CinePile: A Long Video Question Answering Dataset and Benchmark

Ruchit Rawal[♦] Khalid Saifullah[♦] Ronen Basri[♣]
David Jacobs[♦] Gowthami Somepalli[*♦] Tom Goldstein[*♦]
[♦] University of Maryland, College Park [♣] Weizmann Institute of Science

Abstract

Current datasets for long-form video understanding often fall short in providing genuine long-form comprehension challenges, as many tasks derived from these datasets can be successfully tackled by analyzing just one or a few random frames from a video. To address this issue, we present a novel dataset and benchmark, CinePile, specifically designed for authentic long-form video understanding. This paper details our innovative approach for creating a question-answer dataset, utilizing advanced LLMs and building upon human-generated raw data. Our comprehensive dataset comprises 200,000 multiple-choice questions (MCQs), covering a diverse range of visual and multimodal aspects, including temporal comprehension, understanding of human-object interactions, and reasoning about events or actions within a scene. Additionally, we evaluate recent advances in video-centric LLMs, both open-source and proprietary, using the evaluation split of our dataset. The findings reveal that even state-of-the-art vision LLMs significantly lag behind human performance in these tasks, highlighting the challenges inherent to video understanding.

1. Introduction

Large multi-modal models offer the potential to analyze and understand long, complex videos. However, training and evaluating them on video data poses difficult challenges. Most videos contain dialog and pixel data, both essential for a complete scene understanding. Furthermore, existing vision-language models are primarily pre-trained on still frames, while understanding long videos requires identifying interactions and plot progressions over time.

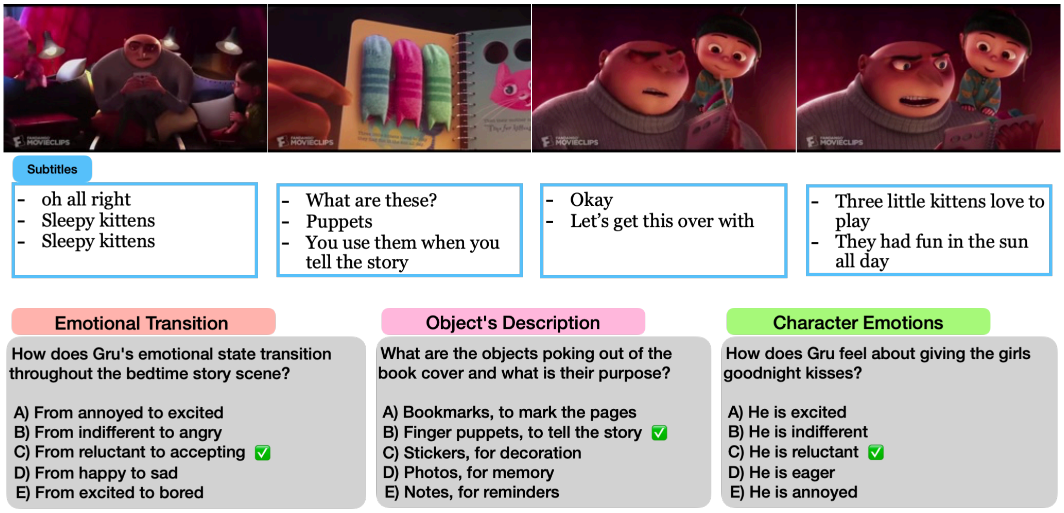
In this paper, we introduce CinePile, a large-scale dataset consisting of over 200,000 question-answer pairs from 8000 videos, split into a train and test set. Our dataset emphasizes question diversity, and topics span temporal understanding, perceptual analysis, complex reasoning, and more. It also emphasizes question difficulty, with humans exceeding the best commercial models by approximately 20%, and exceeding open source models by 50%.

We present a scene and a few question-answer pairs from our dataset in Fig. 1. Consider the first question, `How does Gru's emotional state transition throughout the scene?` For a model to answer this correctly, it needs to understand both the visual and temporal aspects, and even reason about the plot progression of the scene. To answer the second question, `What are the objects poking out of the book cover and what is their purpose`, the model must localize an object in time and space, and use its world knowledge to reason about their purpose.

