CinePile: A Long Video Question Answering Dataset and Benchmark

Supplementary Material

A Additional movie clip & questions examples

We present a few examples from our dataset in figs. 8a, 8b, 9a, 9b, 10a, 10b, 11a and 11b.

B Related Work

Video understanding and question answering. LVU Wu and Krähenbühl [2021], despite being one of the early datasets proposed for long video understanding, barely addresses the problem of video understanding as the main tasks addressed in this dataset are year, genre classification or like ratio prediction. A single frame might suffice to answer the questions and these tasks cannot be considered quite as "understanding" tasks. MovieQA Tapaswi et al. [2016] is one of the first attempts to create a truly understanding QA dataset, where the questions are based on entire plot the movie but not localized to a single scene. On closer examination, very few questions are vision focused and most of them can be answered just based on dialogue. EgoSchema Mangalam et al. [2024] is one of the recent benchmarks, focused on video understanding which requires processing long enough segments in the video to be able to answer the questions. However, the videos are based on egocentric videos and hence the questions mostly require perceptual knowledge, rather than multimodal reasoning. Another recent benchmark, Perception Test Patraucean et al. [2024], focuses on core perception skills, such as memory and abstraction, across various reasoning abilities (e.g., descriptive, predictive, etc) for short-form videos. MAD Soldan et al. [2022] dataset contains subtitles and visual descriptions for full-length movies and is typically used in scene captioning task rather than understanding. Another issue is this dataset does not provide raw visual data, they share only `[CLS]` token embeddings, which makes it hard to use. TVQA Lei et al. [2018] is QA dataset based on short 1-min clips from famous TV shows. The annotators are instructed to ask What/How/Why sort of questions combining two or more events in the video. MoVQA Zhang et al. [2023b] manually curates questions across levels multiple levels—single scene, multiple scenes, full movie— by guiding annotators to develop queries in predefined categories like Information Processing, Temporal Perception, etc. Long video understanding datasets, such as EpicKitchens Damen et al. [2018], tend to concentrate heavily on tasks related to the memory of visual representations, rather than on reasoning skills. While these benchmarks are valuable for gauging the extent of visual representation captured by a model, they fall short in providing insights into video understanding. These datasets mainly test a model's ability to recall and recognize visual elements but do not adequately assess its capability to reason and interpret the context and narrative of videos.

CinePile differs from all the above datasets that have much longer videos and we ask many questions per video to capture the perceptual, temporal, and reasoning aspects of the video. And it is truly multimodal where the person has to watch the video as well as audio/subtitles to answer many questions. Unlike the previous datasets with fixed templates, we automated this process on previously human-generated questions, this let us capture many more question categories compared to previous works. Lastly, our approach to dataset generation is scalable and hence it is easy for anyone to extend our dataset.

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. [2023b], 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

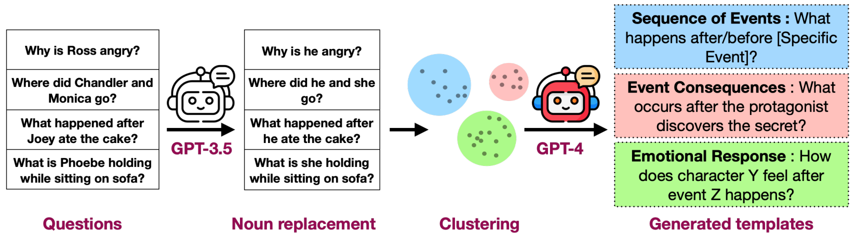


Figure 2: **Question template generation pipeline**: We begin by substituting the first names in human-written source questions and then cluster them. We then feed a selection of questions from each cluster into GPT-4 which in turn outputs "question templates" which are used in the next stage of dataset creation. See methods section for more details.

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 Data Collection Details

Scene localization in AD. The video clips we have gathered are typically 2-3 minutes long, while Audio Descriptions (ADs) cover entire movies. To align descriptions with video, we transcribe the audio from both the YouTube clip and the AD file using an Automatic Speech Recognition (ASR) system Whisper Radford et al. [2023]. More specifically, we use `WhisperX` Bain et al. [2023], an enhanced version of Whisper designed to offer quicker inference and more precise word-level timestamps. We then match the first 3 and last 3 lines of the transcription to the dialog interleaved in the AD files. We do the matching using a sentence embedding model, `WhereIsAI/UAE-Large-V1`, which allows accurate alignment even in cases where there are slight differences between the transcriptions and the dialog. We then extract all AD data that lives between the matched start and end of the clip. For the rest of the paper, we will refer to the human-written description of the scene as "visual description" and the speaking or dialogue part of the video as "dialogue". When combined, we will refer to both data sources as "**annotated-scene-text**".

