CinePile: A Long Video Question Answering Dataset and Benchmark
Appendix

Note: Also included in the zip file are: a) the complete code for loading data, running responses, and evaluating accuracy; b) the Hugging Face dataset objects for the training and test splits, c) the code for running adversarial refinement pipeline, and d) questions generated on longer and different videos.

Contents

A ADDITIONAL MOVIE CLIP & QUESTIONS EXAMPLES

We present a few examples from our dataset in Figs. 14a, 14b, 15a, 15b, 16a, 16b, 17a and 17b.

B ADDITIONAL RELATED WORK

Synthetic data with human in the loop. Training models on synthetic data is a popular paradigm in recent times. We have seen many advances in generation as well as usage on synthetic data in recent times, both in vision Wood et al. (2021); Bordes et al. (2024); Tian et al. (2023); Hemmat et al. (2023) and language Taori et al. (2023); Maini et al. (2024); Li et al. (2023c); Yuan et al. (2024); Wei et al. (2023). For instance, Self-Instruct Wang et al. (2022) proposes a pipeline to create an instruction dataset based on a few instruction examples and categories defined by humans. We mainly derived inspiration and the fact that modern LLMs are quite good at understanding long text and creating question-answer pairs. UltraChat Ding et al. (2023) is another synthetic language dataset which is created by using separate LLMs to iteratively generate opening dialogue lines, simulate user queries, and provide responses. This allows constructing large-scale multi-turn dialogue data without directly using existing internet data as prompts. Additionally, Evol-Instruct Xu et al. (2023), automatically generates a diverse corpus of open-domain instructions of varying complexities by prompting an LLM and applying iterative evolution operations like in-depth evolving (adding constraints, deepening, etc.) and in-breadth evolving (generating new instructions). To our knowledge, we are among the first to apply automated template generation and question synthesis techniques to vision and video modalities using LLMs.

C ADDITIONAL QA GENERATION DETAILS

In addition to the hand-crafted perceptual templates, we also create long-form question and answers based on a scene's visual summary. To achieve this, we first generate a visual summary of a video clip. Then, we prompt the model to create question-answers solely based on that summary.

We create a pure visual summary of the scene by using a vision LLM, similar to some of the recent worksWang et al. (2023); Zhang et al. (2023a). First, we use a shot detection algorithm to pick the important frames[4], then we annotate each of these frames with Gemini vision API (`gemini-pro-vision`). We ablated many SOTA open-source vision LLMs such as Llava 1.5-13B Liu et al. (2023), OtterHD Li et al. (2023a), mPlug-Owl Ye et al. (2023b) and MinGPT-4 Zhu et al. (2023), along with Gemini and GPT-4V (`GPT-4-1106-vision-preview`). While GPT-4V has high fidelity in terms of image captioning, it is quite expensive. Most of the open-source LLM captions are riddled with hallucinations. After qualitatively evaluating across many scenes, we found that Gemini's frame descriptions are reliable and they do not suffer too much from hallucination. Once we have frame-level descriptions, we then pass the concatenated text to Gemini text model `gemini-pro` and prompt it to produce a short descriptive summary of the whole scene. Even though Gemini's scene visual summary is less likely to have hallucinated elements, we however spotted a few hallucinated sentences. Hence all the MCQs generated using this summary are added only to the training split but not to the eval split.

Monetary Costs for Question Generation: We provide a cost estimate of using GPT-4o for generating QA pairs for one particular scene:

- Base prompt (instructions for question-answer generation and templates): 1,167 tokens
- Movie scene (subtitles and visual descriptions): 465 tokens (average; varies across scenes)
- Total Input Tokens per Scene: 1,632 tokens
- Cost per Input Token: \$2.50 per 1M tokens
- Input Cost per Scene**: $\frac{1{,}632}{1{,}000{,}000} \times 2.50 = \$0.00408$
- Average output tokens: 1,582 tokens (average; varies across scenes)
- Cost per Output Token: \$10.00 per 1M tokens

[4] `https://www.scenedetect.com/`

- Output Cost per Scene: $\frac{1,582}{1,000,000} \times 10.00 = \$0.01582$
- Total Cost per Scene: $\$0.00408 + \$0.01582 = \$0.0199$

D Question Template Category Details

Character and Relationship Dynamics: This category would include templates that focus on the actions, motivations, and interactions of characters within the movie. It would also cover aspects such as character roles, reactions, decisions, and relationships.

Narrative and Plot Analysis: This category would encompass templates that delve into the storyline, plot twists, event sequences, and the overall narrative structure of the movie. It would also include templates that explore the cause-and-effect dynamics within the plot.

Thematic Exploration: This category would include templates that focus on the underlying themes, symbols, motifs, and subtext within the movie. It would also cover aspects such as moral dilemmas, emotional responses, and the impact of discoveries.

Setting and Technical Analysis: This category would encompass templates that focus on the setting, environment, and technical aspects of the movie. It would include templates that analyze the location of characters and objects, the use of props, the impact of interactions on the environment, and the description and function of objects.

Temporal: This category pertains to questions and answers that assess a model's comprehension of a movie clip's temporal aspects, such as the accurate counting of specific actions, the understanding of the sequence of events, etc.

Table 3: Sample templates and prototypical questions from each of the categories

Category	Question template	Prototypical question
Character and Relationship Dynamics (CRD)	Interpersonal Dynamics	What changes occur in the relationship between person A and person B following a shared experience or actions?
Character and Relationship Dynamics (CRD)	Decision Justification	What reasons did the character give for making their decision?
Narrative and Plot Analysis (NPA)	Crisis Event	What major event leads to the character's drastic action?
Narrative and Plot Analysis (NPA)	Mysteries Unveiled	What secret does character A reveal about event B?
Setting and Technical Analysis (STA)	Physical Possessions	What is [Character Name] holding?
Setting and Technical Analysis (STA)	Environmental Details	What does the [setting/location] look like [during/at] [specific time/place/event]?
Temporal (TEMP)	Critical Time-Sensitive Actions	What must [Character] do quickly, and what are the consequences otherwise?
Temporal (Temp)	Frequency	How many times does a character attempt [action A]?
Thematic Exploration (TH)	Symbolism and Motif Tracking	Are there any symbols or motifs introduced in Scene A that reappear or evolve in Scene B, and what do they signify?
Thematic Exploration (TH)	Thematic Parallels	What does the chaos in the scene parallel in terms of the movie's themes?

E QA Generation by Different Models

In this section, we present example question-answer (QA) pairs generated by GPT-4 and Gemini across various question categories in Table 4 and Table 5. As alluded to in the main paper, we note that GPT-4 consistently produces high-quality questions in all categories. In contrast, Gemini works well only for a few select categories, namely, Character Relationships and Interpersonal Dynamics

Table 4: Comparing question-answer pairs generated by GPT-4 with those generated by Gemini, for the movie clip: The Heartbreak Kid (3/9) Movie CLIP - Taking the Plunge (2007) HD. TEMP refers to Temporal. Please refer to Table 3 for other acronyms.

Category	**GPT-4 Generated QA**	**Gemini Generated QA**
CRD	Question: What is the significant event that Eddie and Lila are celebrating? - A) Their wedding ✓ - B) Their first date anniversary - C) Lila's birthday - D) Their engagement - E) Eddie's promotion at work	Question: What is Eddie doing at the beginning of the scene? - A) Dancing with Lila - B) Giving a speech - C) Cutting the wedding cake - D) Kissing Lila ✓ - E) Talking to his friends
NPA	Question: What incident leads to the main character's change in attitude towards marriage? - A) His friend's advice ✓ - B) His mother's arrival - C) His bride's beauty - D) His friend's gift - E) His bride's dress	Question: How does Eddie resolve his conflict with his friend? - A) He apologizes for his past behavior. - B) He confronts his friend about their differences. - C) He ignores his friend and moves on. - D) He seeks revenge on his friend. - E) He reconciles with his friend. ✓
TEMP	Question: How long is the couple planning to take off for their road trip? - A) One week - B) Four weeks - C) Five weeks - D) Two weeks - E) Three weeks ✓	Question: What occurs immediately after the wedding ceremony? - A) The couple kisses. - B) The guests congratulate the couple. - C) The bride's mother arrives. ✓ - D) The couple leaves for their honeymoon. - E) The groom gives a speech.
STA	Question: Where is the gift Eddie's friend gives him supposed to end up? - A) With Uncle Tito ✓ - B) With Lila - C) With Eddie - D) With the wedding guests - E) With Eddie's mom	Question: What is the primary color of Lila's dress in the scene? - A) Red - B) Blue - C) Yellow - D) Green - E) White ✓
TH	Question: How does the emotional tone shift from the beginning to the end of the scene? - A) From excitement to disappointment - B) From joy to sorrow - C) From anticipation to regret - D) From happiness to surprise ✓ - E) From nervousness to relief	Question: What does the chaotic atmosphere at the reception symbolize in relation to the film's themes? - A) The unpredictability of life ✓ - B) The challenges of marriage - C) The importance of family - D) The power of love - E) The fragility of relationships

(CDR), and Setting and Technical Analysis (STA). The gap in quality of the QA generated stems not only from the implicitly better and diverse concepts captured by GPT-4, but also from the hallucination tendencies of Gemini. For instance, in Table- 4, Gemini mistakes the dialogue – "Thank you for talking some sense into me, man", between Eddie and his friend as a suggestion for conflict resolution, and forms a narrative question based on it – "How does Eddie resolve his conflict with his friend?". Similarly, in Table 5, Gemini misremembers the temporal sequence and selects a wrong option as the answer choice for the temporal category. We quantify the quality of generated questions across the different choices of question-generation, and template selection models in Tab. 6. Here, we note that while the GPT-4 & GPT-4 combination results in the fewest degenerate questions, the Gemini & GPT-4 pairing also performs well and is cost-efficient on a large scale.

F TRAIN DATA STATISTICS

We present the question category statistics of train split in Fig. 8.

G ADDITIONAL EVALUATION DETAILS

We use two NVIDIA A40 GPUs, each with 48GB of memory, and two NVIDIA A100, each with memory of 82GB, for experiments with open-source models. The model versions and dates are as follows: Gemini 1.5 Pro [gemini-1.5-pro-001] and Gemini 1.5 Flash [gemini-1.5-flash-001], from May 20th to June 1st, 28th. GPT-4o [gpt-4o-2024-05-13] was used on May 14th, 2024; GPT-4 Vision [gpt-4-turbo], Gemini Pro Vision [gemini-pro-vision], and Claude 3 (Opus) [claude-3-opus-20240229] were used from April 29th to May 10th, 2024. The Gemini 1.5 models throw safety-blocking exceptions for a few of the videos, hence we could only evaluate them on $\approx$ 4.2k samples out of 4941. The closed-source models in our evaluations (GPT-4, Gemini, Claude

Table 5: Comparing question-answer pairs generated by GPT-4 with those generated by Gemini, for the movie clip: Ghostbusters: Afterlife (2021) - Muncher Attack Scene (3/7) | Movieclips. TEMP refers to Temporal. Please refer to Table 3 for other acronyms.

