

MLVU: Benchmarking Multi-task Long Video Understanding

Supplementary Material

A. Overview of Appendix

- [B: Evaluation Results on MLVU Dev Set](#)
- [C: Collecting Details of our Universal Long Video Collection \(ULVC\)](#)
- [D: Details of the MLVU Time-Ladder](#)
- [E Detailed Division of Dev and Test Sets in MLVU](#)
- [F: Annotation Details of MLVU](#)
- [G: Details of Baselines and the Evaluation Process.](#)
- [H: Explorations of Video Retrieval Augmented Generation](#)
- [I: More Visualized Examples of MLVU](#)

B. Evaluation Results on MLVU Dev Set

The evaluation results of the baselines on the MLVU dev set are detailed in Table 2. Notably, the multiple-choice questions in the MLVU dev set present four options, whereas the MLVU test set offers six, making the latter more challenging and discriminative.

C. Collecting Details of our Universal Long Video Collection (ULVC)

In the initial stage of our Multi-task Long Video Understanding (MLVU) benchmark creation, we first collected long-form videos from a variety of sources to form our Universal Long Video Collection (ULVC). The entirety of the long videos incorporated into our MLVU benchmark were selected, edited, or synthesized from ULVC.

Specifically, our ULVC includes a diverse set of 986 long videos. This collection features 168 movies from the Movie101 [61] and MovieChat [44] datasets, along with 60 documentaries from MovieChat [44]. It also contains 65 game videos from MineDojo [10], 239 surveillance videos from UCF-Crime [46], and 100 ego-centric videos from Ego4D [15]. Additionally, we independently collected 72 cartoons, 92 TV series, 60 tutorial videos, 60 sports videos, and 70 life records.

It’s important to clarify that the quantity of videos in the ULVC does not directly correspond to the number of videos and questions in our MLVU benchmark, which are 1,730 and 3,102 respectively. For example, a two-hour movie from the ULVC might be utilized in its entirety for the Sub-Scene Captioning task, or it could be segmented into several approximately 10-minute clips for the Video Summarization task, or even used as a background video for synthetic video generation. Moreover, a single video could be annotated with multiple questions simultaneously.

D. Details of the MLVU Time-Ladder

As discussed in Section 3.1, most tasks in our MLVU are subject to segment-level annotation. This approach provides us with the flexibility to adjust the length of the video without requiring additional human annotators. Building on this strategy, as mentioned in Section 4.3, we have generated a derivative dataset, *MLVU Time-Ladder*, which includes videos of varying durations - specifically 3, 6, and 10 minutes. This dataset allows us to investigate how video duration impacts LVU task difficulty.

Specifically, during the annotation process of the VS task, we guided annotators to delineate the summarization in accordance with the initial 3 and 6-minute segments. For the PQA and SSC tasks, we requested annotators to identify the segments within the extended video where the pertinent answers are located. In the case of the ego reasoning task, the Ego4D dataset [15] already comprises the intervals where the answers reside. Lastly, for the synthetic tasks of NQA, AO, and AC, we possess the capability to directly generate the necessary video lengths.

E. Detailed Division of Dev and Test Sets in MLVU

Our MLVU comprises a total of 3,102 questions, divided into a dev set with 2,593 questions and a test set with 509 questions. We present the detailed distribution of questions for each task in Table 1.

Task	Dev	Test	Total
Topic Reasoning	264	91	355
Anomaly Recognition	200	39	239
Video Summarization	217	40	257
Needle QA	355	60	415
Ego Reasoning	352	53	405
Plot QA	539	50	589
Sub-Scene Captioning	201	46	247
Action Order	259	70	329
Action Count	206	60	266

Table 1. Detailed Distribution of Questions in the MLVU Dataset Across Dev and Test Sets for Each Task.

