InfiniBench: A Benchmark for Large Multi-Modal Models in Long-Form Movies and TV Shows

Kirolos Ataallah[*1], Eslam Abdelrahman[*1], Mahmoud Ahmed[1], Chenhui Gou[2], Khushbu Pahwa[3], Jian Ding[1], Mohamed Elhoseiny[1]

[*]Equal contribution

[1]KAUST [2]Monash University [3]RICE University

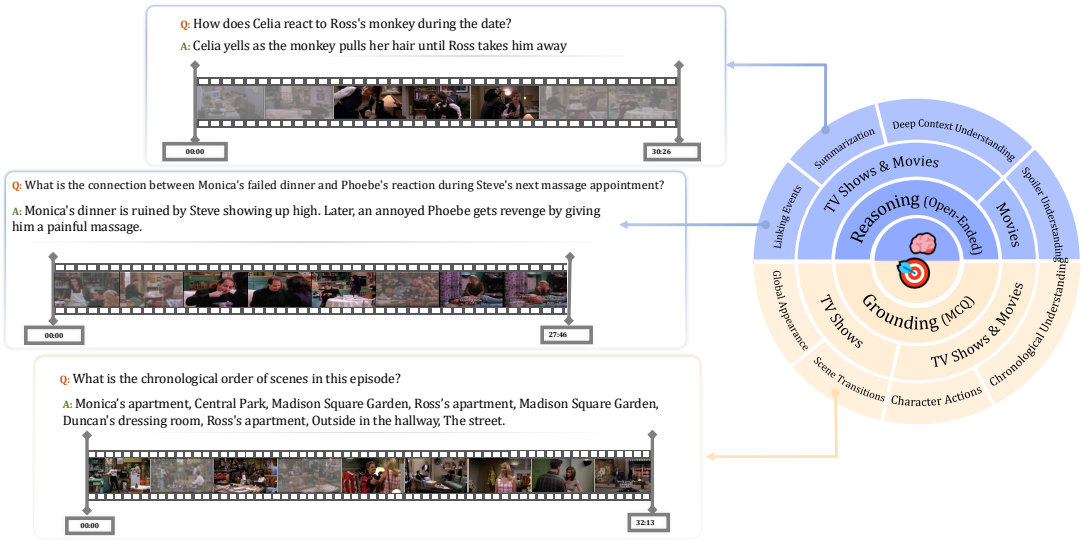


Figure 1: Overview of InfiniBench skills: covers eight core skills, grouped into grounding-based (MCQ) and reasoning-based (open-ended) categories.

Abstract

Understanding long-form videos, such as movies and TV episodes ranging from tens of minutes to two hours, remains a significant challenge for multi-modal models. Existing benchmarks often fail to test the full range of cognitive skills needed to process these temporally rich and narratively complex inputs. Therefore, we introduce InfiniBench, a comprehensive benchmark designed to evaluate the capabilities of models in long video understanding rigorously. InfiniBench offers: ***(1) Over 1,000 hours of video content***, with an average video length of 53 minutes. ***(2) The largest set of question-answer pairs*** for long video comprehension, totaling around 87.7K. ***(3) Eight diverse skills*** that span both grounding-based (e.g., scene transitions, character actions) and reasoning-based (e.g., deep context understanding, multi-event linking). ***(4) Rich annotation formats***, including both multiple-choice and open-ended questions. We conducted an in-depth evaluation across both commercial (GPT-4o, Gemini 2.0 Flash) and most recent open-source vision-language models such as Qwen2.5-VL, InternVL3.0). Results reveal that: (1) Models struggle across the board: Even the best model, GPT-4o, achieves only 47.1% on grounding-based skills, with most models performing near or just above random chance. (2) Strong reliance on world knowledge: Models achieve surprisingly high scores using only metadata (e.g., video titles), highlighting a tendency to rely on pre-trained knowledge rather than actual visual or temporal understanding. (3) Multi-Modal Importance: When provided with full video and subtitle context, however, models show substantial improvements, confirming the critical role of multimodal input in video understanding.

Our findings underscore the inherent challenges in long-video comprehension and point to the need for substantial advancements in both grounding and reasoning capabilities in MLLMs. InfiniBench is publicly available at InfiniBench.

Khushbu Pahwa contributed to this work during her internship at KAUST.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19497–19524
November 4-9, 2025

Category	Benchmarks	# Questions	# Videos	Avg Video Duration (mins)	# Hours	Questions Type		QA Source			Annotations	
						MCQ	Open	Video	Transcript	Summary	Auto	Human
Short	TGIF-QA (Jang et al., 2017)	165.2 K	**72.0 K**	0.05	60.00	✗	✓	✓	✗	✗	✓	✓
	MSRVTT-QA (Xu et al., 2017)	**243.6 K**	10.0 K	0.25	416.6	✗	✓	✓	✗	✗	✓	✗
	MV-Bench (Li et al., 2024b)	4.0 K	3.6 K	0.27	16.38	✓	✗	✓	✗	✗	✓	✗
	TVQA (Lei et al., 2019)	152.5 K	21.8K	1.27	461.20	✓	✗	✓	✗	✗	✗	✓
Long	Activity-QA (Yu et al., 2019)	58.0 K	5800	3.00	290.00	✗	✓	✓	✗	✗	✗	✓
	Egoschema (Mangalam et al., 2023)	5.0 K	5063	3.00	253.15	✓	✗	✓	✗	✗	✓	✓
	LongVideoBench (Wu et al., 2024)	6.7 K	3763	7.88	494.21	✓	✗	✓	✗	✗	✗	✓
	Moviechat (Song et al., 2023)	14.0 K	1000	9.40	156.67	✗	✓	✓	✗	✗	✗	✓
	MLVU (Zhou et al., 2024)	3.1 K	1730	15.50	446.92	✓	✓	✓	✗	✗	✓	✓
	MoVQA (Zhang et al., 2023b)	21.9 K	100	16.53	27.55	✓	✗	✓	✗	✗	✗	✓
	Video-MME (Fu et al., 2024)	2.7 K	900	16.97	254.55	✓	✗	✓	✗	✗	✗	✓
Very Long	LVBench (Wang et al., 2024b)	1.6 K	103	**68.35**	117.33	✓	✗	✓	✗	✗	✗	✓
	InfiniBench (Ours)	**87.7K**	1217	52.59	**1066.70**	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between InfiniBench and existing video QA benchmarks and datasets. InfiniBench contains the largest number of QA pairs for long and very long videos, as well as the longest total video duration.

1 Introduction

Recent progress in Multimodal large language models (MLMMs) has enabled impressive performance on image and short-video understanding by jointly processing visual and textual inputs (Ataallah et al., 2024a; Zhu et al., 2023; Zhang et al., 2023a; Chen et al., 2023; Lin et al., 2023; Liu et al., 2023; Maaz et al., 2023). To push toward richer video comprehension, several models have extended the context length of underlying LLMs, enabling them to process longer videos (Bai et al., 2023, 2025; Li et al., 2024a). However, current public benchmarks remain limited in both scope and realism. Most are restricted to short clips (a few seconds to minutes), and even larger-scale efforts like LVBench, with one-hour videos, lack narrative complexity and modality diversity (Wang et al., 2024b). In real-world scenarios, long-form videos especially movies and TV shows present fundamentally different challenges. These formats feature rich, structured storytelling with long-range temporal dependencies, evolving character arcs, and intricate causal relationships (Zhang et al., 2023b; Lei et al., 2019; Song et al., 2023) . Unlike short videos, they demand high-level skills such as intent inference, multi-event linking, and thematic summarization. Humans can easily integrate dispersed cues across modalities and time; current MLMMs cannot.

To bridge this gap, we introduce InfiniBench, a large-scale benchmark purpose-built to evaluate the long-video understanding capabilities of multimodal models. InfiniBench leverages over 1,000 hours of content drawn from movies and TV series with average video length (52.59 minutes) and the largest set of skill-targeted question–answer pairs (87.7K) among existing long and very long video benchmarks (see Table 1).

Crucially, InfiniBench is organized around eight core skills necessary for holistic long-video understanding. These are grouped into two categories: (1) Grounding-based skills, which assess a model's ability to retrieve and organize visual and temporal information, and (2) Reasoning-based skills, which require causal inference, contextual understanding, and abstract reasoning.

Key Insights from InfiniBench: Despite the scale and capabilities of modern MLMMs, all evaluated models struggle on InfiniBench. For instance, GPT-4o, which is the top-performing commercial model, achieves only 47.1% accuracy on grounding-based tasks, just modestly above the random baseline of 20%. This performance gap highlights how far even the best models are from mastering long-video comprehension. To dissect model behavior further, we conduct controlled experiments that isolate the role of world knowledge vs. visual input. Surprisingly, models still achieve non-trivial accuracy when given only metadata (e.g., show title and episode number), indicating heavy reliance on pre-trained knowledge. Yet, the most substantial gains up to +15% in accuracy come from providing visual input, confirming that true video understanding requires visual grounding, not just memorized associations. Finally, we observe consistent underperformance across both skill groups, grounding and reasoning-based skills. Reasoning-based tasks remain especially challenging, with even commercial models failing to produce coherent, causal, or context-aware answers. These findings underscore the unique difficulty and diagnostic power of InfiniBench, which surfaces fundamental limitations in today's multi-modal models and points toward critical future directions.

2 Related Work

In this section, we cover the existing video benchmarks and position our work among them. An

extended version of the related work can be found in Section C.

Recently, many multimodal large language models have aimed to extend their context lengths, such as Qwen2.5-VL (Bai et al., 2025), Qwen2-VL (Wang et al., 2024a), and LongVU (Wu et al., 2024), which can process videos over an hour long. This highlights the growing need for effective benchmarks. In contrast, short video benchmarks have been widely explored, covering tasks like moment retrieval (Liang et al., 2024; Lei et al., 2021), action classification (Caba Heilbron et al., 2015; Carreira and Zisserman, 2017; Soomro et al., 2012), video reasoning (Xie et al., 2024; Xiao et al., 2021), and video QA (Xu et al., 2017; Lei et al., 2019; Jang et al., 2017). Among longer video benchmarks, TVQA (Lei et al., 2019) is an early example with over 152.5 K QA pairs across 21.8 K clips (1.27 min avg.), totaling 461.2 hours. More recent datasets like MovieChat-1K(Song et al., 2023) offer 1,000 movie clips (9.4 min avg.) with 14K annotations. CRAFT(Ates et al., 2020) and GazeVQA (Ilaslan et al., 2023) extend to causal and embodied reasoning. Other long-form efforts include MLVU (Zhou et al., 2024), MoVQA (Zhang et al., 2023b), and Video-MME (Fu et al., 2024), though their average durations peak at 17 minutes. LVBench(Wang et al., 2024b) stands out with 1-hour videos, but includes only 103 clips totaling 117 hours. Domain-specific datasets like Ego4D QA on EgoSchema(Mangalam et al., 2023) focus on first-person views but use short (3-minute) clips. Motivated by these gaps, our benchmark evaluates grounding (temporal/visual anchoring) and reasoning (causality, narrative) skills to reflect human-like video understanding. As shown in Table 1, our dataset supports near-hour-long videos and offers the most significant total duration (∼1K hours) and the largest QA-pair count in terms of long video understanding (87.7K).