Category	GPT-4 Generated QA	Gemini Generated QA
CRD	Question: How does Phoebe's interaction with the ghost change throughout the scene? - A) She goes from hiding to attacking. ✓ - B) She goes from attacking to hiding. - C) She goes from running away to attacking. - D) She goes from communicating to attacking. - E) She goes from hiding to running away.	Question: How does the interaction between Phoebe and the ghost change throughout the scene? - A) Phoebe becomes more aggressive ✓ - B) The ghost becomes more passive - C) Phoebe becomes more fearful - D) The ghost becomes more aggressive - E) They remain neutral towards each other
NPA	Question: What major event prompts Phoebe to take drastic action against the ghost? - A) The ghost scuttles off and pipes burst into flame. - B) The ghost hides behind machinery. - C) The ghost starts eating a piece of pipe. - D) The ghost belches metal fragments that spark and ricochet around them. ✓ - E) The ghost starts searching the ground.	Question: What observation prompts Phoebe to take action? - A) The ghost's fear of the Aztec death whistle - B) The ghost's vulnerability to proton blasts - C) The ghost's search for something on the ground. ✓ - D) The ghost's reaction to Podcast's camera goggles - E) The ghost's belching of metal fragments
TEMP	Question: What happens immediately after the ghost belches metal fragments? - A) Phoebe ducks down. - B) The ghost scuttles off and pipes burst into flame. - C) Podcast blows the Aztec death whistle. - D) Phoebe powers up and fires a steady stream of protons. ✓ - E) Phoebe pokes her head up.	Question: Between which two events does Phoebe duck down? - A) The ghost searches the ground and Phoebe pokes her head up. - B) The ghost chomps on a pipe and Phoebe pokes her head up. - C) Podcast blows the whistle and the ghost belches metal fragments. - D) The ghost scuttles off and pipes burst into flame. ✓ - E) Phoebe fires protons and the ghost pokes its head out.
STA	Question: Where do Podcast and Phoebe hide during the ghost encounter? - A) Inside a car - B) In a building - C) Behind a tree - D) Under a table - E) Behind machinery ✓	Question: What is the primary material of the object that the ghost is chewing on? - A) Wood - B) Metal ✓ - C) Plastic - D) Rubber - E) Fabric
TH	How does the emotional tone shift throughout this scene? - A) From calm to chaotic - B) From fear to courage ✓ - C) From confusion to understanding - D) From excitement to disappointment - E) From sadness to joy	Question: How does the emotional tone shift from the characters' initial fear to their determination? - A) The podcast's calmness inspires Phoebe to become more assertive. - B) The ghost's search for something on the ground creates a sense of urgency. - C) The characters' realization that they have a plan instills confidence. ✓ - D) The ghost's belching of metal fragments intensifies the fear and chaos. - E) The characters' decision to use the trap marks a shift from fear to determination.

Table 6: Comparison of Template Selection and Question Generation Models in generating better questions (lower degenerate questions) for a subset of movie clips. While the GPT-4 GPT-4 combination performs the best, Template Selection model has minimal effect.

Template Selection Model	Question Generation Model	% Degenerate Questions
Gemini	Gemini	25.12
Gemini	GPT-4	18.51
GPT-4	Gemini	21.66
GPT-4	GPT-4	13.88

families) are released by their respective creators under proprietary licenses. In contrast, open-source models are released under various ope-source licenses such as CC BY-NC-SA 4.0, BSD 3-Clause "New" or "Revised" License, etc.

H Additional Evaluation Strategies

As discussed in Sec. 4 of the main paper, we evaluate a model's performance on CinePile 's test-split by computing its accuracy in choosing the correct answer from a set of multiple-choice options. This involves extracting the chosen answer from the model's raw response and mapping it to one of the predefined answer options. While we perform extensive prompt tuning to ensure the model outputs only the option-letter in its response and rigorously post-process responses to separately extract the

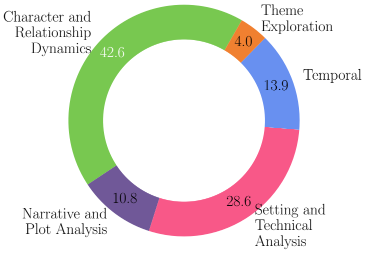


Figure 8: Question category composition in the train split of the dataset.

Table 7: Analyzing raw responses generated by OSS models, scores assigned by our evaluation pipeline, and corresponding failure modes for the movie clip: Area 51 (2015) - Sneaking Onto the Base Scene (4/10)

Question	**Answer Key**	**Model Response**	**Assigned Score**	**Failure Mode**
How does the emotional tone transition during the scene? - A) From despair to hope - B) From fear to acceptance - C) From confusion to understanding - D) From tension to panic - E) From anxiety to excitement	From anxiety to excitement	`“5, 4, 3, 2, 1.\n All right, I think I got it, guys.\n You did it?\n ..... ..... ..... This is the base.”`	0	The model doesn't choose any of the options, rather just repeats all the subtitles.
What action does Darren take after observing the broken sensor? - A) Calls for help - B) Repairs it - C) Panics - D) Ignores it - E) Suggests next steps	Suggests next steps	`"\n ) Darren calls for help.”`	0	While we are able to reliably extract the choice present, it doesn't match the answer key.
Where is Jelena in relation to the helicopter at the end of the scene? - A) Inside - B) In front - C) To the side - D) Directly underneath - E) On it	Directly underneath	`Jelena is standing in front of the helicopter at the end of the scene.`	0	While we are able to reliably extract the choice present, it doesn't match the answer key.
How do Darren and Reed first interact with each other in the scene? - A) They argue - B) They fight - C) They exchange information - D) They greet each other - E) They ignore each other	They exchange information	`In the scene, Darren and Reed are seen walking together, and they exchange information with each other.`	1	We are able to extract the response, and it matches the answer key.

chosen option-letter and the corresponding option-text generated (if generated), there remains a possibility of errors. The model may not always follow these instructions perfectly and could produce verbose responses with unnecessary text snippets, such as "In my opinion," "The correct answer is," or "... is the correct answer."

Therefore, in this section, we compute traditional video-caption evaluation metrics that emphasize the semantic similarity between the answer key text and the raw model response, instead of exact string matching. We focus our evaluation and discussion on open-source models here, as we qualitatively noted that proprietary models, such as GPT-4V, Gemini-Pro, and Claude, strictly adhere to the prompt instructions, producing only the option letter in their response. Specifically, we calculate the following video-captioning metrics – BERTScore (Zhang et al., 2019), CIDEr (Vedantam et al., 2015), and ROUGE-L (Lin, 2004). BERTScore calculates the contextual similarity between the answer key and model response in the embedding space of a pretrained transformer model like BERT-Base. Calculating the similarity between the latent representations, instead of direct string matching, provides robustness to paraphrasing differences in the answer key and model response. In contrast, CIDEr evaluates the degree to which the model response aligns with the consensus of a set

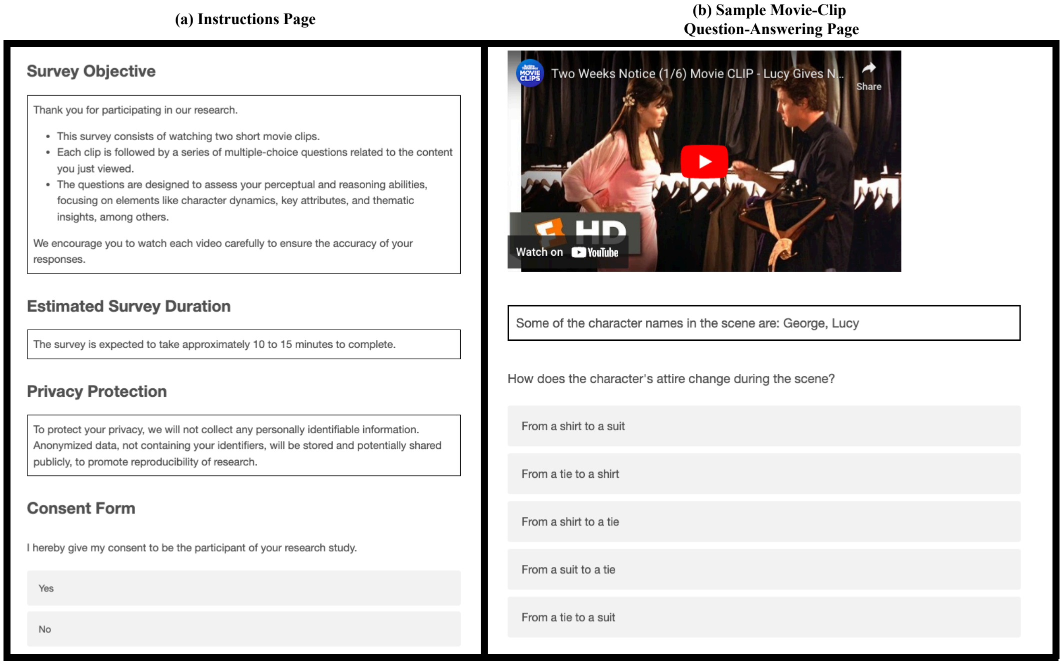
Table 8: Performance of various models on CinePile 's test split, as evaluated using various video captioning metrics – BERTSCoRE (Devlin et al., 2018), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004).

Model	BERTScore↑	CIDEr↑	ROUGE-L↑
mPLUG-Owl Ye et al. (2023a)	0.38	0.74	0.22
Video-ChatGPT Maaz et al. (2023)	0.39	0.63	0.23
Intern-VL-2 (1B) Song et al. (2023)	0.40	1.33	0.28
CogVLM-2 Song et al. (2023)	0.45	1.20	0.31

of reference answer keys. In our setup, each question is associated with only one reference answer. The alignment here is computed by measuring the similarity between the non-trivial n-grams present in the model response and the answer key. Finally, ROUGE-L computes the similarity between the answer key and model response based on their longest common subsequence.

We evaluate four open source models, i.e. mPLUG-Owl, Video-ChatGPT, Intern-VL-2 (1B), and CogVLM2, using the aforementioned metrics and report the results in Table 8. In line with the accuracy trend in the main paper. These findings further support the reliability of our normalization and post-processing steps during accuracy computation.

I Human Study Details



(a) Instructions Page

Survey Objective

Thank you for participating in our research.

- This survey consists of watching two short movie clips.
- Each clip is followed by a series of multiple-choice questions related to the content you just viewed.
- The questions are designed to assess your perceptual and reasoning abilities, focusing on elements like character dynamics, key attributes, and thematic insights, among others.

We encourage you to watch each video carefully to ensure the accuracy of your responses.

Estimated Survey Duration

The survey is expected to take approximately 10 to 15 minutes to complete.

Privacy Protection

To protect your privacy, we will not collect any personally identifiable information. Anonymized data, not containing your identifiers, will be stored and potentially shared publicly, to promote reproducibility of research.

Consent Form

I hereby give my consent to be the participant of your research study.

Yes

No

(b) Sample Movie-Clip Question-Answering Page

Some of the character names in the scene are: George, Lucy

How does the character's attire change during the scene?

From a shirt to a suit

From a tie to a shirt

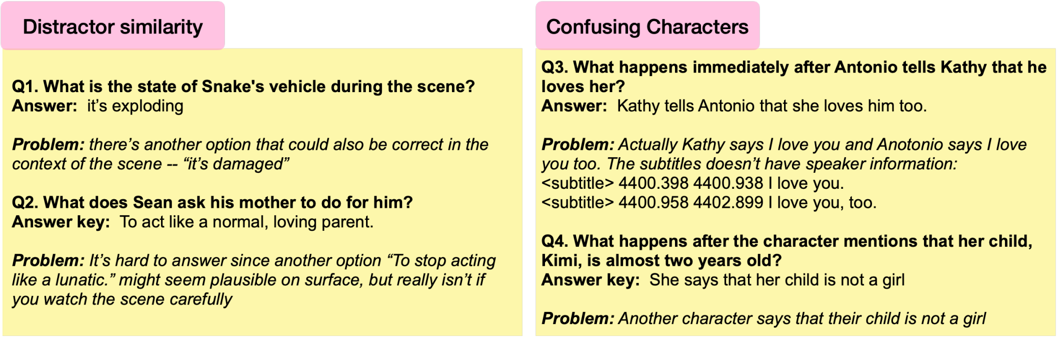
From a shirt to a tie

From a suit to a tie

From a tie to a suit

Figure 9: (*left*) (a) **Instructions Page:** The instructions page at the beginning of the survey, as presented to participants. The participants provide informed consent before viewing any video clip and answering questions. (*right*) (b) **Sample Movie-Clip Question-Answering Page:** An example of one of the movie clips and corresponding question, as presented to the participants. The participants are required to watch the clip and answer the questions by selecting the correct answer choice out of five options.

