VLV-Bench: A Comprehensive benchmark for very long-form videos understanding

Anonymous ACL submission

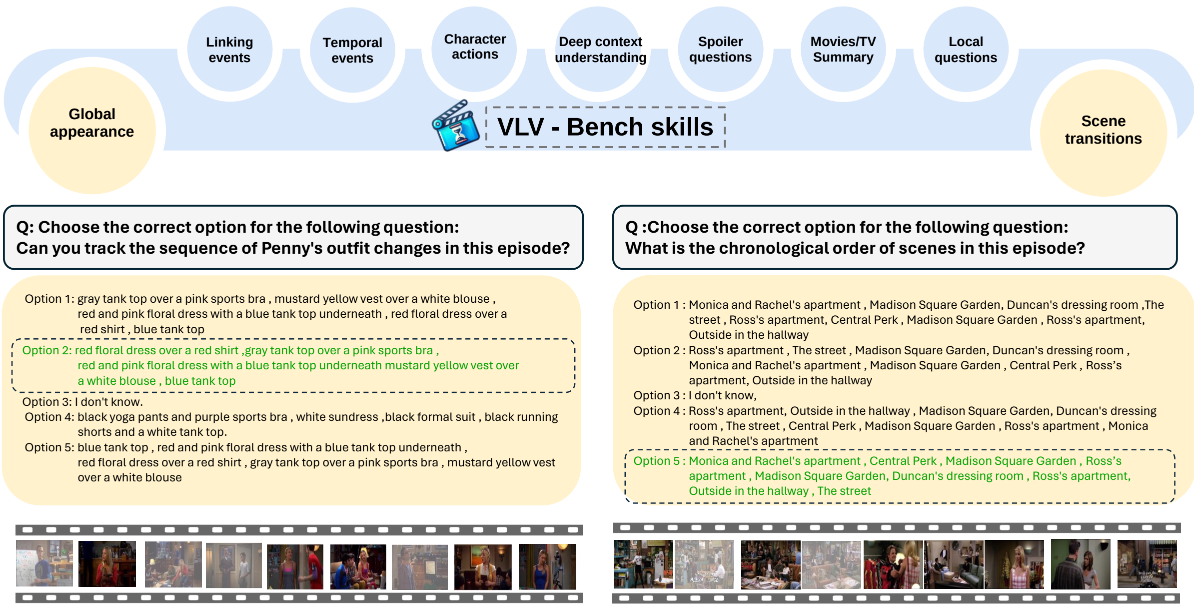

Figure 1: The set of skills introduced by VLV-Benchmark includes a total of 9 skills. The figure includes two question examples for two distinct skills: the left example illustrates the Global Appearance skill, and the right example illustrates the Scene Transition skill.

Abstract

Understanding long videos, ranging from tens of minutes to several hours, presents unique challenges in video comprehension. Despite the increasing importance of long-form video content, existing benchmarks primarily focus on shorter clips. To address this gap, we introduce a comprehensive benchmark for **V**ery **L**ong **V**ideos understanding (**VLV-Bench**), which presents ***1) The longest*** video duration, averaging 76.34 minutes; ***2) The largest*** number of question-answer pairs, 108.2K; ***3) Diversity*** in questions that examine nine different skills and include both multiple-choice questions and open-ended questions; ***4) Human-centric***, as the video sources come from movies and daily TV shows, with specific human-level question designs such as Movie Spoiler Questions that require critical thinking and comprehensive understanding. Using VLV-Bench, we comprehensively evaluate existing Large Multi-Modality Models (LMMs) on each skill, including the commercial model Gemini 1.5 Flash and the open-source models. The evaluation shows significant challenges in our benchmark. Our results show that the best AI models such Gemini struggles to perform well with 42.72% average accuracy and 2.71 out of 5 average score. We hope this benchmark will stimulate the LMMs community towards long video and human-level understanding. Our benchmark can be accessed at VLV-Bench and will be made publicly available.

1 Introduction

Recent Large Language Models (LLMs) (Li et al., 2023a; Achiam et al., 2023) have shown impressive progress in the Natural Language community. Inspired by the strong abilities of LLMs, Large Multi-Modality Models (LMMs) (Ataallah et al., 2024; Zhu et al., 2023; Zhang et al., 2023; Chen et al., 2023; Lin et al., 2023; Liu et al., 2023b; Maaz et al., 2023) which equip the LLMs with visual processors have been developed to solve cross-modality tasks such as image understanding and short video understanding. While current large multi-modality models show some progress in video understanding, their abilities remain unclear for very long-

	Name	#N Q	#N Videos	#T (mins)	Global Q	Type of Q		Source of QA			Annotation	
						MCQ	Open	Video	VS	VSum	Auto	Human
Short	MSRVTT-QA (Xu et al., 2017)	72.8 K	2990	0.25	✗	✗	✓	✓	✗	✗	✓	✗
	TGIF-QA (Jang et al., 2017)	8.5 K	9575	0.05	✓	✗	✓	✓	✗	✗	✓	✓
	MV-Bench (Li et al., 2024)	4.0 K	3641	0.27	✓	✓	✗	✓	✗	✗	✓	✗
Long	Activity-QA (Yu et al., 2019)	8.0 K	800	1.85	✗	✗	✓	✓	✗	✗	✗	✓
	TVQA (Lei et al., 2019)	15.2 K	2179	1.86	✗	✓	✗	✓	✗	✗	✗	✓
	Egoschema (Mangalam et al., 2023)	5.0 K	5063	3.00	✓	✓	✗	✓	✗	✗	✓	✓
	Moviechat (Song et al., 2023)	13.0 K	1000	9.40	✓	✗	✓	✓	✗	✗	✗	✓
Very Long	**VLV-bench (Ours)**	**108.2 K**	**1219**	**76.34**	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Comparison between VLV-bench and existing video understanding benchmarks.** VLV-bench has the largest QA pairs, the most videos, and the longest average duration. (Note: Global Q stands for whether any challenging questions are designed to explain the whole video. VS is the video's script, and VSum is the summary of the video.)

form video understanding.

Long-form video understanding (Song et al., 2023; Regneri et al., 2013; Rohrbach et al., 2014; Awad et al., 2017, 2018, 2020) not only challenges these models by increasing the number of images but also contains more comprehensive information, making it a boundary-pushing task toward human-level intelligence. For example, humans can link multiple events at different times and answer questions requiring a deep understanding of events or characters in a long video. Multi-modal models can address these questions, requiring long-range temporal-spatial reasoning and strong vision-language alignment abilities, potentially serving a wider range of AI applications.While the necessity of a long-video understanding benchmark is evident, there is very limited work (Song et al., 2023; Mangalam et al., 2023) that attempt to develop such benchmarks

However,these benchmarks is either relatively short, up to 10 minutes or lack in diversity such as having only some template questions that are repeated for the whole dataset. To fill this gap in comprehensive long video understanding, we propose VLV-Bench, a comprehensive benchmark for very long-form video understanding. As shown in Tab 1, VLV-Bench is currently the video benchmark that has both the longest length (76.34 minutes) and the largest number of question-answer (QA) pairs (108.2K). The video sources are movies and daily TV shows, and the questions are designed based on multiple sources including video frames, video scripts, and video summaries. As shown in Figure 1 , the QA pairs consist of nine carefully designed types of questions, which mainly focus on human-centric aspects, including Summarization, Global Appearance, Scene Transitions, Sequence of Actions by Each Character, Temporal Questions, Linking Events, Deep Context Understanding, Movie Spoiler Questions, Local Visual and contextual Questions. The questions include multiple-choice questions (MCQs) and open-ended questions. We report the accuracy for MCQs and the GPT-4 rating score for open-ended questions. The annotation process is mainly done by an automatic pipeline using GPT-4, which includes proposing questions and generating answers. To prevent hallucinations and gather sufficient information for generating QA pairs, we collect various sources of information including video frames which are used in the global appearance skill, video transcripts, and video summaries.