CinePile addresses several weakness of existing video understanding datasets. First, CinePile's large size enables it to serve as both an instruction-tuning dataset and an evaluation benchmark. We believe the ability to do instruction tuning for video at a scale comparable to common language-only instruction datasets will lead to large improvements in model performance. Also, the diversity in CinePile makes it a more comprehensive measure of model performance than existing benchmarks. Unlike existing metrics, CinePile puts little emphasis on purely visual questions (e.g., 'What color is the car'), or on classification questions (e.g., 'What genre is the video') that do not require temporal understanding. Rather, CinePile comprehensively evaluates vision, temporal reasoning, and video understanding while still providing a breakdown of question types to help developers identify blind spots in their models.

CinePile's large size is made possible by our novel pipeline for automated question generation using large language models. Our method leverages large existing sets of movie descriptions created to assist the vision impaired. We transcribe these movie descriptions, and align them with publicly available video clips from YouTube. Using this detailed human analysis of scenes, powerful LLMs are able to create complex and difficult questions without relying to video. At test time, models must address these questions from only the dialog and video frames, and without access to the hand-written descriptions used to build the questions.

[*] Equal contribution. Correspondence: ruchitr@umd.edu.



- oh all right
- Sleepy kittens
- Sleepy kittens

- What are these?
- Puppets
- You use them when you tell the story

- Okay
- Let's get this over with

- Three little kittens love to play
- They had fun in the sun all day

How does Gru's emotional state transition throughout the bedtime story scene?

A) From annoyed to excited
B) From indifferent to angry
C) From reluctant to accepting ✓
D) From happy to sad
E) From excited to bored

What are the objects poking out of the book cover and what is their purpose?

A) Bookmarks, to mark the pages
B) Finger puppets, to tell the story ✓
C) Stickers, for decoration
D) Photos, for memory
E) Notes, for reminders

How does Gru feel about giving the girls goodnight kisses?

A) He is excited
B) He is indifferent
C) He is reluctant ✓
D) He is eager
E) He is annoyed

Figure 1. **Example movie clip and multiple-choice questions from CinePile**. The first and second rows depict a selection of image frames extracted from a movie clip from Despicable Me, accompanied by their corresponding subtitles. The next row showcases example questions along with the question template shown in colored headers.

2. Creating a long video reasoning benchmark

Our dataset curation process has four primary components 1) Collection of raw video and related data. 2) Generation of question templates. 3) Automated construction of the question-answer dataset using video and templates, and 4) A novel filtering pipeline to remove malformed questions.

2.1. Data collection and consolidation

We obtain clips from English-language films from the YouTube channel *MovieClips*[1]. This channel hosts self-contained clips, each encapsulating a major plot point, facilitating the creation of a dataset focused on understanding and reasoning. Next, we collected Audio Descriptions from AudioVault[2]. Lastly, we collect movie information for each scene, such as genre, actors, and main plot from IMDB[3].

Getting visual descriptions of video for free. Audio descriptions (ADs) feature a narrator who explains the visual elements crucial to the story during pauses in dialogue. ADs have been created for many films to assist the vision impaired. The key distinction between conventional video caption datasets and ADs lies in the contextual nature of the latter. In ADs, humans emphasize the important visual elements in their narrations, unlike other video caption datasets, which tend to be overly descriptive. We use the audio descriptions as a proxy for visual annotation in the videos for our dataset creation. However, since the video clips we gather are typically 2-3 minutes long, and Audio Descriptions (ADs) cover entire movies, we need to align and extract the *clip relevant part* from the AD. We discuss this process in detail in Appendix Sec. C.

2.2. Automated Question Templates

Mainstream video question-answering benchmarks were written by human annotators. The question-answer pairs are typically curated in one of two ways: 1) Humans are given complete freedom to ask questions about a given scene [25] 2) Humans focus on specific aspects and are trained or provided with examples of questions, encouraging them to write more questions in a similar style [10, 12, 19, 34].

While we use a template-based approach for question generation, rather than confining to a few predefined themes, we propose an automated method to create question templates from existing human-generated questions. We first cluster 30,000 human-generated questions across multiple existing datasets, then use GPT-4 to discern their underlying themes and generate prototypical questions for each template. We discuss the details of the clustering and template discernment process in Appendix Sec. D. In total, we generate 86 unique templates that we categorize into four high-level categories: Character and Relationship Dynamics (CRD), Narrative and Plot Analysis (NPA), Thematic Exploration (TE) and Setting, and Technical Analysis (STA). For a detailed discussion and example questions from each category, please refer to Appendix Sec. E.

