MovieRecapsQA: A Multimodal Open-Ended Video Question-Answering Benchmark

Supplementary Material

A. Copyright

To ensure that MovieRecapsQA is released in full compliance with copyright law and platform terms of service, we adopt a strict non-redistribution policy for all copyrighted material. The public release includes only derived, non-copyrighted metadata such as question–answer pairs, extracted factoids, aligned movie and recap-video timestamps, YouTube identifiers, and links to publicly accessible resources (e.g., IMDb and OpenSubtitles). None of these elements contain audio, video frames, subtitle text, or other proprietary content; instead, they serve solely as pointers that allow researchers to reconstruct the data locally from legally obtained sources. This dataset is to be used for research purpose use only, and users remain responsible for complying with the copyright and platform terms associated with YouTube, IMDb, OpenSubtitles, and all other third-party sources.

Prior to the release of the dataset, we contacted all recap-video creators whose content is referenced by our metadata, informing them of the intended use and the nature of the release. None expressed objections, and we maintain an ongoing opt-out policy under which any creator may request removal of metadata referencing their video at any time. Through this design, MovieRecapsQA supports open, reproducible scientific research while respecting copyright protections and the rights of all content creators involved.

B. Dataset Details – Collection & Generation

In this section, we provide additional details on the construction of our dataset. As described in Section 3, we build MovieRecapsQA around publicly available recap videos for movies that also have publicly accessible movie scripts. This design choice ensures that future extensions of the benchmark can seamlessly incorporate full movie scripts—which offer a richer and more accurate textual source than subtitles—without altering the core data pipeline.

Movie Selection and Collection. We select movies using the open-source, fan-made scripts available on IMSDb. We crawl all available scripts and extract the associated movie titles, and then use the official IMDb API to obtain the corresponding movie metadata. We then obtain movie subtitles by searching and downloading the movie subtitles from the open source website, OpenSubtitles.

Instructions: You are a helpful assistant who can extract atomic claims from a piece of text.
You are trying to create a database of facts for the given text by extracting all atomic claims. To do so, you need to break down a sentence and extract as many fine-grained facts mentioned in the response. Each fact should also be describing either one single event with necessary time and location information. You should focus on the named entities and numbers in the sentence and extract relevant information from the sentence. Recover pronouns, definite phrases (e.g., "the victims" or "the pope"), and so on. Each fact should be understandable on its own and require no additional context. I will provide you with the facts from the previous segment to use as context to do coreference resolution. All entities must be referred to by name but not pronoun. Use the name of entities rather than definite noun phrases (e.g., 'the teacher') whenever possible. If a definite noun phrase is used, be sure to add modifiers (e.g., a embedded clause, a prepositional phrase, etc.). Each fact must be situated within relevant temporal and location whenever needed. Keep each fact to one sentence with zero or at most one embedded clause. You do not need to justify what you extract. Previous Segment's Facts: { *previous_facts* } Extract **atomic** fact from Text: {*text*}

Table S1. **Prompt for Atomic Fact Extraction from Segments.**

Recap Video Collection. We search YouTube using the official API and retrieve the top five results for the query "**recap video for {movie_name}–{movie_release_year}**". We download the video, metadata, and audio for each YouTube video using their official API. We then extract the movie title by applying OCR to sampled 10 frame within the first ten seconds; for certain channels, we instead use a set of rule-based heuristics for title extraction. This step is necessary because many recap creators intentionally avoid including the movie name in the title or description to encourage viewer engagement. After extracting candidate titles, we manually verify the alignment between each recap video and its corresponding movie to ensure that no mismatches occur. Lastly, we obtain the recap video captions through a paid service, Rev.ai.

Recap Video Scene Segmentation As described in Section 3, we segment each recap video using SceneDetect [16]. This step is applied to the video component only. The

Instructions
Generate question with their answers from the atomic facts generated. Make sure the answers are obtained from the atomic facts. Create complex questions that would require the use of one or multiple facts. Segment's Facts: { *segment_facts* }

Table S2. **Prompt for QA Generation.**

scene segmenter operates by first detecting individual shots and then grouping consecutive shots whose visual distance is below a threshold, using the tuned SceneDetect embedding model. We use the publicly available implementation provided by authors of Islam et al. [16].

QA and Fact Generation The dataset is constructed through a three-stage, prompt-based generation pipeline. First, we extract atomic facts from each segmented portion of the recap video, prompt found in Table S1. Next, we generate question–answer pairs grounded in one or more of these facts, prompt found in Table S2. Finally, we simplify the questions to increase their difficulty and reduce dependence on explicit character names, prompt found in Table S3. We provide more examples in Table S11.