Sentence classification. In every AD file, we have text data of both dialogue and visual descriptions. As we aim to develop a category of perceptual-focused questions based solely on visual description data, we do not want to provide dialog to the model at test time. To categorize each sentence as either visual or dialog, we fine-tuned a BERT-Base model Devlin et al. [2018] using annotations from the MAD dataset Soldan et al. [2022], which contains labels indicating whether a sentence is a dialogue or a visual description. We applied a binary classification head for this task. For training this model, we split the MAD dataset annotations into an 80-20 training-evaluation split. The model achieves 96% accuracy in 3 epochs. Qualitatively, we observed that the model accurately classifies sentences in the data we curated, distinguishing effectively between dialogue and visual description content.

Finally, we also augment the hand-written descriptions with visual scene descriptions obtained by feeding key frames to the Gemini API. This ensures a lot of visual information is present, even for scenes for which ADs are lacking details. See the Sec. F in Appendix for details.

D Additional Automated Question Template Details

During early experimentation, we found that providing a range of templates to a VLM helped it create more detailed, diverse, and well-formed questions, so we decided to use a template-based approach for question generation. Rather than confining questions to a few predefined themes, we propose a method to create question templates naturally on top of human-generated questions. We illustrate our automated question template generation pipeline in fig. 2. Our starting point is approximately 30,000 human-curated questions from the MovieQA Tapaswi et al. [2016], TVQA Lei et al. [2018], and Perception Test Patraucean et al. [2024] datasets. We cluster these questions, select a few representatives per cluster, and then use GPT-4 to discern the underlying themes and write a prompt. The whole pipeline is illustrated in fig. 2. First, we preprocess the questions by replacing first names and entities with pronouns, as BERT Reimers and Gurevych [2019] embeddings tend to create clusters with shared names rather than themes. For instance, 'Why is Rachel hiding in the bedroom?' is altered to 'Why is she hiding in the bedroom?'. We used GPT-3.5 to do this replacement, as it handled noun replacement better than open-source and other API alternatives. The modified questions are then embedded using `WhereIsAI/UAE-Large-V1`, a semantic textual similarity model that ranks among the top performers on the MTEB leaderboard[1]. Once the first names are replaced, we observed significant repetition among questions, leading us to deduplicate them, ultimately resulting in 17,575 unique questions. We then perform k-means clustering to categorize the questions into distinct clusters. We experimented with different values of $k = 10, 50, 100$. Qualitatively, we found $k = 50$ to be an optimal number of clusters where the clusters are diverse and at the same time clusters are not too specific. For example, we see a 'high-school dance' specific question cluster when $k = 100$, and these questions are merged into an 'event' cluster when we reduce the number of clusters to 50. The Perception Test questions are less diverse as humans were given a small number of templates, so we used $k = 20$ for this set.

The number of questions in each cluster varied, with counts ranging from 60 to 450. We selected 10 random questions from each, and used them to prompt GPT-4 to create relevant question templates, as illustrated in fig. 3 in the Appendix. This created more general templates than merely using the 10 closest to the cluster center.

We generated four templates for each question cluster, resulting in around 300 templates across three datasets. We then manually reviewed all 300 templates, eliminating those that were overly specific and merging similar ones. overly specific templates included questions like "**Pre-wedding Dilemmas:** What complicates character Z's plans to propose marriage to their partner?" and "**Crime and Consequence:** What is the consequence of the character's criminal actions?". This process resulted in 86 unique templates. Following that, we manually categorized these into four high-level categories: Character and Relationship Dynamics, Narrative and Plot Analysis, Thematic Exploration and Setting, and Technical Analysis.

In the next section, we discuss the definitions and present prototypical questions from each of these categories.

E Question Template Category Examples

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.

We present two question templates per category and their prototypical questions in Tab. 2.

[1] https://huggingface.co/spaces/mteb/leaderboard

Question Template Automation

Generate a few templates based on the following questions

`[Question 1]` Where did the couple meet?
`[Question 2]` What holiday is coming up when they meet?
`[Question 3]` What made him break with her?

Response Based on the nature of these questions, I can identify a few underlying templates that encompass most of them:

`[Template 1]` Catalytic Actions
`[Proto Question 1]` What event acts as a catalyst for the character's next major decision?

`[Template 2]` Setting and Context
`[Proto Question 2]` Where does this interaction take place, and how does the location impact the conversation?
...

Figure 3: **Extracting templates from human-generated questions.** We share 10 samples from each question cluster, and prompt an LLM to create a few templates and a prototypical question. See methods section for more details.

Table 2: 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]?
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?