2.3. Automated QA generation with LLMs

Before creating questions for a scene, we chose the relevant question templates by providing Gemini with the scene-text annotation of a scene, and asking it for the 20 templates most relevant to that scene. From these, we randomly se-

[1] https://www.youtube.com/@MOVIECLIPS
[2] https://audiovault.net/movies
[3] https://www.imdb.com/

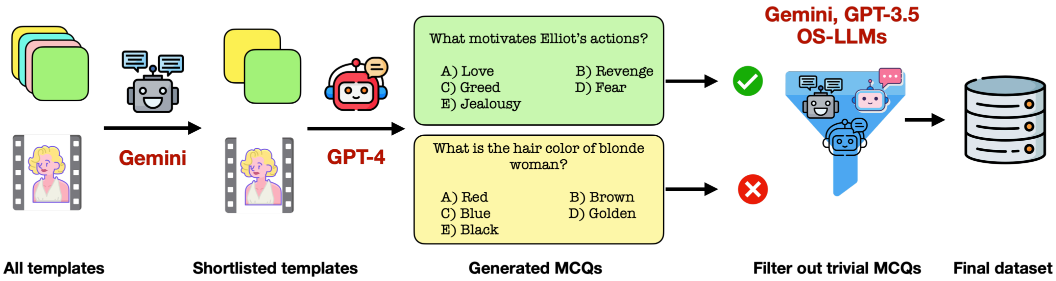


Figure 2. **Automated QA Generation and Filtering** Our process begins with a set of automated templates and scenes. Initially, we filter out the templates relevant to each scene. Next, we pass these templates along with the annotated-scene-text to GPT-4, which is then used to create multiple-choice questions (MCQs). The generated MCQs are then subjected to numerous filters to curate the final dataset. For more detailed information, refer to Sec. 2.3 and Sec. 2.4

lect 5-6 templates. We provide a commercial language model with (i) the audio description of a scene, which includes both visual descriptions and dialog, (ii) the selected question template names (e.g. 'Physical Possession'), (iii) the prototypical questions for the templates (e.g. "What is [Character Name] holding"), and (iv) a system prompt to generate questions. (complete pipeline shown in Fig. 2)

Through rigorous experimentation, we devised a system prompt that makes the model attentive to the entire scene and is capable of generating deeper, longer-term questions as opposed to mere surface-level perceptual queries. We observed that providing the prototypical example prevents GPT-4 from hallucination, and also leads to more plausible multiple-choice question (MCQ) distractors. We also found that asking the model to provide rationale for its answer enhances the quality of the questions. Additionally, we found that including timestamps for dialogues and visual descriptions augments the quality of generated temporal questions. We were able to generate ≈ 26 questions for each video in the dataset. While GPT-4 performs well across all question templates, we found that Gemini excels particularly with perceptual templates. Therefore, we utilized Gemini to generate a segment of questions in the dataset, while using GPT-4 for reasoning templates.

A small proportion of questions produced can be answered directly i.e., without referring to the clip, such as `What's the color of the blonde woman's hair?`. We implemented a few checks to eliminate trivial or poorly framed questions. We discuss these checks and a few axes we evaluate the question-answering dataset next.

2.4. Testing the quality of the dataset

While the process above consistently produces well-formed and answerable questions, we observed that some questions are either trivial, with answers embedded within the question itself, or pertaining to basic world concepts that do not require viewing the clip. To prune these, we evaluated our dataset with the help of a few LLMs on the following axes and we either removed the questions from the dataset or compute metrics that users can use in the downstream tasks.

Degeneracy. A question is considered degenerate if the answer is implicit in the question itself, e.g., `What is the color of the pink house?`. In our dataset, these types of questions constitute only a small fraction. Manually reviewing all questions being impractical, we employed three distinct language models (LMs) to automate this process: Gemini [26], GPT-3.5 [1], and Phi-1.5 [13]. These models vary in their underlying training data and sizes. We presented only the questions and choices to these models, omitting any context, and calculated the accuracy of each question across the multiple models. If all models correctly answer a question, it is likely to be degenerate. We excluded degenerate questions from CinePile's evaluation split.

