of the symbols of the festival that is introduced by the video?
 Answer: (Shamrock.)



(1) Uniform Sampling



(2) RAG-Adapter Sampling

Relevant Frames

Figure 16. Comparison of RAG-Adapter and uniform sampling results: The Shamrock logo appears in the video at 7-8 seconds, 23-25 seconds, 42-44 seconds, 53-55 seconds, and 95-105 seconds. Despite its frequent appearance, uniform sampling fails to capture any of these key frames, whereas RAG-Adapter successfully identifies key frames at 42-43 seconds, 54-55 seconds, and the 101st second.

E. Detailed NIF Statistics

In Section 1, we propose the concept of the Necessary Information Frame (NIF), which represents the average minimum number of frames containing essential information needed to answer each question. Table 1 in the paper presents the NIF values for test benchmarks. To better validate the NIF metric, we manually collect the necessary frames (NIF value for each question) and their corresponding timestamps (in seconds) for all questions across four benchmarks. Figures 17 to 22 illustrate the statistics for Video-MME, including the question IDs and corresponding video IDs. Figures 23 to 27 show similar data for MLVU, including partial question content, as MLVU lacks specific question IDs, along with the corresponding task and video ID. Figures 28 to 34 and Figures 35 and 36 correspond to the relevant data for Perception Test and EgoSchema, respectively.

video_id	question_id	NIF	timestamp
fFYNmVb3NCQ	598-2	4	298;432;494;518
fFYNmVb3NCQ	598-3	2	61;233
101TfTrEnss	895-1	3	48;188;306
101TfTrEnss	895-2	1	160
101TfTrEnss	895-3	1	101
xIWaK92gR1o	896-1	3	317;422;445
xIWaK92gR1o	896-2	4	164;300;793;1254
xIWaK92gR1o	896-3	4	158;330;656;1335
K9MQATj3894	898-1	2	512;1340
K9MQATj3894	898-2	2	19;28
K9MQATj3894	898-3	2	1051;1155
uuCVnqV4cNc	891-1	6	98;1187;1240;1872;1966;2963
uuCVnqV4cNc	891-2	4	1422;2278;2334;2439
uuCVnqV4cNc	891-3	6	11;104;405;1560;1873;2601
5K1S-p5eYH8	900-1	2	1088;1509
5K1S-p5eYH8	900-2	8	1;47;360;913;1043;1500;1973;2020
5K1S-p5eYH8	900-3	5	338;342;1406;1507;2178
1wzgmHrkrys	779-1	1	2015
1wzgmHrkrys	779-2	2	286;2342
1wzgmHrkrys	779-3	1	540

Figure 17. The NIF values for each question in Video-MME.

video_id	question_id	NIF	timestamp
40B1Vzjxu-I	006-1	1	100
40B1Vzjxu-I	006-2	3	1;73;90
40B1Vzjxu-I	006-3	2	0;1
Qyg_91gNHcc	051-1	3	6;73;78
Qyg_91gNHcc	051-2	2	26;108
Qyg_91gNHcc	051-3	1	75
Hv jgQqNOq9A	074-1	4	54;69;79;88
Hv jgQqNOq9A	074-2	1	9
Hv jgQqNOq9A	074-3	1	1
nYLMNQ77FjM	052-1	1	28
nYLMNQ77FjM	052-2	1	42
nYLMNQ77FjM	052-3	1	78
WViSvPFUVd8	079-1	2	0;1
WViSvPFUVd8	079-2	1	39
WViSvPFUVd8	079-3	6	8;12;18;28;45;50
DI6SemRT2iY	357-1	1	29
DI6SemRT2iY	357-2	1	73
DI6SemRT2iY	357-3	1	270
kSBB5PsRV-k	303-1	2	91;96
kSBB5PsRV-k	303-2	3	86;170;251
kSBB5PsRV-k	303-3	3	86;170;251
811NyR6UvxU	307-1	2	40;45
811NyR6UvxU	307-2	2	41;79
811NyR6UvxU	307-3	1	224
t61W12HVvFo	388-1	9	9;12;16;26;34;40;45;56;60
t61W12HVvFo	388-2	1	336
t61W12HVvFo	388-3	1	579
IaTaaNino1U	316-1	1	12
IaTaaNino1U	316-2	2	269;280
IaTaaNino1U	316-3	1	301
j0J-favyUeQ	688-1	4	61;916;1422;2861
j0J-favyUeQ	688-2	3	924;1821;2854
j0J-favyUeQ	688-3	3	731;1422;2226
zxKPjD8urG4	607-1	3	103;209;1379
zxKPjD8urG4	607-2	1	780
zxKPjD8urG4	607-3	2	1643;1646
FQd5bo9nIZs	632-1	3	98;419;1175
FQd5bo9nIZs	632-2	2	150;199
FQd5bo9nIZs	632-3	1	1175
9bbWYVrQgZ8	636-1	5	103;158;337;538;1755
9bbWYVrQgZ8	636-2	4	1452;1453;1454;1455
9bbWYVrQgZ8	636-3	6	18;887;1022;1656;1745;2054
y2kg3M0k1sY	680-1	5	326;739;825;1706;2281
y2kg3M0k1sY	680-2	4	327;741;2149;2826
y2kg3M0k1sY	680-3	2	1451;1452
tGdL-34L-GE	118-1	1	52
tGdL-34L-GE	118-2	1	40
tGdL-34L-GE	118-3	1	33
43wqf_KhiUo	121-1	8	43;44;45;46;47;48;49;50
43wqf_KhiUo	121-2	1	1

Figure 18. The NIF values for each question in Video-MME.