F 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 works Wang et al. [2023], Zhang et al. [2023a]. First, we use a shot detection algorithm to pick the important frames[2], 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

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

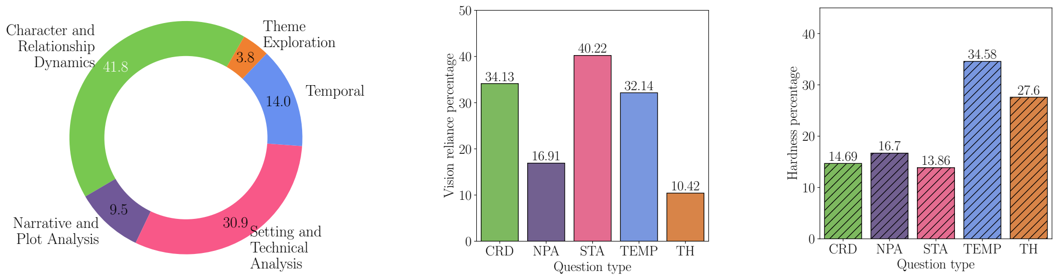


Figure 4: **Left:** Question category composition in the dataset. **Middle:** Percentage of vision-reliant questions across categories. **Right:** Percentage of hard questions per question category type. TEMP refers to Temporal. Please refer to table 2 for other acronyms. The colors correspond to the same categories across the plots.

sentences. Hence all the MCQs generated using this summary are added only to the training split but not to the eval split.

Table 3: Comparing our dataset, CinePile against the existing video-QA datasets. Our dataset is both large and diverse. Multimodal refers to whether both the video and audio data is used for question creation and answering.

Dataset	**Annotation**	**Num QA**	**Avg sec**	**Multimodal**	**QA Type**			
					Temporal	Attribute	Narrative	Theme
TGIF-QA Jang et al. [2017]	Auto	165,165	3	✗	✓	✗	✗	✗
MSRVTT-QA Xu et al. [2017]	Auto	243,690	15	✗	✗	✓	✗	✗
How2QA Li et al. [2020]	Human	44,007	60	✗	✓	✓	✗	✗
NExT-QA Xiao et al. [2021]	Human	52,044	44	✗	✓	✓	✗	✗
EgoSchema Mangalam et al. [2024]	Auto	5,000	180	✗	✓	✓	✓	✗
MovieQA Tapaswi et al. [2016]	Human	6,462	203	✓	✓	✓	✓	✗
TVQA Lei et al. [2018]	Human	152,545	76	✓	✓	✓	✓	✗
Perception Test Patraucean et al. [2024]	Human	44,000	23	✓	✓	✓	✗	✗
MoVQA Zhang et al. [2023b]	Human	21,953	992	✓	✓	✓	✓	✗
CinePile (Ours)	Human + Auto	200,000	160	✓	✓	✓	✓	✓

G Additional Dataset Statistics

In the initial phase of our dataset collection, we collected ∼15,000 movie clips from channels like MovieClips on YouTube. We filtered out clips that did not have corresponding Audiovault recordings. We also excluded clips with low alignment scores when comparing the YouTube clip's transcription with the localized scene's transcription in the Audio Description (AD) file as discussed in methods section. This resulted in a refined dataset of ∼8,000 movie clips. Our dataset's **average video length is ∼160 sec**, significantly longer than many other VideoQA datasets and benchmarks.

We split 8000 videos into train and test splits of 7700 and 300 videos each. We made sure both the splits and the sampling preserved the dataset's diversity in terms of movie genres and release years. We follow the question-answer generation and filtering pipeline which was thoroughly outlined in section 2. We ended up with **200,000 training points and 7,800 test-set points**. Each MCQ contains a question, answer, and four distractors. As a post hoc step, we randomized the position of the correct answer among the distractors for every question, thus eliminating any positional bias. We filtered out the degenerate questions from the test split, however, we left them in the train set, since those questions are harmless and might even teach smaller models some helpful biases the larger multimodal models like Gemini might inherently possess. Our dataset is large and varied because we used a wide variety of movie clips and different prompting strategies about diverse question types. Each strategy zeroes in on particular aspects of the movie content. We present 1 scene and example MCQs from different question templates in fig. 1. In fig. 4 (Left), we provide a visual breakdown of the various categories of questions generated in our dataset. A significant portion of the questions falls under "Character Relationship Dynamics". This is attributed to the fact that a large number of our automated question templates, which were derived from human-written questions, are categorized here. Following this, we have “Setting and Technical Analysis” questions, which predominantly require visual interpretation. We display the metrics for vision reliance and question hardness, as discussed in methods section in the main paper, at the category level in fig. 4 (Middle, Right). As anticipated, questions in the “Setting and Technical Analysis” category exhibit the highest dependency on visual elements, followed by those in the “Temporal” category. In terms of the hardness metric, the “Temporal” category contains the most challenging questions, with “Character Relationship Dynamics” following closely behind. Finally, we compare our dataset with other existing datasets in this field in table 3, showing its superiority in both the number of questions and average video length compared to its counterparts.