The authors conducted a small human study with 25 graduate student volunteers to evaluate the quality of the CinePile dataset questions. Each participant answered ten randomly sampled multiple-choice questions about two video clips. Our human study survey was granted an exemption by our institute's Institutional Review Board (IRB), and all participants gave their informed consent before viewing the videos and responding to the questions. For full instructions and consent questions given to participants, please refer to Fig. 9-(a). Additionally, we did not collect any personally identifiable information from the participants. It's important to note that our dataset consists of English movies produced in the United States. These films are likely certified by the



Distractor similarity

Q1. What is the state of Snake's vehicle during the scene?
Answer: it's exploding

Problem: *there's another option that could also be correct in the context of the scene -- "it's damaged"*

Q2. What does Sean ask his mother to do for him?
Answer key: To act like a normal, loving parent.

Problem: *It's hard to answer since another option "To stop acting like a lunatic." might seem plausible on surface, but really isn't if you watch the scene carefully*

Confusing Characters

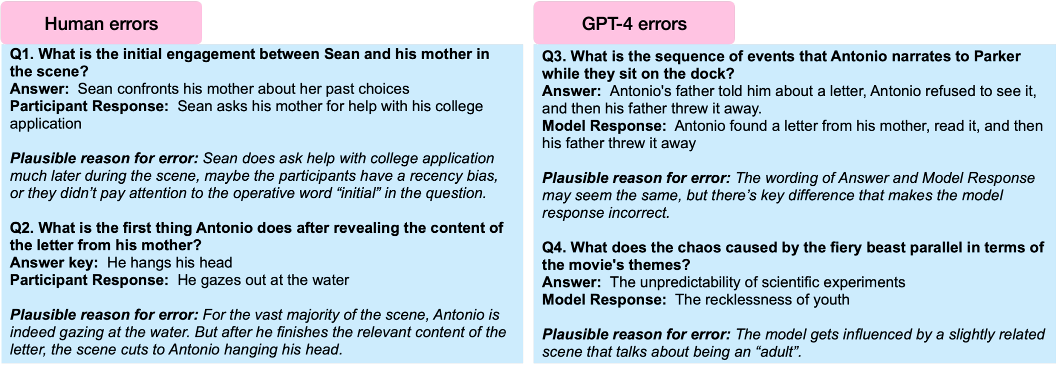
Q3. What happens immediately after Antonio tells Kathy that he loves her?
Answer: Kathy tells Antonio that she loves him too.

Problem: *Actually Kathy says I love you and Anotonio says I love you too. The subtitles doesn't have speaker information:*
<subtitle> 4400.398 4400.938 I love you.
<subtitle> 4400.958 4402.899 I love you, too.

Q4. What happens after the character mentions that her child, Kimi, is almost two years old?
Answer key: She says that her child is not a girl

Problem: *Another character says that their child is not a girl*

Figure 10: **Sample failure cases from human study**: We conducted a human study to check the quality of questions and we found a few systemic issues. We fixed all systemic issues in the final version of the dataset. The movie clip for Q1 can be found here; for Q2, here; for Q3, here; and for Q4, here.



Human errors

Q1. What is the initial engagement between Sean and his mother in the scene?
Answer: Sean confronts his mother about her past choices
Participant Response: Sean asks his mother for help with his college application

Plausible reason for error: *Sean does ask help with college application much later during the scene, maybe the participants have a recency bias, or they didn't pay attention to the operative word "initial" in the question.*

Q2. What is the first thing Antonio does after revealing the content of the letter from his mother?
Answer key: He hangs his head
Participant Response: He gazes out at the water

Plausible reason for error: *For the vast majority of the scene, Antonio is indeed gazing at the water. But after he finishes the relevant content of the letter, the scene cuts to Antonio hanging his head.*

GPT-4 errors

Q3. What is the sequence of events that Antonio narrates to Parker while they sit on the dock?
Answer: Antonio's father told him about a letter, Antonio refused to see it, and then his father threw it away.
Model Response: Antonio found a letter from his mother, read it, and then his father threw it away

Plausible reason for error: *The wording of Answer and Model Response may seem the same, but there's key difference that makes the model response incorrect.*

Q4. What does the chaos caused by the fiery beast parallel in terms of the movie's themes?
Answer: The unpredictability of scientific experiments
Model Response: The recklessness of youth

Plausible reason for error: *The model gets influenced by a slightly related scene that talks about being an "adult".*

Figure 11: **Hard questions according to humans and GPT-4 V**: After conducting the human study, we looked at the questions which human got wrong and the questions which GPT-4 got wrong. Some of these questions are difficult and can only be answered by paying careful attention to the video. The movie clip for Q1 can be found here; for Q2 and Q3, here; and for Q4, here.

Motion Picture Association of America (MPAA), which means they adhere to strict content standards and classification guidelines. As a result, they're expected to contain minimal offensive content. An example of the question-answering page can be found in Fig. 9-(b).

Post the study, we interviewed each participant after the survey to ask if they found any systematic issues in any of the questions they were asked to answer about the video. Later, a panel of authors audited all questions where humans got the answer wrong. We noticed that most of the time when a human got a question wrong it was likely due to one of the following reasons (i) due to their inability to attend over the entire clip at once, (ii) due to their inability to understand the dialogue or understand cultural references (iii) carelessness in answering, as the correct answer was indeed present in the video. We did notice some problematic patterns with a small subset of questions. The main issue is distractor similarity, where humans found two plausible answers and they chose one randomly. We present a few such examples in Fig. 10. We removed the questions from the test set for which we found ambiguous answers.

We again conducted a second human study on the test set's final version, and the human accuracy is 73%. The authors have independently taken the survey, and the corresponding accuracy is 86%. Once again, a careful investigation by a team of authors indicates that even most of these wrong answers are due to human error and confusion over the many events in a scene. We conclude from this study that many of the questions are answerable but difficult. We present the question category-level performance in Sec. 4 in the main paper.

J EXAMPLE DEGENERATE QUESTIONS

As discussed in Section 2.4 of the main paper, most question-answers generated are well-formed and include challenging distractors. However, a small minority are degenerate in that they can be

Table 9: **Example degenerate questions.** Examples of degenerate questions filtered from CinePile. These questions can be categorized as degenerate for various reasons, including: being answerable through common sense (rows one to three) and the models possibly memorizing the movie scripts (rows four and five)

Movie Clip	Degenerate Questions
Scream (1996) - Wrong Answer Scene (2/12) | Movieclips	Question: Where does the conversation between the characters take place? - A) In a restaurant - B) In a car - C) In a classroom - D) At a party - E) Over the phone ✓
The Godfather: Part 3 (8/10) Movie CLIP - Michael Apologizes to Kay (1990) HD	Question: What thematic element is paralleled in the character's dialogue about his past and his destiny? - A) The theme of revenge - B) The theme of fate and free will ✓ - C) The theme of betrayal - D) The theme of lost innocence - E) The theme of love and sacrifice
The Croods (2013) - Try This On For Size Scene (6/10) | Movieclips	Question: What happens right before Grug slips on a banana? - A) Sandy helps Guy hand bananas out to all the monkeys. - B) The saber-toothed cat roars at them from the bottom of a gorge. - C) Grug throws a banana down angrily. ✓ - D) Grug puts up his dukes and so does the monkey. - E) Guy gives Grug a banana.
Rugrats in Paris (2000) - We're Going to France! Scene (1/10) | Movieclips	Question: What event prompts Kira Watanabe to call Mr. Pickles? - A) The robot's destruction of the village. - B) The robot's popularity among the villagers. - C) The malfunction of the giant robot. ✓ - D) The villagers' protest against the robot. - E) The robot's successful performance.
Bottle Rocket (3/8) Movie CLIP - Future Man and Stacy (1996) HD	Question: What happens immediately after Anthony and Dignan finish eating their sandwiches on the patio? - A) Anthony chews a nut. - B) A guy in a brown shirt approaches them. ✓ - C) Stacey Sinclair introduces herself. - D) Anthony tells his story about the beach house. - E) Anthony goes to clean the pool.

answered directly, i.e., without viewing the movie video clip. To automatically filter out such questions, we formulate a degeneracy criterion. If a question can be answered by a wide variety of models without any context—that is, all models select the correct answer merely by processing the question and the five options—we label it as a degenerate question. In this section, we present and discuss some of these degenerate questions in Table 9. We note that a question can be categorized as degenerate due to multiple possible reasons. For instance, consider the questions, "Where does the conversation between the characters take place?", and "What happens right before Grug slips on a banana?". The answer key for these corresponds to the most common-sense response, and the models are able to reliably identify the correct choices ("Over the phone", "Grug angrily throws a banana down") from among the distractions. There's another type of question that models might answer correctly if they've memorized the movie script. For example, the question, "What event prompts Kira Watanabe to call Mr. Pickles?" from the movie Rugrats in Paris, is accurately answered. This likely happens because of the memorization of the script and the distinct character names mentioned in the question.

K Additional Evaluation Results

K.1 Frame Rate Ablation

In this section we perform an ablation to investigate the utility of visual frames (from a model's perspective) by completely remove the visual frames and experiment solely with the provided

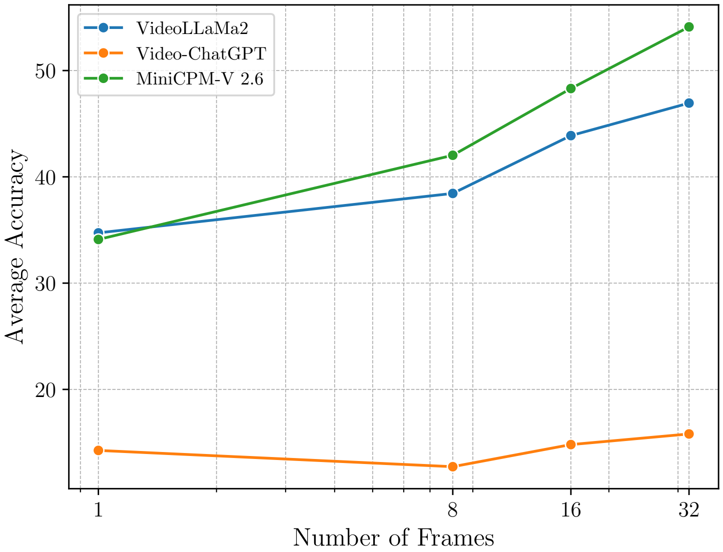


Figure 12: Effect of varying number of samples on overall performance of Video-ChatGPT, VideoLLaMA2, and MiniCPM-V 2.6 on a subset of questions from CinePile.

dialogue when evaluating Video-LLMs. We do exactly this in Table 10, and observe that for all models, except Video-ChatGPT, performance significantly declines when evaluated with "only subtitles." This effect is more pronounced in commercial models compared to open-source ones. It appears that better overall models also tend to utilize visual information more effectively. To further investigate the impact of temporal sampling, we also examine model performance when varying the number of sampled frames: [1,8,16,32] on a subset of CinePile questions and plot the results in Fig. 12. Due to the high cost of running these ablations on closed-source models like Gemini, we focused primarily on open-source models from our earlier experiments, adding a new model, MiniCPM-V 2.6. Our findings show that model performance consistently improves as the number of frames increases, except for Video-ChatGPT, which shows no consistent gains. The improvement is proportional to the model's overall ranking in our benchmarks. MiniCPM-V 2.6 shows the most significant performance gains with additional frames, followed by VideoLLaMa2, while Video-ChatGPT's performance remains relatively unchanged, underscoring its limited reliance on visual inputs.