video_id	question_id	NIF	timestamp
43wqf_KhiUo	121-3	6	18;19;20;21;22;23
y6ReUXtm_VE	126-1	2	3;18
y6ReUXtm_VE	126-2	1	4
y6ReUXtm_VE	126-3	1	21
drbi6HK1gSc	093-1	4	3;7;11;21
drbi6HK1gSc	093-2	1	16
drbi6HK1gSc	093-3	1	76
PU-XOFIJMlg	098-1	3	11;17;27
PU-XOFIJMlg	098-2	2	91;96
PU-XOFIJMlg	098-3	1	11
xr_nln2ZQw8	412-1	2	240;243
xr_nln2ZQw8	412-2	3	37;159;260
xr_nln2ZQw8	412-3	6	63;90;326;573;649;717
-XpJeDGh8No	395-1	2	124;131
-XpJeDGh8No	395-2	4	270;338;358;413
-XpJeDGh8No	395-3	3	565;586;598
V6ui161NyTg	394-1	3	10;53;127
V6ui161NyTg	394-2	5	145;152;166;170;316
V6ui161NyTg	394-3	4	145;151;160;167
dTUaWnvIOp4	427-1	2	9;33
dTUaWnvIOp4	427-2	3	11;84;90
dTUaWnvIOp4	427-3	1	146
dcYgBU4t98E	393-1	3	41;42;43
dcYgBU4t98E	393-2	1	199
dcYgBU4t98E	393-3	1	32
p_4UPdFqgIQ	725-1	2	627;686
p_4UPdFqgIQ	725-2	2	627;723
p_4UPdFqgIQ	725-3	8	141;313;416;463;627;921;964;109
1NH5dJ9VRvU	730-1	3	2343;2344;2406
1NH5dJ9VRvU	730-2	3	66;473;1549
1NH5dJ9VRvU	730-3	1	225
qd2ivr-5oEM	721-1	3	775;830;2093
qd2ivr-5oEM	721-2	3	536;594;603
qd2ivr-5oEM	721-3	2	173;1666
Q8AZ16uBhr8	697-1	1	412
Q8AZ16uBhr8	697-2	1	1418
Q8AZ16uBhr8	697-3	1	2851
xGcfBRkJSWQ	708-1	2	1331;1337
xGcfBRkJSWQ	708-2	3	1090;1102;1280
xGcfBRkJSWQ	708-3	3	2357;2461;2592
rj6rJzs029A	142-1	1	3
rj6rJzs029A	142-2	7	17;18;19;20;21;22;23
rj6rJzs029A	142-3	4	16;18;20;22
zNxi2s36tS0	163-1	1	35
zNxi2s36tS0	163-2	1	3
zNxi2s36tS0	163-3	1	40
PJHkJJZGwKA	143-1	2	10;15
PJHkJJZGwKA	143-2	4	21;22;23;25
PJHkJJZGwKA	143-3	1	4
ya2IXAREZho	155-1	5	1;5;6;32;74

Figure 19. The NIF values for each question in Video-MME.

video_id	question_id	NIF	timestamp
ya2IXAREZho	155-2	3	26;29;75
ya2IXAREZho	155-3	4	14;39;63;88
fkJv7LRa6Pc	179-1	2	4;22
fkJv7LRa6Pc	179-2	2	22;30
fkJv7LRa6Pc	179-3	1	50
OY9-MQ44MdU	445-1	2	34;125
OY9-MQ44MdU	445-2	4	1422;2278;2334;2439
OY9-MQ44MdU	445-3	3	193;437;573
Mxkg3qLIPC8	451-1	1	10
Mxkg3qLIPC8	451-2	2	130;156
Mxkg3qLIPC8	451-3	2	286;290
H54zMD-9Q-8	437-1	1	588
H54zMD-9Q-8	437-2	1	33
H54zMD-9Q-8	437-3	1	35
ZQIZx50qw88	464-1	3	36;62;92
ZQIZx50qw88	464-2	2	257;294
ZQIZx50qw88	464-3	2	373;375
2Gg40Qo7-zA	456-1	3	193;585;798
2Gg40Qo7-zA	456-2	2	492;498
2Gg40Qo7-zA	456-3	1	94
0k2ey_okQ4E	754-1	3	79;1416;2197
0k2ey_okQ4E	754-2	4	170;237;273;338
0k2ey_okQ4E	754-3	4	562;1119;1522;2617
S_5v1PXLCRc	731-1	2	243;435
S_5v1PXLCRc	731-2	1	231
S_5v1PXLCRc	731-3	1	1186
GV5CuB4zPTY	774-1	1	22
GV5CuB4zPTY	774-2	3	1832;1865;2276
GV5CuB4zPTY	774-3	1	566
jdQ-20JEmgc	757-1	2	739;1491
jdQ-20JEmgc	757-2	2	319;354
jdQ-20JEmgc	757-3	7	435;1557;2041;2166;2316;2563;2705
4H8hcvNeWtg	206-1	1	49
4H8hcvNeWtg	206-2	2	51;58
4H8hcvNeWtg	206-3	1	50
fRf2aYYPkrc	194-1	2	1;23
fRf2aYYPkrc	194-2	2	5;22
fRf2aYYPkrc	194-3	3	1;5;66
k74LDvXSnHM	201-1	1	1
k74LDvXSnHM	201-2	2	25;29
k74LDvXSnHM	201-3	1	29
1q-5I IyZL20	197-1	1	6
1q-5I IyZL20	197-2	1	72
1q-5I IyZL20	197-3	1	93
dh81--46j6s	214-1	1	64
dh81--46j6s	214-2	1	43
dh81--46j6s	214-3	1	11
QopYbLq-zIQ	508-1	3	526;597;685
QopYbLq-zIQ	508-2	1	747
QopYbLq-zIQ	508-3	1	0

Figure 20. The NIF values for each question in Video-MME.