Test split. As mentioned previously, our test split comprises 300 video clips, derived from movies distinct between training and testing to avoid information leakage. Additionally, we have eliminated all degenerate questions from this split, which constituted 4.5% of the generated questions. Following several rounds of manual cleanup and thorough testing, our final count stands at 5,500 questions. Of all the test questions, 34.30% are reliant on visual information.

H Post-processing for Accuracy Evaluation

Our evaluation method incorporates a two-stage process to address these variations. In the first stage, we employ a normalization function to parse the model’s response, extracting the option letter (A-E) and the accompanying option text if present. This normalization handles various response formats, such as direct letter responses (e.g., “A”) or more verbose forms (e.g., “Answer: D, The Eiffel Tower”), ensuring that both the option letter and text are accurately identified, if present. Following normalization, the second-stage entails comparing the normalized model response with the correct answer key. This comparison involves checking for both the option letter and text in the model response. If both elements are present and match the answer key, a score of one is awarded. However, if only the option letter or text appears, the comparison is limited to the relevant part, and the score is assigned accordingly.

Table 4: Performance of various open source video-LLMs 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]	0.46	1.36	0.31
Video-ChatGPT Maaz et al. [2023]	0.48	1.14	0.32
MovieChat Song et al. [2023]	0.40	0.66	0.18

I Evaluation on Captioning Metrics

As discussed in Sec. 3 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 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

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 5: **Hard questions according to humans and GPT-4**: 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.

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 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 three open source models, i.e. mPLUG-Owl, Video-ChatGPT, and MovieChat using the aforementioned metrics and report the results in Table 4. In line with the accuracy trend in the main paper, we observe that while mPLUG-Owl and Video-ChatGPT don't have a substantial difference in their performance, they both significantly outperform MovieChat. These findings further support the reliability of our normalization and post-processing steps during accuracy computation, and underscore the need for improving open-source video-LLMs to close the gap with proprietary models.

J 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 5 and Table 6. 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 (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- 5, 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 6, Gemini misremembers the temporal sequence and selects a wrong option as the answer choice for the temporal category.

K Example Degenerate Questions

As discussed in methods section 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 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 7. 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

Table 5: 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 2 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

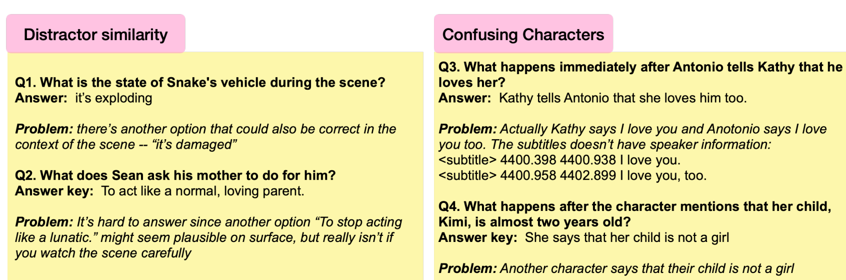


Figure 6: **Failure cases found in human study**: Example of systemic issue identified and fixed post the human study. The movie clip for Q1 can be found here; for Q2, here; for Q3, here; and for Q4, here.

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.

Table 6: 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 2 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.

L Human Study Details

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. 7-(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 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. 7-(b). Each participant (graduate student volunteers) answered 10 questions about two different randomly chosen videos. We randomly sampled the 10 questions from the list of all the questions we generated for the scene. We had 25 participants excluding authors. Human performance was approximately 60%. We also 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. We present a few of the question-answers that

Table 7: **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.

humans got wrong and GPT-4 got wrong and plausible reasons for errors in Fig. 5. 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 their inability to attend over the entire clip at once, as the correct answer was indeed present in the video. We did notice some problematic patterns with a small subset of questions. For example, due to the misalignment of AD with a few clips, some questions were created about events that happened before or after the clip, making the question unanswerable to the participant. This alignment issue was later fixed and the dataset was repaired. Another issue is distractor similarity, where humans found two plausible answers and they chose one randomly. We present a few such examples in fig. 6. We removed questions from the test set with ambiguous answers. We also found that a majority of such questions were written by Gemini, and as a result we decided to use only GPT-4 questions in the test set. We conducted a second human study on the test set's final version, and the new 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 section 3.