C. Dataset Details – Alignment

A distinctive property of recap videos is that their visual content alone functions as a condensed summary of the movie. Recap creators typically stitch together the key shots that convey the major plot points and narrative transitions. As a result, it becomes possible to establish a meaningful alignment between the time-frames in the recap video and those in the full movie. This alignment, in turn, enables efficient matching between recap time-frames and movie subtitles, since subtitles are inherently time-stamped. [5]

For optimal alignment, we first detect all shots in both the full movie using a standard shot-detection pipeline. We encode each shot and segment by extracting SlowFast-50 embeddings from the first and last three seconds. To establish correspondences, we compute the maximum softmax similarity between the start–end embeddings of each recap shot and those of the movie, selecting the highest-confidence matches as candidate alignments while enforcing continuity and chronological consistency. Figure S3 shows example alignments from the dataset, where the x-axis denotes movie time-stamps and the y-axis denotes recap-video time-stamps.

However, because recap videos often reorder shots to maintain narrative flow, the resulting alignment is not always strictly chronological and can occasionally drift. This effect is visible in Figure S3a, where the alignment briefly moves out of order near the top-left region, reflecting cases in which recap creators introduce characters or events earlier than they appear in the movie. To address such issues, we apply a series of heuristic filtering and smoothing techniques to enforce temporal consistency and remove implausible matches.[6]

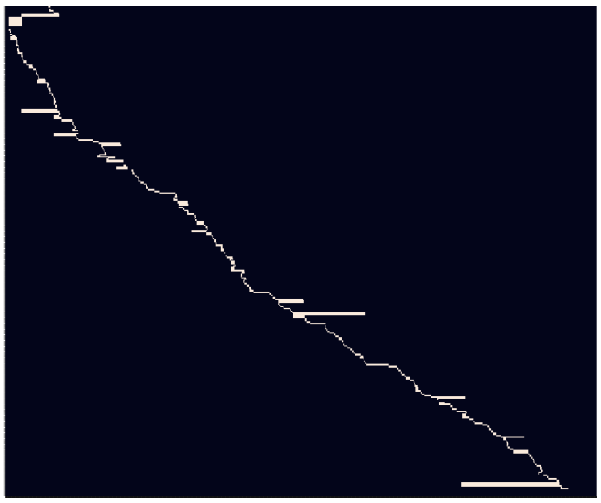
An additional observation about recap videos is that some movie scenes require disproportionately longer summaries. As shown in Figure S3b, a long stretch along the y-axis (recap time) aligns to a much shorter segment on the x-axis (movie time), indicating that a small cluster of shots in the movie contains dense, plot-critical information that warrants a more detailed recap. This behavior is especially common in action or spy movies (e.g., *Mission: Impossible*), where pivotal sequences—such as heists or reveal moments—contain numerous important events compressed into a short timeframe.

Instructions
Simplify and make a list of questions with their answers more ambiguous by removing unnecessary conditions or specifics. Your goal is to create a version of each question where key details are generalized or removed. If you remove a condition from the question, make sure that the answer reflects that change. Do not change ambiguous the answer. **## Steps:** 1. **Identify Key Details**: For each question, pinpoint specific details or conditions that could be removed or generalized. 2. **Generalize or Remove**: Simplify by either making the question more general or eliminating specific conditions, without losing the main intent. 3. **Ensure Ambiguity**: Aim to increase the ambiguity in each question so that it can be interpreted in multiple ways. 4. **Align Answers**: Adjust the answers as needed to reflect the removal or generalization of conditions. 5. **Remove Names**: Adjust the questions by removing all names. **## Output Format:** Provide simplified versions of the original questions, each on a new line, ensuring that answers are adjusted accordingly. Retain the order of the original list. It should be a list of json objects. **## Example:** Original Question: What did Alex use as bait on the day his son was born? Original Answer: Alex used his gold wedding ring as bait. Simplified Questions: What did he use as bait on an important day? Simplified Answer: He used something precious as bait. { *question_answer_pair* }

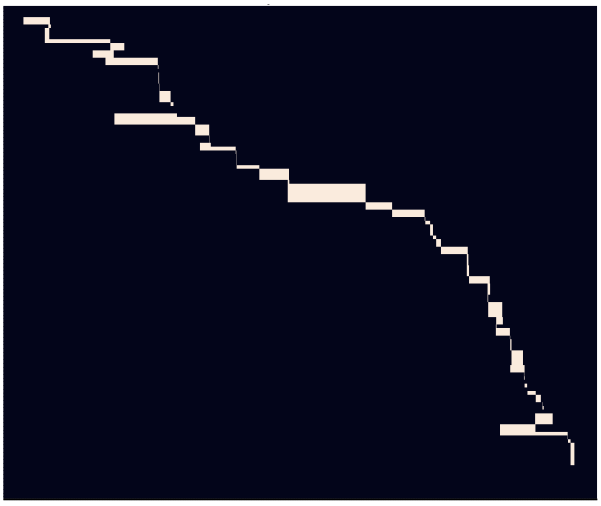
Table S3. **Prompt for QA Simplification.**

[5] We obtain the movie videos using Amazon Prime subscriptions.

[6] All alignment code and heuristics used in this process will be released with the dataset.