Methods	Date	Input	Holistic			Single Detail				Multi Detail		M-Avg	G-Avg
			TR	AR	VS*	NQA	ER	PQA	SSC*	AO	AC		
Full mark	–	–	100	100	10	100	100	100	10	100	100	100	10
Random	–	–	25	25	–	25	25	25	–	25	25	25	–
<i>Image MLLMs</i>													
Otter-I [24]	2023-05	16 frm	25.0	25.0	2.18	25.1	25.0	24.9	4.12	13.1	25.2	23.3	3.15
LLaVA-1.6 [30]	2024-01	16 frm	60.6	41.0	2.11	43.1	38.4	41.0	4.35	25.5	25.7	39.3	3.23
Claude-3-Opus [†] [2]	2024-03	16 frm	67.2	43.5	3.11	21.6	40.2	47.8	3.66	18.2	16.7	36.5	3.39
Qwen-VL-Max [†] [4]	2024-01	16 frm	67.4	63.5	2.71	40.3	40.9	43.3	5.21	25.0	14.8	42.2	3.96
<i>Short Video MLLMs</i>													
Otter-V [24]	2023-05	16 frm	24.6	26.0	2.38	28.2	27.6	22.3	4.23	15.1	26.7	24.4	3.31
mPLUG-Owl-V [57]	2023-04	16 frm	28.0	25.0	2.36	24.5	31.8	27.3	5.31	21.2	23.3	25.9	3.84
VideoChat [26]	2023-05	16 frm	33.0	32.0	2.31	27.0	32.1	27.6	5.01	24.3	28.6	29.2	3.66
Video-LLaMA-2 [62]	2024-08	16 frm	54.5	41.5	2.34	39.4	33.5	35.4	5.22	18.5	25.7	35.5	3.78
VideoChat2-HD [27]	2024-06	16 frm	74.6	51.5	2.57	42.0	47.4	43.8	5.04	22.8	29.6	44.5	3.81
Video-LLaVA [29]	2023-11	8 frm	71.6	57.0	2.43	53.2	45.2	48.4	5.25	20.1	35.9	47.3	3.84
ShareGPT4Video [7]	2024-05	16 frm	75.8	51.5	2.52	47.6	43.2	48.4	5.02	34.0	23.3	46.4	3.77
VideoLLaMA2 [9]	2024-06	16 frm	74.6	64.5	2.79	49.9	43.8	45.1	5.18	34.0	27.4	48.5	3.99
<i>Long Video MLLMs</i>													
MovieChat [44]	2023-07	2048 frm	29.5	25.0	2.33	24.2	24.7	25.8	3.23	28.6	22.8	25.8	2.78
Movie-LLM [45]	2024-03	1 fps	30.0	29.0	2.88	29.6	24.7	24.1	5.00	20.5	24.8	26.1	3.94
TimeChat [42]	2023-12	96 frm	23.1	27.0	2.54	24.5	28.4	25.8	4.29	24.7	32.0	30.9	3.42
LLaMA-VID [28]	2023-11	1 fps	50.8	34.5	3.22	30.1	32.7	32.5	5.22	23.9	27.8	33.2	4.22
MA-LMM [17]	2024-04	1000 frm	51.9	35.5	2.12	43.1	38.9	35.8	4.80	25.1	24.3	36.4	3.46
MiniGPT4-Video [3]	2024-04	90 frm	70.9	52.5	2.64	49.0	48.6	44.5	4.07	23.2	23.0	44.5	3.36
LongVA [63]	2024-06	256 frm	83.3	58.5	3.39	69.3	50.0	67.2	5.26	38.6	27.2	56.3	4.33
Video-CCAM [11]	2024-08	96 frm	84.9	66.0	2.84	73.2	60.5	66.1	5.19	42.1	38.4	63.1	4.01
Video-XL [43]	2024-09	256 frm	80.3	54.5	3.25	73.8	57.4	67.9	5.02	68.3	40.3	64.9	4.14
GPT-4o [†] [39]	2024-05	0.5 fps	87.4	74.5	4.90	64.8	57.1	65.1	6.69	56.7	46.3	64.6	5.80

Table 2. The overall performances on MLVU dev set, including the holistic LVU tasks (TR: Topic Reasoning, AR: Anomaly Recognition, VS: Video Summary), the single-detail LVU tasks (NQA: Needle QA, ER: Ego Reasoning, PQA: Plot QA, SSC: Sub-Scene Captioning), and multi-detail LVU tasks (AO: Action Order, AC: Action Count). M-Avg: the average performance of multiple-choice tasks; G-Avg: the average performance of generation tasks (marked by *). Two input strategies are used by the MLLMs in evaluation: Uniform Sampling (**N frm**), which evenly samples N frames from the video; Frame Rate Sampling (**N fps**), which samples N frames per second. [†] denotes proprietary models.

F. Annotation Details of MLVU

F.1. Topic Reasoning (TR).

The questions and corresponding answers for the TR task were meticulously annotated by human annotators, following the specific guidelines illustrated in Figure 1. We required the annotators to design questions related to the reasoning of the video topic, rather than focusing on the creation of questions about minor details. More visualized examples of TR task can be found in Figure 10.

F.2. Anomaly Recognition (AR).

The anomaly recognition task did not involve manual annotation. We utilized videos exceeding three minutes in duration, extracted from the UCF-Crime dataset [46]. We also modified the original labels to fit a multiple-choice format.

F.3. Video Summarization (VS).

The ground truth data for the VS task were derived from manual annotations. We instructed the annotators to use pronouns instead of specific character names in all annotations. This guideline stemmed from the inherent constraints of most existing MLLMs, which generally lacked the capacity to process audio or subtitles. This made it difficult for

Annotation Guidelines for Topic Reasoning

1. Task Description: Your task is to formulate a question that pertains to the genre and key content of a given long video, and then provide the corresponding answer.

2. Question Requirements:

- Your questions should be centered around the core content of the video, rather than focusing on minor details.
- Suitable topics for questions include the genre of the video, the main events or themes, the primary environmental setting, the depicted weather conditions, and the time period or timeline.

3. Question Format:

- Questions should be structured in a multiple-choice format. Each question should have one correct answer and three plausible, yet incorrect, distractor options.