3 InfiniBench

3.1 Essential Skills for Long-Video Understanding

Understanding long videos requires more than event recognition, it demands tracking narratives, entities, and interactions over time. We categorize the core skills into two groups: 1- Grounding-Based Skills, focused on retrieving and organizing information directly from the video; and 2-Reasoning-Based Skills, which require inference, contextual understanding, and causal analysis. This division mirrors human comprehension: grounding what happens, then reasoning about why. InfiniBench leverages this structure to offer a comprehensive evaluation of long-video understanding.

3.1.1 Grounding-Based Skills

Grounding-based skills evaluate a model's ability to retrieve, order, and structure video content without requiring inference. They reflect the foundation of narrative comprehension, recognizing what happened, in what order, and to whom.

Chronological Understanding: Measures a model's ability to sequence events. This mirrors a core human faculty: understanding stories requires ordering moments in time. Without temporal grounding, narrative comprehension breaks down.

Character Actions Tracking: Assess whether the model can group and order actions performed by a specific character. Understanding "who did what, and when" is key to following character arcs and maintaining narrative coherence.

Scene Transitions: Focuses on recognizing and organizing scene-level shifts. Accurate modeling of transitions enables a higher-level understanding of narrative flow, pacing, and structural progression across long videos.

Global Appearance: Tests the ability of the model to track changes in character appearance throughout the video. This requires long-term visual consistency and awareness of visual storytelling cues, e.g., "What is the change in outfit for (character-name) in this video?".

3.1.2 Reasoning-Based Skills

These skills assess deeper narrative understanding, where models must infer relationships, motivations, and implications not explicitly stated. They reflect how humans interpret stories by connecting context, prior knowledge, and unstated cues.

Linking Multiple Events: Evaluates the ability to connect distant or seemingly unrelated events by inferring causality. Example: "How does Chandler's previous experience with his own parents' separation influence his decision to break up with Janice?" A: "Chandler's experience with his parents' separation strongly influences his decision. Remembering how he hated the man who came between his parents, Chandler decides to step aside for the sake of Janice's family's happiness, which leads him to break up with her."

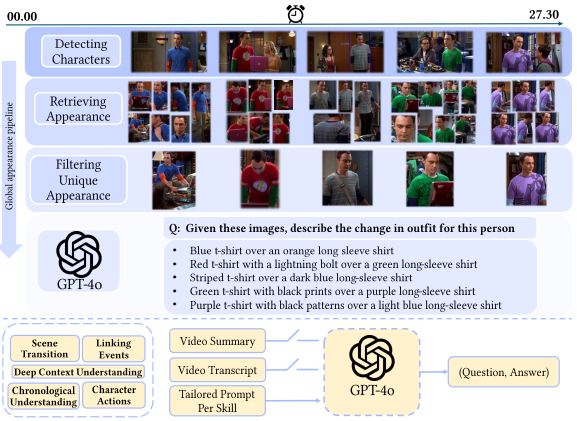


Figure 2: Full annotation pipeline for InfiniBench skill set. The upper section depicts the global appearance pipeline, while the lower section illustrates the question generation using GPT-4o. The gates for video summary and video transcript indicate that some skills utilize only the summary, others use only the transcript, and some use both.

Deep Context Understanding: Tests whether a model can interpret hidden motivations and social cues. Example: Q: How did Ted's co-workers' attitudes towards him change after he became their boss? A: Ted's co-workers clammed up whenever he walked into the room and became distant judging him on everything he did or said.

Spoiler Questions: Determines if a model can recognize when a question reveals a major plot twist or ending. Example: Q: Why did Roth betray Michael? A: It is implied that Roth wanted revenge for the death of Moe Greene, a friend of his who was assassinated on Michael's orders.

Summarization: Assesses the model's ability to distill narratives into summaries that capture events, emotional arcs, and thematic depth. Unlike short clips, movies and episodes demand multi-layered abstraction, not just factual recall.

For more visual samples for each skill, see C.3 in the supplementary.

3.2 Data Collection Pipeline

Building a rigorous long-video benchmark demands data that mirrors real-world storytelling, with multi-character dynamics, evolving plots, and implicit causal links. Our pipeline ensures diversity, depth, and relevance through three steps: 1- Video Selection (Sec. 3.2.1): Choose movies and TV episodes that exhibit complex narratives. 2- QA Generation (Sec. 3.2.2): Craft question–answer pairs aligned with each of the eight skills. 3- Refinement (Sec. 3.2.3): Remove redundancies, and the questions that can be easily answered without the visual signal, e.g., from the subtitles only or from the model knowledge.

3.2.1 Why Movies and TV-Shows

Unlike vlogs or surveillance footage, which are often repetitive and low in narrative complexity, movies and TV shows offer the rich, structured storytelling essential for long-video understanding. They feature: 1- Diverse, Long-Term Contexts: Multi-layered plots and evolving character arcs require memory and integration across time. 2- Non-Linear Storytelling: Flashbacks, subplots, and temporal shifts challenge models to reason over fragmented, distant events. These properties make them ideal for evaluating both grounding and reasoning skills in extended narratives.

3.2.2 Generating Question-Answer Pairs

We construct structured question–answer pairs tailored for long-video understanding, where models must reason over extended contexts, not just recognize isolated moments. Our pipeline, as shown in Figure 2, leverages the video visual-signal combined with transcripts and summaries.

Why Transcripts? Transcripts offer richer context than subtitles, including scene descriptions and character actions (e.g., "Monica and Chandler enter the café"). These cues allow for generating deeper, temporally grounded questions that reflect both visual and structural aspects of the story. For more details, see B.4 in the supplementary. For the TV shows, transcripts were sourced from reputable platforms such as (Forever Dreaming) and manually verified and aligned with the video. For movies, we utilize the transcripts provided by MovieNet (Huang et al., 2020), which have been pre-processed to ensure accurate alignment with both dialogue and subtitles, as detailed in Section A.5 (Huang et al., 2020).

Chronological Understanding & Linking Events: We employ GPT-4o to chronologically extract key events from the transcript. Then, we generate multiple-choice questions that assess the understanding of the full or partial chronological sequence. For linking events, we input GPT-4o the extracted events and asked the model to connect non-adjacent events by identifying causal relations and generate the Q/A pairs.

Scene Transitions: We prompted GPT-4o to identify all scene locations appearing in the transcript in chronological order. Based on this information, we generated multiple-choice questions designed

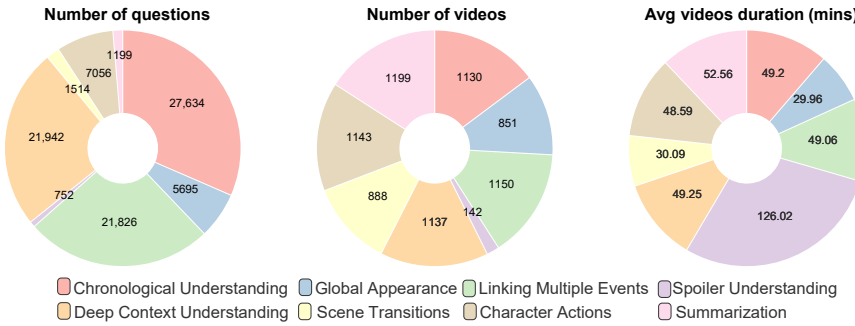


Figure 3: InfiniBench skills statistics. (A) Number of questions per skill, (B) Number of videos per skill, and (C) Average video duration per skill

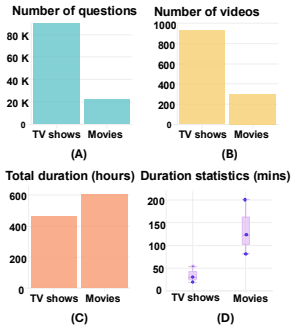


Figure 4: TV shows and Movies statistics.

to assess the correct sequential ordering of these locations in the video.

Character Tracking: To track character actions in the video, we input the transcript and the video summary into GPT-4o, which generates question-answer pairs focused on identifying the sequence of actions performed by the main characters. **Deep Context Understanding:** For the deep context understanding question-answer pairs, we employed GPT-4o with the transcript and the summary to generate questions about the hidden motivations and social cues. **Global Appearance:** Our pipeline, as shown in Figure 2, integrates person detection (Khanam and Hussain, 2024) and face-matching using InsightFace (Guo et al., 2021) to identify characters in video frames. Reference images of the main cast were collected from IMDb and matched with detected faces, using an empirically determined threshold to filter out unmatched individuals. Cropped character images were saved chronologically in timestamped folders. To identify unique outfits, we extracted visual features using DINO-v2 (Oquab et al., 2024) and removed redundant frames based on similarity scores, reducing approximately 200 frames to 20 while conservatively preserving visual diversity. The remaining images were then processed with GPT-4o (OpenAI, 2024) to generate outfit descriptions. Multiple-choice questions were constructed by varying the sequence of outfits. In cases where a character wore a single outfit throughout, distractor options featuring incorrect outfits were introduced to maintain question complexity. **Summarization & Spoiler Questions:** We curate expert-written IMDb (IMDB) summaries for human-level summarization. Spoiler questions are sourced from flagged plot discussions, focusing on twists or endings.

3.2.3 Data Integrity and Diversity

To ensure the reliability and quality of our benchmark, the collected dataset undergoes several refinement steps aimed at eliminating redundancy, mitigating data leakage, and reinforcing multimodal reasoning.

1) Redundancy Removal. To avoid question duplication, we compute pairwise textual similarity using BGE-M3 embeddings (Chen et al., 2024a) and remove highly similar questions (see Section B.5 for details). Additionally, since movies often contain over 100 events, many of which are redundant, we apply the same similarity-based filtering to retain only 20 diverse and representative events per movie.