video_id	question_id	NIF	timestamp
kq9Q9-U0vrc	505-1	2	148;151
kq9Q9-U0vrc	505-2	2	82;243
kq9Q9-U0vrc	505-3	1	3
azZZZbSwLQght	490-1	2	195;265
azZZZbSwLQght	490-2	1	58
azZZZbSwLQght	490-3	1	366
a7jZszvFpTY	515-1	3	131;163;182
a7jZszvFpTY	515-2	2	186;192
a7jZszvFpTY	515-3	1	3
323v_FtWqvo	483-1	2	86;220
323v_FtWqvo	483-2	3	79;81;129
323v_FtWqvo	483-3	2	142;220
eQGSbBANFVg	803-1	2	505;521
eQGSbBANFVg	803-2	2	1603;1620
eQGSbBANFVg	803-3	2	1934;1936
yh-EHgkFci4	812-1	3	143;191;311
yh-EHgkFci4	812-2	2	705;906
yh-EHgkFci4	812-3	1	1161
D97vMwfWxvI	795-1	8	451;734;1057;1117;1677;1811;1944;2250
D97vMwfWxvI	795-2	6	520;542;1160;1692;2385;2431
D97vMwfWxvI	795-3	1	69
XDq081k5GnQ	794-1	4	821;1296;1472;2496
XDq081k5GnQ	794-2	2	1546;1621
XDq081k5GnQ	794-3	2	2443;2456
P69idA8J098	786-1	3	152;621;816
P69idA8J098	786-2	2	1762;1774
P69idA8J098	786-3	2	1884;1937
Kn10Jf1x24Q	273-1	1	1
Kn10Jf1x24Q	273-2	3	22;23;25
Kn10Jf1x24Q	273-3	1	10
RP1AL2DU6vQ	252-1	3	45;51;92
RP1AL2DU6vQ	252-2	1	24
RP1AL2DU6vQ	252-3	1	55
F1cVBKTSjRs	222-1	1	34
F1cVBKTSjRs	222-2	1	34
F1cVBKTSjRs	222-3	4	1;14;36;38
VnvG08masio	239-1	1	4
VnvG08masio	239-2	1	94
VnvG08masio	239-3	1	22
s-1M2uwiwyQ	257-1	2	10;11
s-1M2uwiwyQ	257-2	3	52;53;58
s-1M2uwiwyQ	257-3	1	73
cFqLEwAvaHI	575-1	1	35
cFqLEwAvaHI	575-2	1	123
cFqLEwAvaHI	575-3	1	159
ZBKUqc_ICpg	534-1	3	33;288;309
ZBKUqc_ICpg	534-2	1	52
ZBKUqc_ICpg	534-3	2	18;19
-c8eATXUui8	536-1	2	58;78
-c8eATXUui8	536-2	3	381;385;416

Figure 21. The NIF values for each question in Video-MME.

video_id	question_id	NIF	timestamp
-c8eATXUui8	536-3	3	512;521;582
tCRpDpDgBcE	553-1	3	72;77;79
tCRpDpDgBcE	553-2	1	315
tCRpDpDgBcE	553-3	1	509
A30IuIjQYYg	561-1	2	9;114
A30IuIjQYYg	561-2	3	24;35;66
A30IuIjQYYg	561-3	2	30;88
k3zNTrWrbOU	827-1	2	1967;2033
k3zNTrWrbOU	827-2	5	26;291;758;1150;1464
k3zNTrWrbOU	827-3	3	423;1308;1838
wxff_4tDaou	825-1	1	8
wxff_4tDaou	825-2	4	7;630;1261;1977
wxff_4tDaou	825-3	4	28;267;355;825
wT1ERUE8LVw	836-1	2	28;62
wT1ERUE8LVw	836-2	3	791;796;800
wT1ERUE8LVw	836-3	3	686;772;2297
aqFfjJrLkBA	834-1	7	20;21;24;28;34;36;40
aqFfjJrLkBA	834-2	1	11
aqFfjJrLkBA	834-3	3	386;941;1419
elprD1hnDyU	886-1	3	861;902;937
elprD1hnDyU	886-2	6	514;640;694;1024;1085;1700
elprD1hnDyU	886-3	2	2212;2610
AIs3LoU4JUo	297-1	2	22;117
AIs3LoU4JUo	297-2	2	2;93
AIs3LoU4JUo	297-3	2	74;117
nb0Xpuv7K4Q	294-1	3	25;28;29
nb0Xpuv7K4Q	294-2	4	17;20;28;42
nb0Xpuv7K4Q	294-3	1	36
PYZSjin_Pe8	293-1	2	59;60
PYZSjin_Pe8	293-1	1	5
PYZSjin_Pe8	293-1	2	79;103
bHtORiqz0qo	296-1	1	5
bHtORiqz0qo	296-2	3	35;54;64
bHtORiqz0qo	296-3	2	9;84
nVtTVt9csBc	291-1	5	14;17;19;38;90
nVtTVt9csBc	291-2	3	13;20;21
nVtTVt9csBc	291-3	3	0;3;80
OJjgoicpYkU	599-1	1	20
OJjgoicpYkU	599-2	4	159;217;436;504
OJjgoicpYkU	599-3	1	527
cVIfe0Gxa64	595-1	1	0
cVIfe0Gxa64	595-2	1	249
cVIfe0Gxa64	595-3	1	245
InHaW59CmDw	592-1	4	16;168;271;509
InHaW59CmDw	592-2	2	391;659
InHaW59CmDw	592-3	2	512;538
Gr05sxp3n0E	594-1	1	28
Gr05sxp3n0E	594-2	2	399;409
Gr05sxp3n0E	594-3	1	534
fFYNmVb3NCQ	598-1	3	193;236;422

Figure 22. The NIF values for each question in Video-MME.