4. Question Examples (for reference only, not limited):

- What genre does this movie/video fall into?
- Where does the main scene in the video take place?
- What is the main event being narrated in the video?
- What is the protagonist in the video accomplishing?

Figure 1. Annotation Guidelines for the Topic Reasoning Task.

these models to identify specific characters. The annotation instructions and examples provided to the annotators are elaborated in Figure 2. More visualized examples of VS task can be found in Figure 10.

F.4. Needle Question-Answering (NQA).

We leveraged the GPT-4 [1] and the detailed video caption data from the WebVid dataset [5] to facilitate a semi-automated generation of annotated questions and answers for the NQA task. Initially, we selected video clips from WebVid, which we referred to as *needle* clips. The corresponding captions of these needle clips were then fed into GPT-4, which generated question-answer pairs based on the information encapsulated in the captions. The specific prompt provided to GPT-4 is depicted in Figure 3. The generated questions were carefully crafted to focus on a particular detail within the needle clip. These questions were structured to incorporate the maximum number of hints to effectively guide MLLMs in grounding the content of the needle within the context of the longer video. Following this, we randomly selected longer background videos from our ULVC and manually ensured that the scene indicated by the needle’s question did not feature in these background videos. The final step involves integrating the needle into the longer video, thereby producing the final needle question video. More visualized examples of NQA task can be found in Figure 11.

F.5. Ego Reasoning (ER).

The video resources, questions, and correct responses used in the ER task were derived from the Natural Language Queries (NLQ) task within the Ego4D dataset [15]. This data was restructured to fit a multiple-choice question format.

F.6. Plot Question-Answering (PQA).

The PQA task’s questions and answers were annotated by human annotators, following specific guidelines illustrated in Figure 4. We instructed the annotators to craft questions that probe into the intricate plot details encapsulated within the videos. These questions were designed to encompass both perception and reasoning aspects. We stipulated that both questions and their corresponding answers should avoid the use of specific character names or any objective hints, and should instead utilize pronouns. This approach was strategized to prevent potential information leakage, given that MLLMs often demonstrate a familiarity with the storylines of well-known movies and TV series. Such common-sense knowledge could potentially allow the MLLMs to answer questions correctly without the essential requirement of analyzing the input video.

Nonetheless, the complexity of character interactions and actions in longer videos poses a challenge to conveying plot details using only pronouns and feature descriptions. Previous datasets for plot question answering that avoided the use of character names often resulted in compromised

Annotation Guidelines for Video Summarization

1. Task Description: Your task is to provide a comprehensive summary of the key events occurring within a video clip that ranges from 3 to 15 minutes in length.

2. Annotation Requirements:

- The annotation should encapsulate the principal events portrayed in the video, structured in chronological order.
- Refrain from using specific character names in the annotation. Instead, all characters should be referred to using pronouns and identified by their unique attributes or roles, such as attire, age, profession, etc. For instance, characters could be described as an "elderly individual" (age), a "medical professional" (profession), among others.
- Disregard audio-related information, such as dialogues between characters. The summaries should be derived exclusively from the visual content presented in the video.

3. Annotation Template:

- Initiate your summary by outlining the overall content of the video: the event being narrated or the video's main theme.
- Subsequently, chronologically depict the key events that unfold in the video. The aim is to provide a clear and concise description of the main content, events, and scenes exhibited in the video.

4. Annotation Examples:

- **Cartoon:** This is a video about a cartoon sponge's whimsical adventures. The video begins with a cartoon sponge rushing into a house to converse with a cartoon starfish on a rocking chair. The sponge then heads to a concert hall where he watches a performance, during which a cartoon animal on a throne reprimands a cartoon octopus who continues his act. Later, the cartoon sponge and a cartoon squirrel are seen flying and conversing in the air. The sponge also encounters a cartoon shark preparing to drink coffee and a cartoon lobster sailing on a sponge, after which the lobster chases the sponge away.
- **Movie / TV Series:** This is a video depicting a dramatic narrative. The video starts with a man singing into a microphone, with a few other men playing instruments behind him. The scene changes to someone pushing open a door and walking into a room where others are resting. She then opens another door, enters a room and starts arguing with the singing man, which results in a fight. Next, the woman drives the man away, which results in a car crash. The car then falls off a bridge and gets hit by another car. The screen goes black and then lights up again, revealing a bookshelf filled with books at the end.
- **Documentary:** This is a documentary about forest animals and ecology. The video begins by showing scenes of fish, butterflies, orangutans, and birds in the forest. Then, the video depicts two birds cooperatively building a nest on a rock. As it starts to rain in the forest, a hatchling is born. The two birds catch bugs and frogs in the forest and feed them to the newborn. The camera follows the direction of the flowing river, which converges to form a spectacular waterfall. The video ends with a calm sea and beach, with a large flock of seabirds flying over the sea, hunting for prey close to the water.