Based on VLV-Bench, we evaluate the current state-of-the-art MLLMs capable of handling very long videos, including the open-source models Movie-Chat, Llama-Vid, Large World Model, and the only commercial model capable of handling long videos, Gemini 1.5 Flash.

We summarize the key experimental findings here: (1) All existing models struggle with VLV-benchmark, showing the unique challenges of our benchmark. (2) Experiments show that Gemini outperforms all open-source models on each skill with a large gap. (3) The most difficult skills are Deep Context Understanding and Movie Spoiler questions, which require both visual and contextual understanding. Especially, Movie Spoiler questions are designed to relate to human understanding, posing specific challenges.

By introducing this comprehensive VLV-Bench, we hope to:

- Help bridge the gap of lacking a large-scale long-form video understanding benchmark.
- Boost the development of current open-source LMMs.
- Push LMMs towards human-centric and human-level long video understanding.

2 Related Work

Existing Video understanding Benchmarks. Here we refer to the average time of less than 1 minute as a short video, 1-10 as a long video, and >10 minutes as a very long video. The previous short and long videos are listed in Table 1. Our benchmark is the only one that includes very long videos. Short video benchmarks have been extensively studied (Jang et al., 2019; Xu et al., 2017; Lei et al., 2019). MSRVTT-QA (Xu et al., 2017) has a large number of questions, but it does not support global questions, and the annotations are automatically generated without human verification. TGIF-QA (Jang et al., 2017) and MV-Bench (Li et al., 2024) are short video benchmarks that support global questions, but their scale is limited. Activity-QA (Yu et al., 2019), TVQA (Lei et al., 2019), do not support global and have only local questions. Egoschema (Mangalam et al., 2023) is a human-annotated Long-form Video understanding Benchmark and video length are three-minute-long.

The most relevant dataset to our work is MovieChat-1K (Song et al., 2023), a benchmark for long video understanding. MovieChat-1K is based on videos with an average duration of 9.4 minutes and includes 1,000 video clips from different genres, with 14,000 annotations for diverse visual narratives and question-answering pairs.

Our dataset has several advantages over previous benchmarks: (1)We support the very long videos; (2) Our scale is significantly larger; (3) Ours support both MCQ and open-ended evaluations; (4) Ours include the script of the video and a summary of the video as sources for QA.

Long Video Models. Google Gemini-Flash 1.5 model (Gemini) is currently the only available commercial model capable of processing extremely long videos, boasting an unprecedented context window of 1 million tokens. This extensive context window allows Gemini-Flash 1.5 to effectively handle both video frames and subtitles simultaneously. In contrast to the commercial solutions, LLama-vid (Li et al., 2023b) is a recent open-source model that comprehends long videos due to its excellent efficiency in representing each frame using only two tokens. The Large World Model (LWM) (Liu et al., 2024) is another open-source model capable of processing millions of tokens using the innovative ring attention mechanism (Liu et al., 2023a). Consequently, Moviechat (Song et al., 2023) processes

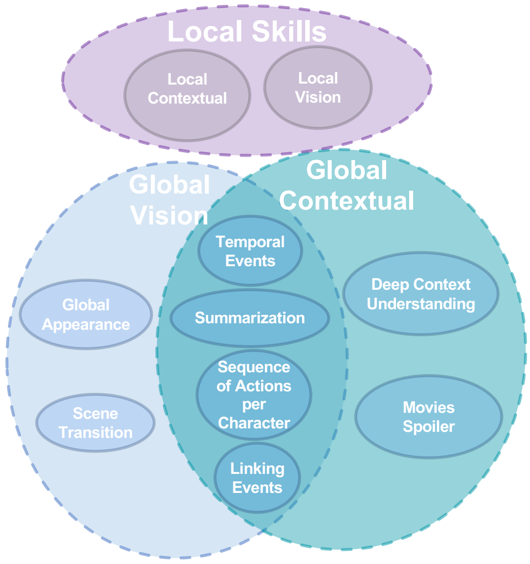

Figure 2: Skills mind-map. An abstract overview of the skills covered in our benchmark, grouped based on relevancy.

long videos but without subtitles and operates in two modes: global and breakpoint. The global mode exclusively utilizes long-term memory, and the breakpoint mode additionally incorporates the current short-term memory as part of the video representation. The breakpoint mode allows for understanding the video at a specific moment in time.

3 VLV-Benchmark

In this section, we first dissect the skills definition, grouped in Figure 2 (Sec 3.1), then the data collection pipeline (Sec 3.2), and finally, the benchmark statistics (Sec 3.3).

3.1 Skills

To create a robust benchmark for long video understanding, the questions should encompass local and global events throughout the video. Additionally, the questions should address the video's visual and contextual content. Based on these considerations, we defined a long video understanding covering nine skills through four critical aspects, as shown in Figure 2.

3.1.1 Global Vision Skills

Global Appearance. In this skill, we focused on generating questions that require continuous visual understanding, which cannot be answered from short video segments but necessitate watching the entire video. We selected changes in outfits as the basis for these continuous vision questions. To cre-

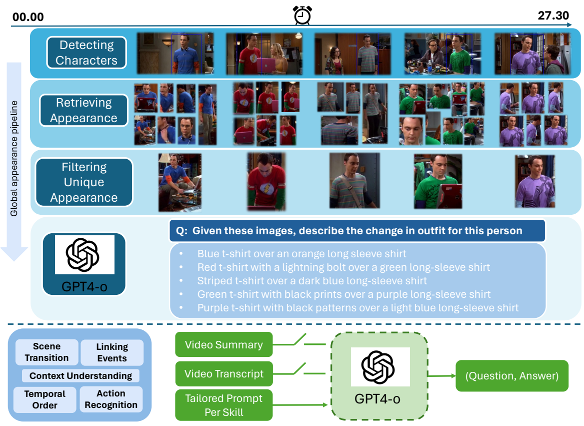

Figure 3: Full annotation pipeline for VLV-Bench skill set. The upper section depicts the global appearance pipeline, while the lower section illustrates the question generation using GPT-4. The gates for video summary and video transcript indicate that some skills utilize only the summary, others use only the transcript, and some use both.

ate this type of question, we developed the global appearance pipeline, as shown in Figure 3. The TVQA+ (Lei et al., 2020) dataset was used, providing bounding boxes for each character in one of the six TV shows in TVQA (Lei et al., 2019), specifically *The Big Bang Theory*. Images were cropped using these bounding boxes, and all images of each character in the episode were collected. Manual filtering was performed to select each character's best and most unique outfits. GPT-4 described the outfit for each unique image and generated a sequence list of the outfits For evaluation, multiple-choice questions were formulated by altering the sequence of outfits. For example: "Choose the correct option for the following question: In what order does Leonard change outfits in this episode?" The correct option is (a) a red T-shirt under a beige jacket with a green hood, a white t-shirt with a green print under a grey jacket and black vest, or a white dress shirt with a patterned tie under a brown blazer. Other options present the outfits in the incorrect order. In special cases where a character's outfit does not change throughout the episode, distractor options with incorrect outfits were added as alternative choices.