task	video_id	question	NIF	timestamp
l_plotQA	movie101_55	What is the first expression of	1	151
l_plotQA	movie101_55	Why is the person in the gray s	2	3;5
l_plotQA	movie101_56	What color is the man's clothes	1	164
l_plotQA	movie101_56	What is on the woman's table th	1	188
l_plotQA	movie101_56	What is the man's emotion at th	1	99
l_plotQA	movie101_10	What is the girl's reaction aft	1	4
l_plotQA	movie101_10	What is the girl with glasses'	2	60;61
l_plotQA	movie101_10	Why does the girl with glasses	3	276;300;314
l_plotQA	xiaoliyu_3	What did the cartoon dragon do	3	211;258;268
l_plotQA	xiaoliyu_3	What did the cartoon dragon tur	1	301
l_plotQA	xiaoliyu_3	What did the cartoon snake plac	1	36
l_plotQA	haimian_7	What does the Cartoon Sponge do	1	123
l_plotQA	haimian_7	What do the other cartoon anima	1	191
l_plotQA	haimian_7	Why did the Cartoon Sponge turn	2	180;181
l_plotQA	haimian_7	Why is the Cartoon Octopus angr	1	394
l_plotQA	movie101_14	What color are the pants worn b	1	160
l_plotQA	movie101_14	What color is the bag on the ta	1	155
l_plotQA	movie101_14	What color is the long skirt wo	1	108
l_plotQA	movie101_14	What color is the suit worn by	1	19
l_plotQA	movie101_15	What color is the animal that a	1	1
l_plotQA	movie101_15	What color is the hair of the p	1	44
l_plotQA	movie101_15	What color is the suit the girl	1	395
l_plotQA	movie101_15	What color is the table in the	1	106
l_plotQA	movie101_34	How many candles are lit in the	1	260
l_plotQA	movie101_34	What color is the girl's hair i	1	66
l_plotQA	movie101_34	What color is the hair of the b	1	622
l_plotQA	movie101_34	What color is the hoodie that t	1	187
l_plotQA	movie101_36	In the scene where two people a	1	120
l_plotQA	movie101_36	What color is the hat worn by t	1	136
l_plotQA	movie101_36	What color is the scarf worn by	1	58
l_plotQA	movie101_36	What color is the top worn by t	1	2
l_plotQA	tomjerry_10	What does the cartoon big mouse	2	148;154
l_plotQA	tomjerry_10	What is the cartoon big mouse t	2	176;243
l_plotQA	tomjerry_10	Why did the cartoon little mous	2	246;248
l_plotQA	tomjerry_10	Why does the cartoon little mou	3	33;56;61
2_needle	needle_1	What is the backdrop of the bas	1	148
2_needle	needle_1	Where is the basketball court	1	148
2_needle	needle_13	What are the volunteers doing i	1	340
2_needle	needle_13	What are the volunteers searchi	1	340
2_needle	needle_13	What are the volunteers doing t	2	340;342
2_needle	needle_41	What animal is sitting very sti	1	416
2_needle	needle_41	What is the American toad doing	2	352;358
2_needle	needle_41	What is the nature of the den w	1	351
2_needle	needle_41	What is the state of movement o	3	351;368;376
2_needle	needle_51	What animals are tied beside th	1	475
2_needle	needle_51	What are the two kids doing in	1	476
2_needle	needle_51	What nationality are the kids h	1	478
2_needle	needle_78	What part of the doctor's face	1	506
2_needle	needle_86	What is the direction of the bo	1	132
2_needle	needle_86	What is the kid doing with the	1	132

Figure 23. The NIF values for each question in MLVU.

task	video_id	question	NIF	timestamp
2_needle	needle_86	What is the mood of the boy wal	2	134;147
2_needle	needle_109	What does the engineer begin to	1	803
2_needle	needle_113	What are the volunteers searchi	1	5264
2_needle	needle_119	What is happening to the net on	1	603
2_needle	needle_119	What is the backdrop of the bas	1	603
2_needle	needle_119	What is the weather condition o	1	603
2_needle	needle_119	Where is the basketball court	1 1	603
2_needle	needle_149	What logo is displayed on the s	1	306
2_needle	needle_149	Where is the Goldman Sachs Grou	1	306
3_ego	ego_10	What colour is the stool I sat	1	408
3_ego	ego_13	How many green cups were on the	1	310
3_ego	ego_13	How many picture frames were on	1	293
3_ego	ego_13	In what location did I see the	1	205
3_ego	ego_13	What color is the towel on the	1	319
3_ego	ego_13	Where is the pack of Jenga game	1	57
3_ego	ego_16	Did I leave the door of the sec	2	474;477
3_ego	ego_16	In what room did I see the blac	1	453
3_ego	ego_16	Where was the blue chair?	1	377
3_ego	ego_16	Where was the blue poly bag?	1	377
3_ego	ego_16	Where was the water bottle?	1	427
3_ego	ego_17	Did I cut the wood plank?	2	367;373
3_ego	ego_17	What color was the measuring ta	1	307
3_ego	ego_21	Where did I put the jenga box?	1	49
3_ego	ego_21	Where was the dust pan?	3	168;171;178
3_ego	ego_28	Did I attached the drill into t	2	458;461
3_ego	ego_28	Did I throw the drill on the gr	2	453;454
3_ego	ego_28	How many boxes did I pick up?	1	389
3_ego	ego_28	Where did I keep the drill?	2	454;457
3_ego	ego_28	Where was the square bucket bef	1	387
3_ego	ego_35	Did I leave the front door open	2	183;184
3_ego	ego_35	What colour was the bottle I pr	1	304
3_ego	ego_35	What did I put in the orange tr	2	169;171
3_ego	ego_35	Where did I put the blue helmet	1	193
3_ego	ego_35	Where did I put the leftover pa	2	270;277
3_ego	ego_35	Where was the cat after I put f	2	348;359
3_ego	ego_39	Did I leave the car bonnet open	1	34
3_ego	ego_39	What did I remove from the box?	2	260;261
3_ego	ego_53	What did I put in Dustbin ?	2	132;135
3_ego	ego_76	What the green t-shirt man was	1	272
3_ego	ego_76	Where was the wooden bamboo?	1	252
3_ego	ego_76	Who did I talk to at the garage	1	285
4_count	count_1	In this video, how many times d	3	222;267;469
4_count	count_14	In this video, how many instanc	2	5;295
4_count	count_69	In this video, how many instanc	1	193
4_count	count_91	In this video, how many instanc	5	47;169;221;359;477
4_count	count_102	Throughout this video, what is	4	207;409;1220;1381
4_count	count_112	In this video, how many times d	3	84;304;321
4_count	count_115	In this video, how many instanc	1	94
4_count	count_117	In this video, how many instanc	2	212;314
4_count	count_130	In this video, how many instanc	2	216;455

Figure 24. The NIF values for each question in MLVU.