2) Filtering Internal Model's Knowledge. Large video-language models are trained on massive datasets and may memorize content from widely available media, raising the risk of data leakage. To address this, we evaluate three powerful models—GPT-4o, Gemini 2, and Qwen2.5-VL—on our dev and test splits using only the movie/TV show metadata (e.g., title, genre, year) as input. If all three models correctly answer a question using this metadata alone, we consider the question likely answerable from internal knowledge and remove it. This step filters out approximately 16% of the questions, resulting in a subset we refer to as the *Knowledge-Filtered Set*.

3) Making Vision Matters. To ensure that our benchmark requires visual understanding—beyond textual reasoning, we further filter out questions answerable using subtitles alone. Using the same three models, we provide only the subtitle stream and explicitly instruct the models to rely solely on this modality. As in the previous step, questions correctly answered by all three models are excluded. This removes an additional 16%, lead-

ing to a total reduction of 32% across both filtering stages. The remaining set constitutes our final benchmark, on which all evaluations and ablations are performed unless otherwise noted.

4) **Human Verification.** To further ensure the validity, clarity, and correctness of the dataset, we conducted manual verification on the test set, covering 20% of the entire benchmark. This verification process revealed a 90.11% accuracy rate for valid question–answer pairs across different skills, demonstrating high overall reliability. Importantly, any identified errors during this process were filtered out, ensuring that the final test subset reflects a clean and trustworthy evaluation set. Additional details can be found in Section B.6. The custom verification interface used in this process is shown in Figure 6.

3.3 Data Statistics

InfiniBench ultimately comprises 87.7K questions over videos averaging 53 minutes, with some running as long as 201 minutes, making it the largest and most challenging benchmark for long-video understanding. InfiniBench includes 923 episodes from six TV shows (Lei et al., 2019) based on the long-form version in GoldFish (Ataallah et al., 2024b) and filtered 296 movies that have subtitles from MovieNet (Huang et al., 2020). These movies offer a broad and challenging foundation for benchmarking long-video understanding. Figures 3 and 4 illustrate the per-skill distribution, source breakdown (TVQA (Lei et al., 2019), MovieNet (Huang et al., 2020)), and duration statistics across the dataset. We humanly verified 20% of the data, ensuring fair coverage across TV shows and movies. The remaining 80 % is available for training video models. The verified portion is further divided into test and validation splits, 10% each. All reported results and ablation studies are conducted using the test split.

4 Experiments

4.1 Models

To assess the capabilities and limitations exposed by InfiniBench, we evaluate 13 state-of-the-art long-video MultiModal Large Language Models (MLLMs) spanning both open-source and commercial ecosystems.

Open-Source Models We benchmark eleven leading open-source MLLMs, each evaluated using their official implementations and following their official best practices and configuration, e.g., recommending sampling strategy and frame-rate. These models vary in architecture, context length, and vision-language alignment strategies, offering a diverse perspective on the current research landscape: Qwen2.5-VL (Bai et al., 2025), Qwen2-VL (Wang et al., 2024a), InternVL3 (Zhu et al., 2025b), InternVL2.5 (Chen et al., 2024b), InternVL2(InternVL2), LLaVA-OneVision (Li et al., 2024a), LongVU (Shen et al., 2024),VideoChat-Flash (Li et al., 2025), InternLM-XComposer-2.5 (Zhang et al., 2024), Goldfish (Ataallah et al., 2024b), and MiniGPT4-video(Ataallah et al., 2024a)

Commercial Models We also evaluate two cutting-edge proprietary models via their official APIs: GPT-4o (OpenAI, 2024) and Gemini 2.0 Flash (Gemini, 2024). These models represent the frontier of closed-access MLLMs with advanced capabilities in video, text, and multi-turn interaction. Their inclusion offers a valuable benchmark ceiling for current open-source efforts.

We followed the standard best practices for all the models based on their official implementation. More details can be seen in Section B.7.

4.2 Evaluation Setup

We adopt distinct evaluation protocols for grounding-based and reasoning-based skills, reflecting their differing task formats: multiple-choice (MCQ) and open-ended, respectively.

Multiple-Choice (MCQ). For grounding-based skills, model responses are evaluated using standard accuracy. Models are prompted to output the correct option number, and accuracy is computed based on an exact match with the correct answer.

Open-Ended. Reasoning-based skills are assessed using a GPT-4o mini, assigning a score from 0 to 10. This score reflects the overall quality of the response, based on five criteria: 1- Factual correctness. 2- Relevance to the question. 3- Proximity to the expected answer. 4- Hallucination avoidance. 5- Completeness. The scoring prompt is provided in Figure 8.

4.3 Findings and Insights

We analyze the results across 13 models to uncover key patterns in long-video understanding. Our findings highlight substantial performance gaps, modality dependencies, and emerging limitations in current MLLMs, even at the frontier.

Models	Video	Subtitles	Meta data	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score
				Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events		
Random Per.	–	–	–	20.0	20.0	20.0	20.0	N/A	N/A	N/A	N/A	20.0	N/A
w/o Filtering	✓	✓	✗	51.8	42.9	43.5	77.2	6.3	6.8	6.7	7.5	53.9	6.8
	✗	✓	✗	29.1	39.6	46.8	82.0	7.2	7.0	5.9	7.5	49.4	6.9
	✗	✗	✓	25.7	26.1	39.1	56.5	3.4	5.2	3.9	7.4	36.8	5.0
w Filtering	✓	✓	✗	49.7	37.7	39.9	61.0	6.3	6.4	6.6	6.8	47.1	6.5
	✗	✓	✗	26.5	34.2	42.2	68.5	7.1	6.5	5.8	6.8	42.8	6.5
	✗	✗	✓	22.2	19.9	31.9	35.4	3.3	4.8	3.5	6.6	27.4	4.6

Table 2: Analysis of the impact of filtering our benchmark by excluding the models' internal knowledge and emphasizing the visual questions. The reported numbers are on GPT4o. Table 6 in the Appendix, demonstrates the full results across more models. Random Per. is random performance.

Models	Frame Rate	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score	Overall
		Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events			
Random Per.	N/A	20.0	20.0	20.0	20.0	N/A	N/A	N/A	N/A	20.0	N/A	N/A
GPT-4o	450 FPV	**49.7**	37.7	39.9	**61.0**	**6.3**	**6.4**	**6.6**	**6.8**	**47.1**	**6.5**	**56.0**
Gemini-2.0	1 FPS	45.1	**39.3**	**50.0**	50.1	5.7	6.0	4.3	5.4	46.1	5.4	49.9
InternVL3	128 FPV	**34.3**	27.7	20.5	31.1	**3.8**	3.7	3.3	5.3	28.4	4.0	**34.4**
Qwen2.5VL	768 FPV	30.0	25.3	22.7	20.3	3.3	4.3	3.4	5.4	24.6	4.1	32.8
Qwen2VL	768 FPV	23.5	28.2	30.2	27.4	2.2	4.3	**3.5**	5.0	27.3	3.8	32.5
Goldfish	60 FPW	16.2	22.9	21.3	25.4	3.0	**4.9**	3.4	**5.6**	21.5	**4.2**	31.9
VideoChat-Flash	1000 FPV	20.5	**29.5**	**34.9**	**37.4**	2.6	3.4	2.2	4.2	**30.6**	3.1	30.9
InternVL2	128 FPW	26.6	24.9	21.5	26.6	2.9	3.5	3.0	5.0	24.9	3.6	30.4
LLava-OneVision	128 FPV	21.0	24.7	23.3	30.3	2.0	3.7	2.7	5.1	24.8	3.4	29.4
InternVL2.5	128 FPV	27.1	25.1	21.2	29.2	2.4	2.8	2.1	4.2	25.6	2.9	27.4
InternLM-XComposer	16 FPW	20.1	27.9	26.6	26.4	1.6	2.6	2.2	4.0	25.2	2.6	25.7
LongVU	512 FPV	27.67	21.0	27.9	19.0	1.7	2.9	2.7	3.6	23.9	2.7	25.6
MiniGPT4-video	60 FPV	18.1	23.1	25.9	23.1	2.0	2.9	2.0	3.3	22.5	2.6	24.1

Table 3: InfiniBench leaderboard across eight skills. Models are ranked by their equally weighted performance for Grounding and reasoning skills using this formula $\left(0.5 \cdot \frac{\text{acc}}{100} + 0.5 \cdot \frac{\text{score}}{10}\right)$. FPV (Frames Per Video), FPS (Frames Per Second), and FPW (Frames Per Window) are reported. All models use both video and subtitles as inputs. Random Per. is random performance.

4.3.1 World Knowledge & Data Leakage

To isolate the influence of pre-trained knowledge from visual understanding, we conducted a controlled experiment where models were prompted using only video metadata—such as movie titles, release years, or TV episode identifiers—without access to any visual or subtitle input. As shown in Tables 2 and 6, several models perform surprisingly well under this setting. For example, GPT-4o achieves 36.8% and 5.0 in grounding and reasoning scores, respectively, indicating a substantial reliance on memorized knowledge from pretraining. To mitigate this issue, we apply a filtering strategy described in Section 3.2.3. Post-filtering, model performance under metadata-only input drops significantly—e.g., GPT-4o's accuracy decreases from 36.8% to 27.4%, approaching the random baseline of 20%. This validates the effectiveness of our filtering and suggests that the remaining questions require genuine understanding of the visual and textual input. Moreover, when full video and subtitle context is provided, model performance increases substantially (e.g., GPT-4o improves from 36.8% to 53.9%), reinforcing the role of multimodal input in successful reasoning. This finding raises an important question: ***"Do current models inherently favor memorized knowledge over true multi-modal understanding?"*** Our results suggest that overcoming this bias is key to advancing long-video reasoning.

4.3.2 Models Are Struggling

Despite recent progress in vision-language modeling, none of the evaluated models—commercial or open-source—achieve strong performance on grounding-based skills. As shown in Table 3, the best-performing model, GPT-4o, reaches only 47.1% accuracy, approximately 2.5 times the random baseline of 20%. This substantial gap highlights a key limitation: even state-of-the-art models struggle to maintain accurate visual grounding over long temporal contexts. Reasoning-based performance is similarly low. GPT-4o achieves the highest reasoning score among commercial models with an average of only 6.5, followed by Gemini-2 at 5.4, and the best open-source model, InternVL3, at just 4.0. These results underscore the difficulty of our benchmark and point to the need for further advancements in long video understanding and multi-step reasoning capabilities.