Scene Transitions. Scene transition skills necessitate continuous visual comprehension and cannot be adequately addressed using short video segments; they require viewing the entire video. To assess this skill, questions concerning transitions between scenes were generated. It was observed that the locations of each scene are mentioned in the transcript. Utilizing GPT-4 by inputting the transcript of the TV shows as in Figure 3. We extracted these locations and created a list in the correct sequence. Then, for evaluation, we follow a template-based approach to collect multiple-choice questions to assess the correct sequence of these scene transitions.

3.1.2 Global contextual questions

Deep Context Understanding. For this skill, we aim to test the model's ability to answer hard and tricky questions requiring a deep understanding of the full video. We utilized GPT-4 to generate challenging and nuanced questions about the video. We did not restrict GPT-4 to a specific skill set, allowing the advanced AI model to autonomously generate questions. We provided GPT-4 with comprehensive information about the video, including the transcript and summary as in Figure 3, enabling it to create complex questions that require a profound understanding of the context and the main topic of the movie or the TV show. These open-ended questions were developed for the Long TVQA we created and MovieNet(Huang et al., 2020) datasets.

Movies Spoiler Questions. Spoiler questions are inquiries that reveal critical plot points, twists, or specific details that could potentially spoil the experience for viewers who have not yet seen the movie. These questions are crucial for evaluating long videos because they delve into significant, often pivotal moments in the narrative, requiring a deep and comprehensive understanding of the entire storyline. These questions are important for long video evaluation for several reasons:

- *Comprehensive Understanding*: Answering spoiler questions necessitates a thorough comprehension of the entire video, as they often reference events from various points in the narrative. This ensures that the evaluator has engaged with the content meaningfully and sustainably.
- *Critical Thinking*: These questions require viewers to think critically about the plot and its developments, analyzing character actions and narrative resolutions.
- *Detail Orientation*: Spoiler questions often focus on specific, detailed aspects of the plot, ensuring that the evaluator has paid close attention to the video.

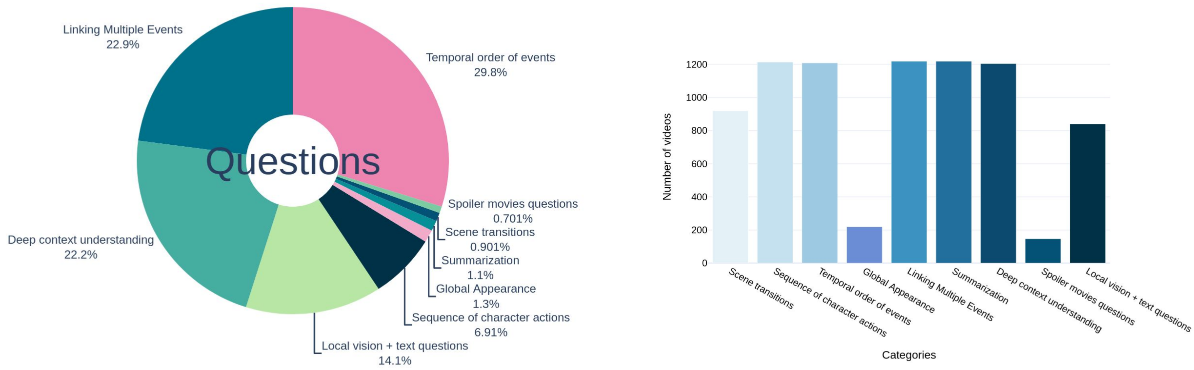

Figure 4: Left) Number of questions distribution for each skill set. Right) Number of videos for each skill.

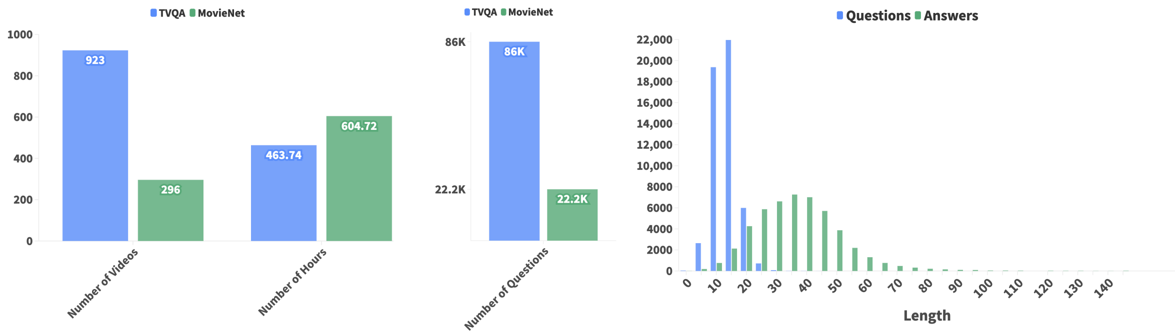

Figure 5: Data statistics. On the left, we report the number of videos and their length in hours from each data source: TVQA and MovieNet datasets. In the middle, we demonstrate the number of questions. On the right, we show the histogram of the lengths of the questions and answers.

3.1.3 Global vision and contextual questions

Sequence of Actions by Each Character. This skill involves generating questions about each character's actions, encompassing both contextual and visual actions, which can often be identified in the transcript where scene actions are described. For example, "Rachel serving coffee to her friends in Central Park." To create these questions, we utilized GPT-4 by inputting both the video summary and the transcript as in Figure 3. This approach ensures that the questions accurately reflect the sequence of actions depicted in the video. To evaluate this skill, we formulated multiple-choice questions regarding the correct order of actions performed by each character. These questions were generated for both the Long TVQA and MovieNet datasets.

Temporal Questions. This skill assesses the temporal understanding of long videos by generating questions about the correct sequence of events in movies or TV series, and these events cover both visual and contextual events. We ask questions regarding which event occurred first or the correct order of adjacent events. For instance, "Is event A before event B?" or "What is the correct sequence of these events: event A, event B, or event C?" To generate these questions, we utilized GPT-4 by inputting the episode's transcript as in Figure 3. We used the transcript instead of the summary, as the correct order of events can only be accurately extracted from the detailed transcript.