(a) Alignment between recap video *7zJ0nvgWpgk* and the movie *Year One*



(b) Alignment between recap video *pB6ULQIgmfg* and the movie *Mission Impossible*

Figure S3. **Alignment Between Recap Videos and Full Movies.** We show alignment examples for selected videos in the dataset. Each plot maps recap-video time-stamps (y-axis) to movie time-stamps (x-axis) using our segment-shot similarity procedure. While most alignments follow a near-diagonal structure, indicating chronological correspondence, recap videos occasionally reorder scenes for narrative flow (e.g., character introductions), resulting in local misalignments.

D. Alternative Problem Settings

Primary in this study, we evaluated the performance of MLLMs on MovieRecapsQA benchmark using the recap segment, $s_r \subset v_r$ (typically $\sim$ 73 seconds) as the video context to the questions. However, the benchmark provides deeper video–text alignments that enable several alternative and increasingly challenging evaluation modes.

Alignment Units. For each QA pair, the benchmark supplies: (1) the full recap video v_r (8–15 minutes), (2) the specific recap segment s_r from which the question is constructed, (3) the aligned movie segment s_m from the original film, and (4) the complete movie v_m. Textual components are also aligned across levels: the movie subtitles u_m align with s_m. The pair is also combined with the fact set $F = \{f_1, \ldots, f_k\}$ obtained from s_r. Each datapoint in MovieRecapsQA provides a structured multimodal tuple

$$\{\text{Question}, \text{Answer}, v_r,\ s_r,\ v_m,\ s_m,\ u_m,\ F\},$$

which supports multiple interchangeable input configurations.

Default Recap-Segment QA. The main benchmark setting evaluates models using only the recap segment:

$$\langle s_r,\ u_m,\ \text{Question}\rangle \rightarrow \text{Answer}.$$

This is the current setting designed in this study as it is the most cost effictive.

Movie-Segment QA. The benchmark also includes the movie segment s_m temporally aligned to the same narrative moment as s_r. This allows a formulation where models answer from original film footage rather than recap edits:

$$\langle s_m,\ u_m,\ \text{Question}\rangle \rightarrow \text{Answer}.$$

This setting isolates how models behave when provided full video context which differs in from the compressed video recap summaries.

Full-Recap QA. Instead of using a short clip, models may be evaluated on the entire recap video:

$$\langle v_r,\ u_m,\ \text{Question}\rangle \rightarrow \text{Answer}.$$

This introduces long-range multimodal reasoning and requires models to track character arcs, causality, and scene transitions across a 10–15 minute video summary.

Full-Movie QA. The benchmark further supports a long-video setting in which the full movie v_m and its subtitles u_m are used as input:

$$\langle v_m,\ u_m,\ \text{Question}\rangle \rightarrow \text{Answer}.$$

This is the most challenging configuration, demanding temporal grounding and narrative understanding across an entire 1.5–2.5 hour film.

Instructions: You are a helpful assistant who classifies movie questions into semantic categories.
Your task is to analyze each question from the movie "{MOVIE_NAME}" and assign it to **exactly one** semantic category. The categories capture different dimensions of reasoning required to answer the question. **TEMP (Temporal)**: questions involving time, sequence of events, chronology, duration, or ordering of actions. **CRD (Character and Relationship Dynamics)**: questions about character motivations, emotions, intentions, interactions, or interpersonal relationships. **NPA (Narrative and Plot Analysis)**: questions about story structure, plot developments, causal reasoning, or narrative elements. **STA (Setting and Technical Analysis)**: questions focusing on location, environment, staging, cinematography, or production-related elements. **TH (Thematic Exploration)**: questions addressing themes, symbolism, messages, or deeper conceptual meanings. For each provided question, output only its corresponding category label. **Questions**: {questions} **Categories**:

Table S4. **Prompt for Movie Question Categorization.**

Summary. These four configurations—recap segment, movie segment, full recap, and full movie—provide a continuum of evaluation difficulty, all derived from the same aligned QA pairs. While the benchmark's default setting emphasizes short, efficient VideoQA, its multi-granular alignment structure enables a broad family of alternative tasks including long-video comprehension, cross-modal alignment (recap ↔ movie), scene retrieval, and fact verification over extended multimodal context. This flexibility allows MovieRecapsQA to serve not only as a short-video QA benchmark but also as a testbed for long video context.

E. Question Categorization and Modality Prompt Templates

We categorize each question into one of five semantic reasoning categories, using the prompt shown in Table S4. The categories are:

- **NPA** (Narrative and Plot Analysis): questions concerning events, causality, and overall story progression.
- **CRD** (Character and Relationship Dynamics): questions involving character traits, emotions, intentions, or interpersonal relationships.
- **TH** (Thematic Exploration): questions about themes, moral lessons, symbolism, or overarching narrative messages.
- **STA** (Setting and Technical Attributes): questions focused on locations, visual style, cinematography, or production-related elements.
- **TEMP** (Temporal Reasoning): questions that involve ordering, duration, or timing of events within the segment.