Vision Reliance. When generating the multiple-choice questions (MCQs), we considered the entire scene without differentiating between visual text and dialogue. Consequently, some questions in the dataset might be answerable solely based on dialogue, without needing the video component. For this analysis, we utilized the Gemini model. The model was provided with only the dialogue, excluding any visual descriptions, to assess its performance. If the model correctly answers a question, it is assigned a score of 0 for the visual dependence metric; if it fails, the score is set at 1. In later sections, we present the distribution of the visual dependence scores across different MCQ categories.

Hardness. We developed a metric to gauge the difficulty of questions for the models, even when provided with full context. For this purpose, we selected the Gemini model, given its status as one of the larger and more capable models. This metric differs from accuracy; during evaluation, the models are only supplied with videos and dialogue information, excluding visual descriptions. However, in calculating the hardness metric, we include visual descriptions as part of the context given to the model.

Additionally, authors regularly verified the quality of questions across multiple scenes and corrected any systemic errors that arose in the pipeline. We also conducted a human study to identify weaknesses in the dataset, further discussed in Appendix Sec. L.

3. A look at the dataset

Our dataset consists of 8000 video clips with average length of ≈160 seconds, split into train and test splits of 7700 and 300 videos each. Following the pipeline outlined in Sec. 2, we ended up with over 200,000 training points and 7,800 test-set points (before degeneracy filtration). Each MCQ contains a question, answer, and four distractors. After filtration of the degenerate questions from the test split, we are left with 5,500 questions. Of all the test questions, 34.30% are reliant on visual information. We present additional dataset statistics including distribution of questions, hardness scores across different categories, in Appendix Sec. G.

4. Model evaluation

In this section we discuss the evaluations of various closed and open source video LLMs on our dataset, some challenges, and the model performance trends. Given that our dataset is of type multiple-choice question answers (MCQs), we evaluate a given model's performance on our benchmark questions by measuring its ability to accurately select the right answer from a set of options, containing only one correct answer and four distractors. One key challenge is reliably parsing the model's response to extract its chosen answer, and mapping it to one of the predefined answer choices. Model's responses may vary in format, including additional markers, or may only contain the option letter, or have a combination of the option letter and its corresponding text, etc. Such variations necessitate a robust post-processing step to accurately extract and match the model's response to the correct answer. Due to space constraints, we discuss the process in detail in Appendix Sec. H.

During the evaluation, we specifically instruct the model to be concise and only output the option letter. Qualitatively we see that most commercial models are good at following these instructions, and we can map these responses well. Some OSS models are very verbose in their response, and poor at following instructions. Hence, we also computed traditional video-caption evaluation metrics like BertScore [42], CIDEr [28], and ROUGE-L [14] for open-source models, and present results in Appendix.

We evaluate various commercial and open-source LLM models and we present their performance in Tab. 1. We also present human numbers (author and non-author) for comparison. On average, VLM models both commercial and OSS, are behind human performance on our dataset. While commercial VLMs perform reasonably well, the OSS models perform quite poorly showing the gap in their capabilities. Among the question categories, GPT-4V model performed poorest in "Thematic Exploration" category followed by "Temporal". Surprisingly, GPT-4V did well in "Setting and Technical Analysis" which pertains to purely visual questions. One possible reason for this is, GPT-4V is aware of famous movies and it can answer the questions even without any context. To understand the extent of this memorization, we computed the degeneracy metrics for the commercial models, we see GPT-4 stands at 33.4%, Gemini at 39.64% and Claude at 34%. This states that the true performance of these models might be much lower than the numbers reflected in the table! However, we hope that training OSS models on our data will help them match the performance of commercial models in the future.

Table 1. **Evaluations on CinePile.** Accuracy of various video LLMs on the test split, along with Human performance for comparison. Chance performance is 20%. TEMP refers to Temporal. See Sec. 2.2 for other acronyms.