These questions are presented in a multiple-choice format and generated for both the Long TVQA we created and MovieNet (Huang et al., 2020) datasets.

Linking Events. This skill involves generating a set of questions that link multiple events together, such as events from the beginning of an episode that affect later events, to ensure the questions comprehensively cover the entire video. Examples of such questions include:

- What is the influence of event A on event B?
- How does event A lead to event B?
- What is the relationship between event A and event B?
- What is the impact of event A on event B?

We generated these questions by inputting the video summary into GPT-4 and instructing GPT-4 to create this type of question as in Figure 3. These open-ended questions were developed for the Long TVQA we created and MovieNet(Huang et al., 2020) datasets.

Summarization. Summarization is a critical skill for evaluating long sequence data, such as long text understanding in NLP, and is equally crucial for assessing long video comprehension. Our benchmark includes human-generated summaries for movies and TV shows sourced from IMDb. These summaries, created by humans, encapsulate visual and contextual events in the videos, making it a strong skill for evaluating a long video understanding.

3.1.4 Local vision and contextual questions

Local Vision and Text Questions. In this skill, we will talk about the importance of local questions besides the global questions. Local questions in our benchmark are responsible for testing the model's ability to localize the questions in a long video. If the model can answer these questions, it can focus on the fine-grained details in the video.

3.2 Data Collection

We utilized two sources to obtain very long videos: Movies and TV shows. For Movies, we employed the MovieNet dataset (Huang et al., 2020). However, no dataset is available for complete TV shows, as TVQA(Lei et al., 2019) provides only short clips. To address this limitation, we transformed the TVQA dataset from a collection of short clips into a long video dataset by gathering and sequencing the clips corresponding to each episode, thereby reconstructing the full episode frames. We obtained 924 full-length episodes from six different TV shows through this modification. Consequently, MovieNet dataset(Huang et al., 2020), it is found that only 296 movies had shots aligned with subtitles. Therefore, only these movies are included; we excluded the rest from our benchmark. In addition, we relied on two extra data sources: video summaries and transcripts. For the TVQA dataset (Lei et al., 2019), the summaries from IMDB and the transcripts were scraped for the 924 episodes. For the filtered MovieNet (Huang et al., 2020), we obtained transcripts from the MovieNet annotations. However, since MovieNet (Huang et al., 2020) annotations do not include complete movie summaries, the missing summaries are scrapped from IMDB to obtain comprehensive movie summaries and transcripts for all filtered movies.

For spoiler skill, out of 296 movies in the MovieNet (Huang et al., 2020) dataset, we identified 147 movies with associated spoiler questions available on IMDb, totaling 806 questions. These questions were meticulously collected and integrated into our benchmark dataset. Consequently, we directly adopted TVQA questions for the local skills by aggregating questions corresponding to clips from the same episode, ensuring multiple questions per episode. Notably, these questions in TVQA (Lei et al., 2019) exhibit a dual property encompassing visual and contextual dimensions. It's pertinent to mention that these questions are exclusive to the TVQA dataset and have hitherto remained unutilized for long video evaluation solely for analyzing short clips.

3.3 Benchmark statistics

The VLV-Benchmark is the largest long video question-answering benchmark, containing 108.2K questions covering nine distinct skills. Figure 4 (left) illustrates the distribution of the number of questions for each skill. Additionally, our benchmark includes the largest collection of long videos, with a total of 1,219 videos, as detailed in Table 1. Figure 4 (right) depicts the distribution of these videos across the different skills. Figure 5, on the left, shows the detailed distribution of the number of questions for each skill in our benchmark. On the right, we discuss the number of videos that have been used for each skill. For more benchmark statistics details see A.3 in the supplementary.

4 Experiments

4.1 Evaluation Metrics

We employed distinct evaluation metrics appropriate for the two questions types: open-ended and multiple-choice (MCQs). For MCQs, accuracy was the chosen metric, while for open-ended questions, we utilized a scoring system based on GPT-4, ranging from 0 to 5. For MCQ, GPT-4 is used to match the predicted answer with one of the options or to match with the "I don't know option", that indicates there is no match. See Sec. A for more details. For open-ended questions, GPT-4 evaluated the LLMs' predictions based on multiple criteria: correctness, meaningfulness, proximity to the expected answer, presence of hallucinations, and completeness. Based on these criteria, GPT-4 generates a score ranging from 0 to 5, reflecting the overall quality of the response.

(a) Global Appearance

Rank	Model	Acc
1	**Gemini-Flash 1.5**	33.31
2	LLama-vid	9.47
3	Large world Model (LWM)	7.35
4	Moviechat	6.59

(b) Scene transition

Rank	Model	Acc
1	**Gemini-Flash 1.5**	29.48
2	Moviechat	6.41
3	Large world Model (LWM)	5.54
4	LLama-vid	3.6

(c) Sequence of character actions

Rank	Model	Acc
1	**Gemini-Flash 1.5**	35.48
2	LLama-vid	6.52
3	Large world Model (LWM)	6.41
4	Moviechat	4.51

(d) Temporal order of events

Rank	Model	Acc
1	**Gemini-Flash 1.5**	54.92
2	LLama-vid	40.52
3	Large world Model (LWM)	38.44
4	Moviechat	36.99

(e) Local visual+context questions

Rank	Model	Acc
1	**Gemini-Flash 1.5**	60.41
2	LLama-vid	25.65
3	Large world Model (LWM)	21.92
4	Moviechat	17.76

(f) Summarization

Rank	Model	GPT4-score(0-5)
1	**Gemini-Flash 1.5**	2.85
2	LLama-vid	1.19
3	Moviechat	0.14
4	Large world Model (LWM)	0.03

(g) Deep context understanding

Rank	Model	GPT4-score(0-5)
1	**Gemini-Flash 1.5**	2.70
2	LLama-vid	2.02
3	Large world Model (LWM)	0.88
4	Moviechat	0.55

(h) Movies Spoiler questions

Rank	Model	GPT4-score(0-5)
1	**Gemini-Flash 1.5**	1.93
2	LLama-vid	1.32
3	Large world Model (LWM)	0.55
4	Moviechat	0.34

(i) Linking Multiple events

Rank	Model	GPT4-score(0-5)
1	**Gemini-Flash 1.5**	3.34
2	LLama-vid	2.36
3	Large world Model (LWM)	1.2
4	Moviechat	0.85

(j) Average results over the Nine skills

Rank	Model	AVG Accuracy (%)	AVG Score (0-5)
1	Gemini-Flash 1.5	42.72	2.71
2	LLama-vid	17.15	1.72
3	Large World Model (LWM)	15.93	0.67
4	MovieChat	14.45	0.47

Table 2: VLV-Benchmark Leaderboard over the Nine Skills. Also, the statics of options in MCQ and random accuracy are provided in Supplementary Table 5

Rank	Model Name	Global visual questions	Global contextual questions	Global vision and context		Local vision and context
		AVG-accuracy	AVG-score	AVG-accuracy	AVG-score	AVG-accuracy
1	Gemini-Flash 1.5	31.395	2.315	45.2	3.095	60.41
2	LLama-vid	6.535	1.67	23.52	1.775	25.65
3	Large World Model(LWM)	6.445	0.715	22.425	0.615	21.92
4	MovieChat	6.5	0.445	20.75	0.495	17.76

Table 3: Average results for the high level 4 skills: Global appearance and scene transitions are Global visual questions. Movie spoiler questions and deep context understanding are global contextual questions. Linking multiple events, Character actions, summarization, and temporal order of events are Global visual and contextual questions together. Local vision and context skills contain local vision and contextual questions.