Instructions: You are a helpful assistant who classifies movie questions into semantic categories.
Analyze the following questions from the movie "{MOVIE_NAME}" and classify each one into exactly one modality type. **Dialogue-based**: can be answered solely from dialogue (subtitles or spoken lines), without requiring any visual information. **Scene-based**: requires visual scene information (characters, actions, objects, locations) and cannot be answered from dialogue alone. **Multimodal**: requires both dialogue and visual information; neither modality alone is sufficient. Here are the questions to classify:

Table S5. **Prompt for Question Modality Classification (Dialogue, Scene, Multimodal).**

In addition, we annotate each question with a modality type. To determine this, we provide the model with the question, the associated dialogue, and a visual scene description, stored collectively as `context_pairs`. The prompt for this task appears in Table S5, and the modality types are:

- **Dialogue-based**: answerable solely from dialogue (spoken lines or subtitles), without requiring visual cues.
- **Scene-based**: requiring information from the visual scene (characters, actions, objects, locations) that cannot be inferred from dialogue alone.
- **Multimodal**: requiring both dialogue and visual information, where neither modality alone is sufficient.

F. Evaluation Metric Prompts

In this section, we provide the full prompts used to evaluate model responses in our benchmark. All evaluations are executed using the OpenAI Batch API, which allows us to scale the assessment of thousands of model outputs in a cost-efficient and environmentally responsible manner.

Atomic Claim Extraction. Given a model-generated answer, we first extract its underlying atomic claims using the prompt shown in Table S6. This step decomposes the model's answer into fine-grained, verifiable units of meaning, enabling consistent downstream factuality and relevance evaluation. This is inspired by work on factuality evaluation in text question-answering.

Factuality Evaluation. To evaluate factual correctness, we supply the evaluator with: (i) all atomic facts extracted

Instructions: You are a helpful assistant who can extract atomic claims from a piece of text.

You are trying to verify how factual a response to a question or request is. To do so, you must break down the model's answer into as many fine-grained, atomic facts as possible. Each fact must describe a single event, state, or relation, including necessary temporal or location information when relevant.

Focus on named entities and numbers, and extract all relevant information expressed in the sentence. Do **not** extract claims from the question itself; the question serves only as context to resolve pronouns, definite noun phrases (e.g., "the victims", "the pope"), and other referring expressions. Each fact must be understandable on its own, without requiring additional context.

All entities should be referred to by explicit name rather than pronoun. When using definite noun phrases, include modifiers (e.g., embedded clauses, prepositional phrases) to ensure specificity. Each fact should be one sentence long, with zero or at most one embedded clause. You do not need to justify the extracted facts.

Extract **atomic** facts.

Question: {question}
Model Answer: {answer}

Facts:

Table S6. **Prompt for Atomic Fact Extraction from Model Answers.**

from the corresponding video segment, (ii) the model's answer claims, (iii) the question, and (iv) the aligned SRT dialogue. This design reflects the core principle of factuality: a model's answer must not introduce information that contradicts, hallucinates beyond, or misrepresents the input evidence. Because the segment-level facts serve as a textual representation of the visual content, and the SRT dialogue captures additional narrative cues, these two sources together provide comprehensive grounding for judging factual accuracy. The exact factuality evaluation prompt is provided in Table S12.

Relevance Evaluation. For relevance, we evaluate whether each extracted claim meaningfully contributes to answering the user's question. We provide the evaluator with: (i) the question, (ii) the model's answer claims, (iii) the SRT dialogue, and (iv) the *aligned facts used to generate that question*. These aligned facts encode the semantic intent of the question, allowing the evaluator to determine whether a claim "belongs" to the same underlying evidence that motivated the question. Unlike factuality, correctness plays no role here—a claim can be factually wrong yet still relevant if it attempts to answer the question. The full relevance evaluation prompt is shown in Table S13.

G. Evaluation Metric – Coherence

Beyond factuality and relevance, coherence is frequently used as an auxiliary evaluation dimension in text-based question answering, as seen in metrics such as G-Eval Fluency and HELMET Fluency. Motivated by this, we also incorporate a coherence assessment in our benchmark.

To evaluate coherence, we provide the evaluator with: (i) the model's extracted answer claims, and (ii) the question. Coherence here measures the internal logical consistency of the response: a coherent answer should not contain claims that contradict one another or repeat the same information unnecessarily. The full coherence evaluation prompt is provided in Table S14.

However, unlike long-form text generation tasks, coherence is significantly less informative in open-ended VideoQA. Answers in VideoQA are typically short and contain very few distinct claims, which greatly limits the possibility of internal contradictions. Consequently, coherence scores exhibit extremely low variance across models. We report these results on the 118 manually evaluated questions in Table S7, where the near-uniform scores confirm that coherence is not a meaningful discriminative dimension for short-form VideoQA responses. For this reason, while we compute coherence for completeness, it does not play a substantial role in evaluating model performance in our setting.