Table 10: Performance of models with video and subtitles (base case), and when only with subtitles on a subset of CinePile. TEMP - Temporal, CRD - Character and Relationship Dynamics, NPA - Narrative and Plot Analysis, STA - Setting and Technical Analysis, TH - Thematic Exploration.

Model	**Average**	**CRD**	**NPA**	**STA**	**TEMP**	**TH**
Gemini 1.5 Pro	51.72	51.61	56.25	55.45	40.62	50.00
(Only Subtitles) Gemini 1.5 Pro	34.53	35.87	44.44	31.35	32.60	36.36
GPT-4o	50.45	51.14	66.66	52.54	34.78	45.45
(Only Subtitles) GPT-4o	37.23	45.03	44.44	29.66	28.26	45.45
Video-LLaMA2	38.44	45.80	40.74	36.44	19.56	54.54
(Only Subtitles) Video-LLaMA2	33.33	41.22	40.74	27.11	17.39	45.45
Video-ChatGPT	12.92	16.80	3.70	12.82	6.52	20.00
(Only Subtitles) Video-ChatGPT	16.16	22.04	11.53	12.71	13.04	9.09

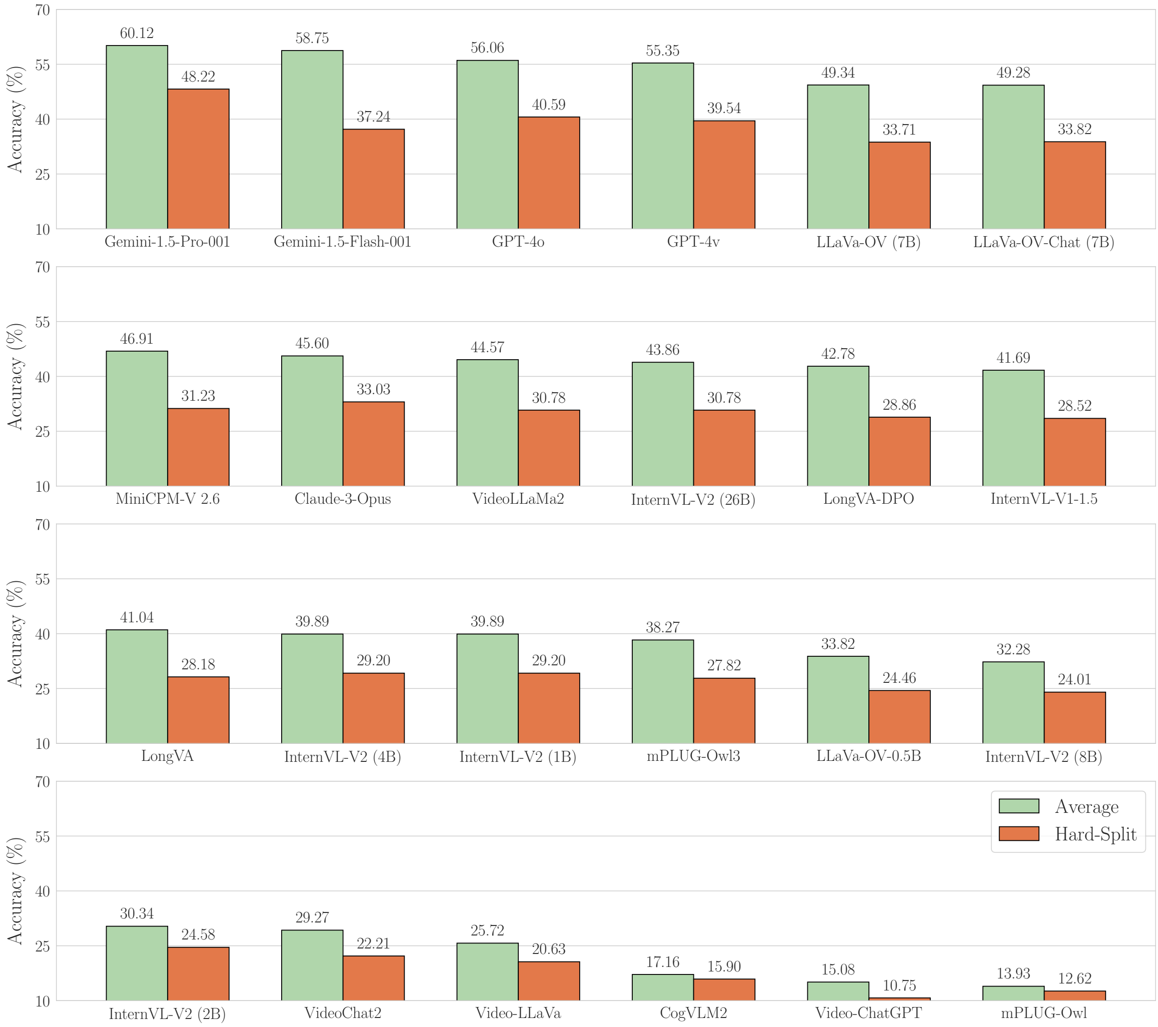


Figure 13: Models' performance on CinePile test split, all questions vs hard questions.

K.2 Performance on Hard-Split

L QA Generation Prompt

As the curator of an advanced cinema analysis quiz, your expertise lies in designing intricate and diverse multiple-choice questions with corresponding answers that span the entire spectrum of film analysis.

- **Objective:** Create diverse and challenging questions based on the film analysis spectrum templates provided below. This spectrum is divided into five subcategories, each comprising several templates. Each template includes a title and a corresponding prototypical question or guideline. Avoid directly replicating the template title and these prototypical questions. Instead, your questions should reflect these elements' essence, even if not explicitly using the category titles in the question's wording.

Mandatory Guidelines:
- **Template Use:** Use the provided question templates as a strict guide, ensuring that your questions are both relevant to the scene and varied in their analytical perspective. The prototype question in each template is for inspiration and should not be copied. Your questions should subtly reflect the prototype's essence, tailored to the specifics of the scene.

- **Sub-Category Balance:** Ensure to generate an equal number of questions from each subcategory. This balance is crucial to cover a wide range of analytical perspectives.

- **Question and Answer Format:**
- **Selected Template:** Indicate the film analysis Sub-Category and corresponding template your question is inspired by, without restricting the question's phrasing to the template's title.

- **Questions:** Limited to one or two lines, formulated to be insightful and not overtly indicative of the answer. Avoid using direct template titles or overly descriptive language that could hint at the correct answer.

- **Answers:** Five options per question, formatted as "**- A**), **- B**), **- C**), **- D**), and **- E**)", concise and reflective of the question's depth.

- **Answer Key:** Specify the correct answer clearly with the formatting, "**Correct Answer:**", in the line following all the answer options.

- **Rationale:** Write a rationale explaining the correctness of the "Answer Key" based on the scene's context in the next line.

Input Information Format:
- Movie scene details will be provided in a structured format comprising two distinct categories, and the relevant scene description. The two categories are as follows:

- **<subtitle>** for character dialogues (to be used only for identifying character presence, not actions or dialogue content).

- **<visual descriptions>** for noting characters' presence, attributes, thematic elements, etc., within the scene.

Movie Scene: {MOVIE_SCENE_TS}

- **Spectrum of Film Analysis with Templates:**
Sub-Category: Character Analysis
{TEMPLATES_CHAR}

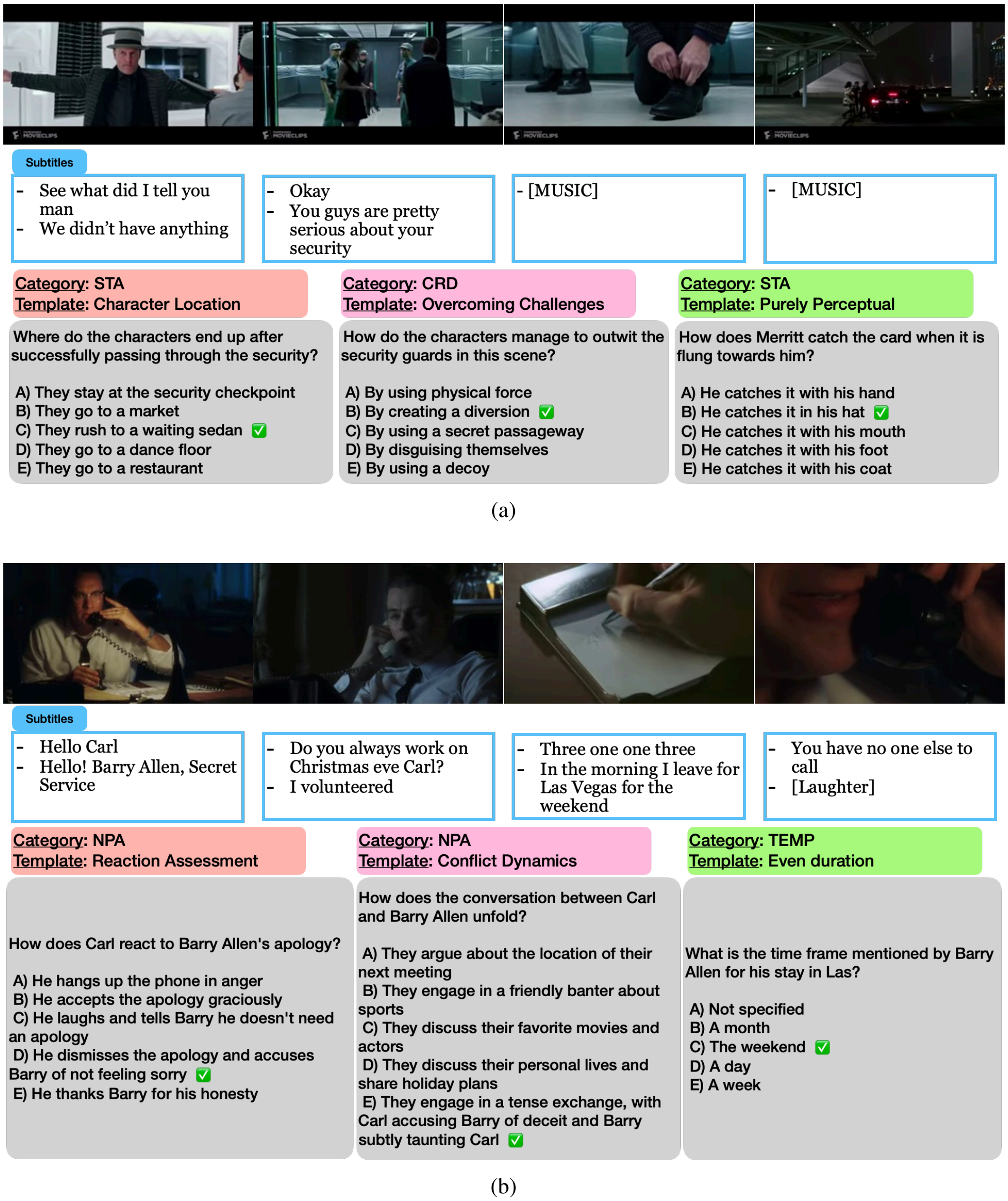
Sub-Category: Narrative Understanding
{TEMPLATES_NARV}

Sub-Category: Scene Setting
{TEMPLATES_SETTING}

Sub-Category: Temporal
{TEMPLATES_TEMPORAL}

Sub-Category: Theme
{TEMPLATES_THEME}

Instructions: Your task is to generate clear, unique, and insightful question-answer pairs strictly following the provided templates. Ensure the distribution of questions covers all subcategories evenly. Strictly avoid using words in the questions that give a strong hint about the answer. You can achieve this by keeping the questions concise and not using too many adjectives or adverbs in the question. Incorrect answers must be plausible and closely mirror the correct answer in length and form. The correct answer should not be deducible solely from the question and/or the wrong answers. After presenting all the options, the correct answer must be distinctly specified, but separate from the list of choices. Additionally, provide a concise rationale about why the question-answer falls into one of the selected templates from the Spectrum of Film Analysis by giving verbatim evidence from the subtitles and/or visual descriptions in the movie scene information.