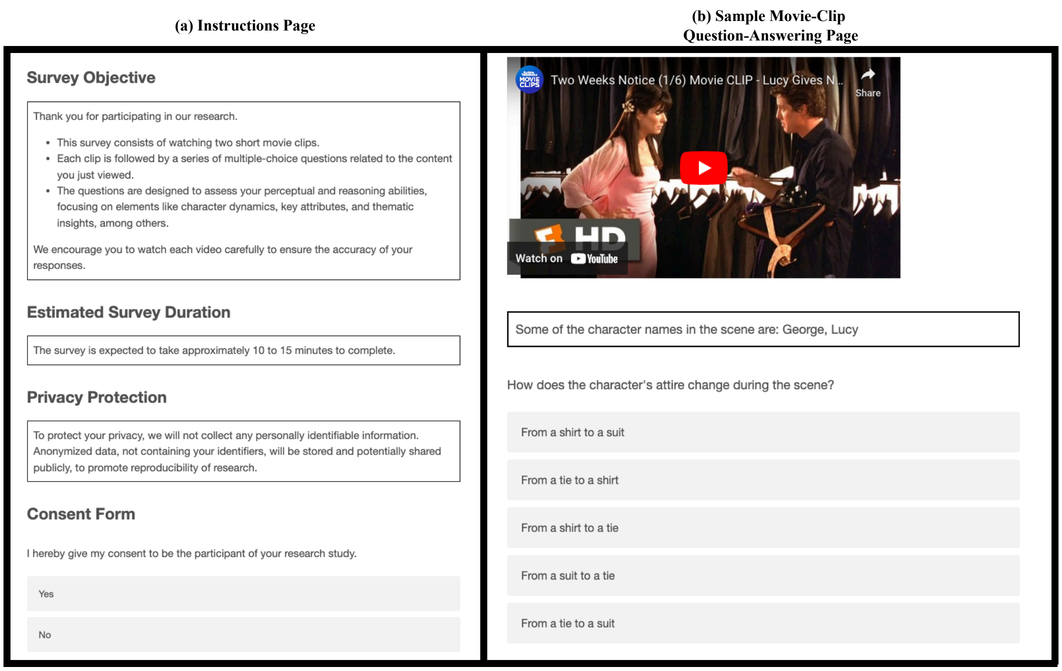
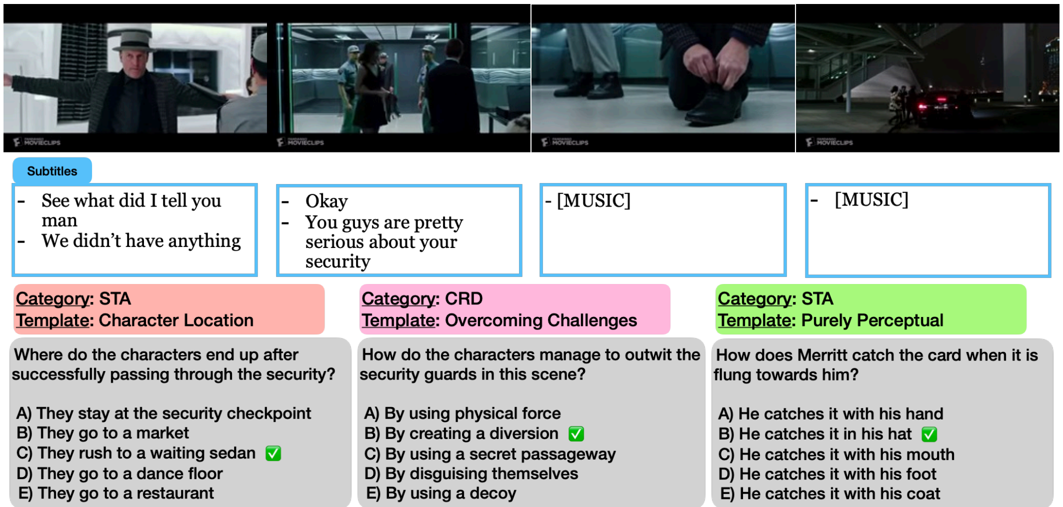
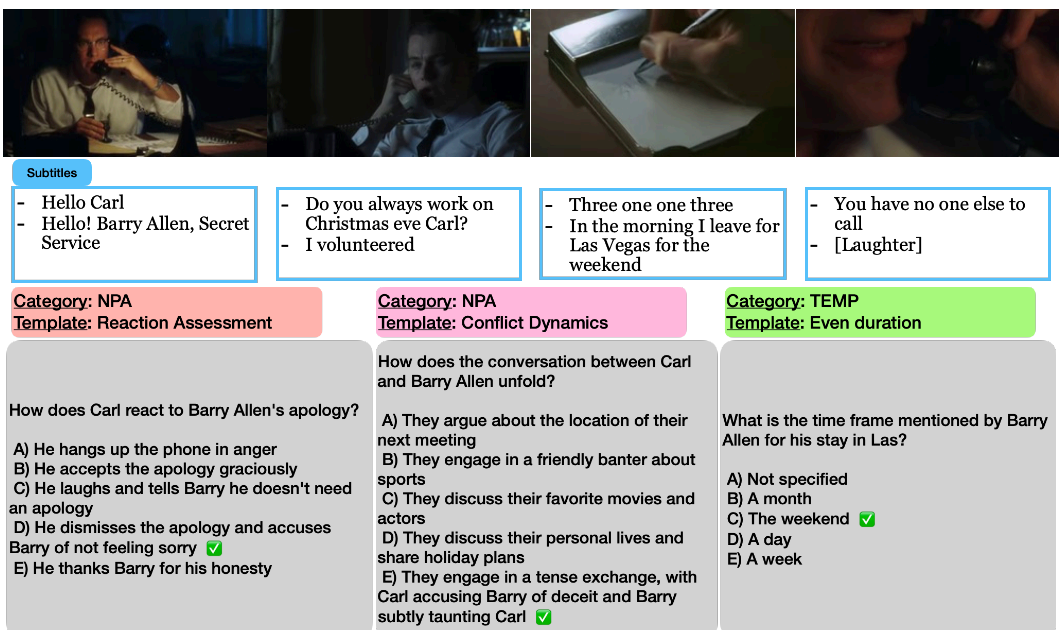