Model	**Average**	**CRD**	**NPA**	**TEMP**	**STA**	**TH**
Human	73.21	82.92	75.00	75.52	73.00	64.93
Human (authors)	**86.00**	**92.00**	**87.5**	**100**	71.20	**75.00**
GPT-4 Vision [1]	66.75	68.14	76.54	65.33	**76.04**	57.19
Gemini Pro Vision [26]	57.68	59.25	70.37	56.80	63.54	46.15
Claude 3 (Opus) [2]	46.34	46.15	64.19	44.82	47.91	39.46
mPLUG-Owl [37]	15.93	15.92	14.19	14.88	15.78	18.85
Video-ChatGPT [16]	14.70	16.44	11.72	11.83	29.03	12.20
MovieChat [23]	7.12	6.96	6.78	8.05	11.32	5.08

5. Discussion and Conclusion.

In this paper, we introduced CinePile, a unique long video understanding dataset and benchmark, featuring over 200k questions in the training set and 5500 in the test split. We detailed a novel method for curating and filtering this dataset, which is both scalable and cost-effective. Additionally, we benchmarked various recent commercial multimodal LLMs and conducted a human study to gauge the achievable performance on this dataset. To our knowledge, CinePile is the only large-scale dataset that focuses on multi-modal understanding, as opposed to the purely visual reasoning addressed in previous datasets. We intend to make the questions and answers from the training set of CinePile publicly available. Additionally, we will set up a leaderboard for the test set, providing a platform for new video LLMs to assess and benchmark their performance on CinePile.

Despite its strengths, there are still a few areas for improvement in our dataset, such as the incorporation of character grounding in time. While we believe our dataset's quality is comparable to or even better than that of a Mechanical Turk annotator, we acknowledge that a motivated human, given sufficient time, can create more challenging questions than those currently generated by an LLM. Our goal is to narrow this gap in future iterations of CinePile.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 4

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 4

[3] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023. 7

[4] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 7

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 8, 11

[7] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023. 7

[8] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023. 7

[9] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 11

[10] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2, 7, 8, 11

[11] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 10

[12] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2, 11

[13] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023. 3, 7

[14] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 4, 11

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 10

[16] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023. 4, 11

[17] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024. 7

[18] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 7, 11

[19] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 7, 8, 11

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 7

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 8

[22] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 7, 8

[23] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 4, 11

[24] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 7

[25] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2, 7, 8, 11

[26] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 4

[27] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023. 7

[28] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4, 11

[29] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 7

[30] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 9

[31] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023. 7

[32] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 7

[33] Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *CVPR*, 2021. 7

[34] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2, 11

[35] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023. 7

[36] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 11

[37] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 4, 11

[38] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 10

[39] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024. 7

[40] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 9

[41] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*, 2023. 7, 11

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 4, 11

[43] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 10

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. 9a, 9b, 10a, 10b, 11a, 11b, 12a and 12b.

B. Related Work

Video understanding and question answering.
LVU [33], 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 [25] 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 [18] 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 [19], focuses on core perception skills, such as memory and abstraction, across various reasoning abilities (e.g., descriptive, predictive, etc) for short-form videos. MAD [22] 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 [10] 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 [41] 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 [5], 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 [4, 8, 27, 32] and language [13, 17, 24, 31, 39]. For instance, Self-Instruct [29] 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 [7] 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 [35], 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 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 [20]. More specifically, we use `WhisperX` [3], 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 [6] using annotations from the MAD dataset [22], 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. 3. Our starting point is approximately 30,000 human-curated questions from the MovieQA [25], TVQA [10], and Perception Test [19] 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. 3.

First, we preprocess the questions by replacing first names and entities with pronouns, as BERT [21] 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[4]. 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. 4 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