Subtitles

- See what did I tell you man
- We didn't have anything

- Okay
- You guys are pretty serious about your security

- [MUSIC]

- [MUSIC]

Category: STA
Template: Character Location

Where do the characters end up after successfully passing through the security?

A) They stay at the security checkpoint
B) They go to a market
C) They rush to a waiting sedan ✅
D) They go to a dance floor
E) They go to a restaurant

Category: CRD
Template: Overcoming Challenges

How do the characters manage to outwit the security guards in this scene?

A) By using physical force
B) By creating a diversion ✅
C) By using a secret passageway
D) By disguising themselves
E) By using a decoy

Category: STA
Template: Purely Perceptual

How does Merritt catch the card when it is flung towards him?

A) He catches it with his hand
B) He catches it in his hat ✅
C) He catches it with his mouth
D) He catches it with his foot
E) He catches it with his coat

(a)

Subtitles

- Hello Carl
- Hello! Barry Allen, Secret Service

- Do you always work on Christmas eve Carl?
- I volunteered

- Three one one three
- In the morning I leave for Las Vegas for the weekend

- You have no one else to call
- [Laughter]

Category: NPA
Template: Reaction Assessment

How does Carl react to Barry Allen's apology?

A) He hangs up the phone in anger
B) He accepts the apology graciously
C) He laughs and tells Barry he doesn't need an apology
D) He dismisses the apology and accuses Barry of not feeling sorry ✅
E) He thanks Barry for his honesty

Category: NPA
Template: Conflict Dynamics

How does the conversation between Carl and Barry Allen unfold?

A) They argue about the location of their next meeting
B) They engage in a friendly banter about sports
C) They discuss their favorite movies and actors
D) They discuss their personal lives and share holiday plans
E) They engage in a tense exchange, with Carl accusing Barry of deceit and Barry subtly taunting Carl ✅

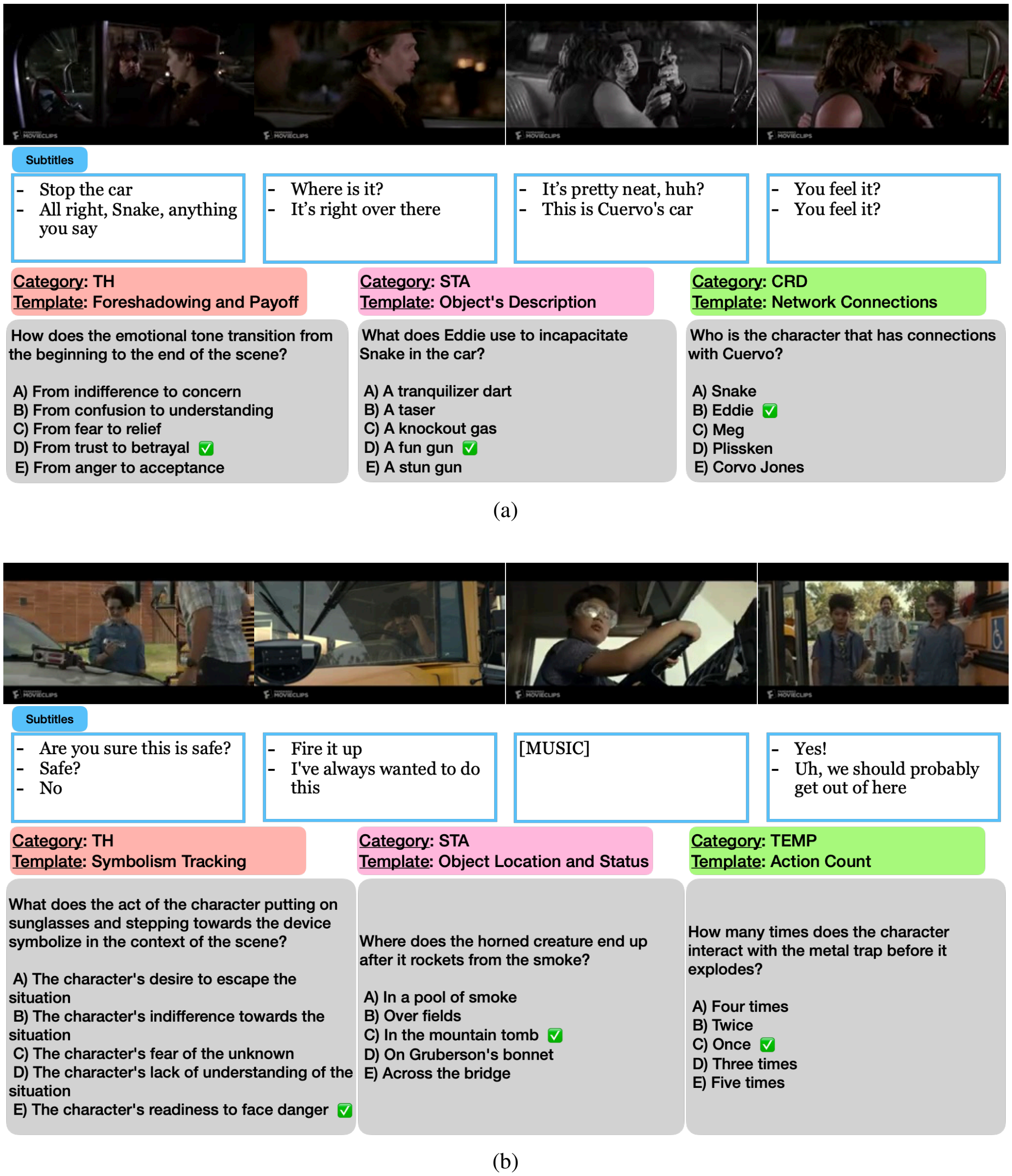
Category: TEMP
Template: Even duration

What is the time frame mentioned by Barry Allen for his stay in Las?

A) Not specified
B) A month
C) The weekend ✅
D) A day
E) A week

(b)

Figure 14: **Example movie clip and multiple-choice questions from CinePile**. The first and second rows depict a selection of image frames extracted from movie clips from (a) Now You See Me 2, and (b) Catch Me if You Can, accompanied by their corresponding subtitles. The next row showcases example questions along with the question template shown in colored headers. TEMP refers to Temporal. Please refer to Table 3 for other category acronyms.



Subtitles

- Stop the car
- All right, Snake, anything you say

- Where is it?
- It's right over there

- It's pretty neat, huh?
- This is Cuervo's car

- You feel it?
- You feel it?

Category: TH
Template: Foreshadowing and Payoff

How does the emotional tone transition from the beginning to the end of the scene?

A) From indifference to concern
B) From confusion to understanding
C) From fear to relief
D) From trust to betrayal ✓
E) From anger to acceptance

Category: STA
Template: Object's Description

What does Eddie use to incapacitate Snake in the car?

A) A tranquilizer dart
B) A taser
C) A knockout gas
D) A fun gun ✓
E) A stun gun

Category: CRD
Template: Network Connections

Who is the character that has connections with Cuervo?

A) Snake
B) Eddie ✓
C) Meg
D) Plissken
E) Corvo Jones

(a)

Subtitles

- Are you sure this is safe?
- Safe?
- No

- Fire it up
- I've always wanted to do this

[MUSIC]

- Yes!
- Uh, we should probably get out of here

Category: TH
Template: Symbolism Tracking

What does the act of the character putting on sunglasses and stepping towards the device symbolize in the context of the scene?

A) The character's desire to escape the situation
B) The character's indifference towards the situation
C) The character's fear of the unknown
D) The character's lack of understanding of the situation
E) The character's readiness to face danger ✓

Category: STA
Template: Object Location and Status

Where does the horned creature end up after it rockets from the smoke?

A) In a pool of smoke
B) Over fields
C) In the mountain tomb ✓
D) On Gruberson's bonnet
E) Across the bridge

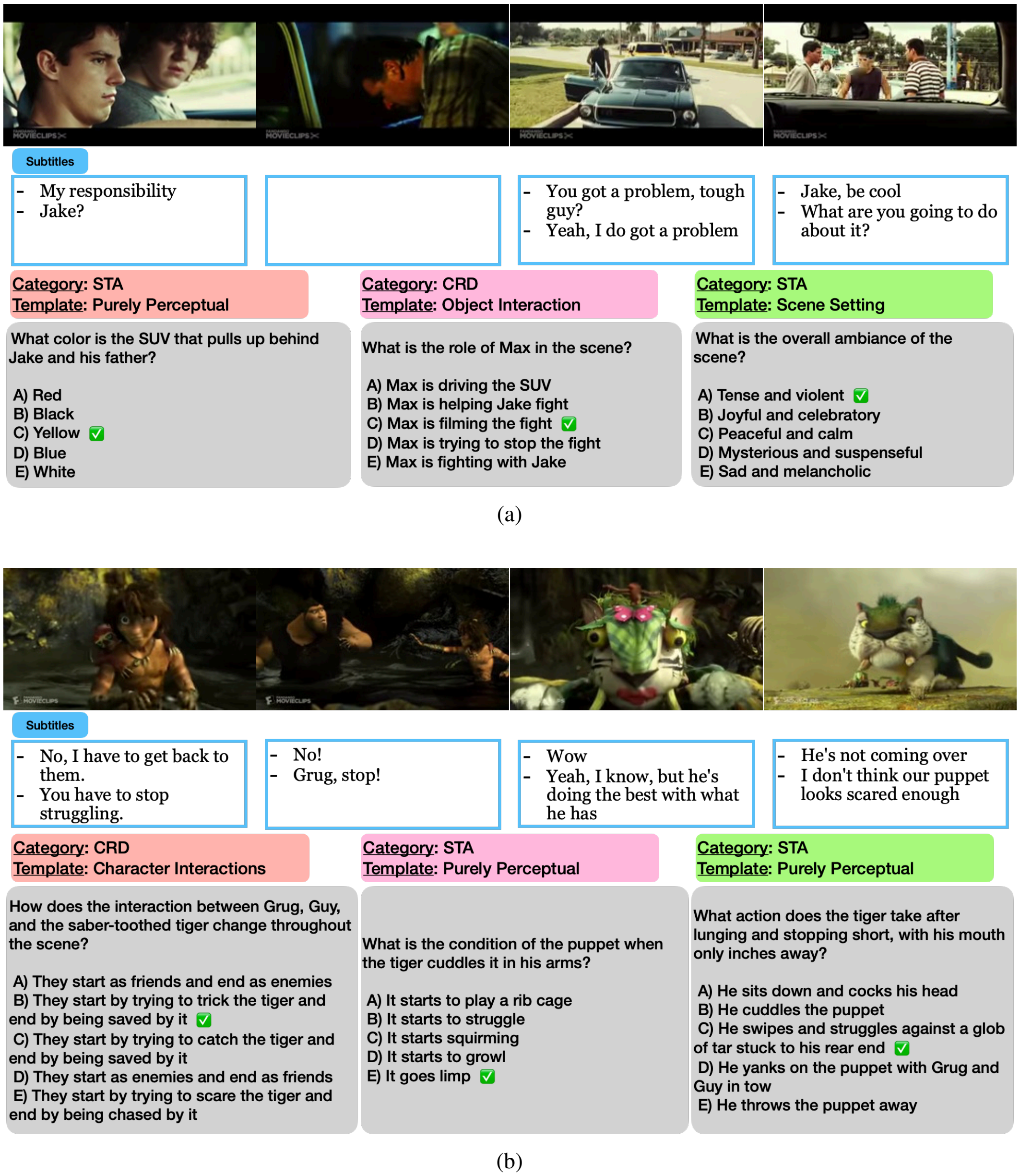
Category: TEMP
Template: Action Count

How many times does the character interact with the metal trap before it explodes?