Figure 2. Annotation Guidelines for the Video Summarization Task.

question diversity and tended towards generalized queries. We illustrate this through a comparative analysis of TVQA [22], Moviechat [44], and our PQA dataset's question word clouds in Figure 5. While TVQA provides a diverse range of questions, it does so by employing specific character names. In contrast, Moviechat avoids character names, but its questions are frequently overly broad, lack specific plot details, and exhibit diminished diversity. Our PQA dataset successfully navigates these challenges, offering a diverse

range of questions without resorting to the use of character names. More visualized examples of PQA task can be found in Figure 11.

F.7. Sub-Scene Captioning (SSC).

In the development process of the SSC task, we employed human annotators to generate both prompts and standard caption data. The specific guidelines provided to annotators are illustrated in Figure 6. Initially, the annotators identi-

Prompt for Generating Needle Questions

You are a question setter. Your task is to evaluate the participants' ability to capture detailed information from an extremely long video. The participants will receive a lengthy and content-rich video, and you are required to ask a question about a specific piece of information from the video.

I will provide you with a description of the segment that needs to be questioned at the end. Your question must include as much contextual information as possible to help the participants locate the source of the information. The description I provide generally contains multiple clues, and you should ask questions targeting different clues. Your question should be in a multiple-choice format, necessitating the provision of at least four choices, including the correct answer. Depending on the depth of information in the segment description, you can craft between 1 to 3 distinct questions.

Please provide the questions in the JSON format as follows...

Here is the description of the segment that needs to be questioned...

Figure 3. The prompt provided to GPT-4 in the process of creating the question-answer pair for the Needle Question-Answering task.

fied a specific, easily referable sub-scene within a lengthy movie. Subsequently, they crafted a prompt replete with adequate clues to reference this scene, ensuring the uniqueness of these clues throughout the entire film. To prevent any leakage of information, the prompt was designed to exclude any character-specific names or objective hints, instead incorporating rich descriptive details to allude to the plot. Following this, the annotators produced a detailed caption for this sub-scene, and deconstructed the caption into multiple, non-redundant "scoring points" to facilitate quantitative assessment (the details of the evaluation metric can be found in Appendix G.3). More visualized examples of PQA task can be found in Figure 12.

F.8. Action Order (AO).

The videos, questions, and answers for the action order task were all synthetically generated. In order to maintain the high quality of our evaluation data, we adopted a dual-strategy approach. Firstly, we selected actions for the *probe* videos that were not commonly seen in most films, such as making jewelry and water skiing. Secondly, in the selection of background videos, we conducted a cursory review of the video content to further ensure that the actions referenced in the questions were not present in the video. This rigorous methodology ensured the reliability of our data.

F.9. Action Count (AC).

The process of data acquisition and annotation for the action count task closely mirrored that of the action order task. All videos, questions, and answers were synthetically generated.

We employed a strategy consistent with the action order task to ensure the validity and reliability of our evaluation data.

G. Details of Baselines and the Evaluation Process

G.1. Baselines

In this section, we outline the primary baselines evaluated on our MLVU. For image-based MLLMs, most models lack multi-image inference capabilities. Therefore, we select Otter-I, LLaVA-1.6, and InternVL, which have official multi-image implementations. Additionally, we include two proprietary models—Claude-3-Opus and Qwen-VL-Max—that offer APIs for multi-image inference. For the available models, we determine the maximum input frames based on their LLM context length. Claude and Qwen support a maximum of approximately 20 images, so we choose 16 frames to ensure fair comparisons. Regarding video MLLMs, we use default frame sampling strategies. For example, VideoChat2 uniformly samples 16 frames, while LLaMA-Vid samples 1 frame per second. Specifically, GPT-4o can handle up to approximately 500 images at a resolution of 512×512 pixels. Thus, we select a sampling rate of 0.5 fps to accommodate most of our videos.

G.2. Inference Details

We have developed two templates specifically for Multiple-Choice and Generation tasks, as illustrated in Figure 7. Distinct system prompts were designed to accommodate the differences between video-based and image-based MLLMs. Considering the variances in task requirements, we incorporated "option prediction guidance" into the Multiple-Choice template to aid in option extraction. Conversely, in Generation tasks, we do not implement any additional interventions but employ fixed-question guidance to enable models to respond to diverse task questions. In our evaluation, the templates are seamlessly integrated into the evaluation code of open-release models or available API of proprietary models.

G.3. Evaluation Metrics

For the evaluation of Multiple Choice tasks, we directly compute absolute accuracy by matching the predicted option with the ground truth. In Generation tasks, we develop multiple criteria for assessment and employ GPT-4 to rank the alignment between generated texts and the provided answers. As illustrated in Figure 8, we use "Accuracy" and "Relevance" to benchmark Sub-scene Captioning, and "Completeness" and "Reliability" to evaluate the capabilities of Video Summary.

1. Task Description: Your task is to generate questions and answers based on the plot events depicted in various media, including movie, TV series, and cartoon animations.

- The questions should target specific details or events within the given video. Both factual and inferential questions are encouraged.
- Avoid using specific character names in the questions. Instead, use pronouns or identify characters by unique attributes or roles (e.g., attire, age, profession).
- Ensure that the plot referred to in your question is unique within the long video. Avoid using vague descriptions that can apply to multiple instances (like "eating"). Instead, refer to unique scenes or add enough details to specify the exact event.