task	video_id	question	NIF	timestamp
4_count	count_165	Throughout this video, what is	1	31
5_order	order_2	Arrange the following events fr	4	33;51;93;160
5_order	order_10	Arrange the following events fr	6	188;414;422;523;542;549
5_order	order_14	Arrange the following events fr	5	378;431;482;486;565
5_order	order_41	Arrange the following events fr	6	243;251;255;266;279;362
5_order	order_114	Arrange the following events fr	5	377;380;425;566;592
5_order	order_128	Arrange the following events fr	4	98;130;148;214
5_order	order_133	Arrange the following events fr	5	525;529;591;635;636
5_order	order_160	Arrange the following events fr	8	4572;4573;4576;4582;4625; 4630;4645;4691
5_order	order_254	Please identify the option that	4	7;205;375;433
5_order	order_260	Can you tell me which option re	4	382;598;1761;2093
6_anomaly_reco	surveil_0	Does this surveillance footage	3	216;424;504
6_anomaly_reco	surveil_28	Is there any abnormality in thi	6	94;136;177;324;754;872
6_anomaly_reco	surveil_98	Are there any irregularities in	3	53;72;301
6_anomaly_reco	surveil_101	Does this surveillance footage	1	106
6_anomaly_reco	surveil_103	Does this surveillance footage	6	154;196;864;2980;3290;353 7
6_anomaly_reco	surveil_107	Does this surveillance footage	3	79;137;163
6_anomaly_reco	surveil_115	Is there any abnormality in thi	2	108;522
6_anomaly_reco	surveil_125	Does this surveillance footage	3	36;52;126
6_anomaly_reco	surveil_140	Is there any abnormality in thi	3	84;90;272
6_anomaly_reco	surveil_159	Are there any irregularities in	6	3;30;53;114;194;400
7_topic_reasoning	203	What is the main scene in the v	3	35;73;177
7_topic_reasoning	203	What type of video is this?	2	133;477
7_topic_reasoning	231	What scenery is mainly shown in	5	102;128;135;190;216
7_topic_reasoning	AWA-16	What is the setting of the vide	2	8;37
7_topic_reasoning	AWD-1	What is the background of the v	3	29;79;340
7_topic_reasoning	AWD-1	What type of video is this?	4	75;222;226;345
7_topic_reasoning	en_tv_15	In what kind of setting does th	3	46;136;198
7_topic_reasoning	en_tv_15	What genre of movie is the clip	2	325;341
7_topic_reasoning	game_14	What is the object being built	1	10

Figure 25. The NIF values for each question in MLVU.

task	video_id	question	NIF	timestamp
7_topic_reasoning	movie101_10	What color is the hat worn by t	1	15
7_topic_reasoning	movie101_10	What color is the scarf worn by	1	251
7_topic_reasoning	movie101_26	What is the genre of this movie	6	32;41;56;195;410;747
7_topic_reasoning	tomjerry_11	What character appears the most	7	56;74;107;109;116;131;169
7_topic_reasoning	xiaoliyu_2	What is the background of the v	4	85;271;279;293
7_topic_reasoning	xiaoliyu_2	What is the protagonist in the	3	37;40;343
8_sub_scene	subPlot_new_all_0	Please describe what happened w	7	842;843;844;860;862;866;898
8_sub_scene	subPlot_new_all_51	Please describe the process of	8	149;163;171;175;180;196;199;203
8_sub_scene	subPlot_new_all_97	Please describe the situation a	0	N/A
8_sub_scene	subPlot_new_all_103	Please describe the situation w	3	605;608;612
8_sub_scene	subPlot_new_all_108	Please describe the action of t	3	232;233;236
8_sub_scene	subPlot_new_all_113	Please describe in detail what	3	37;39;115
8_sub_scene	subPlot_new_all_130	Please describe what the man in	4	1731;1733;1737;1740
8_sub_scene	subPlot_new_all_136	What did the girl do after hail	4	168;169;177;179
8_sub_scene	subPlot_new_all_172	Can you describe how the man wi	3	308;329;341
8_sub_scene	subPlot_new_all_196	Please describe the situation w	3	22;26;27
9_summary	9	Can you summarize the main cont	17	0;48;57;91;115;136;241;242;302;306;316;347;412;416;442;447;468
9_summary	AWD-5	Could you provide a summary of	14	0;12;20;68;89;108;115;140;186;195;257;308;313;365
9_summary	AWH-4	Can you provide a summary of th	14	11;15;20;23;28;64;123;137;164;296;324;376;378;437
9_summary	en_tv_27	Can you summarize the main cont	10	1;4;55;57;87;89;234;237;408;415
9_summary	en_tv_29	Could you provide a summary of	9	0;81;106;174;355;364;410;424;425
9_summary	movie101_1	Can you summarize the main cont	10	14;18;84;89;156;172;181;190;199;212
9_summary	movie101_58	Can you provide a summary of th	15	17;44;57;68;212;239;290;364;421;443;459;468;487;523;547

Figure 26. The NIF values for each question in MLVU.

task	video_id	question	NIF	timestamp
9_summary	movie101_108	Can you provide a summary of th	11	0;14;33;43;120;121;178;181;263;311;318
9_summary	tomjerry_2	Can you summarize the main cont	12	32;36;38;120;127;128;154;168;224;235;255;262
9_summary	xiaoliyu_5	Could you provide a summary of	20	0;1;8;55;57;58;77;83;143;149;189;203;219;260;343;350;352;382;406;412

Figure 27. The NIF values for each question in MLVU.