A) Four times
B) Twice
C) Once ✓
D) Three times
E) Five times

(b)

Figure 15: **Example movie clip and multiple-choice questions from CinePile**. The first and second rows depict a selection of image frames extracted from movie clips from (a)Escape From L.A., and (b)Ghostbusters: Afterlife, accompanied by their corresponding subtitles. The next row showcases example questions along with the question template shown in colored headers. TEMP refers to Temporal. Please refer to Table 3 for other acronyms.



Subtitles

- My responsibility
- Jake?

- You got a problem, tough guy?
- Yeah, I do got a problem

- Jake, be cool
- What are you going to do about it?

Category: STA
Template: Purely Perceptual

What color is the SUV that pulls up behind Jake and his father?

A) Red
B) Black
C) Yellow ✓
D) Blue
E) White

Category: CRD
Template: Object Interaction

What is the role of Max in the scene?

A) Max is driving the SUV
B) Max is helping Jake fight
C) Max is filming the fight ✓
D) Max is trying to stop the fight
E) Max is fighting with Jake

Category: STA
Template: Scene Setting

What is the overall ambiance of the scene?

A) Tense and violent ✓
B) Joyful and celebratory
C) Peaceful and calm
D) Mysterious and suspenseful
E) Sad and melancholic

(a)

Subtitles

- No, I have to get back to them.
- You have to stop struggling.

- No!
- Grug, stop!

- Wow
- Yeah, I know, but he's doing the best with what he has

- He's not coming over
- I don't think our puppet looks scared enough

Category: CRD
Template: Character Interactions

How does the interaction between Grug, Guy, and the saber-toothed tiger change throughout the scene?

A) They start as friends and end as enemies
B) They start by trying to trick the tiger and end by being saved by it ✓
C) They start by trying to catch the tiger and end by being saved by it
D) They start as enemies and end as friends
E) They start by trying to scare the tiger and end by being chased by it

Category: STA
Template: Purely Perceptual

What is the condition of the puppet when the tiger cuddles it in his arms?

A) It starts to play a rib cage
B) It starts to struggle
C) It starts squirming
D) It starts to growl
E) It goes limp ✓

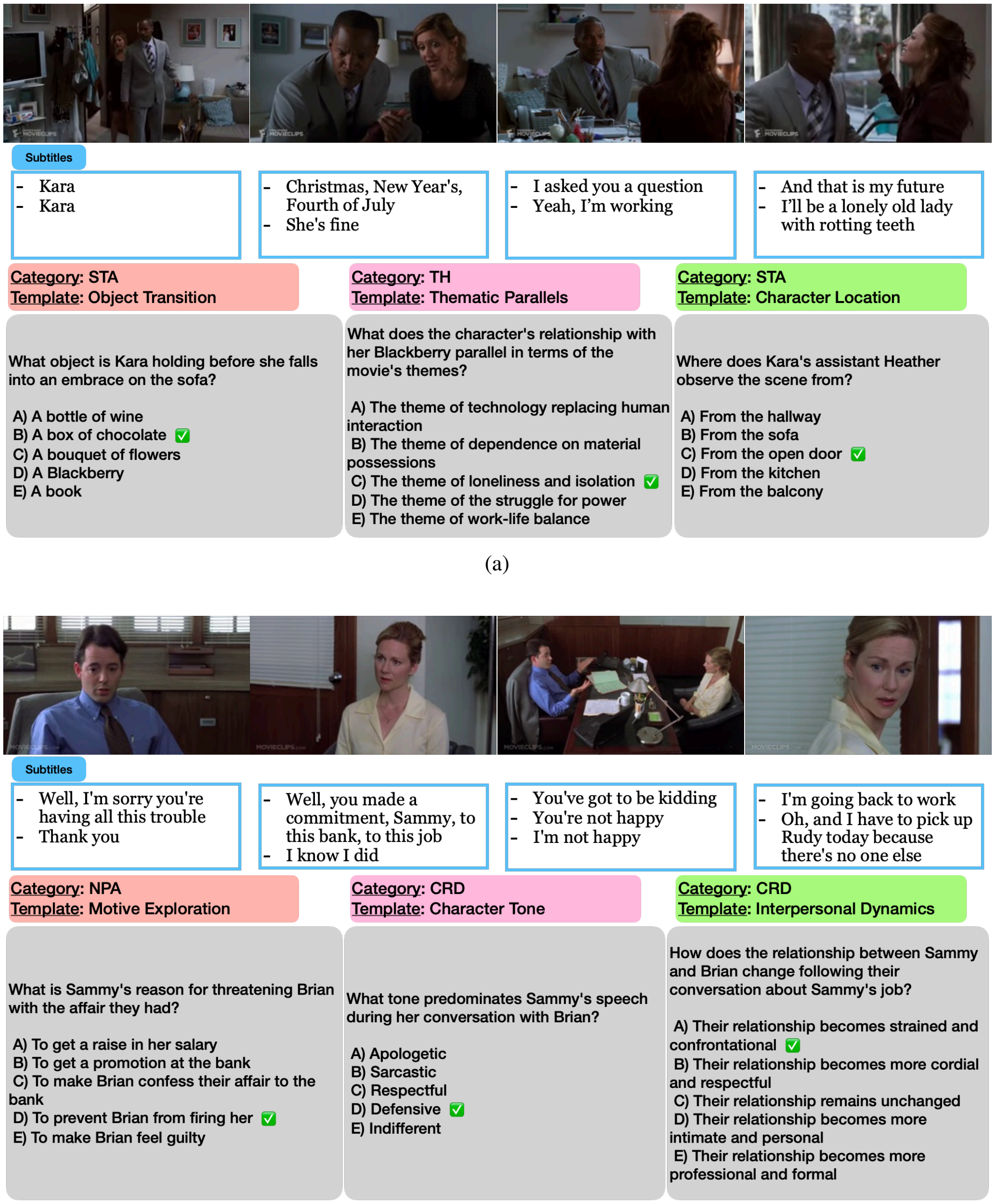
Category: STA
Template: Purely Perceptual

What action does the tiger take after lunging and stopping short, with his mouth only inches away?

A) He sits down and cocks his head
B) He cuddles the puppet
C) He swipes and struggles against a glob of tar stuck to his rear end ✓
D) He yanks on the puppet with Grug and Guy in tow
E) He throws the puppet away

(b)

Figure 16: **Example movie clip and multiple-choice questions from CinePile**. The first and second rows depict a selection of image frames extracted from movie clips from (a) Never Back Down, and (b) The Croods, accompanied by their corresponding subtitles. The next row showcases example questions along with the question template shown in colored headers. TEMP refers to Temporal. Please refer to Table 3 for other acronyms.



- Kara
- Kara

- Christmas, New Year's, Fourth of July
- She's fine

- I asked you a question
- Yeah, I'm working

- And that is my future
- I'll be a lonely old lady with rotting teeth

Category: STA
Template: Object Transition

What object is Kara holding before she falls into an embrace on the sofa?

A) A bottle of wine
B) A box of chocolate ✓
C) A bouquet of flowers
D) A Blackberry
E) A book

Category: TH
Template: Thematic Parallels

What does the character's relationship with her Blackberry parallel in terms of the movie's themes?

A) The theme of technology replacing human interaction
B) The theme of dependence on material possessions
C) The theme of loneliness and isolation ✓
D) The theme of the struggle for power
E) The theme of work-life balance

Category: STA
Template: Character Location

Where does Kara's assistant Heather observe the scene from?

A) From the hallway
B) From the sofa
C) From the open door ✓
D) From the kitchen
E) From the balcony

(a)

- Well, I'm sorry you're having all this trouble
- Thank you

- Well, you made a commitment, Sammy, to this bank, to this job
- I know I did

- You've got to be kidding
- You're not happy
- I'm not happy

- I'm going back to work
- Oh, and I have to pick up Rudy today because there's no one else

Category: NPA
Template: Motive Exploration

What is Sammy's reason for threatening Brian with the affair they had?

A) To get a raise in her salary
B) To get a promotion at the bank
C) To make Brian confess their affair to the bank
D) To prevent Brian from firing her ✓
E) To make Brian feel guilty

Category: CRD
Template: Character Tone

What tone predominates Sammy's speech during her conversation with Brian?

A) Apologetic
B) Sarcastic
C) Respectful
D) Defensive ✓
E) Indifferent

Category: CRD
Template: Interpersonal Dynamics

How does the relationship between Sammy and Brian change following their conversation about Sammy's job?

A) Their relationship becomes strained and confrontational ✓
B) Their relationship becomes more cordial and respectful
C) Their relationship remains unchanged
D) Their relationship becomes more intimate and personal
E) Their relationship becomes more professional and formal

(b)

Figure 17: **Example movie clip and multiple-choice questions from CinePile**. The first and second rows depict a selection of image frames extracted from movie clips from (a) Valentine's Day, and (b) You Can Count on Me, accompanied by their corresponding subtitles. The next row showcases example questions along with the question template shown in colored headers. TEMP refers to Temporal. Please refer to Table 3 for other acronyms.

M Adapting CinePile to Longer and Different Videos

While we primarily focused on ≈ 160 seconds movie clips as the data source for generating question answers in CinePile, as future models with improved temporal resolution get released, they will require even longer and diverse videos for training and evaluation. To meet this need, CinePile was developed not only as a dataset and benchmark but also as a reproducible, scalable, and efficient pipeline for curating long-form video datasets. In this section, we demonstrate this adaptability by experimenting with three longer videos from diverse domains: Survive 100 Days Trapped, Win $500,000 (1620 seconds, YouTube Challenge-Reward), How Hansi Flick's Tactics Are Revolutionizing Barcelona (540 seconds, soccer tactical analysis), and Eminem - Stan (Long Version) ft. Dido (480 seconds, music video). These videos, vastly different from CinePile's movie clips, were transcribed using Whisper, with key visual descriptions annotated by the authors. Additionally, we slightly revised the question generation prompt to reduce the emphasis on general video analysis (e.g., changing "Create diverse and challenging questions based on the film analysis. . . " to "Create diverse and challenging questions based on the video analysis. . . "). We utilized the same question template bank (86 total templates) without adding or removing any. Feeding "video scene information" into our pipeline generated high-quality questions. For instance:

"What are the strong points of conflict between the characters in the video?" (video: *Survive 100 Days Trapped, Win $500,000*)

With options:

- *A)* Hot water running out, disinterest in playing board games, rave at 3 a.m.
- *B)* Hot water running out, disinterest in video games, rave at 3 a.m.
- *C)* Essential food running out, hygiene in the bathroom, snoring at night.
- *D)* Essential food running out, disinterest in video games, hygiene in the bathroom.
- *E)* Essential food running out, disinterest in playing board games, hygiene in the bathroom.

Answering this required analyzing the entire clip to identify key conflicts and select the correct option.