- Questions should be structured in a multiple-choice format. Each question should have one correct answer and three plausible, yet incorrect, distractor options.

- How does the character in the small boat end up?
- How did the warship and the small boat approach each other?
- Why didn't the old man buy the chicken?
- What mode of transportation did the old man take in the end?
- What was the young woman doing when she drove to the airport?

[illegible]

H. Explorations of Video Retrieval Augmented Generation

decrease is noted in Action Order and Overall Reasoning. This is primarily because RAG facilitates the retrieval of detail-oriented video clips, which makes models more likely to focus on answer-related cues in specific single-detail reasoning tasks. However, RAG exhibits limited capabilities in multi-detail reasoning and holistic understanding tasks, which require global perception and knowledge aggregation.

The pipeline of our video retrieval augmented generation is illustrated in Figure 9. Initially, a long video is uniformly divided into N video clips, each containing C frames. Subsequently, we employ a robust video feature extraction tool, LanguageBind [65] to extract clip embeddings $F_I \in \mathbb{R}^{N \times d}$, where d represents the dimension of each clip embedding.

Annotation Guidelines for Sub-Scene Captioning

1. Task Description: You are required to provide a detailed **caption** for a specific scene in a long movie and clearly provide a unique **prompt** that can point to this scene.

2. Prompt Requirements:

- The clue in the prompt should direct to a specific and singular scene in the movie.
- Ensure that the prompt does not contain specific character names or movie-specific terms.
- The scene to be described should generally not exceed 1 minute.

3. Caption Requirements:

- Avoid using specific character names in the captions. Instead, use pronouns or identify characters by unique attributes or roles (e.g., attire, age, profession).
- Provide a caption and a list of unique plot details as scoring points, ensuring there's no repetition of details already present in the prompt.

4. Examples:

- Example (1):
 - Prompt: Please describe the situation after the man at the door takes off his hat and throws it away.
 - Caption: The hat flies into the room and is kicked into the large clock by the man in black who stands up.
 - Scoring points: "The hat flies into the room", "is kicked into the large clock", "by the man in black who stands up"
- Example (2):
 - Prompt: Please describe the reaction of the short-haired man when the long-haired man took out the urn.
 - Caption: The short-haired man stood up, held the urn in his hands, and pressed his forehead against the mouth of the urn, unable to hold back his tears.
 - Scoring points: "The short-haired man stood up", "held the urn in his hands", "pressed his forehead against the mouth of the urn", "unable to hold back his tears"

Figure 6. Annotation Guidelines for the Sub-Scene Captioning Task.

We then compute the similarities between F_I and the text embedding F_T , concatenating the top K clips to enhance the model’s capability for question-answering. Given that many Video MLLMs are limited to processing only 16 frames, we have adjusted the settings for C and K to accommodate video retrieval in 16-second intervals. As discussed below, the RAG strategy excels in detail-oriented tasks but shows limitations in global understanding tasks. Moreover, it is relatively inefficient, requiring more than one minute to complete the process. Consequently, more effective approaches need to be developed for long video understanding tasks, and we aim to address this in future work.

I. More Visualized Examples of MLVU.

We present additional visualizations of our MLVU annotation examples in Figures 10, 11, and 12.