Model	Coherence Evaluation
LLaVA-NeXT-Video	$4.79_{\pm 0.41}$
MiniCPM-o	$4.75_{\pm 0.28}$
Qwen	$4.82_{\pm 0.28}$
Amazon Nova	$4.71_{\pm 0.40}$
Claude	$4.65_{\pm 0.40}$
Gemini 2.5 Flash	$4.81_{\pm 0.37}$
GPT-4o	$4.77_{\pm 0.36}$
Avg. Human*	$4.86_{\pm 0.31}$
Best Human*	$5.00_{\pm 0.00}$

Table S7. **Coherence Scores (1–5).** Mean ± variance coherence scores for all models and human annotators, rounded to two decimal places. The consistently low variance confirms that coherence is not a discriminative metric for short-form VideoQA.

H. Detailed Ablation Tables

In this section, we provide the complete ablation results for each model under all evaluation settings. Table S8 reports the relevance scores across question types and categories, while Table S9 presents the corresponding factuality scores. These tables complement the main results by showing how each model behaves when provided with different input modalities and ablated forms of the video–text context.

Summary of Model Behaviors. Tables S8–S9 reveal consistent trends across both relevance and factuality. Proprietary models (GPT-4o, Claude 3.5 Sonnet, Amazon Nova Lite) outperform open-source systems across nearly all question types and categories when given full context. Dialogue-based questions are the easiest for all models, while scene-based and multimodal questions expose clear gaps in visual grounding, especially for open-source models.

Effect of Input Ablations. Dialogue-only inputs yield the strongest gains for most models—often surpassing the full-context baseline by large margins. GPT-4o and Claude 3.5 improve by up to +0.4–0.7 in both relevance and factuality, indicating that subtitles carry the dominant grounding signal. In contrast, frames-only inputs provide limited benefit: Qwen2.5VL is the only model that consistently improves under visual-only conditions, particularly for temporal and spatial reasoning. For other models, removing subtitles significantly degrades performance.

Category-Level Observations. CRD, NPA, and TH categories benefit most from dialogue, reflecting their reliance on explicit narrative cues. STA and TEMP remain the most challenging: even strong models show reduced factual grounding, and multimodal fusion rarely helps. Notably, Gemini-2.5-Flash exhibits the weakest temporal grounding (e.g., TEMP factuality 2.53), while Qwen2.5VL achieves the strongest temporal improvements under frames-only input.

Overall. The evaluation highlights three core findings: (i) subtitles dominate model grounding on recap-derived QA; (ii) visual understanding remains model-dependent and uneven; and (iii) multimodal fusion remains an open challenge, with models frequently performing better when one modality is suppressed.

	Question Types			Question Categories				
Model	**Dialogue**	**Scene**	**Multimodal**	**CRD**	**NPA**	**STA**	**TEMP**	**TH**
LLaVA-NeXT-Video	3.36	3.35	3.33	3.30	3.31	3.37	3.54	3.52
(only frames)	3.37↑	3.41↑	3.43↑	3.45↑	3.32↑	3.41↑	3.43↓	3.53↑
(only dialogue)	3.59↑	3.40↑	3.48↑	3.56↑	3.45↑	3.42↑	3.43↓	3.54↑
Mini-CPM-o	3.54	3.55	3.52	3.52	3.50	3.56	3.66	3.74
(only frames)	3.70↑	3.72↑	3.71↑	3.71↑	3.65↑	3.68↑	3.72↑	3.84↑
(only dialogue)	3.86↑	3.65↑	3.82↑	3.81↑	3.73↑	3.72↑	3.74↑	4.00↑
Qwen2.5VL	3.93	3.69	3.72	3.78	3.75	3.80	3.90	3.91
(only frames)	3.92↓	3.88↑	3.87↑	3.88↑	3.80↑	3.88↑	3.95↑	4.15↑
(only dialogue)	3.52↓	3.27↓	3.51↓	3.49↓	3.45↓	3.17↓	3.42↓	3.65↓
Amazon Nova Lite	4.12	3.82	3.99	3.97	3.95	3.81	3.94	4.23
(only frames)	3.81↓	3.53↓	3.77↓	3.78↓	3.66↓	3.74↓	3.39↓	3.73↓
(only dialogue)	–	–	–	–	–	–	–	–
Claude 3.5 Sonnet	3.88	3.71	3.83	3.86	3.72	3.61	3.99	3.82
(only frames)	3.48↓	3.68↓	3.68↓	3.58↓	3.60↓	3.62↑	3.77↓	3.54↓
(only dialogue)	4.09↑	3.84↑	4.07↑	4.08↑	3.99↑	3.77↑	3.87↓	4.10↑
Gemini-2.5-Flash	3.66	3.45	3.67	3.67	3.58	3.38	3.41	3.62
(only frames)	3.55↓	3.38↓	3.51↓	3.49↓	3.47↓	3.38↑	3.44↑	3.58↓
(only dialogue)	4.07↑	3.77↑	4.03↑	4.05↑	3.95↑	3.67↑	3.73↑	4.02↑
GPT-4o	3.71	3.55	3.84	3.78	3.73	3.32	3.59	3.76
(only frames)	3.83↑	3.99↑	3.93↑	3.87↑	3.91↑	3.82↑	3.89↑	3.91↑
(only dialogue)	4.20↑	3.82↑	4.13↑	4.16↑	4.05↑	3.80↑	3.85↑	4.12↑