4.3.3 Modality Benefits vs. Context Limitations

To further analyze the contribution of each modality, we compare model performance across three input settings: (1) metadata only, (2) subtitles only, and (3) video combined with subtitles. As shown in Tables 2 and 6, subtitles alone offer only marginal improvements over metadata for certain visually grounded skills such as *Global Appearance*, increasing accuracy from 22.2% to 29.1%. However, for multi-modal skills that require both visual and textual reasoning, e.g., *Chronological Understanding*, the performance improves substantially, from 35.4% to 68.5%. Surprisingly, adding video can be inconsistent: it significantly boosts performance on skills like *Global Appearance* and *Spoiler Questions*, but offers no benefit, or even degrades performance, for tasks such as *Linking Events* and *Character Actions*. We hypothesize that these inconsistencies stem from the tension between visual diversity and context length. Many skills require long-range contextual reasoning, which subtitles provide in a continuous form. In contrast, video inputs, sampled uniformly across the full episode, introduce fragmented and scattered visual information that can overwhelm the limited context window of current models. This fragmentation may distract the model, leading to worse performance than when using subtitles alone. This highlights a fundamental trade-off: while some skills benefit from multi-modal inputs, others rely more heavily on extended, coherent textual context, which can be disrupted by the inclusion of sampled visual frames. Overall, the results underscore a key limitation in existing models: constrained context windows hinder their ability to jointly reason over extended textual narratives and dispersed visual evidence.

4.3.4 Grounding & Reasoning: Both Are Challenging

Across the eight skill types, both grounding and reasoning tasks remain far from solved. While commercial models outperform open-source ones by a notable margin, all models struggle across both skill categories, with no single model dominating in either domain. This reflects the intrinsic difficulty of the benchmark and underscores the need for significant model innovation, particularly in handling extended temporal structure, entity tracking, entity linking, and causal reasoning. Our results suggest that long-video understanding is far from solved and demands focused future efforts.

4.3.5 Impact of Different Frame Rates

To assess the influence of input frame rates, we conducted an ablation study using four open-source models (QwenVL2.5, QwenVL2.0, InternVL3.0, and Video-Flash). As shown in Table 4, increasing the number of input frames generally leads to better accuracy, although the degree of improvement varies between models. For example, both QwenVL2.5 and QwenVL2.0 show substantial improvements with higher frame rates. On the MCQ set, QwenVL2.5 performance increases from 23.6% to 27.4%, while QwenVL2.0 improves from 24.6% to 28.7% as the number of frames increases from 16 to 128. In contrast, Video-Flash and InternVL3.0 exhibit minimal improvement even when the number of frames is increased by a factor of eight in InternVL3.0 and by a factor of 62 in the case of VideoChat-Flash. The performance of QwenVL2.5 and QwenVL2.0 begins to degrade when entering the 768 frames, indicating that these models struggle when the context window is fully saturated. A likely explanation is that, these models are constrained to a certain number of input frames, limiting their capacity to exploit longer temporal context. In particular, reasoning-based skills benefit the most from higher frame rates, as the additional visual context reduces the likelihood of missing critical shots, thus improving model performance. In summary, while higher frame rates tend to improve accuracy, the degree of benefit is closely related to the architectural capacity of the model and the training strategy.

4.4 Evaluation Robustness

4.4.1 Self-enhancement Bias

The main results in Table 3 use GPT-4o as the judge model, which raises the potential concern of self-enhancement bias (Zheng et al., 2023), since GPT-4o serves as both the task model and evaluator. However, we argue that such bias is unlikely in our setup due to the clear separation between generation and evaluation. Specifically, during inference, the model accesses both video and subtitles to generate answers, whereas the judge model evaluates responses using only the generated text and reference answers. This separation in goals, modalities, and inputs minimizes the risk of any model-specific advantage transferring across tasks.

To further assess evaluation objectivity, we conducted a multi-judge study using two strong and di-

Models	Frame Rate	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score
		Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events		
Baseline Random	N/A	20.0	20.0	20.0	20.0	N/A	N/A	N/A	N/A	20.0	N/A
QwenVL2.5	768	30.0	25.3	22.7	20.3	3.3	4.3	3.4	5.4	24.6	4.1
	400	31.3	27.1	24.0	23.3	3.5	4.3	3.3	5.4	26.4	4.1
	350	30.9	27.8	24.1	23.9	3.4	4.2	3.7	5.6	26.6	**4.2**
	256	33.3	29.8	24.6	23.8	3.3	4.1	3.4	5.4	**27.9**	4.1
	128	32.0	29.7	22.1	25.8	2.9	3.7	3.3	5.2	27.4	3.7
	16	28.3	25.1	21.2	19.8	2.2	3.2	2.9	4.8	23.6	3.3
QwenVL2.0	768	23.5	28.2	30.2	27.4	2.2	4.3	**3.5**	5.0	27.3	3.8
	256	27.5	30.9	30.2	29.9	2.2	4.0	3.1	5.1	**29.6**	**3.6**
	128	27.9	29.0	29.2	28.9	2.2	3.7	3.4	4.8	28.7	3.5
	16	26.4	22.6	23.9	25.4	1.8	3.0	3.1	4.6	24.6	3.1
InterVL3	128	34.3	27.8	20.5	31.1	3.8	3.7	3.3	5.3	**28.4**	**4.0**
	16	35.7	26.8	18.8	28.2	3.0	3.2	3.7	5.0	27.4	3.7
VideoChat-Flash	1000	20.5	**29.5**	**34.9**	**37.4**	2.6	3.4	2.2	4.2	**30.6**	3.1
	128	18.0	30.0	33.5	36.8	2.4	2.8	2.2	4.0	**29.6**	**2.9**
	16	18.4	32.9	29.5	34.1	2.1	2.5	1.9	3.8	28.7	2.6

Table 4: Impact of varying frame rates on model performance.

verse alternatives: Qwen2.5-VL-7B-Instruct (open-source) and Gemini 2.0-flash (commercial). Pairwise Pearson and Spearman correlations revealed strong agreement: 1) GPT-4o-mini & Gemini 2.0-flash: Pearson 0.97, Spearman 0.94 2) GPT-4o-mini & Qwen2.5-VL: Pearson 0.95, Spearman 0.92 We also report a Krippendorff's alpha of 0.8014 across all judges, indicating high inter-rater reliability. These results demonstrate that GPT-4o is well-aligned with independent judges, supporting the robustness and impartiality of our evaluation.

Model / Judge	GPT-4o	Gemini 2.0	Qwen 2.5
GPT-4o	6.77	7.67	6.76
Gemini-2.0	5.61	6.76	5.78
Goldfish (Mistral)	4.64	4.53	5.06
Qwen2.5VL	4.56	4.58	5.36
InternVL3	4.46	4.89	5.21
Qwen2VL	4.11	4.00	4.50
InternVL2	3.90	4.01	4.50
LLava-Onevision	3.89	3.84	4.38
Video-Flash	3.46	3.84	3.73
InternVL2.5	3.26	3.29	3.97
MiniGPT4-video (Mistral)	3.22	2.65	3.16
LongVU	3.00	3.24	3.41
InternLM-XComposer	2.99	3.06	3.28

Table 5: Ablation study demonstrates our LLM-based judge robustness.

4.4.2 Reliability of the reasoning skills scoring

To assess the alignment between our LLM-based judge and human evaluators, we conducted a detailed comparison on 10% of the test set using standard statistical measures. Specifically, we collected ratings from three independent human annotators per sample and computed the average human score. We then measured the Pearson correlation between the LLM scores and the averaged human ratings, obtaining a score of 0.92, indicating a high degree of linear agreement. Additionally, we report Spearman correlation of 0.89, confirming that the LLM's ranking of outputs closely matches human preferences.

To contextualize this alignment, we computed inter-annotator agreement using Krippendorff's alpha, which yielded a value of 0.86, showing that the LLM aligns with human consensus nearly as well as humans agree among themselves. We also measured per-annotator Pearson correlations with the LLM, which ranged from 0.88 to 0.93, further supporting the strong consistency between the LLM and individual human raters. These results demonstrate that our LLM judge exhibits near-human-level alignment, validating its use as a reliable proxy for human evaluation in our setting.

5 Conclusion

Our evaluation of current Multimodal Large Language Models (MLLMs) on the InfiniBench benchmark reveals key limitations in long-video understanding. Top models like GPT-4o perform modestly on grounding tasks and struggle with reasoning-based skills that require causal inference and context-aware understanding. Models show significant reliance on pre-trained knowledge, performing better with metadata alone, but improve substantially when provided full video and subtitle context, emphasizing the importance of multimodal inputs. We also observe that visual input diversity often conflicts with the limited context windows of current models, affecting performance on certain tasks. Overall, InfiniBench highlights critical gaps in existing models and directs future research toward improving visual grounding and multi-step reasoning for long-video comprehension.

6 Limitations

Despite its strengths, our question–answer generation pipeline currently depends on the availability of transcripts, limiting its use to movies and TV shows. To broaden applicability, we will develop a more general pipeline that can handle arbitrary videos without preexisting transcripts. We also plan to leverage our dense, skill-based annotations to create multi-video benchmarks, such as season-level evaluations, that extend each skill across related clips. Furthermore, our training split can support retrieval-based MLLM methods that focus on relevant frames (inspired by Video-XL (Liu et al., 2025)), and we aim to explore long-video reasoning approaches, such as those based on the GRPO algorithm (Shao et al., 2024), using samples from our dataset.

References

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

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. 2024b. Goldfish: Vision-language understanding of arbitrarily long videos. *Preprint*, arXiv:2407.12679.

Tayfun Ates, M Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. 2020. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Forever Dreaming. Tv show transcripts. `http://transcripts.foreverdreaming.org`. Accessed: 2025-05-20.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Google Gemini. 2024. Gemini technical report.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*.

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.

Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. 2023. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10462–10479.

IMDB. Imdb website.

InternVL2. Internvl2.

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.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *Preprint*, arXiv:2410.17725.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *Preprint*, arXiv:2107.09609.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

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. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

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

Renjie Liang, Li Li, Chongzhi Zhang, Jing Wang, Xizhou Zhu, and Aixin Sun. 2024. Tvr-ranking: A dataset for ranked video moment retrieval with imprecise queries. *Preprint*, arXiv:2407.06597.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. 2024. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6577–6602, Miami, Florida, USA. Association for Computational Linguistics.