Figure 7: (*left*) (a) **Instructions Page:** The instructions page at the beginning of the survey, as presented to participants. The participants provide informed consent before viewing any video clip and answering questions. (*right*) (b) **Sample Movie-Clip Question-Answering Page:** An example of 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.

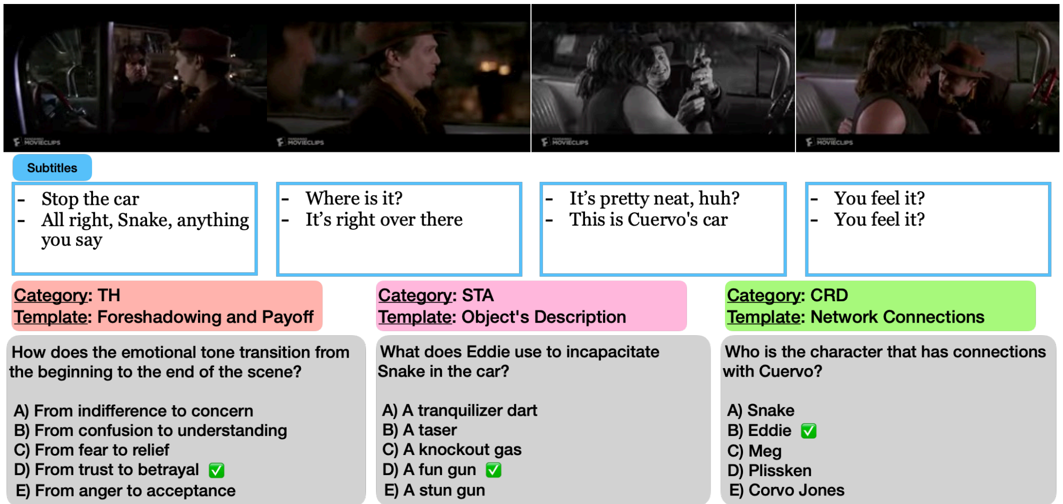


(a)