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

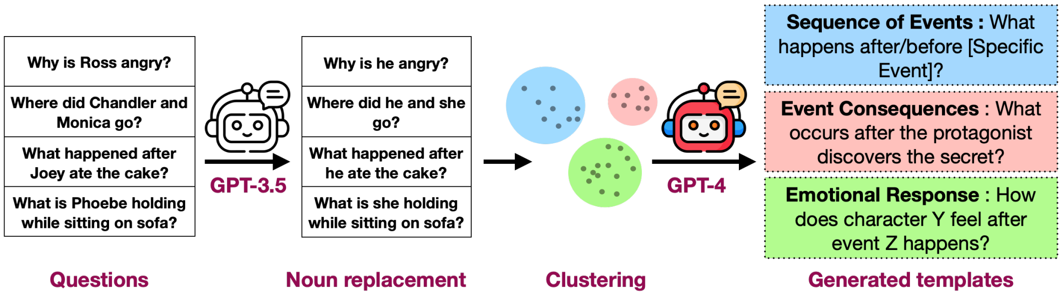

Why is Ross angry?
Where did Chandler and Monica go?
What happened after Joey ate the cake?
What is Phoebe holding while sitting on sofa?
GPT-3.5
Why is he angry?
Where did he and she go?
What happened after he ate the cake?
What is she holding while sitting on sofa?
GPT-4
Sequence of Events : What happens after/before [Specific Event]?
Event Consequences : What occurs after the protagonist discovers the secret?
Emotional Response : How does character Y feel after event Z happens?
Questions
Noun replacement
Clustering
Generated templates

Figure 3. **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 Sec. 2.2 for more details.

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.

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 4. **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 Fig. 3 and Sec. 2.2 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[30, 40]. First, we use a shot detection algorithm to pick the important frames[5], 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 [15], OtterHD [11], mPlug-Owl [38] and MinGPT-4 [43], 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.

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 Sec. 2.1. 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 Sec. 2.4. 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. 5 (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 Sec. 2.4, at the category level in Fig. 5 (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 Tab. 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.

[5] https://www.scenedetect.com/

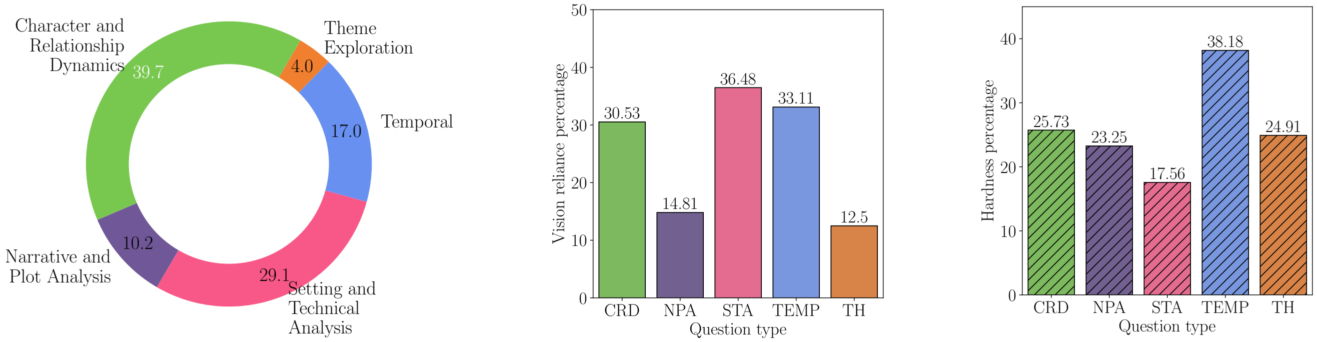


Figure 5. **Left:** Question category composition in the dataset. **Middle:** Percentage of vision-reliant questions across categories. **Right:** Percentage of hard questions per question category type. TEMP refers to Temporal. Please refer to Tab. 2 for other acronyms. The colors correspond to the same categories across the plots.

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. For details of different QA types, refer to Sec. 2.3

Dataset	**Annotation**	**Num QA**	**Avg sec**	**Multimodal**	**QA Type**			
					Temporal	Attribute	Narrative	Theme
TGIF-QA [9]	Auto	165,165	3	✗	✓	✗	✗	✗
MSRVTT-QA [36]	Auto	243,690	15	✗	✗	✓	✗	✗
How2QA [12]	Human	44,007	60	✗	✓	✓	✗	✗
NExT-QA [34]	Human	52,044	44	✗	✓	✓	✗	✗
EgoSchema [18]	Auto	5,000	180	✗	✓	✓	✓	✗
MovieQA [25]	Human	6,462	203	✓	✓	✓	✓	✗
TVQA [10]	Human	152,545	76	✓	✓	✓	✓	✗
Perception Test [19]	Human	44,000	23	✓	✓	✓	✗	✗
MoVQA [41]	Human	21,953	992	✓	✓	✓	✓	✗
CinePile (Ours)	Human + Auto	200,000	160	✓	✓	✓	✓	✓