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*.

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

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*.

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*.

Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu, and Yong Zhu. 2022. Dynamic multistep reasoning based on video scene graph for video question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3894–3904.

OpenAI. 2024. Gpt-4o technical report.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge

Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *Preprint*, arXiv:2410.17434.

Carina Silberer and Manfred Pinkal. 2018. Grounding semantic roles in images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2626.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and 1 others. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Preprint*, arXiv:1212.0402.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and 1 others. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. *Preprint*, arXiv:2306.14899.

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. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, and 8 others. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *Preprint*, arXiv:2407.03320.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *Preprint*, arXiv:2406.04264.

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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Extended Experimental Results

A.1 Results on the Validation-set

We are organizing a challenge based on our benchmark; therefore, we will not release the test set. Accordingly, the follow-up works that will leverage our benchmark can use the results reported on the validation set reported in Table 7. As shown in 7, the same insights and conclusions are seen, similar to those reported on the test-set in Table 3.

A.2 Results on the Full Test-set Before Filtration

As discussed in Section 3.2.3, we have applied two types of filtration to make our benchmark more reliable and robust: 1) Filtering the internal model's knowledge by removing the questions that can be answered from the metadata solely. 2) Making vision matters by removing the questions that can be answered from the subtitles only. The main results reported in Tables 3 and 7 are reported on the filtered version of our benchmark.

To demonstrate the importance of the filtration process, we demonstrate the results on the test-set before filtration in Table 8. As shown in Table 8 and in comparison to Table 3, all the models achieve a higher performance before filtration, due to the knowledge leakage, and the intensive dependency on the input text modality, subtitles. So by filtering these two factors, the performance of all the models drops significantly. We believe the filtered version is more reliable and truly assesses the model's capabilities.

A.3 Impact of the "I Don't Know" Option

As shown in Table 9, removing the "I don't know" option leads to an increase in model accuracy. Without this option, models are forced to select an answer, even when uncertain, which can occasionally result in correct guesses due to chance. In contrast, the inclusion of the "I don't know" option requires the model to explicitly acknowledge uncertainty when it lacks sufficient evidence to answer correctly, thereby increasing task difficulty and realism. Additionally, excluding the "I don't know" option reduces the number of choices per question, which increases the expected performance of random guessing. This factor should be considered when comparing results across configurations with and without this option.

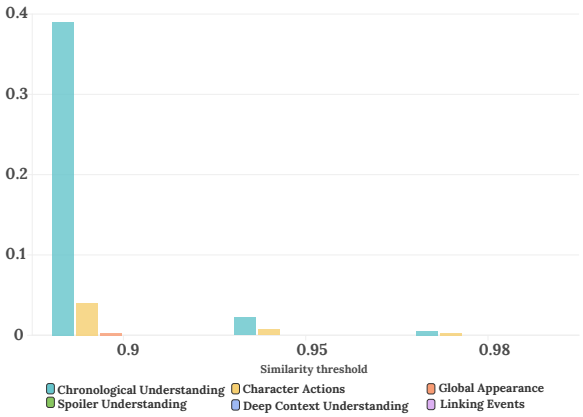


Figure 5: Duplication analysis for different thresholds using cosine similarity of text vector embeddings.

B InfiniBench Further Details

B.1 More Clarification about the generated questions

Leveraging the grounding information obtained during the verification process, we generated additional event-based questions. Specifically, we utilized the chronological events in conjunction with other grounding skills such as global appearance, scene transitions, and character actions to enrich the verified set. For example, in scene transitions skill, we can formulate systematic questions such as "What are the scene transitions that occur before the event "Ross watches TV with Rachel"?" This is feasible because we can extract the relevant transitions occurring before the event timestamp from the complete set of scene transitions using temporal annotations. The same approach is applied with global appearance and character actions; we formulated questions that refer to moments immediately preceding or following specific events in the video. These questions are only added to the verified set.

B.2 Data Diversity

Our benchmark is not narrowly scoped around human characters or interactions. Instead, it is built to holistically assess a wide range of long-video understanding capabilities:

1. Diversity in Evaluated Skills Our benchmark explicitly targets a broad spectrum of cognitive and perceptual skills, not just character-centric ones. While certain tasks—such as Global Appearance and Character Actions—do emphasize characters, a substantial portion of the benchmark is designed around generic, high-level understanding of visual narratives, events, and environ-

	Models	Video	Subtitles	Meta data	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score
					Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events		
	Random Per.	–	–	–	20.0	20.0	20.0	20.0	N/A	N/A	N/A	N/A	20.0	N/A
GPT4o	w/o Filtering	✓	✓	✗	51.8	42.9	43.5	77.2	6.3	6.8	6.7	7.5	53.9	6.8
		✗	✓	✗	29.1	39.6	46.8	82.0	7.2	7.0	5.9	7.5	49.4	6.9
		✗	✗	✓	25.7	26.1	39.1	56.5	3.4	5.2	3.9	7.4	36.8	5.0
	w Filtering	✓	✓	✗	49.7	37.7	39.9	61.0	6.3	6.4	6.6	6.8	47.1	6.5
		✗	✓	✗	26.5	34.2	42.2	68.5	7.1	6.5	5.8	6.8	42.8	6.5
		✗	✗	✓	22.2	19.9	31.9	35.4	3.3	4.8	3.5	6.6	27.4	4.6
Qwen2.5VL	w/o Filtering	✓	✓	✗	32.8	30.1	30.0	41.4	3.4	4.8	3.7	6.5	33.6	4.6
		✗	✓	✗	22.7	36.2	31.0	56.9	4.8	5.2	3.6	6.6	36.7	5.1
		✗	✗	✓	19.0	26.3	28.9	43.5	2.2	3.7	3.5	6.2	29.4	3.9
	w Filtering	✓	✓	✗	30.0	25.3	22.7	20.4	3.3	4.3	3.4	5.4	24.6	4.1
		✗	✓	✗	18.4	24.2	21.5	23.9	4.2	4.1	3.7	5.4	21.9	4.3
		✗	✗	✓	15.7	19.6	19.8	17.2	2.2	3.4	3.1	4.9	18.1	3.4
InternVL 3.0	w/o Filtering	✓	✓	✗	35.9	27.8	24.8	41.2	3.9	4.1	3.6	6.2	32.4	4.4
		✗	✓	✗	30.8	34.3	24.8	46.5	4.2	4.2	3.8	6.4	34.0	4.7
		✗	✗	✓	25.0	25.6	26.5	38.9	1.9	3.6	3.9	6.1	29.0	3.9
	w Filtering	✓	✓	✗	34.3	27.8	20.5	31.1	3.8	3.7	3.3	5.2	28.4	4.0
		✗	✓	✗	29.2	31.2	20.8	33.9	4.2	3.9	3.5	5.4	28.8	4.2
		✗	✗	✓	23.4	23.3	23.3	24.9	2.0	3.3	3.6	5.0	23.7	3.5
LlaVa-OneVision	w/o Filtering	✓	✓	✗	23.2	27.6	27.0	41.1	2.0	4.0	3.2	6.1	29.7	3.9
		✗	✓	✗	20.8	27.8	29.8	43.4	4.0	5.1	3.7	6.6	30.4	4.8
		✗	✗	✓	16.7	27.8	28.5	39.3	0.8	3.7	3.4	5.8	28.1	3.4
	w Filtering	✓	✓	✗	21.1	24.7	23.3	30.3	2.0	3.7	2.7	5.1	24.8	3.4
		✗	✓	✗	18.5	24.5	25.7	28.9	3.9	4.5	3.3	5.6	24.4	4.3
		✗	✗	✓	14.8	24.5	26.2	28.4	0.8	3.4	2.9	4.8	23.5	3.0

Table 6: Analysis of the impact of filtering our benchmark by excluding the models' internal knowledge and emphasizing the visual questions. Random Per. is random performance.

Models	Frame Rate	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score	Overall
		Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events			
Random Per.	N/A	19.48	16.67	16.52	40.56	N/A	N/A	N/A	N/A	23.31	N/A	N/A
GPT-4o	450 FPV	54.26	50.0	43.96	49.66	6.2	6.49	5.61	7.00	49.47	6.33	56.36
Gemini-2.0	1 FPS	50.39	50.79	56.61	36.60	5.70	5.40	3.41	5.94	48.60	5.11	49.86
InternVL3	128 FPV	33.33	28.57	23.46	36.29	3.91	3.79	3.41	5.28	30.41	4.10	35.69
Qwen2VL	768 FPV	31.01	23.81	32.59	48.41	2.26	4.23	2.93	5.14	33.96	3.64	35.18
Qwen2.5VL	768 FPV	37.98	22.22	22.53	27.17	3.22	4.13	3.67	5.34	27.48	4.09	34.19
LLava-OneVision	128 FPV	37.21	25.40	20.11	43.91	1.99	3.66	3.28	5.21	31.66	3.54	33.50
Goldfish	60 FPW	20.93	12.70	21.97	38.48	2.88	4.85	3.61	5.48	23.52	4.21	32.79
InternVL2	128 FPV	24.03	15.87	22.16	42.16	2.74	3.37	2.87	5.02	26.06	3.50	30.53
VideoChat-Flash	1000 FPV	28.68	20.63	33.33	29.73	2.45	3.24	2.35	4.20	28.09	3.06	29.35
InternVL2.5	128 FPV	34.88	20.63	21.6	32.79	2.60	2.78	2.35	4.20	27.48	2.98	28.65
InternLM-XComposer	16 FPW	32.56	14.29	28.49	30.11	1.72	2.50	2.11	4.13	26.36	2.62	26.26
MiniGPT4-video	60 FPV	17.83	11.11	27.00	40.1	2.04	2.87	2.15	3.39	24.01	2.61	25.07
LongVU	512 FPV	20.93	15.87	21.42	29.86	1.70	2.84	3.07	3.53	22.02	2.79	24.94

Table 7: InfiniBench leaderboard on the validation-set, across eight skills. Models are ranked by their equally weighted performance for Grounding and reasoning skills using this formula $\left(0.5 \cdot \frac{\text{acc}}{100} + 0.5 \cdot \frac{\text{score}}{10}\right)$. FPV (Frames Per Video), FPS (Frames Per Second), and FPW (Frames Per Window) are reported. All models use both video and subtitles as inputs. Random Per. is random performance.