Table S8. **Model performance (mean relevance) across question types and categories.** Arrows indicate increase/decrease relative to the full model baseline.

	Question Types			Question Categories				
Model	**Dialogue**	**Scene**	**Multimodal**	**CRD**	**NPA**	**STA**	**TEMP**	**TH**
LLaVA-NeXT-Video	2.99	2.88	2.88	2.99	2.90	2.65	3.04	2.78
(only frames)	2.69↓	2.82↓	2.78↓	2.79↓	2.77↓	2.64↓	2.73↓	2.64↓
(only dialogue)	3.14↑	2.93↑	3.02↑	3.14↑	3.02↑	2.81↑	2.92↓	2.91↑
Mini-CPM-o	3.15	3.00	3.09	3.14	3.10	2.76	3.02	3.02
(only frames)	3.11↓	3.10↑	3.07↓	3.13↓	3.09↓	2.84↑	3.10↑	2.87↓
(only dialogue)	3.37↑	3.17↑	3.35↑	3.41↑	3.37↑	2.94↑	3.13↑	3.07↑
Qwen2.5VL	3.50	3.28	3.35	3.42	3.40	3.07	3.39	3.27
(only frames)	3.42↓	3.49↑	3.41↑	3.43↑	3.43↑	3.36↑	3.45↑	3.29↑
(only dialogue)	3.10↓	3.10↓	3.24↓	3.21↓	3.25↓	2.58↓	2.92↓	3.16↓
Amazon Nova Lite	3.73	3.35	3.58	3.59	3.60	3.15	3.51	3.37
(only frames)	2.98↓	2.78↓	3.17↓	3.22↓	3.02↓	2.62↓	2.69↓	2.87↓
(only dialogue)	–	–	–	–	–	–	–	–
Claude 3.5 Sonnet	3.69	3.17	3.58	3.65	3.42	3.12	3.30	3.44
(only frames)	3.19↓	3.15↓	3.24↓	3.23↓	3.25↓	2.81↓	3.19↓	3.05↓
(only dialogue)	4.17↑	3.93↑	4.21↑	4.17↑	4.13↑	3.78↑	3.95↑	4.18↑
Gemini-2.5-Flash	3.34	2.65	3.03	3.15	3.00	2.57	2.53	3.16
(only frames)	2.99↓	2.78↑	2.87↓	2.94↓	2.92↓	2.50↓	2.68↑	2.82↓
(only dialogue)	4.03↑	3.77↑	4.11↑	4.13↑	4.00↑	3.48↑	3.66↑	3.87↑
GPT-4o	3.76	3.43	3.66	3.73	3.64	3.10	3.58	3.55
(only frames)	3.47↓	3.54↑	3.58↓	3.59↓	3.57↓	3.26↑	3.32↓	3.26↓
(only dialogue)	4.07↑	3.72↑	4.06↑	4.14↑	4.00↑	3.41↑	3.63↑	3.86↑

Table S9. **Model performance (mean factuality) across question types and categories.** Arrows indicate increase/decrease relative to the full model baseline.

	Verbose Question	**Question**	**Answer**	Category	Type
1	What indication is there that Jim is unaware of the circumstances in the city?	What shows that Jim is unaware of the situation?	Jim has no idea what happened, indicating his lack of knowledge about the situation in the city.	NPA	multimodal
2	What activities does he engage in as part of recalling his earlier years?	What activities are linked to his fond memories?	He reminisces about his past glory days as part of his daily routine.	CRD	dialogue-based
3	Where is he hoping to find other people?	Where is he hoping to find others?	He is hoping to find other people in the city.	CRD	scene-based
4	In what circumstances does Neo find himself after swallowing the red pill, and what does this reveal about his previous existence?	What situation does someone find themselves in after making a choice, and what does this suggest about their past?	After taking the red pill, Neo awakens in a pod among countless humans, revealing his previous life was a simulation.	TH	multimodal
5	Describe the sequence of events that occurs during the encounter between Deckard and Chris.	What happens during the encounter between two characters?	Deckard is ambushed by Chris, but he manages to kill her just as Roy returns.	NPA	multimodal
6	How does Dr. Chilton's behavior towards Clarice reflect the dynamics of their interaction?	How does one person's behavior towards another reflect their interaction dynamics?	Dr. Chilton's suggestive remarks show personal interest, contrasting with the professional setting and hinting at manipulation.	CRD	dialogue-based
7	What are the final moments witnessed by Deckard after Roy delivers his monologue?	What does one character witness after a speech from another?	Roy dies in front of Deckard, who watches silently as Gaff arrives and calls to him.	CRD	scene-based
8	What does Valentina believe is the only choice in the situation?	What does Valentina think is the choice to make?	Valentina states that the only choice is to help.	CRD	dialogue-based
9	What emotions does Andrew experience after stepping out of the party, and what does he do in the parking lot?	What feelings does Andrew go through after leaving an event, and what does he do outside?	Andrew cries as he cleans dirt off his camera after leaving the party.	CRD	multimodal