Similarly:

"How does the video develop the theme of Barcelona's tactical variations in attack from start to finish?" (video: *How Hansi Flick's Tactics Are Revolutionizing Barcelona*)

With options:

- *A)* Dynamic-1: utilizing pace of the attacking wingers, Dynamic-2: slowing the tempo with tiki-taka, Dynamic-3: center-back pinning by the center forward.
- *B)* Dynamic-1: counter-attacks using wingers, Dynamic-2: tiki-taka in possession, Dynamic-3: center forward making constant in-behind runs.
- *C)* Dynamic-1: utilizing the depth created by the full back, Dynamic-2: diagonal runs by the midfielders, Dynamic-3: center-back pinning by the center forward.
- *D)* Dynamic-1: inverted full-backs that come into midfield, Dynamic-2: long balls behind for runs by forwards, Dynamic-3: center defensive midfielder dropping into the backline.
- *E)* Dynamic-1: overlapping full-backs, Dynamic-2: center-back dropping into midfield to push the midfielders up, Dynamic-3: wingers constantly swapping wings to confuse the defense.

Answering this involved identifying and mapping out the tactical variations discussed throughout the video.

These examples demonstrate our pipeline's ability to generalize effectively across different video sources and contexts. Additionally, we evaluated several models on questions generated from these longer videos. The results were as follows: Gemini-Pro-1.5: 41.67% accuracy, GPT-4V: 33.33%, GPT-4o: 41.67%, and LLaVa-OV: 33.33%. This shows that the trend in model performance remains similar; however, as expected, there is a substantial drop in performance compared to the 160-second clips.

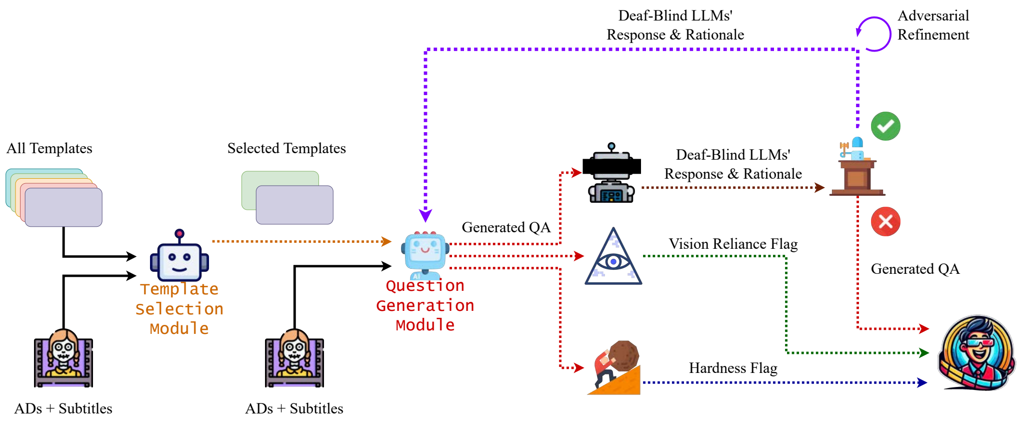


Figure 18: Pipeline demonstrating steps involved in generation, filtration, and refinement of question-answer pairs in CinePile.

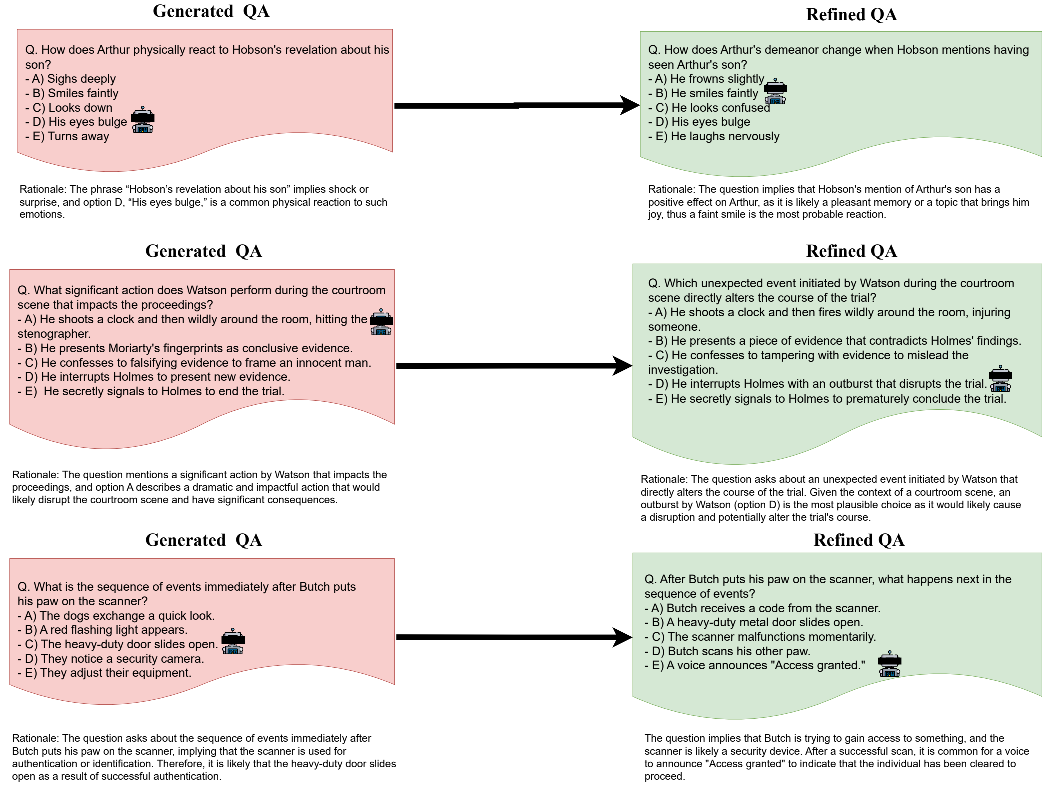
N Additional Adversarial Refinement Details

Adjusting for chance performance: While refining questions in our adversarial refinement pipeline, one concern was that the deaf-blind LLM might only get the right answer by chance. Since our problem involves a multiple-choice QA setup, there is a 25% chance that questions could be answered correctly by a random baseline. Similarly, it was possible that the LLM got the wrong answer due to chance, even though it would be expected to answer correctly the majority of the time. To address this, we devised a methodology where the LLM's response was tested five times using different permutations of the choice order, rotating the options clockwise. We considered the refinement successful only if the LLM failed to answer the question correctly in the majority of cases, i.e., at least three out of five times. If the refinement failed, we repeated the process up to five times, although this is a hyperparameter that can be adjusted based on available computational resources.

Monetary costs for adversarially refining QAs: For adversarial refinement, we use GPT-4o for question rephrasing and the free-tier of LLaMA 3.1 70B API provided by Groq. The cost per question fix is only dependent on rephrasing by GPT-4o, and can be calculated as follows:

- Base prompt (instructions for fixing the question): 709 tokens
- Movie scene (subtitles and visual descriptions): 465 tokens (average; varies across scenes)
- Deaf-blind LLM response and rationale: 102 tokens (average; varies across scenes)
- Total Input Tokens per Attempt: 1,276 tokens
- Cost per Input Token (GPT-4o): \$2.50 per 1M tokens Input Cost per Attempt: $\frac{1{,}276}{1{,}000{,}000} \times 2.50 = \$0.00319$
- Output Tokens: 74 tokens (average)
- Cost per Output Token: \$10.00 per 1M tokens
- Output Cost per Attempt: $\frac{74}{1{,}000{,}000} \times 10.00 = \$0.00074$
- Total Cost per Attempt: $\$0.00319 + \$0.00074 = \$0.00393$
- Number of Attempts per Question Fix: Up to 5 (Upper bound, average ≈ 3)
- Total Cost per Question Fix: $\$0.00393 \times 5 = \$0.01965$

Refined QA Examples: We present a few examples of the weak QAs and the corresponding refined QAs along with the deaf-blind LLM's responses and rationale in Fig. 19.



Generated QA

Q. How does Arthur physically react to Hobson's revelation about his son?
- A) Sighs deeply
- B) Smiles faintly
- C) Looks down
- [✓ D) His eyes bulge]
- E) Turns away

Rationale: The phrase "Hobson's revelation about his son" implies shock or surprise, and option D, "His eyes bulge," is a common physical reaction to such emotions.

Refined QA

Q. How does Arthur's demeanor change when Hobson mentions having seen Arthur's son?
- A) He frowns slightly
- [✓ B) He smiles faintly]
- C) He looks confused
- D) His eyes bulge
- E) He laughs nervously

Rationale: The question implies that Hobson's mention of Arthur's son has a positive effect on Arthur, as it is likely a pleasant memory or a topic that brings him joy, thus a faint smile is the most probable reaction.

Generated QA

Q. What significant action does Watson perform during the courtroom scene that impacts the proceedings?
- [✓ A) He shoots a clock and then wildly around the room, hitting the stenographer.]
- B) He presents Moriarty's fingerprints as conclusive evidence.
- C) He confesses to falsifying evidence to frame an innocent man.
- D) He interrupts Holmes to present new evidence.
- E) He secretly signals to Holmes to end the trial.

Rationale: The question mentions a significant action by Watson that impacts the proceedings, and option A describes a dramatic and impactful action that would likely disrupt the courtroom scene and have significant consequences.

Refined QA

Q. Which unexpected event initiated by Watson during the courtroom scene directly alters the course of the trial?
- A) He shoots a clock and then fires wildly around the room, injuring someone.
- B) He presents a piece of evidence that contradicts Holmes' findings.
- C) He confesses to tampering with evidence to mislead the investigation.
- [✓ D) He interrupts Holmes with an outburst that disrupts the trial.]
- E) He secretly signals to Holmes to prematurely conclude the trial.

Rationale: The question asks about an unexpected event initiated by Watson that directly alters the course of the trial. Given the context of a courtroom scene, an outburst by Watson (option D) is the most plausible choice as it would likely cause a disruption and potentially alter the trial's course.

Generated QA

Q. What is the sequence of events immediately after Butch puts his paw on the scanner?
- A) The dogs exchange a quick look.
- B) A red flashing light appears.
- [✓ C) The heavy-duty door slides open.]
- D) They notice a security camera.
- E) They adjust their equipment.

Rationale: The question asks about the sequence of events immediately after Butch puts his paw on the scanner, implying that the scanner is used for authentication or identification. Therefore, it is likely that the heavy-duty door slides open as a result of successful authentication.

Refined QA

Q. After Butch puts his paw on the scanner, what happens next in the sequence of events?
- A) Butch receives a code from the scanner.
- B) A heavy-duty metal door slides open.
- C) The scanner malfunctions momentarily.
- D) Butch scans his other paw.
- [✓ E) A voice announces "Access granted."]

The question implies that Butch is trying to gain access to something, and the scanner is likely a security device. After a successful scan, it is common for a voice to announce "Access granted" to indicate that the individual has been cleared to proceed.

Figure 19: Examples of the weak QAs and the corresponding refined QAs along with the deaf-blind LLM's responses and rationale

O Additional Dataset Characteristics Details

O.1 Within-Dataset Analysis

Distribution of Dataset Choices. One way models can perform well on multiple-choice-based benchmarks is if the correct answer consistently appears in certain positions within the choice order, allowing the model to leverage this information rather than relying on actual understanding. To address this, we randomized all the choices so that the distribution of correct answer positions is approximately uniform. Specifically, the distribution is: "A" (18.72%), "B" (21.35%), "C" (20.18%), "D" (20.26%), and "E" (19.49%), indicating no significant position bias.