Models	Frame Rate	Grounding Skills				Reasoning Skills				Avg. Acc.	Avg. Score	Overall
		Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	Summarization	Deep Context Understanding	Spoiler Understanding	Linking Events			
Random Per.	N/A	20.00	20.00	20.00	20.00	N/A	N/A	N/A	N/A	20.00	N/A	N/A
GPT-4o	450 FPV	51.83	42.91	43.49	77.15	6.32	6.82	6.73	7.47	53.85	6.84	61.10
Gemini-2.0	1 FPS	46.73	44.59	55.94	67.62	5.80	6.56	4.36	6.19	53.72	5.73	55.50
Qwen2.5VL	768 FPV	32.78	30.10	30.04	41.36	3.35	4.84	3.71	6.46	33.57	4.59	39.74
InternVL3	128 FPV	35.93	27.76	24.75	41.17	3.86	4.06	3.57	6.19	32.40	4.42	38.30
Qwen2VL	768 FPV	25.37	31.37	34.91	36.68	2.23	4.87	3.59	6.03	32.08	4.18	36.94
Goldfish	60 FPW	17.28	24.45	24.03	29.32	3.02	5.45	3.61	6.56	23.77	4.66	35.19
VideoChat-Flash	1000 FPV	21.48	31.37	37.48	48.14	2.64	3.93	2.38	5.10	34.62	3.51	34.87
LLava-OneVision	128 FPV	23.15	27.57	27.04	41.07	2.04	4.04	3.18	6.12	29.71	3.85	34.08
InternVL2	128 FPV	27.72	26.18	23.89	30.47	2.89	3.74	3.16	5.96	27.07	3.94	33.22
InternVL2.5	128 FPV	28.58	27.10	25.18	34.00	2.46	3.10	2.30	5.10	28.72	3.24	30.56
InternLM-XComposer	16 FPW	22.53	30.33	30.33	34.48	1.63	2.86	2.43	5.02	29.42	2.99	29.63
LongVU	512 FPV	27.04	22.38	25.89	21.49	1.70	3.26	3.00	4.19	24.20	3.04	27.29
MiniGPT4-video	60 FPV	18.52	25.84	29.18	25.79	2.09	3.09	2.21	3.89	24.83	2.82	26.52

Table 8: InfiniBench before filtration, Section 3.2.3, leaderboard across eight skills. Models are ranked by their equally weighted performance for Grounding and reasoning skills using this formula $\left(0.5 \cdot \frac{\text{acc}}{100} + 0.5 \cdot \frac{\text{score}}{10}\right)$. FPV (Frames Per Video), FPS (Frames Per Second), and FPW (Frames Per Window) are reported. All models use both video and subtitles as inputs. Random Per. is random performance.

Models	# Frames	I don't know option	Grounding Skills				Avg. Acc.
			Global Appearance	Scene Transitions	Character Actions	Chronological Understanding	
Qwen2.5-VL	**768**	✓	30.0	25.3	22.7	20.3	24.6
		✗	31.4	23.3	23.9	24.3	**25.7**
InternVL3.0	**128**	✓	34.3	27.8	20.5	31.1	28.4
		✗	35.5	30.9	22.9	33.5	**30.7**
Qwen2-VL	**768**	✓	23.5	28.2	30.2	27.4	27.3
		✗	25.2	31.8	29.9	32.3	**29.8**
Video-flash	**1000**	✓	20.5	29.6	34.9	37.4	**30.6**
		✗	20.8	27.4	31.9	36.8	29.2
InternLM-XC	**16 FPW**	✓	20.1	27.9	26.6	26.4	25.3
		✗	22.6	30.1	25.5	26.6	**26.2**
InternVL2.5	**128**	✓	27.1	25.1	21.2	29.2	25.6
		✗	27.7	27.7	22.2	28.9	**26.7**
LongVU	**512**	✓	27.7	21.0	28.0	19.0	23.9
		✗	27.7	23.5	29.5	23.5	**26.0**

Table 9: Analysis of the impact of incorporating the "I don't know" option on model performance.

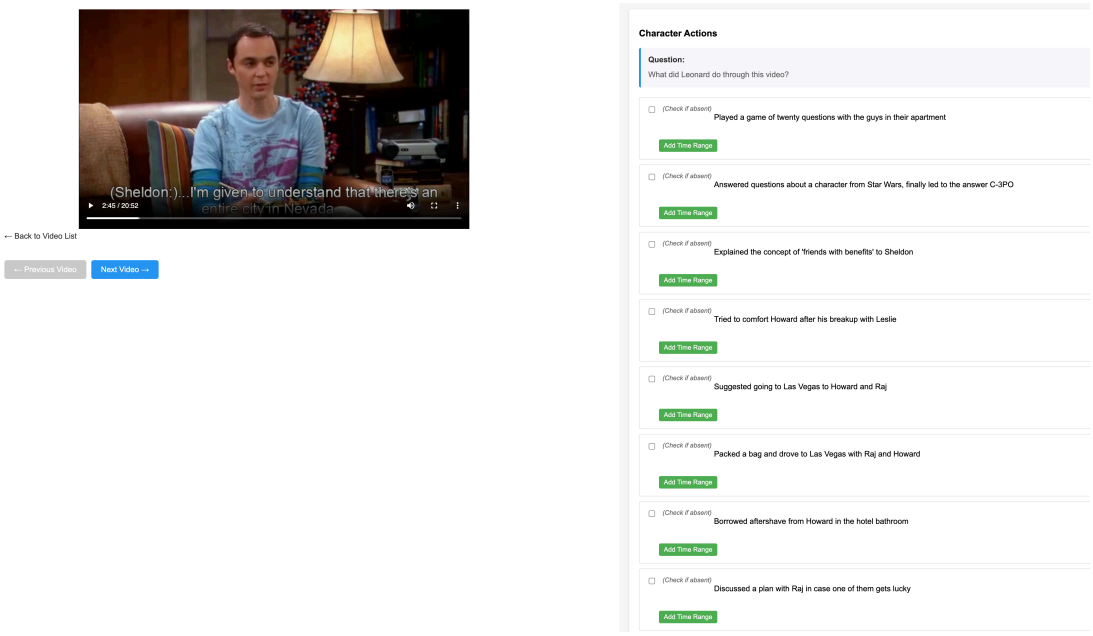


Figure 6: Data verification Website.

ments. These include: Chronological Understanding, Deep-Context Understanding, Spoiler Questions, Linking Multiple Events, Scene Transitions. These skills collectively test a model's capability in grounding, temporal reasoning, and narrative comprehension far beyond surface-level character recognition.

2. Diversity and Richness of Video Sources Our dataset comprises a large, balanced mix of TV shows and over 296 movies spanning multiple decades, cultures, genres, and visual settings. This includes diverse environments such as nature, abstract spaces, and sci-fi/fantasy worlds, demonstrating that the benchmark is not narrowly human-centric. For example : Nature & Survival: Movies like "Into the Wild" and "Life of Pi" focus on wilderness and animal interaction. Abstract & Visual Reasoning: Movies such as "2001: A Space Odyssey", "Inception", and "Donnie Darko" challenge models with surreal, symbolic visuals. Space & Environmental Focus: Movies such as "Gravity" and "The Thing" take place in isolated, non-social settings. Sci-Fi & Fantasy Worlds: Movies such as "Avatar" and "The Matrix Reloaded" emphasize non-human environments and abstract reasoning. Minimal Dialogue / Object-Centric: "Cube", "Moon", and "Minority Report" require visual comprehension beyond character interaction.

Additionally, we argue that movies and TV shows are inherently more challenging than other long-form video domains, such as vlogs or surveillance footage. This is due to: 1) Multi-layered narratives require long-term memory and reasoning. 2) Non-linear structures involving flashbacks and fragmented timelines. 3) Visually rich environments that go beyond typical human-centric interactions.

These factors make them ideal testbeds for evaluating true long-video understanding.

B.3 Data Copyrights

Our benchmark provides only annotations and uses two widely adopted, publicly available video datasets that provide legally compliant, downsampled video content (e.g., 3 FPS, scene shots). Specifically, we use: (1) six full-season TV shows from the TVQA (Lei et al., 2019) dataset, based on the long-form version in GoldFish (Ataallah et al., 2024b), and (2) movies from the MovieNet (Huang et al., 2020) dataset. For the TV shows, transcripts were sourced from reputable platforms such as foreverdreaming and manually verified by the authors for alignment with the video. This usage is consistent with fair use for academic research. For movies, we use transcripts provided by MovieNet (Huang et al., 2020), which are already filtered and accurately aligned with dialogue and subtitles.

B.4 Transcripts vs. Subtitles

Figure 7 illustrates the distinction between transcripts and subtitles.

Transcripts are comprehensive documents authored by screenwriters, containing not only spoken dialogue but also detailed scene descriptions, setting and location information, character actions, and camera directions (e.g., angles, cuts, and shot composition). Effectively, transcripts serve as blueprints for both the visual and narrative structure of a movie or TV show. In our benchmark, transcripts are used exclusively during the data construction phase to generate rich and diverse questions. They are *not* provided to the model during inference or evaluation.

Subtitles, by contrast, are limited to the spoken dialogue extracted from the video's audio track. These are optionally provided to the model at inference time, serving as an auxiliary textual modality alongside the video frames. If available, subtitles can enrich the input context and help improve performance, particularly for dialogue-centric questions.

Transcript of episode 1 season 1 of Friends TV shows:

[Scene: Central Perk, Chandler, Joey, Phoebe, and Monica are there.]
Monica: There's nothing to tell! He's just some guy I work with!
...
(Ross gestures his consent.)
Joey: Strip joint! C'mon, you're single! Have some hormones!
(Rachel enters in a wet wedding dress and starts to search the room.)

Subtitle of the same episode
1
00:00:54,012 --> 00:00:57,641
There's nothing to tell.
It's just some guy I work with.

2
00:00:57,892 --> 00:00:59,962
There's gotta be something wrong with him.

Figure 7: Difference between transcript and subtitles.

B.5 Duplications Filtering

To ensure that our dataset is free of duplicate or near-duplicate questions, we implemented a semantic similarity filtering pipeline.