4.2 Detailed Models Setting

There are a limited number of accessible models, both commercial and open-source, that are capable of handling very long video understanding. In our evaluation, we assessed one commercial model and three open-source models.

Gemini-Flash 1.5. The Gemini-Flash 1.5 model, developed by Google (Gemini), is currently the only commercial model capable of processing extremely long videos, boasting an unprecedented context window of 1 million tokens. This extensive context window allows Gemini-Flash 1.5 to effectively handle both video frames and subtitles simultaneously.

LLama-vid. The LLama-vid model (Li et al., 2023b) accepts both video frames and subtitles. For our evaluation of the movies, we utilized our dataset with one frame per second, accompanied by aligned subtitle shots. The model was evaluated using the default settings without any modifications to the inference parameters.

Large World Model (LWM). LWM is efficiently optimized for execution on Google TPUs, and have another version for GPUs.Our evaluation is done by using (NVIDIA A100), which allows for the processing of a maximum of 8 frames per video. While this setup does not represent the optimal configuration for LWM, it was the most feasible setting.LWM can accept only the video frames without the subtitles.

Moviechat. The Moviechat model (Song et al., 2023) processes video frames without subtitles and operates in global and breakpoint modes. Our evaluation focused on the global mode, utilizing the default inference settings without any modifications.

4.3 Results

In this section, we first evaluate the existing SOTA open-source long-video understanding models and the state-of-the-art commercial model, Gemini, which is the only commercial model currently capable of handling very long videos. The overall performance averaged across all 9 skills, and Specific skill performance is detailed in Table 2, and the average results on four types of questions are illustrated in Table 3, investigating how visual and contextual information affects long video understanding.

Overall performance. The overall performance of different models on the VLV-bench is shown in Table 2 (j). Three findings can be observed: (1) All models' performance is relatively lower compared to other benchmarks (e.g., Movie-chat benchmark), highlighting the unique challenges of our benchmark, such as longer duration. (2) Gemini-Flash 1.5 achieves the best performance on both multiple-choice and open-ended questions, with 47.72 accuracy (0-100) and 2.70 GPT4-score (0-5). There is also a large performance gap between Gemini and other open-source models. (3) For open-source models, LLama-vid achieves the best result. with 17.15 accuracy and 1.7 GPT4-score. One reason may be that LLama-vid is pre-trained with longer duration QA-pairs, which helps handle longer sequences.

Performance on specific skills. Table 2 (a)-(i) shows the performance of SOTA long video understanding models on each skill. The performance varies significantly among different skills, highlighting the unique challenges introduced by each one. Obeservation of the results: (1) scene transition is the most difficult MCQ question type, with Gemini achieving only 29.48% accuracy. The potential reason for the low performance is that this question requires global reasoning across the entire hour-long video instead of one clip. (2) all models struggle with Movie Spoiler questions in open-ended questions. The difficulty lies in the need for deeper understanding and reasoning to get the correct answer. Since Movie Spoiler questions are meaningful for human-centric video understanding, current model capabilities need improvement. (3) All open-source models' results on MCQ are below random choice, except for the Local visual+context questions. This shows that the main challenge for existing models is long-sequence global reasoning.

Performance on Four Types of Questions. As introduced in Section 3.1, in the VLV-benchmark, questions for each skill can be identified as one of four high-level types: Global visual, Global contextual, Global vision + text, and Local vision + context. The results for each type of question are provided in Table 3. Only two models, Gemini Flash 1.5 and LLama-VID accept both video and video subtitles among these SOTA models. The table clearly shows that LLama-VID outperforms the other two open-source models for questions requiring context understanding. The main reason for the poor performance of LWM and MovieChat is that these two models make predictions from video only, missing important text information. This highlights the importance of long video understanding models handling both modalities. Additionally, global contextual questions are challenging for all models, requiring complex reasoning.

5 Conclusion

We introduced VLV-Bench, a comprehensive benchmark for very long-form video understanding, featuring the longest average video duration (76.34 minutes) and the largest number of question-answer pairs (108.2K). Our diverse and human-centric questions evaluate nine distinct skills, posing significant challenges to current Large Multi-Modality Models (LMMs).

Evaluations reveal that all existing models, including the commercial Gemini 1.5 Flash and various open-source models, struggle with VLV-Bench, particularly in tasks requiring deep context understanding and critical thinking. Despite these challenges, Gemini 1.5 Flash outperforms open-source models across all skills. VLV-Bench aims to bridge the gap in long-form video understanding benchmarks, promoting the development of LMMs toward achieving human-level comprehension and reasoning.

6 Limitations

This section outlines the limitations of our work: **Restricted Video Sources:** The video sources utilized in this study are limited exclusively to movies and television shows. Consequently, the benchmark lacks a broader spectrum of general videos encompassing various aspects of human life or the diverse field of wildlife. **Dependency on Transcripts:** The generation pipeline of questions and answers employed in this benchmark is inherently dependent on the availability of transcripts. This reliance confines its applicability to movies and television shows where such transcripts are readily available. For more general videos, the absence of transcripts poses a significant challenge, thereby limiting the pipeline's utility in those contexts. We hope to overcome these limitations in the future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan F Smeaton, Yvette Graham, Wessel Kraaij, et al. 2018. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*.

George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, et al. 2020. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *arXiv preprint arXiv:2009.09984*.

George Awad, Asad A Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Willie Mcclinton, Martial Michel, Alan F Smeaton, Yvette Graham, Wessel Kraaij, et al. 2017. Trecvid 2017: evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking. In *TREC Video Retrieval Evaluation (TRECVID)*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Google Gemini. Gemini technical report.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127:1385–1412.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Preprint*, arXiv:1704.04497.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering. *Preprint*, arXiv:1809.01696.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. *Preprint*, arXiv:1904.11574.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023a. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. Llama-vid: An image is worth 2 tokens in large language models. *Preprint*, arXiv:2311.17043.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. World model on million-length video and language with blockwise ringattention. *Preprint*, arXiv:2402.08268.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *Preprint*, arXiv:2310.01889.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings, Part II 36*, pages 184–195. Springer.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Evaluation Details

A.1 Evaluation metric details

For MCQs, large language models (LLMs) do not consistently provide direct responses. The output may vary, sometimes giving the option number, other times the option sentence, or occasionally providing additional clarifications for the selected option. For example, an LLM might produce a response such as: "I think option 1 is close, but my final answer will be option 2." Additionally, some responses may include hallucinations not found in the given options. To address this variability, we implemented a standardized evaluation method using GPT-4 to match the LLM's prediction with one of the provided options. Specifically, we input the set of options and the LLM's prediction into GPT-4, which then attempts to match the predicted answer with one of the given options. If no matching option is found or if the response includes hallucinations, GPT-4 matches the prediction with an "I don't know" option. Using the prediction option number and the ground truth option number, we then calculate the accuracy. For open-ended questions, GPT-4 assessed the LLMs' predictions based on several criteria: correctness, meaningfulness, alignment with the expected answer, presence of hallucinations, and completeness. Using these criteria, GPT-4 assigned a score from 0 to 5 to indicate the overall quality of each response.