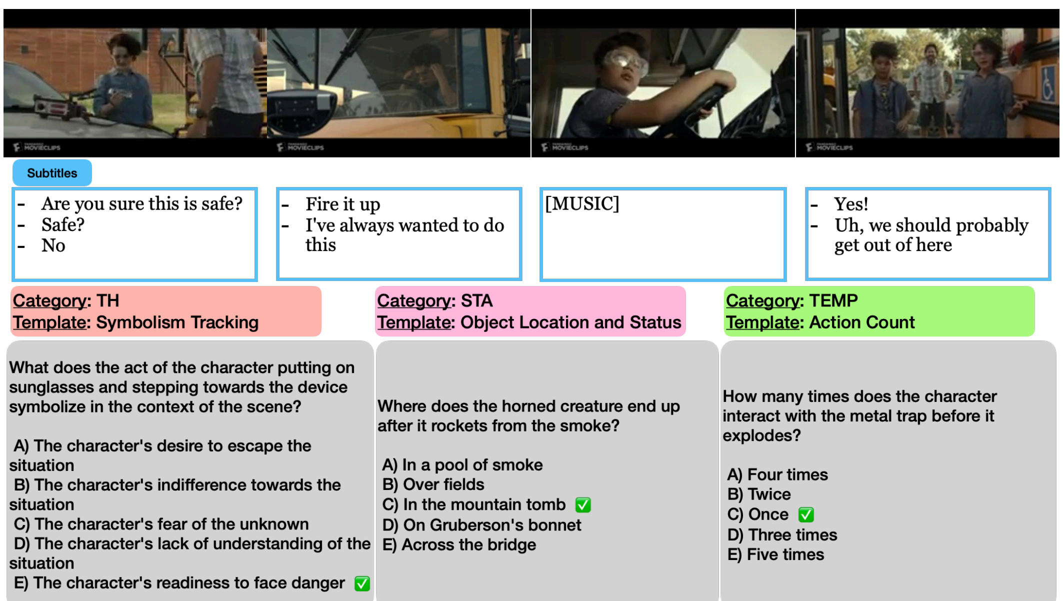


(b)

Figure 8: **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 2 for other category acronyms.

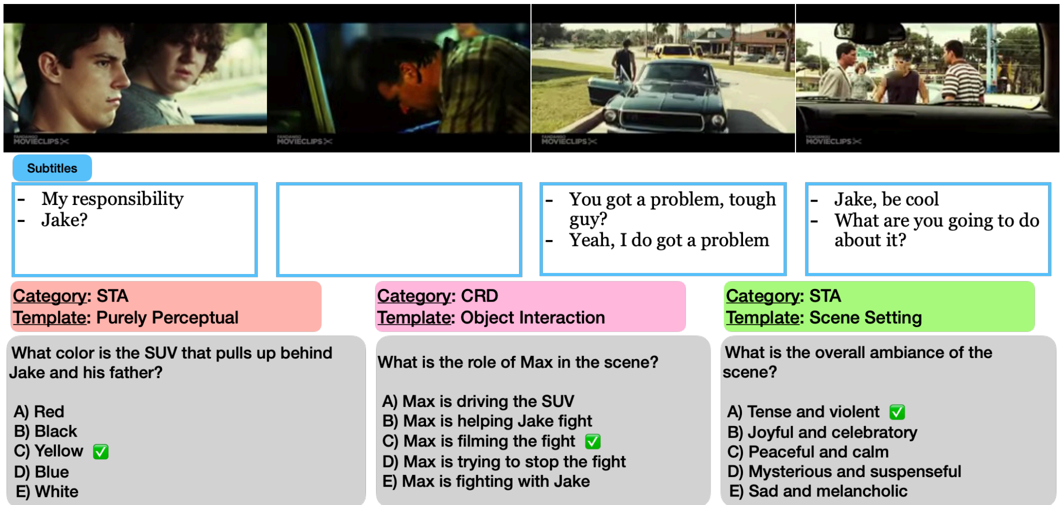


(a)

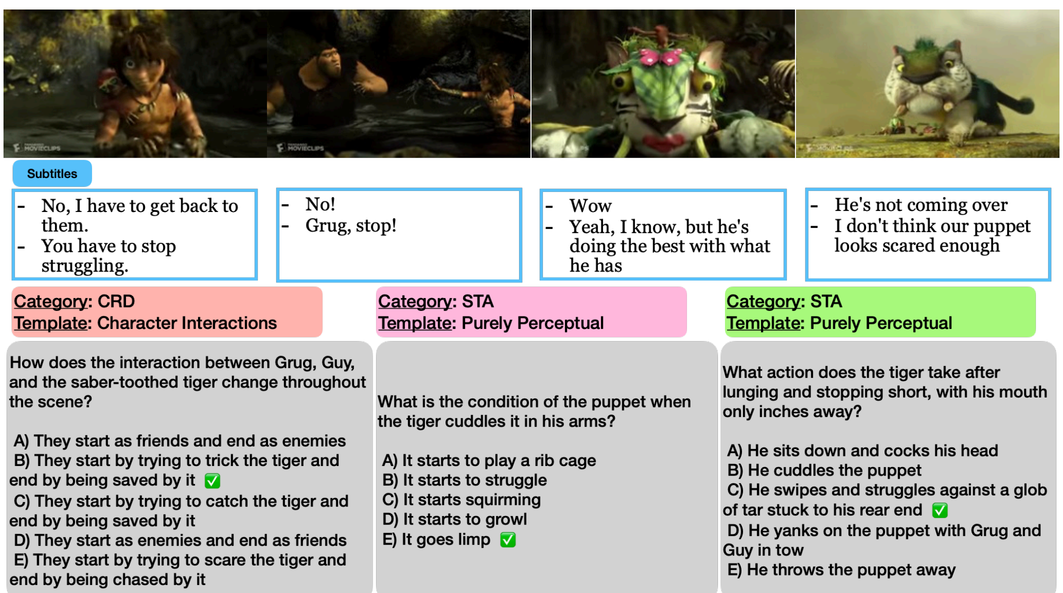


(b)

Figure 9: **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 2 for other acronyms.