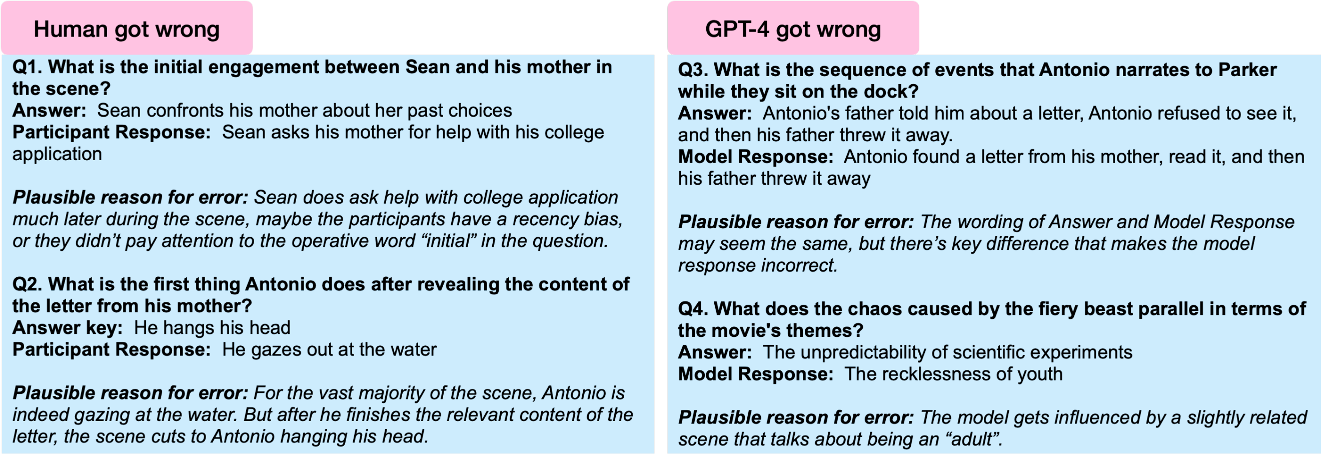
Table 4. Performance of various open source video-LLMs on CinePile 's test split, as evaluated using various video captioning metrics – BERTSCoRE [6], CIDEr [28], ROUGE-L [14].

Model	BERTScore↑	CIDEr↑	ROUGE-L↑
mPLUG-Owl [37]	0.46	1.36	0.31
Video-ChatGPT [16]	0.48	1.14	0.32
MovieChat [23]	0.40	0.66	0.18

I. Evaluation on Captioning Metrics

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 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 [42], CIDEr [28], and ROUGE-L [14]. 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



Human got wrong

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 got wrong

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 6. **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.

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 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 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 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 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

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 Tab. 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

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.

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. 8-(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. 8-(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 humans got wrong and GPT-4 got wrong and plausible reasons for errors in Fig. 6. 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

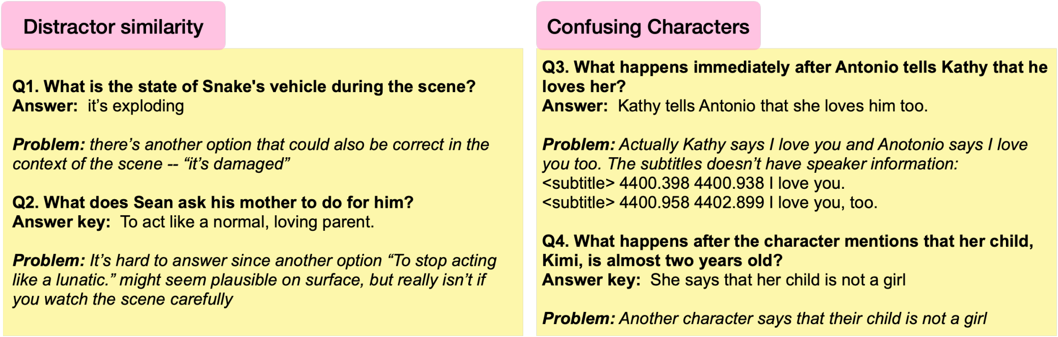
Table 6. Comparing question-answer pairs generated by GPT-4 with those generated by Gemini, for the movie clip: Ghostbusters: Afterlife (2021) - Muncher Attack Scene (3/7) — Movieclips. TEMP refers to Temporal. Please refer to Tab. 2 for other acronyms.