A.2 Evaluation prompts details

In this section we will discuss the details for the prompts that have been used for evaluation for both the open ended questions and multiple choices. Figure. 6 show the detailed prompt used for the results matching.Figure 7 show the detailed prompt for the GPT-4 scores.

A.3 Extra statistics

Table. 4 shows the video durations for our various video sources, such as (Lei et al., 2019) and (Huang et al., 2020). The VLV-Benchmark includes some videos with a maximum duration of 201 minutes (3.35 hours).

Table. 5 provides details about the number of options for each multiple-choice question (MCQ) skill, including Global Appearance, Scene Transitions, Sequence of Character Actions, Temporal Order of Events, and Local Vision and Context Questions. The table also reports the weighted random accuracy for each skill.

video source	Minimum (min)	Maximum (min)	Average (min)
TVQA (Lei et al., 2019)	17.81	53.32	30.11
MovieNet (Huang et al., 2020)	81.04	201.82	122.57

Table 4: VLV-Benchmark videos duration analysis

Skill Name	Number of options				Weighted
	2	5	6	7	Random accuracy
Global Appearance	0	9	1447	0	0.17
Scene transitions	0	0	920	0	0.17
Character actions	0	0	5829	1665	0.16
temporal order of events	24056	0	8208	0	0.42
Local vision + text questions	0	15246	0	0	0.2

Table 5: Detailed calculations for the random accuracy for the whole MCQ skills

B Extra Benchmark Examples

Here in this sections, we are showing more examples of our benchmark skills such as the temporal order of events in Fig. 8, linking multiple events in Figure.9, deep context understanding in Figure. 10 , local questions in Figure.11 , and summarization in Figure. 12.

C Benchmark Generation Details

This section elaborates on the specific prompts employed to generate questions for each skill category. The prompts, utilized within the GPT-4 framework, are depicted in Figures 13, 15, 14, 16,17.These figures provide the exact phrasing and structure used for question generation, ensuring reproducibility and clarity in the benchmarking creation process.

MCQ matching prompt:

System prompt:

```
You are an intelligent chatbot designed to evaluate the correctness of generative
outputs for multiple-choice questions (MCQs).
Your task is to match the predicted answer with one of the provided options, which
include an 'I don't know' option. If there is no match between the predicted answer
and the options, choose the option that says, 'I don't know'. Here's how you can
accomplish the task:
------
## INSTRUCTIONS:
- Focus on finding a meaningful match between the predicted answer and the correct
option.
- Consider synonyms or paraphrases as valid matches.
- Choose an option only if you believe there is sufficient evidence to directly derive the
answer from the predicted information or indirectly with minimal reasoning. If there
isn't enough evidence to support any option, simply select the option with 'I don't
know.'
- Provide only the integer that represents the option number for your evaluation
decision.
- Evaluate as a human would, considering context and meaning, not just exact words.
- Provide your answer in the form of a Python dictionary string with the key 'decision',
such as {'decision': 3}.
```

User prompt:

```
Please evaluate the following question-answer pair:
Options: {options}
Predicted Answer: {pred}
Provide your evaluation as a decision with the matched option number.
Generate the response in the form of a Python dictionary string with the key 'decision'.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the
Python dictionary string.
For example, your response should look like this: {'decision': 1}.
Do not include any other information in your response such as ```python```.
```

Figure 6: Detailed prompt for MCQ evalaution

~~~~
Scoring evaluation prompt:
System prompt:
You are an intelligent chatbot designed to evaluate the correctness of generative outputs for question-answer pairs.
Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:
------
## INSTRUCTIONS:
- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.
- Provide a score between 0 and 5, where 5 indicates the highest meaningful match.
- Penalize the score if the predicted answer contains hallucinations or is missing key parts of the correct answer.
- Assign your score based on how far the predicted answer is from the correct answer.
- Evaluate as a human would, not as a machine.
- Provide your score in the form of a Python dictionary string with the key 'score', such as {'score': 3.7}.
User prompt:
Please evaluate the following video-based question-answer pair:
Question: {question}
Correct Answer: {answer}
Predicted Answer: {pred}
Provide your evaluation only as a score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match.
Generate the response in the form of a Python dictionary string with the key 'score'.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.
For example, your response should look like this: {'score': 4}.
Do not include any other information in your response such as ``` python ```.
~~~~

Figure 7: Detailed prompt for Scoring system evaluation

Temporal questions:

Q: Choose the correct option for the following question: Looking at these events : [Chandler reluctantly agrees to return to his old job after negotiation, Monica gets disappointed by the lost job opportunity, Phoebe gives Steve a painful massage as payback, Monica's audition dinner is ruined by Steve being stoned], how do they unfold in the episode?

Option 1: [Phoebe gives Steve a painful massage as payback ,Chandler reluctantly agrees to return to his old job after negotiation, Monica gets disappointed by the lost job opportunity ,Monica's audition dinner is ruined by Steve being stoned],
Option 2: [Monica's audition dinner is ruined by Steve being stoned ,Chandler reluctantly agrees to return to his old job after negotiation, Phoebe gives Steve a painful massage as payback ,Monica gets disappointed by the lost job opportunity],
Option 3: I don't know,
Option 4: [Monica gets disappointed by the lost job opportunity ,Chandler reluctantly agrees to return to his old job after negotiation, Phoebe gives Steve a painful massage as payback ,Monica's audition dinner is ruined by Steve being stoned],
Option 5: [Chandler reluctantly agrees to return to his old job after negotiation ,Monica's audition dinner is ruined by Steve being stoned, Monica gets disappointed by the lost job opportunity ,Phoebe gives Steve a painful massage as payback]

Figure 8: Example for the temporal order of events skill

Linking multiple events :

Q: What is the connection between Monica's failed dinner and Phoebe's reaction during Steve's next massage appointment?

Monica's dinner for Steve fails due to his stoned condition and disruptive behavior. Phoebe, out of frustration with Steve's behavior and the ruined dinner, takes out her anger on him during his next massage appointment by giving him a painful massage.