Answer-Distractor Length Similarities. Models can perform well on multiple-choice-based benchmarks if the correct answer consistently differs in its linguistic features from the distractor options. For example, the correct answer may often be longer than the distractors. To investigate this, we conducted quantitative experiments analyzing whether the correct option tends to differ in length. Our findings show that the correct answer is the longest option in only 14.18% of the questions, indicating that this occurs in a minority of cases. Similarly, the correct answer is the shortest option in just 5.14% of the questions, demonstrating that no reverse bias exists either. We plot the word count distributions in Fig. 20 for correct answer and distractor options, and in Fig. 21 for the question, correct answer, and different distractor options. We find that, while there is variation across question categories, the answer and distractor options share similar characteristics within each category and, consequently, overall. On average, correct answers have a length of 4.84 words, while distractor options average 4.59 words.

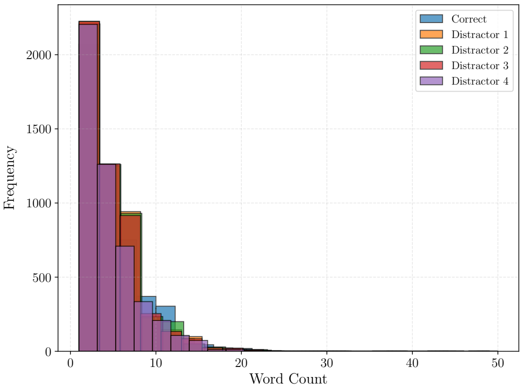


Figure 20: Histograms showing word count distributions for the "correct answer", and the four "distractor" options.

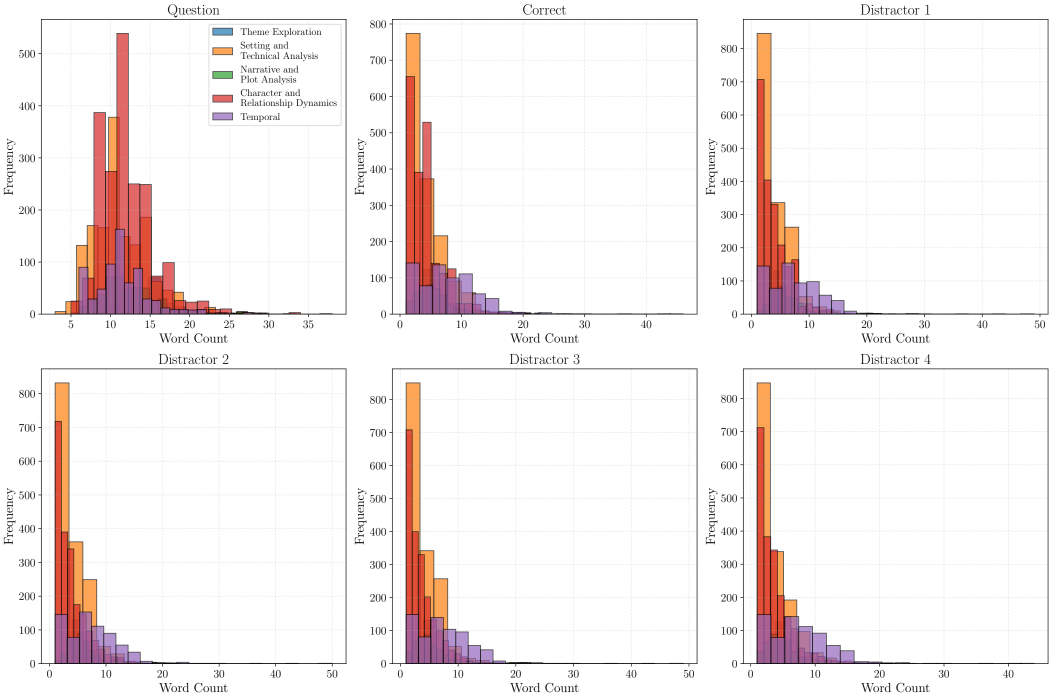


Figure 21: Histograms showing word count distributions for "question", "correct answer", and the four "distractor" options, across different question categories.

O.2 Comparison with Other Datasets

O.2.1 Question Diversity

To ensure that the questions in our dataset capture a wide range of aspects, we take the following steps. Firstly, rather than applying fixed templates for every video, we automatically select relevant ones from a diverse bank of 86 templates tailored to various aspects, such as Character Reaction Insight, Event Sequence Ordering, and Moral Dilemma Exploration. Thus, different videos receive different templates, ensuring diversity across the dataset. Secondly, the question generation process is guided by detailed prompts that incorporate both the chosen template and the specific video clip context. As a result, even when the same template is used, the questions vary significantly based on the unique characters, actions, and environments in each video. For example, the questions “How does the decision to buy the coffee machine and the Harry Potter collection lead to a significant consequence in the video?” and “What early tactical trait of Barcelona hinted at their ultimate attacking strategy?” both stem from the "Causal Chain Analysis" template but differ greatly in wording and focus due to the distinct video contexts. This approach contrasts with other datasets relying on human annotators, which often limit template categories (e.g., Perception Test uses four template areas) for human labeling feasibility.

To quantify question diversity, we conducted an experiment to measure the average semantic diversity of questions both within a video clip and across different video clips in our dataset.

Within-Video Diversity

For a video clip v_i, assume it has j questions $\{q_{i1}, q_{i2}, \ldots, q_{ij}\}$. Using an embedding model, we encoded each question into the embedding space and measured their semantic similarity using cosine similarity $\text{cosim}(q_{ik}, q_{il})$ for all pairs where $1 \leq k, l \leq j$ and $k \neq l$. Since question diversity is inversely related to similarity, we computed the pairwise cosine distance as $1 - \text{cosim}(q_{ik}, q_{il})$. The within-video diversity score for a clip v_i is then given by the expected pairwise cosine distance:

$$D_{\text{within}}(v_i) = \mathbb{E}_{q_{ik}, q_{il} \sim v_i} \left[1 - \text{cosim}(q_{ik}, q_{il})\right]$$

We aggregated this across the dataset by sampling clips $v_i \sim \mathcal{D}$, where $\mathcal{D}$ represents the distribution of video clips in CinePile:

$$D_{\text{within}} = \mathbb{E}_{v_i \sim \mathcal{D}} \left[D_{\text{within}}(v_i)\right]$$

Across-Video Diversity:

To measure diversity across different video clips, we considered the pairwise cosine distances between questions from different videos. For two different video clips v_i and v_j ($i \neq j$), with their associated questions $\{q_{ik}\}$ and $\{q_{jl}\}$, we computed:

$$1 - \text{cosim}(q_{ik}, q_{jl})$$

The across-video diversity score is given by the expected pairwise cosine distance between questions from different videos:

$$D_{\text{across}} = \mathbb{E}_{v_i, v_j \sim \mathcal{D}} \left[\mathbb{E}_{q_{ik} \sim v_i, q_{jl} \sim v_j} \left[1 - \text{cosim}(q_{ik}, q_{jl})\right]\right], \quad i \neq j$$

Combined Diversity Score:

To obtain an overall measure of diversity, we computed the harmonic mean of the within-video and across-video diversity scores:

$$\text{Diversity Score} = 2 \times \frac{D_{\text{within}} \times D_{\text{across}}}{D_{\text{within}} + D_{\text{across}}}$$

The harmonic mean is appropriate in this context because it balances both aspects of diversity by emphasizing the smaller of the two values, and ensuring that neither within-video nor across-video diversity disproportionately influences the combined score. We compute the diversity score on 50 randomly sampled video clips, and share the results in the table below. CinePile achieves a diversity score of 0.45. For context, we computed the same metric on other datasets: Video-MME: 0.45, MV-Bench 0.42, and IntentQA: 0.37. These comparisons demonstrate the strong semantic diversity of questions in CinePile that is greater or on-par with other (even purely human-curated) datasets.

Table 11: Diversity analysis across datasets based on Within-Video Diversity, Across-Video Diversity, and overall Diversity-Score.

Dataset	Within-Video Diversity	Across-Video Diversity	Diversity-Score
CinePile	0.55	0.38	0.45
Video-MME	0.53	0.40	0.45
MVBench	0.57	0.33	0.42
IntentQA	0.45	0.32	0.37

O.2.2 Model Ranking Correlations

In this subsection, we compute the Spearman rank correlation (ρ) between model ranks on CinePile and their ranks on other datasets, including Video-MME, MV-Bench, and EgoSchema. For each dataset, we use the model ranks provided in their official publications and calculate correlations based on the ranks of models common to both CinePile and the respective dataset. Our results show strong correlations: $\rho = 0.964$ for Video-MME (7 common models, i.e., Gemini 1.5 Pro-001, GPT-4o, Gemini 1.5 Flash-001, GPT-4 Vision, Intern VL-V1.5-25.5, VideoChat2-7B, Video LLaVa-7B), $\rho = 1.000$ for MV-Bench (3 common models, i.e., VideoChat2, Video-ChatGPT-7B, mPLUG-Owl), and $\rho = 1.000$ for EgoSchema (2 common models, i.e., mPLUG-Ow, InternVideo). While CinePile evaluates 26 state-of-the-art models, the number of models evaluated by other benchmarks is often smaller, with limited overlap. For example, MV-Bench assesses only 6 models, of which 3 overlap with CinePile, making some correlations less robust. However, these strong correlations suggest that models performing well on CinePile also perform well on manually curated benchmarks, underscoring CinePile's validity as a reliable test set. That said, performance levels naturally vary due to differences in dataset characteristics and task difficulty. For instance, Gemini-1.5 Pro achieves 81.3% on Video-MME but only 60% on CinePile, highlighting the unique challenges CinePile presents.

P Open-Source Failure Modes

We had previously discussed one of the reasons for why are (some) OSS models so far behind in Sec. 4 of the main paper, where we found that, for extremely poorly performing models (sub 20% overall performance), it was partly due to their inability to follow instructions as we both qualitatively and quantitatively discussed such failure cases in Fig. 7a in the main paper and Appendix Sec. H (Tab. 8). In this section, we discuss a few additional failure modes of open-source models.

Does Scale (In Parameter Space) Alone Lead to Better Performance? There is a lot of focus on model scale these days, so we were curious whether scale alone can lead to better performance (ignoring the architecture, training data, etc). So we computed the Pearson-r correlation between the model scale and overall performance and found it to be weakly positively correlated i.e., 0.157. Obviously, there are alot of confounders across different models like different training data, architecture, etc, so this is not definitely saying that scale would not improve significantly performance, rather it alone is not enough. If we control for everything else by only analyzing one particular model family i.e., InternVL, we see a positive correlation of 0.72.

Poor ability to utilize visual information; and overdependence on LLM-priors Another possible reason for the performance gap in open-source models could be their weaker reliance on visual information and over-reliance on language priors (Tong et al., 2024; Lin et al., 2023). In our experiments (see Appendix Sec. K.1) examining the effect of model performance on the number of sampled frames, we observe that while models improve with additional frames, the extent of this improvement correlates with the model's overall performance. Specifically, better-performing models tend to utilize visual information more effectively, showing greater performance gains with more frames, whereas weaker models exhibit minimal to no improvement.

Gap with closed-source models The performance advantage of closed-source models likely stems from a combination of factors rather than a single artifact. State-of-the-art models like Gemini-1.5-Pro and GPT-4o operate at scales of hundreds of billions of parameters, significantly outpacing the 7B-26B parameter range of the best open-source models we evaluated. Additionally, while these closed-source models do not disclose details about their training data mixtures or the GPU hours spent, it is reasonable to assume they adhere to scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) and are trained on datasets that are substantially larger and more diverse than those available to open-source models. The lack of transparency from closed-source models also means there are no ablation studies to pinpoint the optimal combinations of data mixtures or architectural choices contributing to their performance. This makes it challenging to draw precise comparisons.Despite these gaps, open-source models are rapidly catching up, with only about a $\approx$ 10% performance difference in our evaluations. We are optimistic that this gap will continue to shrink in the coming months, and CinePile's training set can be helpful in advancing the capabilities of open-source models.