We first embedded each question and its answer choices using M3-Embedding (Chen et al., 2024a), then computed pairwise cosine similarity to detect potential duplicates. We experimented with similarity thresholds of 90%, 95%, and 98%. As shown in Figure 5, the vast majority of questions are unique across all skill categories. Only two skills—*Temporal Questions* and *Character Actions*—produced potential duplicates. However, upon manual inspection, these were found to be false positives. For example, the following two questions were flagged as duplicates due to high embedding similarity, despite a critical semantic difference:

Q1: Did the event flashback to Phoebe completing a mile on a hippity-hop before turning thirty happen **before** the event Monica makes breakfast with chocolate-chip pancakes?
Q2: Did the event flashback to Phoebe completing a mile on a hippity-hop before turning thirty happen **after** the event Monica makes breakfast with chocolate-chip pancakes?

Although these questions differ by a single word, that word fully reverses the temporal meaning, highlighting the importance of semantic context when evaluating similarity. Based on this analysis, we conclude that the dataset contains no true duplicates. The filtering pipeline successfully identifies and flags near-duplicates, and the observed false positives illustrate the subtlety and complexity of semantic variation in question phrasing.

B.6 Human Verification

To ensure the reliability of our benchmark, we conducted a human evaluation on 20% of the dataset. Annotators were tasked with verifying the correctness of the question–answer (Q/A) pairs across each skill category. Figure 6 shows the custom-built verification interface developed to streamline the annotation process.

During verification, annotators assessed each video using the following criteria:

- **Global Appearance and Character Actions:** Annotators were provided with the character's name and a list of appearance or action events extracted by our pipeline. Each event was marked as *absent* if not observed in the video; if present, the annotator recorded its corresponding timestamp.
- **Linking Events and Deep Context Understanding:** Annotators confirmed the validity of each Q/A pair, ensuring the answer could only be derived from the visual or textual content of the video. Temporal grounding was also recorded when applicable.
- **Scene Transitions and Chronological Understanding:** Annotators reviewed the sequence of collected events for each video, verifying their presence and providing timestamps for valid instances.
- **Spoiler Questions:** Annotators ensured that each Q/A pair relied solely on information present in the video, excluding any external or prior knowledge.

The collected timestamps were cross-referenced with the dataset to validate both the extracted events and their chronological consistency. This also enabled the generation of *temporal certificates* for each Q/A pair, following the methodology introduced in EgoSchema (Mangalam et al., 2023). A temporal certificate represents the number of video segments required to answer a given question. Following (Mangalam et al., 2023), video clips shorter than 0.1 seconds were discarded, and consecutive segments within a 5-second window were merged into a single clip. The total duration of these merged segments defines the certificate length.

Our human-verified test set contains an average of 3.23 temporal certificates per Q/A pair, with an average certificate length of 202.97 seconds, underscoring the benchmark's emphasis on comprehensive long-video understanding. The presence of timestamps for each answer further enhances the traceability and trustworthiness of the verified set.

Skill Name	Number of Questions	Per Question Acc.(%)
Character Actions	1370	89.76
Global Appearance	750	93.21
Scene Transitions	224	85.23
Chronological understandings	224	90.37
Deep Context Understanding	5574	94.14
Linking Events	4587	95.92
Spoiler Questions	151	82.11

Table 10: Human verification for the keypoints used in the generation pipeline. The number of questions corresponds to the ones shown to the annotators, not the final count in our dataset.

Accuracy statistics across skill categories are reported in Table 10, with an overall verification accuracy of 90.11%. Incorrect or ambiguous Q/A pairs were removed, and only validated examples were retained for the final test set.

Note that *Summarization* questions do not require human verification, as they are sourced from the official IMDb website (IMDB). According to IMDb's contribution guidelines, all plot summaries must adhere to strict formatting and factual standards, and each submission is manually reviewed and approved by the IMDb team. This ensures the summaries are reliable, accurate, and effectively pre-verified by human annotators.

B.7 Models Configuration

GPT-4o (OpenAI, 2024): GPT-4o does not support direct ingestion of `.mp4` video files. To accommodate this, we sample up to 450 frames per video and provide the model with the corresponding subtitle text and question.

Gemini-Flash 2.0 (Gemini, 2024): Developed by Google, Gemini-Flash 2.0 supports native video input. For TV shows, we supply the model with the full video and the associated question. For movies—which lack audio—we additionally provide the subtitle file alongside the video and question.

Qwen2.5-VL (Bai et al., 2025) and Qwen2-VL (Wang et al., 2024a): These models accept up to 768 input frames, fitting within the memory limits of an NVIDIA A100 GPU (80 GB) under the authors' recommended inference settings. Subtitles are concatenated with the input question to provide multimodal context.

Video-Chat-Flash (Li et al., 2025): Although the model supports up to 10,000 frames, we limit input to 1,000 frames due to memory constraints on the A100 GPU (80 GB). Other configurations follow the default setup. Subtitles are concatenated with

the input question to enrich contextual understanding.

InternVL3.0 (Zhu et al., 2025a), InternVL2.5 (Chen et al., 2024b), and InternVL2 (InternVL2): For the InternVL family, we evaluate models using both 16 and 128 frames, the latter being the maximum supported without hallucination on an A100 GPU. Although these models can technically process more frames, exceeding 128 often introduces unstable outputs. Subtitles corresponding to the selected frames are appended to the input question.

LLaVA-OneVision (Li et al., 2024a): This model supports both image and video inputs. We evaluate performance at 16 and 128 frames—found to be the upper bound for stable performance on an A100 GPU. Subtitles for selected frames are appended to the question.

LongVU (Shen et al., 2024): LongVU is evaluated using 512 frames, staying within the memory and context window limits of the A100 GPU. Subtitles aligned with the selected frames are appended to the input question.

InternLM-XComposer-2.5 (Zhang et al., 2024): We use the default configuration with a 16-frame window and increase the number of clips per video to 120. Subtitles aligned with these clips are appended to the input question. All evaluations are conducted on an A100 GPU.

Goldfish and MiniGPT-4-Video (Ataallah et al., 2024b): We use the Mistral backbone with default parameters, retrieving $k = 3$ video clips per query. Each 80-second video yields 60 frames (approx. 1.3 FPS). Although the model supports up to 90 frames, we limit it to 60 to reserve space for the question, subtitles, and output tokens—reducing hallucinations due to context overflow. Subtitles are interleaved with the frames, following the MiniGPT-4-Video configuration. For MiniGPT-4-Video, we similarly sample 60 frames per video, interleave them with subtitles, and prompt the model with the question for generation.

B.8 Benchmark Examples

Here in this section, we are showing more examples of our benchmark skills, such as the Chronological Understanding in Fig. 9, linking multiple events in Figure.13, and deep context understanding in Figure. 11, spoiler questions in Figure.15, Character Actions Figure.16,Scene transitions in Figure. 12, Global Appearance in Figure 10, and summarization in Figure. 14.

B.9 Success and Failure Cases

In this section, we present examples of both success and failure cases in question generation using GPT-4o. Figure 18 illustrates cases involving the generation of chronological understanding questions, while Figure 17 showcases examples related to Linking Multiple Events questions. As highlighted in the human evaluation section B.6, such failure cases are infrequent, with 90.11% of the generated data verified as accurate.

B.10 Exact Prompts

This section elaborates on the specific prompts employed to generate questions for each skill category. The prompts, utilized within the GPT-4o framework, are depicted in Figures 19, 21, 20, 22,23.These figures provide the exact phrasing and structure used for question generation, ensuring reproducibility and clarity in the benchmarking creation process.

C Extended Related Work

C.1 Semantic Roles in Vision and Language.

Early work like *Grounding Semantic Roles in Images* (Silberer and Pinkal, 2018) and *Grounded Situation Recognition* (Pratt et al., 2020) showed how semantic roles can be aligned with visual regions in images. This concept was extended to videos by *Video Object Grounding* (Sadhu et al., 2020) and *Visual Semantic Role Labeling* (Sadhu et al., 2021), emphasizing semantic alignment for event understanding. Building on these ideas, HERO (Li et al., 2020) proposed a hierarchical encoder that fuses visual and textual modalities, setting the stage for video-language pretraining. Recent models like VSGR (Mao et al., 2022) further reinforce this trajectory by explicitly building video scene graphs to reason over semantic elements and their relations. In this context, our work introduces a benchmark that evaluates models' ability to track and reason about evolving semantic roles across long-form, multimodal narratives.

C.2 Long Video Models.

Recent advancements in commercial AI models have introduced native support for long-form video understanding. For instance, Google's Gemini Flash family (Gemini, 2024) that provides a one million-token context window for coherent multi-hour understanding, and GPT-4o (OpenAI, 2024) that processes video frames in conjunction with

subtitles within its 128K-token context window. In contrast to the commercial solutions, Open-source efforts rely on compression, memory, and retrieval techniques. Qwen2.5-VL (Bai et al., 2025) extends dynamic resolution temporally to spot key events in hours-long streams and ground them visually. Qwen2-VL (Wang et al., 2024a) uses Multimodal Rotary Position Embedding (M-RoPE) to encode up to 768 frames, preserving spatial and temporal structure. LongVU (Shen et al., 2024) applies cross-modal queries and inter-frame dependencies for adaptive compression, trimming redundancy while keeping salient content. InternLM-XComposer-2.5-OmniLive (Zhang et al., 2024) pairs a short-term memory buffer with on-the-fly compression into long-term storage for later retrieval. Goldfish (Ataallah et al., 2024b) segments videos into clips and retrieves the top-k most relevant, and LLaMA-VID (Li et al., 2023) represents each frame with only two tokens for maximal efficiency. VideoINSTA (Liao et al., 2024) enables long video understanding without pretraining by combining event-based temporal reasoning with LLMs and self-reflective spatial-temporal queries, achieving strong results on EgoSchema and NExT-QA. Despite these innovations, truly unrestricted long-form video understanding, with deep temporal dependencies and narrative complexity, remains an open challenge.

C.3 Video benchmarks.