video_id	question_id	NIF	timestamp
video_222	0	1	13
video_312	0	3	11;17;23
video_312	1	3	11;17;23
video_312	2	1	23
video_312	3	1	11
video_312	4	2	2;8
video_312	5	1	13
video_557	0	1	0
video_557	1	3	11;12;13
video_557	2	3	11;12;13
video_557	0	3	10;13;14
video_557	1	1	32
video_557	2	1	27
video_557	3	3	10;13;14
video_766	0	1	11
video_766	1	2	10;12
video_812	0	1	29
video_812	1	1	29
video_812	2	3	4;9;15
video_812	3	3	4;9;15
video_921	0	7	2;7;11;16;20;25;31
video_921	1	7	2;7;11;16;20;25;31
video_921	2	1	2
video_921	3	1	7
video_921	4	1	11
video_980	0	2	0;1
video_980	1	5	11;13;17;19;28
video_1200	0	2	11;35
video_1200	1	1	35
video_1200	2	1	35
video_1200	3	1	21
video_1200	4	3	2;8;11
video_1261	0	1	33
video_1261	1	1	33
video_1261	2	2	6;33
video_1261	3	1	33
video_1261	4	6	8;9;10;12;13;14
video_1261	5	1	33
video_1261	6	1	13
video_1262	0	1	33
video_1262	1	1	22
video_1599	0	1	0
video_1830	0	1	18
video_2246	0	1	29
video_2246	1	1	29
video_2250	0	3	0;28;31
video_2250	1	2	0;31
video_2250	2	1	31
video_2250	3	1	0
video_2250	4	1	24

Figure 28. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_2250	5	1	24
video_2280	0	1	0
video_2298	0	1	7
video_2369	0	2	30;32
video_2369	1	3	3;14;28
video_2959	0	1	0
video_2959	1	1	20
video_2959	2	1	10
video_2959	3	2	10;20
video_2959	4	2	10;20
video_2959	5	3	14;16;20
video_2959	6	1	10
video_2959	7	1	5
video_2959	8	1	10
video_2959	9	1	10
video_2999	0	1	0
video_2999	1	1	31
video_2999	2	1	31
video_3145	0	3	0;1;2
video_3145	1	1	2
video_3451	0	1	0
video_3451	1	1	18
video_3451	2	1	26
video_3464	0	1	0
video_3464	1	1	5
video_3619	0	2	15;22
video_3619	1	1	22
video_3768	0	3	0;2;4
video_3768	1	2	0;32
video_3768	2	1	32
video_3768	3	1	0
video_3768	4	1	32
video_3768	5	1	32
video_3928	0	1	24
video_4171	0	3	0;1;2
video_4171	1	1	1
video_4306	0	1	0
video_4306	1	1	1
video_4306	2	1	32
video_4306	3	1	0
video_4306	4	1	16
video_4306	5	1	16
video_4459	0	1	33
video_4459	1	1	20
video_4810	0	1	26
video_4810	1	1	26
video_4810	2	1	26
video_4846	0	4	0;5;14;31
video_4846	1	1	5
video_4846	2	1	5

Figure 29. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_4846	3	1	5
video_4859	0	1	0
video_4859	1	1	13
video_4906	0	1	1
video_4991	0	3	0;1;2
video_4991	1	1	1
video_5015	0	3	0;1;2
video_5015	1	1	24
video_5015	2	1	14
video_5015	3	1	24
video_5308	0	1	5
video_5308	1	4	4;5;17;31
video_5308	2	1	5
video_5308	3	1	5
video_5308	4	1	5
video_5439	0	4	4;5;16;31
video_5439	1	1	5
video_5439	2	1	0
video_5439	3	1	5
video_5689	0	3	0;1;2
video_5875	0	1	1
video_5963	0	4	3;4;15;31
video_5963	1	1	4
video_5963	2	1	0
video_5963	3	1	4
video_5970	0	7	1;6;12;17;21;25;31
video_5970	1	7	1;6;12;17;21;25;31
video_5970	2	1	1
video_5970	3	1	1
video_5970	4	1	6
video_5970	5	1	12
video_6164	0	3	0;1;2
video_6164	1	1	1
video_6223	0	1	26
video_6223	1	1	17
video_6223	2	1	26
video_6249	0	8	2;5;9;13;17;23;27;33
video_6249	1	8	2;5;9;13;17;23;27;33
video_6249	2	1	2
video_6249	3	1	5
video_6249	4	1	9
video_6321	0	3	0;4;9
video_6321	1	2	0;9
video_6321	2	1	9
video_6321	3	1	0
video_6321	4	1	18
video_6321	5	1	18
video_6418	0	1	34
video_6418	1	1	0
video_6418	2	1	34

Figure 30. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_6418	3	1	18
video_6418	4	2	18;34
video_6418	5	2	18;34
video_6418	6	5	23;25;29;32;34
video_6418	7	1	18
video_6418	8	1	9
video_6418	9	1	18
video_6418	10	1	10
video_6626	0	3	18;22;25
video_6736	0	6	3;8;15;21;26;31
video_6736	1	6	3;8;15;21;26;31
video_6736	2	1	3
video_6736	3	1	3
video_6736	4	1	8
video_6829	0	5	3;10;19;25;31
video_6829	1	5	3;10;19;25;31
video_6829	2	1	3
video_6829	3	1	3
video_6829	4	1	10
video_6829	5	1	19
video_6842	0	2	5;13
video_6842	1	1	33
video_6842	2	1	7
video_6896	0	7	1;6;11;15;20;25;29
video_6896	1	7	1;6;11;15;20;25;29
video_6896	2	1	1
video_6896	3	1	6
video_6896	4	1	11
video_7092	0	1	0
video_7092	1	1	8
video_7272	0	1	0
video_7318	0	3	0;1;2
video_7318	1	3	0;26;32
video_7318	2	2	0;32
video_7318	3	1	32
video_7318	4	1	0
video_7318	5	1	16
video_7318	6	1	16
video_7453	0	1	0
video_7453	1	1	30
video_7472	0	1	5
video_7609	0	3	0;1;2
video_7609	1	3	3;8;10
video_7609	2	3	3;8;10
video_7609	3	3	3;8;10
video_7609	4	2	3;30
video_7609	5	2	30;31
video_7609	6	1	3
video_7609	7	1	3
video_7704	0	7	1;7;12;17;23;28;33