Figure 9: Example for the linking multiple events skill

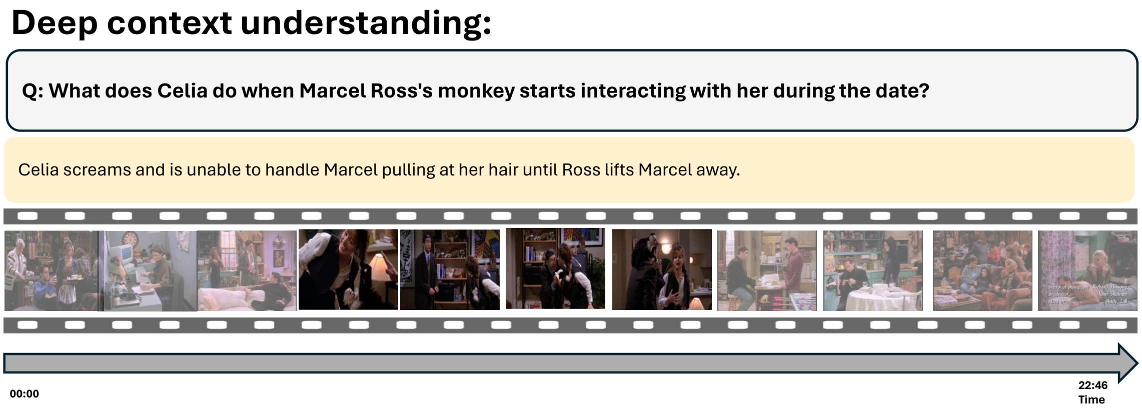

Figure 10: Example for the deep context understanding skill

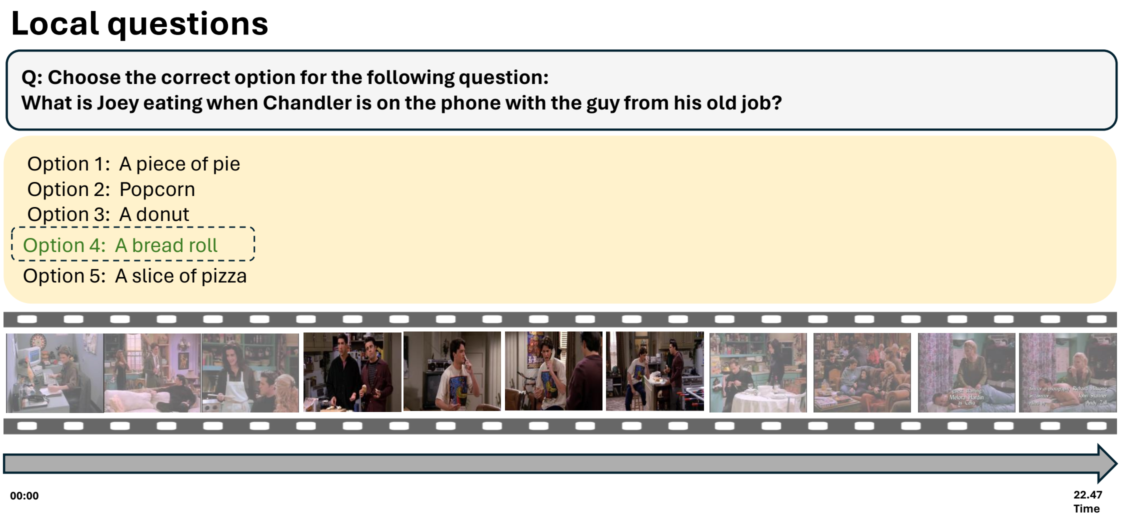

Figure 11: Example for the local questions

Summarization

Q: Please summarize the video with as much detail as possible.

Monica cooks a gourmet meal for Steve (Jon Lovitz), a restaurateur looking for a new head chef. Steve is a massage client for Phoebe, and she makes the introduction between Monica and him. The job is perfect as Steve wants something eclectic and needs someone who can create the entire menu. As an audition, Monica is cooking dinner for him the coming week. She wants Phoebe to be there. Monica hires a professional waitress Wendy (for $10/hr.), which offends Rachel (Monica says that she needed a professional waitress).Wendy bails on Monica at the last minute. Monica begs Rachel and even says that she gave her shelter when she had nowhere else to go.. Eventually she offers Rachel $20/hr. He arrives stoned and wants to eat everything in sight, including taco shells and gummy bears. Phoebe tells Rachel who tries to handle the situation by offering Steve some wine. Eventually Monica realizes that Steve is super stoned. She tries to yank the gummy bears from Steve, and they end up falling in the punch bowl.. Dinner is a total disaster, and the gang tells her that she doesn't want to work for a guy like that. After working as a data processor for five years, Chandler gets promoted to supervisor. Chandler quits, claiming he only intended for his job to be temporary (and Chandler already has been there for over 5 yrs.). Chandler goes to meet a career counselor. After 8 hrs. of aptitude, personality and intelligence tests he learns that he is fit for a career in data processing, for a large multinational corporation. he is disappointed as he always pictured himself doing something cool. When his boss calls and offers more money (& more bonus.. Chandler resists, but the boss keeps throwing more and more numbers), Chandler caves and goes back to work. Chandler gets the corner office, and he shows it off to Phoebe. He has a view and an assistant. But Chandler has more responsibility now and starts spending more time & late nights at work and yelling at his juniors. He doesn't like it. Ross has a date with a beautiful colleague named Celia (Melora Hardin) (curator of insects at the museum) and gives new meaning to the term 'spanking the monkey' when she meets Marcel. The date goes bad when Marcel hands on Celia's hair and pulls it. Eventually Ross takes Celia to bed, and she wants him to talk dirty and he says 'Vulva'. Ross turns to Joey for advice as Celia wants him to talk dirty as foreplay. Joey gets Ross to practice on him.. When Ross talks smack, Chandler overhears and amuses himself at their expense. Ross does well at the next date and talks very dirty (with theme, plot, motif and story-lines. at one point there were villagers), but eventually they get tired and cuddle. Phoebe takes out her anger at Steve at his next massage appointment by treating him to a bad massage (she elbows him on his back and pinches his skin so that it hurts).

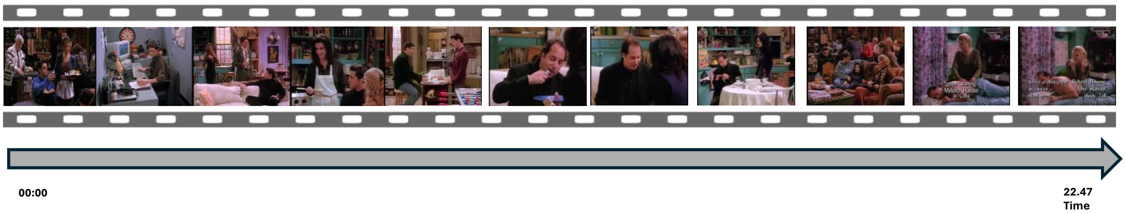

Figure 12: Example for the summarization skill

Linking multiple events:

System prompt :
You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.
Your task is to generate question-answer pairs specifically related to linking multiple events in the video content.
You will first play the role of a human who asks questions that link multiple events together in the video, and then play the role of an AI assistant that provides information based on the video content.
##TASK:
Users will provide information about the video, and you will generate a conversation-like question-and-answer pairs specifically focusing on linking multiple events together in the video to make the questions comprehensive across the video.
Generate TWENTY descriptive and conversational-style questions and their detailed answers based on the given information, specifically related to linking multiple events together in the video.
##INSTRUCTIONS:
- The questions must be conversational, as if a human is asking them, and should directly relate to linking multiple events together in the video.
- The answers must be detailed, descriptive, and should directly reference the information provided.
- The number of events to link together can vary from 2 to any number of events.