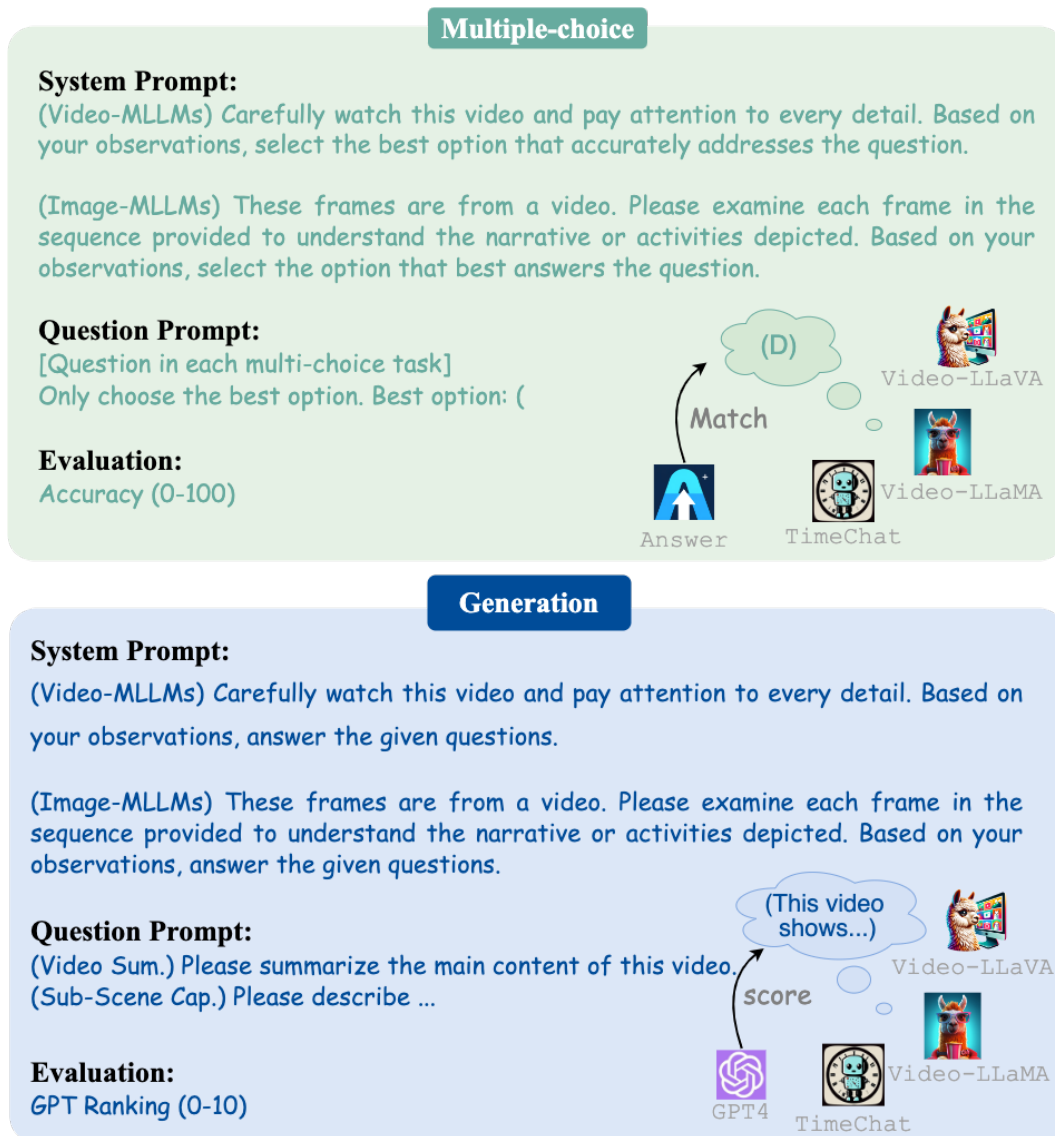


Figure 7. Inference template for our MLVU.

Evaluation Prompt For Sub-Scene Captioning Task

##TASK DESCRIPTION: You are required to evaluate a respondent's answer based on a provided question, some scoring points, and the respondent's answer. You should provide two scores. The first is the accuracy score, which should range from 1 to 5. The second is the relevance score, which should also range from 1 to 5. Below are the criteria for each scoring category.

##ACCURACY Scoring Criteria:

Evaluate the respondent's answer against specific scoring points as follows:

Score 1: The response completely misses the scoring point.

Score 3: The response mentions content related to the scoring point but is not entirely correct.

Score 5: The response accurately addresses the scoring point.

Calculate the average score across all scoring points to determine the final accuracy score.

##RELEVANCE Scoring Criteria:

Assess how the respondent's answer relates to the original question:

Score 1: The response is completely off-topic from the question.

Score 2: The response is partially related to the question but contains a significant amount of irrelevant content.

Score 3: The response primarily addresses the question, but the respondent seems uncertain about their own answer.

Score 4: The response mostly addresses the question and the respondent appears confident in their answer.

Score 5: The response is fully focused on addressing the question with no irrelevant content and demonstrates complete certainty.

##INSTRUCTION:

1. Evaluate ACCURACY: First, assess and score each scoring point based on the respondent's answer. Calculate the average of these scores to establish the final accuracy score. Provide a detailed rationale before assigning your score.

2. Evaluate RELEVANCE: Assess the relevance of the respondent's answer to the question. Note that when evaluating relevance, the correctness of the answer is not considered; focus solely on how relevant the answer is to the question. Provide a comprehensive rationale before assigning your score.

3. Output Scores in JSON Format: Present the scores in JSON format as follows...

Evaluation Prompt For Video Summarization Task

##TASK DESCRIPTION:

You are required to evaluate the performance of the respondent in the video summarization task based on the standard answer and the respondent's answer. You should provide two scores. The first is the COMPLETENESS score, which should range from 1 to 5. The second is the RELIABILITY score, which should also range from 1 to 5. Below are the criteria for each scoring category:

##COMPLETENESS Scoring Criteria:

The completeness score focuses on whether the summary covers all key points and main information from the video.

Score 1: The summary hardly covers any of the main content or key points of the video.

Score 2: The summary covers some of the main content and key points but misses many.

Score 3: The summary covers most of the main content and key points.

Score 4: The summary is very comprehensive, covering most to nearly all of the main content and key points.

Score 5: The summary completely covers all the main content and key points of the video.

##CORRECTNESS Scoring Criteria:

The correctness score evaluates the correctness and clarity of the video summary. It checks for factual errors, misleading statements, and contradictions with the video content. If the respondent's answer includes details that are not present in the standard answer, as long as these details do not conflict with the correct answer and are reasonable, points should not be deducted.