Table S10. Representative examples from the MovieRecapsQA dataset. Each entry shows the question, answer, category, type, and its corresponding verbose formulation used during question generation. Alternating row shading improves readability.

	Model Response	Claims	Factuality score	Relevence Score
Claude 3.5 Sonnet	They skin the Father and make a mask and breastplate out of his treated flesh	(1) The group skins the Father. (2) The group makes a mask out of the treated flesh of the Father. (3) The group makes a breastplate out of the treated flesh of the Father.	(1) Supported (2) Supported (3) Supported **(Overall 5)**	(1) Relevant (2) Relevant (3) Relevant **(Overall 5)**
Gemini-2.5-Flash	The group member Sweetie calls her father from a phone booth. Based on this call, her father, a retired police officer, contacts the police and provides them with the location of the museum where his daughter called from.	(1) Sweetie calls her father from a phone booth. (2) Sweetie's father is a retired police officer. (3) Sweetie's father contacts the police. (4) Sweetie's father provides the police with the location of the museum. (5) The call from Sweetie occurs at the phone booth. (6) The museum is the location from which Sweetie called her father.	(1) Unsupported (2) Unsupported (3) Unsupported (4) Unsupported (5) Unsupported (6) Unsupported **(Overall 0)**	(1) NotRelevant (2) NotRelevant (3) NotRelevant (4) NotRelevant (5) NotRelevant (6) NotRelevant **(Overall 2)**
GPT-4o	They skin Father and make a mask and breastplate out of his treated flesh.	(1) The group skins Father. (2) The group makes a mask out of Father's treated flesh. (3) The group makes a breastplate out of Father's treated flesh.	(1) Supported (2) Supported (3) Supported **(Overall 5)**	(1) Relevant (2) Relevant (3) Relevant **(Overall 5)**
Amazon Nova Lite	The group burns the mask that the father used.	(1) The group burns the mask. (2) The mask was used by the father.	(1) Unsupported (2) Unsupported **(Overall 1)**	(1) NotRelevant (2) NotRelevant **(Overall 2)**
LLaVA-NeXT-Video	The group takes the father's ashes and uses them to make a statue of him.	(1) The group takes the father's ashes. (2) The group uses the father's ashes to make a statue of him.	(1) Unsupported (2) Unsupported **(Overall 1)**	(1) Partially relevant (2) NotRelevant **(Overall 2)**
Qwen2.5VL	The group makes a mask and breastplate out of the father's treated flesh.	(1) The group makes a mask out of the father's treated flesh. (2) The group makes a breastplate out of the father's treated flesh.	(1) Supported (2) Supported **(Overall 5)**	(1) Relevant (2) Relevant **(Overall 5)**
Mini-CPM-o	The group skins the father and makes a mask and breastplate out of his treated flesh.	(1) The group skins the father. (2) The group makes a mask out of the father's treated flesh. (3) The group makes a breastplate out of the father's treated flesh.	(1) Supported (2) Supported (3) Supported **(Overall 5)**	(1) Relevant (2) Relevant (3) Relevant **(Overall 5)**

Table S11. Representative examples for models responses and their evaluation on the question **Q. What does the group do with something they have from the father?** *The group skins the father and makes a mask and breastplate out of his treated flesh.*

Instructions: You are an expert factuality evaluator for video question answering systems. Your task is to evaluate the factual accuracy of claims made in a model's response by comparing them against ground truth atomic facts and dialogue from the same video segment.

Input Information:
Question: {question}
Claims Extracted from Model Response:
{claims}
Ground Truth Atomic Facts from Video Segment:
{facts}
SRT Dialogue Context:
{context}

Evaluation Task:
For each claim in the model response, evaluate its factual accuracy by checking against:
1. The ground truth atomic facts
2. The SRT dialogue context

Scoring Rubric: For each claim, assign one label:
– *SUPPORTED (S)*: Directly supported by the facts or dialogue.
– *PARTIALLY_SUPPORTED (PS)*: Some accurate information, but imprecise or partially incorrect.
– *UNSUPPORTED (U)*: Not supported by any fact or dialogue.
– *CONTRADICTORY (C)*: Directly contradicts facts or dialogue.
– *NOT_CHECKABLE (NC)*: Cannot be verified from the provided sources.