Figure 31. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_7704	1	7	1;7;12;17;23;28;33
video_7704	2	1	1
video_7704	3	1	1
video_7704	4	1	7
video_7704	5	1	12
video_8245	0	1	0
video_8245	1	1	1
video_8245	2	1	33
video_8245	3	1	0
video_8245	4	2	14;20
video_8245	5	1	21
video_8551	0	3	0;1;2
video_8551	1	1	22
video_8551	2	1	22
video_8635	0	3	0;18;33
video_8635	1	2	0;33
video_8635	2	1	33
video_8635	3	1	0
video_8635	4	1	22
video_8635	5	1	22
video_8679	0	1	2
video_8679	1	1	2
video_8679	2	1	2
video_8679	3	1	0
video_8732	0	1	0
video_8732	1	1	0
video_8732	2	2	12;16
video_8732	3	2	5;6
video_8732	4	1	12
video_8735	0	4	6;7;8;9
video_8735	1	3	17;21;26
video_8735	2	3	7;10;16
video_8735	3	1	4
video_8735	4	3	7;10;16
video_8888	0	2	10;11
video_8952	0	3	6;16;28
video_8952	1	1	6
video_8952	2	3	6;16;28
video_8952	3	1	0
video_8952	4	1	0
video_8952	5	1	33
video_8952	6	3	6;16;28
video_8975	0	1	0
video_8975	1	1	29
video_8975	2	1	16
video_8975	3	2	16;29
video_8975	4	2	16;29
video_8975	5	2	16;29
video_8975	6	3	23;26;29
video_8975	7	1	16

Figure 32. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_8975	8	1	6
video_8975	9	1	16
video_8975	10	1	8
video_8995	0	3	0;1;2
video_8995	1	2	19;20
video_9008	0	1	0
video_9008	1	1	33
video_9008	2	1	20
video_9008	3	2	20;33
video_9008	4	2	20;33
video_9008	5	2	27;29
video_9008	6	1	20
video_9008	7	1	9
video_9008	8	1	20
video_9008	9	1	11
video_9342	0	3	0;1;2
video_9342	1	1	2
video_9431	0	3	0;1;2
video_9431	1	1	0
video_9431	2	2	14;20
video_9431	3	5	6;14;17;20;27
video_9431	4	1	24
video_9691	0	3	0;11;32
video_9691	1	2	0;32
video_9691	2	1	32
video_9691	3	1	0
video_9691	4	1	15
video_9691	5	1	15
video_9727	0	4	5;10;18;25
video_9727	1	1	5
video_9727	2	4	5;10;18;25
video_9727	3	1	0
video_9727	4	4	5;10;18;25
video_9727	5	4	5;10;18;25
video_9727	6	4	5;10;18;25
video_9772	0	1	33
video_9772	1	1	33
video_9772	2	1	33
video_9772	3	1	33
video_9772	4	1	33
video_9772	5	1	33
video_9799	0	1	22
video_9799	1	1	26
video_10028	0	3	0;1;2
video_10028	1	3	5;7;16
video_10076	0	1	0
video_10159	0	7	4;8;13;16;19;23;28
video_10159	1	7	4;8;13;16;19;23;28
video_10159	2	1	4
video_10159	3	1	8

Figure 33. The NIF values for each question in Perception Test.

video_id	question_id	NIF	timestamp
video_10159	4	1	13
video_10212	0	3	0;1;2
video_10212	1	1	0
video_10212	2	1	0
video_10212	3	1	0
video_10212	4	2	7;30
video_10212	5	1	30
video_10212	6	1	7
video_10212	7	1	0
video_10307	0	3	0;1;2
video_10307	1	1	17
video_10307	2	1	33
video_10386	0	4	7;15;17;23
video_10505	0	3	0;1;2
video_10505	1	1	0
video_10505	2	1	1
video_10505	3	1	28
video_10600	0	1	0
video_10600	1	1	6
video_10664	0	1	1
video_10664	1	6	1;3;5;7;11;18
video_10664	2	3	27;28;29
video_10690	0	2	9;10
video_10690	1	2	9;10
video_10944	0	1	0
video_10944	1	6	6;8;11;14;19;24
video_10944	2	6	6;8;11;14;19;24
video_10944	3	1	8
video_11253	0	3	0;1;2
video_11306	0	1	2
video_11306	1	1	23
video_11331	0	7	0;7;12;16;20;25;31
video_11331	1	7	0;7;12;16;20;25;31
video_11331	2	1	1
video_11331	3	1	0
video_11331	4	1	7
video_11331	5	1	12
video_11457	0	3	4;10;13
video_11457	1	3	25;30;33
video_11457	2	1	25
video_11457	3	3	4;10;13

Figure 34. The NIF values for each question in Perception Test.