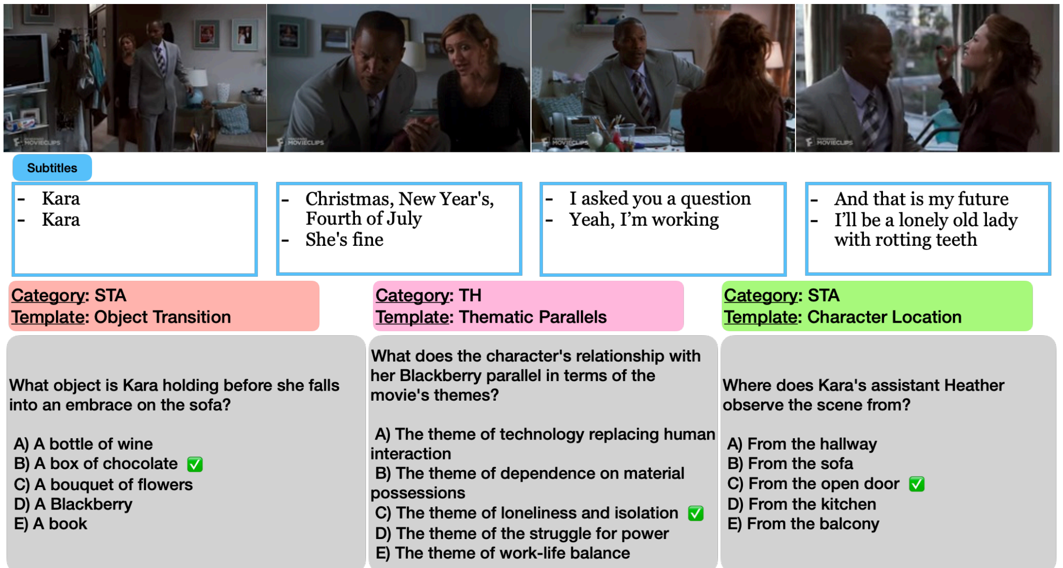


(a)

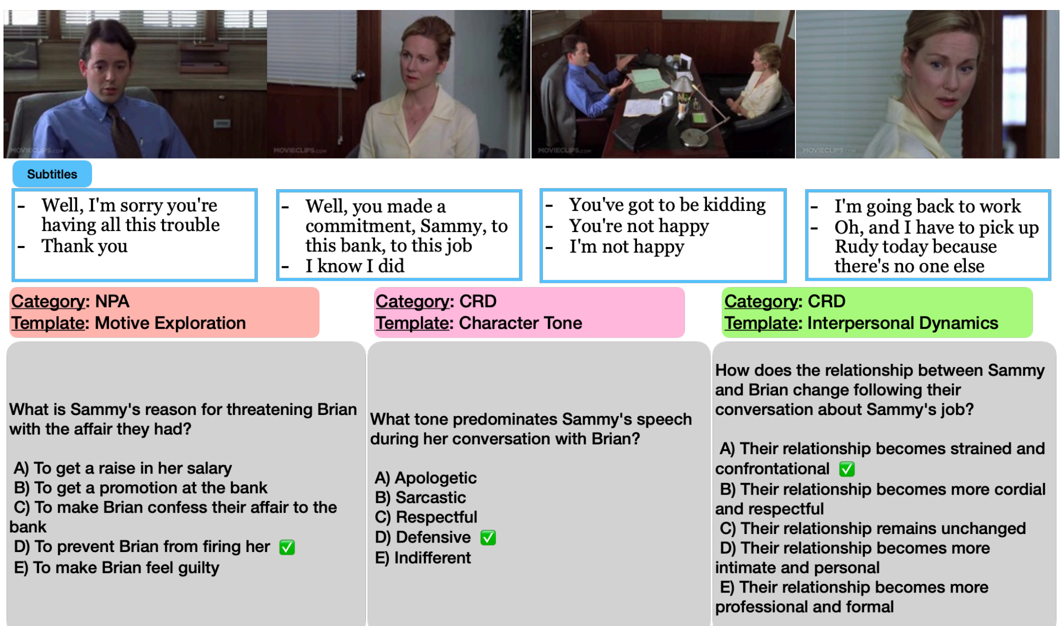


(b)

Figure 10: **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 2 for other acronyms.



(a)



(b)

Figure 11: **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 2 for other acronyms.