Important Guidelines:
1. Penalize hallucinations strictly. Unsupported additions should be marked U or C.
2. Consider the dialogue when verifying the claims. It may contain information not in the atomic facts.

Factuality Score (0–5):
5: All claims supported; no hallucinations.
4: Mostly supported; minor partial issues.
3: Mix of supported and partial; some unsupported but no contradictions.
2: Multiple unsupported claims or one contradiction.
1: Mostly unsupported or contradictory.
0: Entirely incorrect, unsupported, or contradictory.

Examples:
Supported: Claim matches facts and dialogue.
Partially Supported: Claim has some accuracy but adds imprecise details.
Unsupported: Claim adds unverifiable information.
Contradictory: Claim conflicts with facts/dialogue.

Overall Score Examples:
Score 5: All claims supported.
Score 4: Mostly supported with one minor issue.
Score 3: Mix of supported/partial.
Score 2: Includes unsupported or contradictory.
Score 1: Mostly hallucinated.
Score 0: Fully wrong or contradictory.

At the end, evaluate all claims according to these guidelines.

Table S12. **Prompt for Reference-Free Factuality Evaluation.**

Instructions: You are an expert relevance evaluator for video question answering systems. Your task is to evaluate the relevance of claims made in a model's response by comparing them against the user's question. The ground truth facts and dialogue should be used only as context to understand the claims, not to judge correctness.

Input Information:
Question: {question}
Claims Extracted from Model Response:
{claims}
Ground Truth Atomic Facts Used to Answer the Question:
{facts}
SRT Dialogue Context:
{context}

Evaluation Task:
For each claim, evaluate its relevance to the question.
1. Compare the claim's topic to the question's topic.
2. Use the facts and dialogue to understand the meaning of the claim.
3. Do **not** evaluate factual correctness. A claim may be wrong but still relevant.

Scoring Rubric:
RELEVANT (R): Directly answers or is clearly pertinent to the question.
PARTIALLY_RELEVANT (PR): Related to the general topic but tangential or not directly asked for.
NOT_RELEVANT (NR): Off-topic; does not help answer the question.

Important Guidelines:
Relevance is independent of correctness.
Claims answering only part of a multi-part question are still relevant.
Tangential information should be marked PR or NR.

Relevance Score (0–5):
5: All claims are RELEVANT. Perfect focus.
4: Mostly RELEVANT; at most minor PR tangents.
3: Mix of R and PR; no NR claims.
2: At least one NR claim; answer loses focus.
1: Multiple NR claims; mostly off-topic.
0: All claims NR; completely irrelevant.

Examples:
Relevant: Answers "who" and "what he does."
Partially Relevant: Describes the correct entity but not the attribute asked for.
Not Relevant: About a different person, not answering the question.
Relevant but Incorrect: Still relevant if it tries to answer the question.

After evaluating all claims, assign a final relevance score following these rules.

Table S13. **Prompt for Reference-Free Relevance Evaluation.**

Instructions: You are an expert coherence evaluator for video question answering systems. Your task is to evaluate the internal logical coherence of a model's response by comparing the claims it makes against one another. A coherent response should not contain contradictions or excessive redundancy.

Input Information:
Question: {question}
Claims Extracted from Model Response:
{claims}

Evaluation Task:
For each claim, evaluate its logical consistency with all other claims in the same response. The goal is to detect internal contradictions or redundancies; external factual correctness should **not** be considered.

Scoring Rubric:
CONSISTENT (CO): Logically consistent with all other claims. Introduces new information without conflict.
REDUNDANT (R): Repeats information expressed in another claim using different phrasing. A minor coherence flaw.
CONTRADICTORY (C): Directly contradicts one or more other claims.

Important Guidelines:
– Look for direct logical opposites (e.g., "Tony is happy" vs. "Tony is sad").
– Increased specificity is **not** contradiction (e.g., "doctor" vs. "brain surgeon").
– Redundancy occurs when a claim adds no new information.
– For REDUNDANT or CONTRADICTORY labels, reference the related claim number in the justification.

Coherence Score (0–5):
5: All claims CONSISTENT; no redundancy or contradictions.
4: Mostly CONSISTENT; at most one or two REDUNDANT claims; no contradictions.
3: Several REDUNDANT claims but no CONTRADICTORY claims.
2: At least one CONTRADICTORY pair.
1: Multiple CONTRADICTORY claims; logically confusing.
0: Most claims CONTRADICTORY; response is incoherent.

Examples:
CONSISTENT: Introduces new, non-conflicting information.
REDUNDANT: Restates ideas already given (e.g., "in charge of the hospital" vs. "head of the hospital").
CONTRADICTORY: Directly conflicts with previous claims (e.g., "head of the hospital" vs. "junior intern").

After evaluating all claims, assign a final coherence score following these rules.

Table S14. **Prompt for Reference-Free Coherence Evaluation.**