Score 1: Contains multiple factual errors and contradictions; presentation is confusing.

Score 2: Includes several errors and some contradictions; needs clearer presentation.

Score 3: Generally accurate with minor errors; minimal contradictions; reasonably clear presentation.

Score 4: Very accurate with negligible inaccuracies; no contradictions; clear and fluent presentation.

Score 5: Completely accurate with no errors or contradictions; presentation is clear and easy to understand.

##INSTRUCTION:

1. Evaluate COMPLETENESS: First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.

2. Evaluate CORRECTNESS : First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.

3. Output Scores in JSON Format: Present the scores in JSON format as follows...

Figure 8. Detailed prompt for evaluation of generation tasks in MLVU.

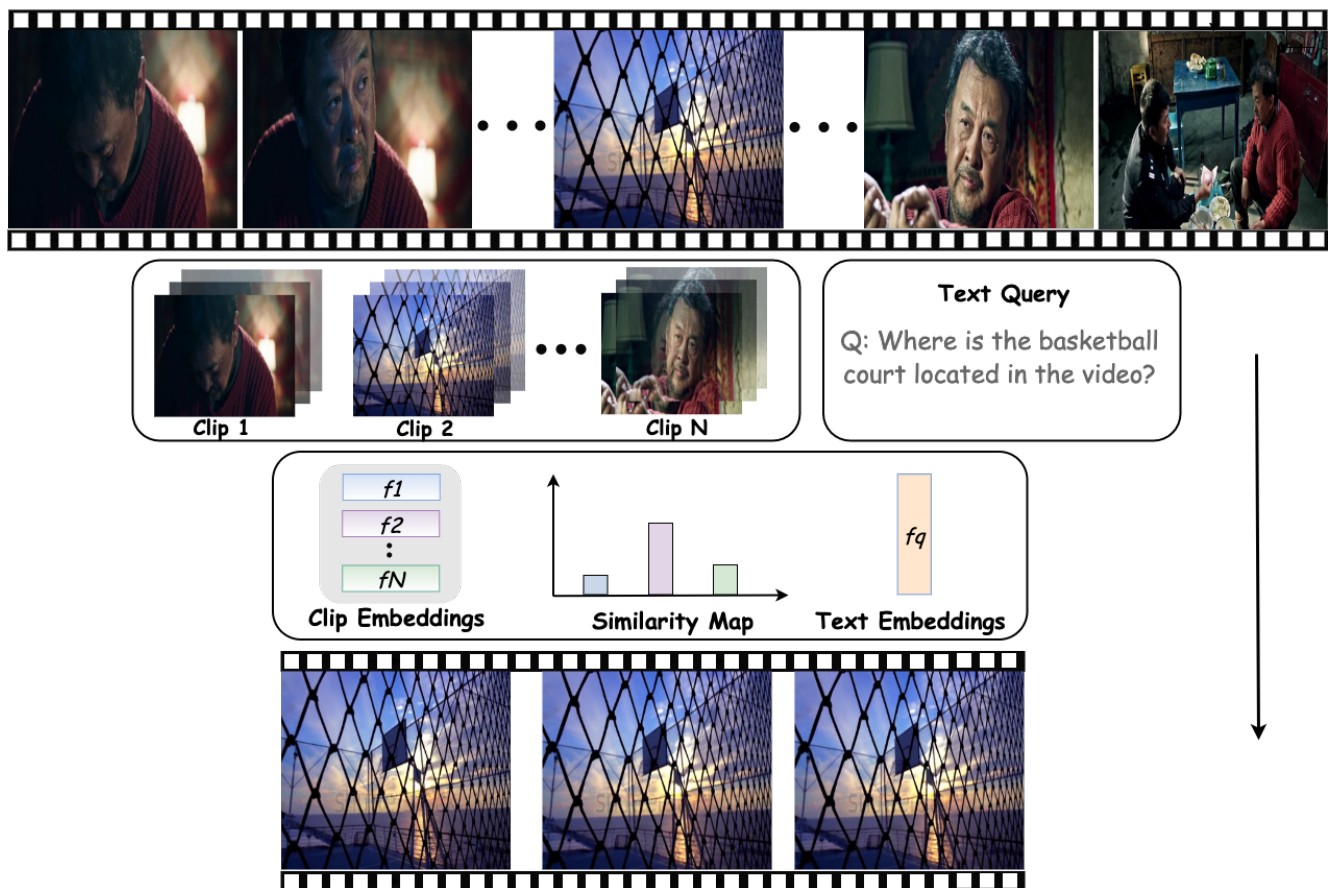


Figure 9. Pipeline of our video retrieval augmented generation strategy.