video_id	NIF	timestamp
0a8109fe-15b9-4f5c-b5f2-993013cb216b	1	138
46853bef-9052-428d-8e61-df684147f4af	2	38;131
b4cc9985-97e8-423a-9737-22e5d9b4dbce	4	4;13;28;118
0a8b2c9d-b54c-4811-acf3-5977895d2445	1	32
e9322513-15b2-4b89-8ddb-7d1432beb8a1	4	20;25;28;32
b1a1280a-1f7f-4796-bca2-ba03f3fb9345	3	0;24;73
05f1fc03-0c9e-4fd4-9d85-bb7be4e69234	2	13;34
9de66400-05ec-4173-93c9-16c2cc9d881d	3	27;56;153
55ea09ed-4590-4e59-8753-40a64d67abd9	1	166
6472e377-b65c-461a-a750-9b28a673dc86	3	19;40;139
901ae1fe-5be2-495b-9506-9ec2a28d8ae0	2	69;138
cd384dae-4229-4fa5-9fed-e4b5a1432f29	3	7;67;169
71d00225-7ea5-46a0-8015-9b5a667f619a	4	3;49;120;143
425f2fb4-a2a7-4925-94b5-25f6b0b85f78	3	17;33;53
340a76ff-7144-4b31-906f-8a43ed866bc0	1	155
b58ab03f-3520-4916-81b8-2c42e3d0d31d	1	39
0aadf5ce-07eb-4934-9e07-317a46bc0b21	1	16
93a92b6f-5ed2-4b2c-9f9c-a6307e1fb256	3	4;12;136
cd2e4351-de59-4511-ab43-36c37b388a8d	1	96
509e6545-5fff-4f73-ae0f-524dfa8b3c2c	2	11;22
24f4b88e-2294-4017-a669-9e27c07d44e7	1	167
86d91c31-1bfe-4f99-a803-7466c8d801d1	1	66
d68218d2-5071-458d-8e4d-87f5707b7fbc	3	34;38;120
a0a00d56-0f2d-4f3d-ae12-ee5bc4c7ba19	1	143
0688f66e-f115-49c6-85ff-712bf4f4a758	3	13;30;125
1eb2f153-055f-4004-ad55-154359af8025	2	1;35
01a144a5-24d2-4a5a-af01-1f318d674bed	1	39
824c2b85-b40b-4bbd-82bf-70468f9c042b	3	16;35;37
5fe9843b-0b74-4505-b864-86eb53c25cc6	1	79
e27e9ec7-aaf3-4e5c-a387-1699fe66ea4f	3	28;116;127
c66fe71d-e9c3-4983-ad77-26c0a8b1c0b9	1	109
e86bb89c-baf4-4463-8941-e296d1d4d62f	1	140
b5867202-c87b-4ffa-8617-e8b2e9eba1a2	1	164
2faa1516-ed55-4a96-a4be-09c402cf2c76	1	74
13da1294-2b42-4ef6-8dd6-ff651ef4571f	3	10;14;20
7ad240de-34ab-4694-a6be-05a47e14793f	2	40;41
e614f1a5-7c7b-468f-9840-d7373f740255	1	95
bcf9aba0-a4b6-4210-8823-025f43f2631f	1	20
2e22aafd-1fbb-4e73-ab6f-d8f628b66ba1	1	29
84bd1e04-370a-4e4a-9255-776f8d8e38ad	1	56

Figure 35. The NIF values for each question in EgoSchema.

video_id	NIF	timestamp
90c3f31b-3b44-4b9a-a684-cef313a45c32	2	35;78
217fe8d0-dfc8-407b-86be-269378c5259a	2	2;104
631682a5-5574-41a8-904f-7ee96fc93683	2	2;52
628e252f-e743-4054-92fd-b0ed6983571d	2	22;91
51688142-10e7-48ab-edef-2caa5448b456	1	180
45e313dd-9b30-444e-9577-b15a21cb59b4	4	10;31;66;91
86de4235-953e-458f-a951-40314e92a33b	1	83
aebf3455-59b7-4071-a7d2-18d053c38f8f	1	1
dfb6c468-e124-40f6-9c4e-c13ee45a2ad9	1	76
d0bbd7fa-2a15-4c25-ab06-adc7d04bce7b	3	9;19;64
e48b8359-d35e-45f1-aa3b-eb1417e10dc8	3	11;12;29
97dc6bb7-7eed-45f7-bdb0-269ab8c2f639	4	1;60;71;99
81ba0fd6-cc69-410d-9e2d-8317fd22cce8	3	12;37;60
b5d7c421-2b86-4ed0-b314-ce810c778c47	1	7
2b8d1e50-3ba7-492a-8a0b-104eb659c27b	2	110;170
4b7495f1-e2c2-4d69-bbf4-c2dabeb5e634	2	13;104
0d173aa3-9a94-4ba4-84bc-949d3254a63d	4	41;45;48;134
cd882b7a-0766-4582-8388-3990b009b11b	4	3;23;101;131
4070a4e2-14d5-4618-9889-dd18a416e2b5	2	4;148
68e0bb20-414d-42ef-a0a0-e821efbe8e06	4	20;22;37;106
9796f529-40ca-4e74-89ed-6a25efb24c8c	3	76;78;79
223164c8-abed-4f1a-8f7c-4088c89d3ece	1	26
97f80e1c-164d-4072-8ee9-980e366eec6c	2	0;28
ece69b04-2e67-434a-b923-1329feed590d	2	5;31
e00836f3-1506-4479-a028-e17f19cff0bf	1	79
94f6e8bd-b65d-4fd0-b2a9-75b69397fe2e	2	16;31
e04cd624-17e1-4986-b344-55aa92d7c0c3	1	50
5782e464-d9e-4bee-88f7-6c2f6727854b	2	9;15
86cb525a-d61c-46f1-9c5f-cc7284184900	3	85;136;156
b3aaaae6-e6f6-499f-8b03-ddf85b3f62c4	1	12
22b86648-340c-4338-b46e-5eaba3a44b06	2	106;119
cc3abb84-1317-4718-b414-75f899c20ee3	2	12;30
e67de76c-1058-49a7-a47e-12736da4ffc0	2	0;27
f95e7f60-0f9a-40e7-bb60-55ecb287b2dc	2	20;22
1bf1fdea-6f44-4d3a-b5c0-6852aaada71b	2	35;36
9d83c60a-5d84-4bea-8325-56beed585df2	1	34
0e4804e0-85fa-48bc-ada3-a94167b06e53	1	4
6f42ba6a-cb2a-428f-bffb-7a15abd94727	3	9;22;114
04c51dba-1dcb-4b8f-a62c-efc363561d7b	1	31
ee227b56-c12b-4725-89e9-aa29e0b4dbe8	1	14
803c8ecf-9448-48b6-8bf2-debd052dbe43	1	8
1bd933df-3575-4fa7-839f-765c7108259e	2	20;106
c04da37a-b98f-4796-afe2-1b7d3af20911	6	104;110;117;124;129;160
9748f410-2316-4a2f-9893-56f8f240dc67	3	34;88;152
0f181d98-5036-4990-b397-62e934e168ef	3	2;12;21
640ad606-0376-48cb-bc6a-14bc34ec4eaa	3	10;21;35
36424b90-8a32-44b3-8db1-ad8bc3b222d0	4	17;40;76;104
fcf8719c-b32d-463f-aa4c-6ac4149bb1c0	2	67;166
6c901d03-3413-41a8-9aa0-8f9f6ed6b6f1	3	2;6;97
5d15c716-f179-462c-84a6-fa94e9d14e94	1	86

Figure 36. The NIF values for each question in EgoSchema.