Please generate the response in the form of a list of Python dictionaries as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.
For example, your response should look like this: [{\Q\: \Your question here...\, \A\: \Your answer here...\},{\Q\: \Your question here...\, \A\: \Your answer here...\}].
Make sure to avoid to put double quotes inside string with double quotes, use single quotes instead. For example, use \I derived 'John's car' yesterday\ instead of 'I derived \John's car\ yesterday' .
please only output the required format, do not include any additional information.
Remember well the output format of ONLY a PYTHON LIST as output and DON'T output the python shell because I will use python ast library to parse your output list.
Few shot examples about the questions:
- What is the influence of event A on event B?
- How does event A lead to event B?
- What is the relationship between event A and event B?
- What is the impact of event A on event B?
- What is the connection between event A, event B, and event C?

User prompt:
The user input is {summary}.
Please generate the response in the form of a PYTHON LIST OF DICTIONARIES as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.
For example, your response should look like this: [{'Q': 'Your question here...', 'A': 'Your answer here...'},{'Q': 'Your question here...', 'A': 'Your answer here...'}].
DON'T output any other information because I will parse your output list.

Figure 13: Detailed prompt for Linking multiple events questions generation

Character actions:

System prompt :
You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.
Your task is to generate a question-answer pairs specifically related to each character actions through the whole video content.
Your task is to first play the role of a human who asks questions about each character actions through the whole video content. and then play the role of an AI assistant that provides information based on the video content.

##TASK:
Users will provide information about a video, and you will generate a conversation-like question and answers pair specifically focusing on each character actions through the whole video content.
Generate one question for each character that summarize all the actions did through the whole video content.

##INSTRUCTIONS:
- The questions must be like a human conversation and directly related to each character actions through the whole video content.
- The answer must be detailed and descriptive that summarize all actions for each character in the video and should directly reference the information provided.
- Focus on both the visual and textual actions but focus more on the vision actions as these questions are designed for video understanding.
##SAMPLE QUESTIONS:
- {'Q1': 'What did ross do through this video?', 'A': 'At the beginning of the episode he drank coffee in central park , then went to his apartment then ate some pizza.'}
- {'Q1': 'Summarize all actions that chandler did in this video.', 'A': 'At the beginning of the episode he read a magazine then went to his work by taxi , and finally he went to Monica's apartment to set with his friends.'}
User prompt:
This is the episode summary: {caption}. \n
This is the episode script: {script}. \n
Please generate the response in the form of list of Python dictionaries string with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text, respectively.
For the answer, please make it as a python list of actions in chronological order
For example, your response should look like this: [{'Q': 'Your question here...', 'A': ['Action 1','Action 2',...]},{'Q': 'Your question here', 'A': '['Action 1','Action 2',...]'}].
Please be very accurate and detailed in your response. Thank you!

Figure 14: Detailed prompt for sequence of character actions questions generation

Temporal order of events:

System prompt:
You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.
##TASK:
Users will provide an episode Screenplay Script. Your task is to extract the events from this Screenplay Script. Ensure that the events are listed in chronological order
First read the Screenplay Script and think carefully to extract the all events.

##Few shot samples
Episode Screenplay Script: {user Screenplay Script}
Extract the events from this episode Screenplay Script:
The response should be in the format: ['Event A', 'Event B', 'Event C', 'Event D',...], ensuring that the event B is after event A and before Event C.
Remember well the output format of ONLY a PYTHON LIST of events and DON'T output the python shell because I will use python ast library to parse your output list.
User prompt:
Episode Screenplay Script: {script}
Extract the events from the Screenplay Script in a list
please provide the response in the format of PYTHON LIST of DON'T output any other information because I will parse your output list.
DON'T output any ’ or ' in your response but use /u2019 for ’ and /u2019s for ’s and /u2019t for ’t and s/u2019 for s' or s’

Figure 15: Detailed prompt for Temporal order of events questions generation

Scene transitions:

System prompt:
##TASK:
Users will provide an episode Screenplay Script. Your task is to extract scene transitions in from this script.
First read the Screenplay Script and think carefully to extract the transitions.

##Few shot samples
Episode Screenplay Script: {user Screenplay Script}
Extract the scene transitions from this episode Screenplay Script:
please provide the response in the format of PYTHON LIST of scene transitions like this example : ['scene A name', 'scene B name', 'scene C name',...], ensuring that the scene changed from A to B then C and so on.
Remember well the output format of ONLY a PYTHON LIST of events and DON'T output the python shell because I will use python ast library to parse your output list.
Scene names should be places name or location names where the scene is taking place such as home , cafe , bar , car and so on.
User prompt:
Episode Screenplay Script: {script}
Extract the scene transitions from this Screenplay Script in a list
please provide the response in the format of PYTHON LIST of scene transitions like this example : ['scene A name', 'scene B name', 'scene C name',...], ensuring that the scene changed from A to B then C and so on.
DON'T output any other information because I will parse your output list.

Figure 16: Detailed prompt for scene transitions questions generation

Deep context understanding:

System prompt:
You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.
##TASK:
Your task is to first play the role of a human who asks questions related to deep context understanding in the video and then play the role of an AI assistant that provides information based on the video content.
Users will provide human video summary and the video script,and you will generate a conversation-like question and answers pair specifically focusing on measuring the viewer's context understanding.
##INSTRUCTIONS:
- The questions must be conversational, as if a human is asking them, and should directly relate to deep context understanding for the video content.
- The answers must be detailed, descriptive, and should directly reference the information provided.
- The number of questions should be up to 20 questions and answers.
- The questions should be tricky and hard to answer to measure the viewer's context understanding.
- The answers must be detailed, descriptive, and should directly reference the information provided.
- It will be good if most of the questions are related to the visual content of the video.
-Again, the questions should be very tricky and hard to answer to measure the viewer's context understanding.
Please generate the response in the form of a list of Python dictionaries as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.
For example, your response should look like this: [{'Q': 'Your question here...', 'A': 'Your answer here...'},{'Q': 'Your question here...', 'A': 'Your answer here...'}].
please only output the required format, do not include any additional information.
If you want to type 's or 't and so on, please use \u2019s for 's and \u2019t for 't and so on.
Test your output by using the python ast library to parse your output list.
Remember well the output format of ONLY a PYTHON LIST as output
User prompt:
video summary: {caption}.
video transcript: {script}.
Please generate up to 20 questions and their answers in the form of list of Python dictionaries string with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.
For example, your response should look like this: [{'Q': 'Your question here...', 'A': 'Your answer here...'},{'Q': 'Your question here...', 'A': 'Your answer here...'}].

Figure 17: Detailed prompt for deep context understanding questions generation