Model	Settings	Needle QA	Ego Rea.	Plot QA	Action Or.	Action Co.	Anomaly Rec.	Topic Rea.
LLaVA-B	-	43.1	38.4	41.0	25.5	25.7	41.0	60.6
LLaVA-R	C=2,K=8	50.7	45.7	49.7	26.3	26.7	40.8	59.8
	C=4,K=4	53.5	43.5	50.6	25.9	29.6	39.9	58.5
	C=8,K=2	55.2	42.6	50.3	25.1	30.1	40.6	59.5
InternVL-B	-	52.7	43.5	54.4	32.8	23.8	67.0	78.8
InternVL-R	C=2,K=8	77.2	52.6	61.4	30.1	36.4	57.9	69.2
	C=4,K=4	76.3	51.4	59.9	29.3	36.9	58.3	69.4
	C=8,K=2	77.8	48.9	61.6	31.7	33.0	60.2	62.3
Video-LLaMA-B	-	39.4	33.5	35.4	18.5	25.7	41.5	54.5
Video-LLaMA-R	C=2,K=8	61.4	42.6	38.8	17.4	17.5	35.7	48.5
	C=4,K=4	58.9	42.6	39.1	17.8	23.8	36.0	49.3
	C=8,K=2	62.0	38.4	36.2	25.5	18.0	38.5	51.0
VideoChat2-B	-	42.0	47.4	43.8	22.8	29.6	51.5	74.6
VideoChat2-R	C=2,K=8	72.1	53.7	55.5	21.6	30.1	45.8	68.2
	C=4,K=4	72.4	55.4	53.4	22.4	31.1	45.3	68.9
	C=8,K=2	73.8	53.1	55.3	22.0	31.6	46.6	69.7
MiniGPT4-Video-B	-	49.0	48.6	44.5	23.2	23.0	52.5	70.9
MiniGPT4-Video-R	C=2,K=8	60.6	44.3	47.4	23.2	23.7	42.8	60.9
	C=4,K=4	60.3	44.6	46.9	26.3	23.8	42.6	60.7
	C=8,K=2	56.3	44.6	46.6	27.4	24.8	45.0	47.5

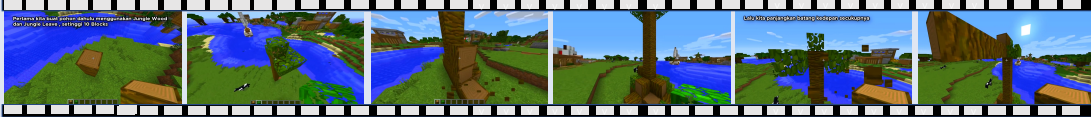
Table 3. Quantitative results on video Retrieval Augmented Generation. “model-B” and “model-R” denote Baseline and RAG models respectively. We evaluate two image MLLMs and three video MLLMs in different settings.

Topic Reasoning



Question: What type of film is this?

- (A) Mystery (B) Action (C) Comedy **(D) Romance**



Question: What is this video about?

- (A) A person in the game taking care of pets **(B) A person in the game building a structure by the lake**
 (C) A person in the game planting trees by the lake (D) A documentary about humans and nature



Question: Where is the main setting of the video?

- (A) Desert (B) Grassland **(C) Outside the house** (D) Inside the house

Video Summarization



Prompt: Please summarize the main content of this video.

The video begins with two men talking in a dimly lit room. After one of the men leaves, he enters another house where an elderly woman is present. They engage in conversation, and the elderly woman appears sad. In another scene, two women are talking, and one of them takes car keys and leaves. She arrives at another location and talks with a woman and a man. Subsequently, one of the women makes a phone call.



Prompt: Please summarize the main content of this video.

The video starts with a man singing into a microphone, with a few other men playing instruments behind him. The scene changes to someone pushing open a door and walking into a room where others are resting. She then opens another door, enters a room and starts arguing with the singing man, which results in a fight. Next, the woman drives the man away, which results in a car crash. The car then falls off a bridge and gets hit by another car. The screen goes black and then lights up again, revealing a bookshelf filled with books at the end.

Figure 10. More Examples of Topic Reasoning and Video Summarization Tasks.

Needle Question-Answering



Question: Where is the senior businessman having a serious conversation on the cell phone?

- (A) In a park **(B) By the sea shore** (C) In his office (D) At a restaurant



Question: What are the little girl and her grandmother doing together?

- (A) Watching TV (B) Playing a game **(C) Reading a children's book** (D) Eating dinner

Plot Question Answering



Question: What happened after the person with the yellow stripe arrived at the camp?

- (A) He went to eat (B) He went hunting
(C) He went to war **(D) He started a fight with the person holding the pipe**




Question: What color is the table lamp in the background of the scene where a man and a women are chatting?

- (A) Black **(B) White** (C) Green (D) Yellow


Figure 11. More Examples of Needle Question Answering and Plot Question Answering Tasks.

Sub-Scene Captioning



Prompt: Please describe the situation after the woman in red walked to the window of the bridal shop.

Answer: The woman in red took a picture with her camera. As the photo slowly slid out, she looked down at it.



Prompt: Please describe the process of a man alone in a room looking for a camera.

The man raises his cue stick to find the angle, then turns around and walks to a statue where he finds the camera.

Figure 12. More Examples of Sub-Scene Captioning.