Recently, many multimodal large language models have aimed to extend their context lengths, such as Qwen2.5-VL (Bai et al., 2025), Qwen2-VL (Wang et al., 2024a), and LongVU (Wu et al., 2024), which can process videos over an hour long. This highlights the growing need for effective benchmarks. In contrast, short video benchmarks have been widely explored, covering tasks like moment retrieval (Liang et al., 2024; Lei et al., 2021), action classification (Caba Heilbron et al., 2015; Carreira and Zisserman, 2017; Soomro et al., 2012), video reasoning (Xie et al., 2024; Xiao et al., 2021), and video QA (Xu et al., 2017; Lei et al., 2019; Jang et al., 2017). Among longer video benchmarks, TVQA (Lei et al., 2019) is an early example with over 152.5 K QA pairs across 21.8 K clips (1.27 min avg.), totaling 461.2 hours. More recent datasets like MovieChat-1K(Song et al., 2023) offer 1,000 movie clips (9.4 min avg.) with 14K annotations. CRAFT(Ates et al., 2020) and GazeVQA (Ilaslan et al., 2023) extend to causal and embodied reasoning. Other long-form efforts include MLVU (Zhou et al., 2024), MoVQA (Zhang et al., 2023b), and Video-MME (Fu et al., 2024), though their average durations peak at 17 minutes. LVBench(Wang et al., 2024b) stands out with 1-hour videos, but includes only 103 clips totaling 117 hours. Domain-specific datasets like Ego4D QA on EgoSchema(Mangalam et al., 2023) focus on first-person views but use short (3-minute) clips. Motivated by these gaps, our benchmark evaluates grounding (temporal/visual anchoring) and reasoning (causality, narrative) skills to reflect human-like video understanding. As shown in Table 1, our dataset supports near-hour-long videos and offers the most significant total duration ($\sim$1K hours) and the largest QA-pair count in terms of long video understanding (87.7K).

Scoring evaluation prompt (GPT4o-mini):

System prompt:

You are an intelligent and fair evaluator AI that specializes in assessing the correctness and semantic alignment
between ground truth answers and predicted responses for question-answering tasks, including those based on video content.
Your role is to evaluate how well a predicted answer matches the correct (reference) answer based on the following detailed criteria:
EVALUATION INSTRUCTIONS:
- Focus on **semantic similarity**, **factual correctness**, and **completeness**.
- Accept paraphrases, synonyms, or rephrasings **as valid**, as long as they preserve the original meaning.
- **Do not penalize** for stylistic differences or changes in tone, unless they impact factual accuracy.
- **Penalize** if:
- The predicted answer omits **key factual elements** present in the correct answer.
- The prediction includes **hallucinated content** or unfounded details.
- The prediction **contradicts** the correct answer.
- Use human-like judgment: apply reasoning beyond surface text similarity.
- When uncertain, provide a **conservative but fair** score.
- Use a scoring scale from **0 (completely incorrect)** to **10 (perfect match)**.
OUTPUT FORMAT:
Return a JSON object with **two fields**:
- "score": an integer from 0 to 10
- "justification": a concise explanation (1-3 sentences) of your reasoning
Example Output:

```
{
  "score": 7,
  "justification": "The predicted answer captures the main idea, but it omits some key details about the setting described in the correct answer."'
}
```

Be fair, consistent, and concise. Follow the format exactly.

User prompt:

Please evaluate the following video-based question-answer pair:
Question: {question}
Correct Answer: {answer}
Predicted Answer: {pred}
Please return your evaluation in the specified JSON format with both a score and a justification.

Figure 8: Detailed prompt for Scoring system evaluation.

Chronological Understanding

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 9: Example for the chronological understandings skill.

Global Appearance

Q: Choose the correct option for the following question:
Can you track the sequence of Penny's outfit changes in this episode?

Option 1: gray tank top over a pink sports bra , mustard yellow vest over a white blouse , red and pink floral dress with a blue tank top underneath , red floral dress over a red shirt , blue tank top

Option 2: red floral dress over a red shirt ,gray tank top over a pink sports bra , red and pink floral dress with a blue tank top underneath mustard yellow vest over a white blouse , blue tank top

Option 3: black yoga pants and purple sports bra , white sundress ,black formal suit , black running shorts and a white tank top.

Option 4: I don't know.

Option 5: blue tank top , red and pink floral dress with a blue tank top underneath , red floral dress over a red shirt , gray tank top over a pink sports bra , mustard yellow vest over a white blouse

Figure 10: Example for the Global appearance skill.

Deep context understanding:

Q: What does Celia do when Marcel Ross's monkey starts interacting with her during the date?

Celia screams and is unable to handle Marcel pulling at her hair until Ross lifts Marcel away.

Figure 11: Example for the deep context understanding skill.

Scenes Transition

Q :Choose the correct option for the following question:
What is the chronological order of scenes in this episode?

Option 1 : Monica and Rachel's apartment , Madison Square Garden, Duncan's dressing room ,The street , Ross's apartment, Central Perk , Madison Square Garden , Ross's apartment, Outside in the hallway
Option 2 : Ross's apartment , The street , Madison Square Garden, Duncan's dressing room , Monica and Rachel's apartment , Madison Square Garden , Central Perk , Ross's apartment, Outside in the hallway
Option 3 : I don't know,
Option 4 : Ross's apartment, Outside in the hallway , Madison Square Garden, Duncan's dressing room , The street , Central Perk , Madison Square Garden , Ross's apartment , Monica and Rachel's apartment
Option 5 : Monica and Rachel's apartment , Central Perk , Madison Square Garden , Ross's apartment , Madison Square Garden, Duncan's dressing room , Ross's apartment, Outside in the hallway , The street

Figure 12: Example for the scenes transitions skills.

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 13: Example for the linking multiple events skill.

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 14: Example for the summarization skill.

Spoiler questions

Q: Why didn't the Arquillian in the jeweler's head simply tell Jay that the galaxy was on his cat's collar?

To add a bit of mystery to the story. If he'd said 'the galaxy in the jewel on the cat's collar', the movie would have ended much faster. Actually, Arquillian was indeed trying to tell Jay that the galaxy was on the cats collar. He just didn't have the correct vocabulary to do so. Note how he stumbles over the word \"war\". He almost certainly thinks \"belt\" is the correct word for \"collar\", which is understandable because the articles of clothing are identical, as the only differences are that one is worn around the waist and the other is worn around the neck. And the cat's name is Orion, so he's being accurately descriptive, not deceitful.It's likely that the Arquillian didn't understand much English and that the Jeweler's body had a translator in it when conversing with humans. It was likely damaged when Edgar stabbed it through the neck.

Figure 15: Example for the spoiler questions in Movies.

Character Actions :

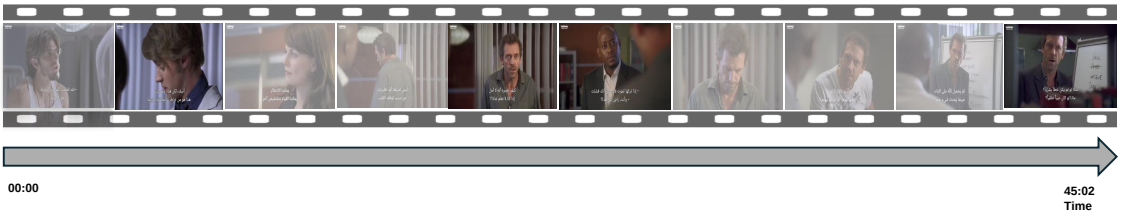
Q: Choose the correct option for the following question: What did Rachel do through this video?

Option 1: enjoying a cappuccino with a dash of cinnamon at a trendy coffee shop , discussing the latest book club read with a friend over lunch , savoring a slice of gourmet pizza with sun-dried tomatoes and arugula at a pizzeria , exploring an art museum's new exhibit on modern sculpture.
Option 2: I don't know
Option 3: Rachel attends an initial interview, accidentally kisses Mr. Zelner. , She calls back, gets another meeting with Mr. Zelner and eventually gets the job after apologizing and explaining her actions. , She practices handshaking with Phoebe and Monica., Accidentally touches Mr. Zelner's crotch while offering a handshake., Rachel gets a second interview call and gets ink on her lips during the second attempt., Rachel enters Central Perk and announces her job interview at Ralph Lauren., Rachel goes home, realizes her mistake regarding the ink., Misinterprets Mr. Zelner's gesture and walks out thinking he's making an advance.

Option 4: Rachel enters Central Perk and announces her job interview at Ralph Lauren., She practices handshaking with Phoebe and Monica. , Rachel attends an initial interview, accidentally kisses Mr. Zelner. , Rachel gets a second interview call and gets ink on her lips during the second attempt. , Misinterprets Mr. Zelner's gesture and walks out thinking he's making an advance. , Rachel goes home, realizes her mistake regarding the ink. , She calls back, gets another meeting with Mr. Zelner and eventually gets the job after apologizing and explaining her actions. , Accidentally touches Mr. Zelner's crotch while offering a handshake.

Figure 16: Example for character actions questions.

Linking multiple events



How does Dr. House's internal conflict towards the end connect to the events of the episode?

Dr. House's internal conflict at the end ties together the various events of the episode. The stress of the almost Sci-fi case, the emotional impact of Foreman's departure, and the unresolved medical mystery all contribute to House's turmoil, leaving him with an immense conflict that sets the stage for the next season.

In what ways do the sea rescue and the medical mystery serve as catalysts for character development within the Diagnostics team?

The sea rescue brings the couple to the team's attention, setting off a series of events that act as catalysts for character development. The challenging medical mystery forces team members to confront their own abilities, resolve conflicts, and cope with Foreman's departure, leading to significant personal and professional growth.

Why the answer is not valid ?

Because the couple traveled a great distance, time and danger to reach the hospital through the Coast Guards. This encourages the House team to do their best for this case.

Figure 17: Examples of success and failure cases in Linking Multiple Events questions.

Temporal questions

Choose the correct option for the following question: Given the events listed ['Rachel apologizes to Ross and suggests a romantic dinner to make it up to him', 'Monica goes to her eye appointment with Dr. Burke', 'Monica and Dr. Burke kiss'], what is the sequential order in this episode?

['Rachel apologizes to Ross and suggests a romantic dinner to make it up to him', 'Monica goes to her eye appointment with Dr. Burke', 'Monica and Dr. Burke kiss']

Choose the correct option for the following question: Looking at these events ['Phoebe offers to waitress for Monica instead of Rachel', 'Monica and Phoebe arrive at Dr. Burke's apartment for a catering job', 'Dr. Burke tells Monica about his divorce', 'Chandler orders a pizza for him and Joey'], how do they unfold in the episode?

['Phoebe offers to waitress for Monica instead of Rachel', 'Monica and Phoebe arrive at Dr. Burke's apartment for a catering job', 'Chandler orders a pizza for him and Joey', 'Dr. Burke tells Monica about his divorce']

Why the answer is not valid ?

Because Dr. Burke tells Monica about his divorce' before 'Chandler orders a pizza for him'

Figure 18: Examples of success and failure cases in Temporal Order of Events questions.

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 19: 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 20: 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 21: Detailed prompt for chronological understanding of events skill.

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 22: 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 23: Detailed prompt for deep context understanding questions generation.