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.

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. 7. 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 Sec. 4.

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.



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

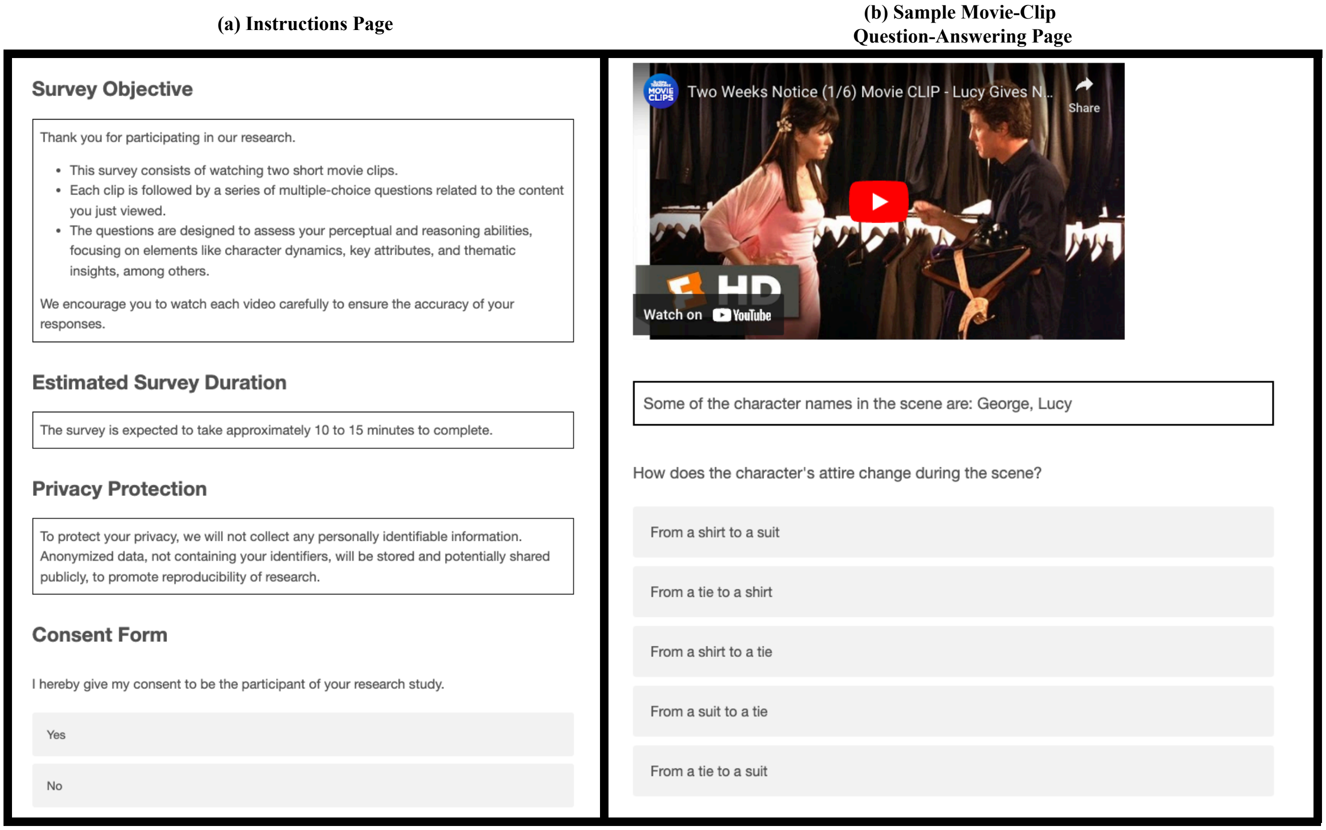
Q3. What happens immediately after Antonio tells Kathy that he loves her?
Answer: Kathy tells Antonio that she loves him too.

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 7. **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.



(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

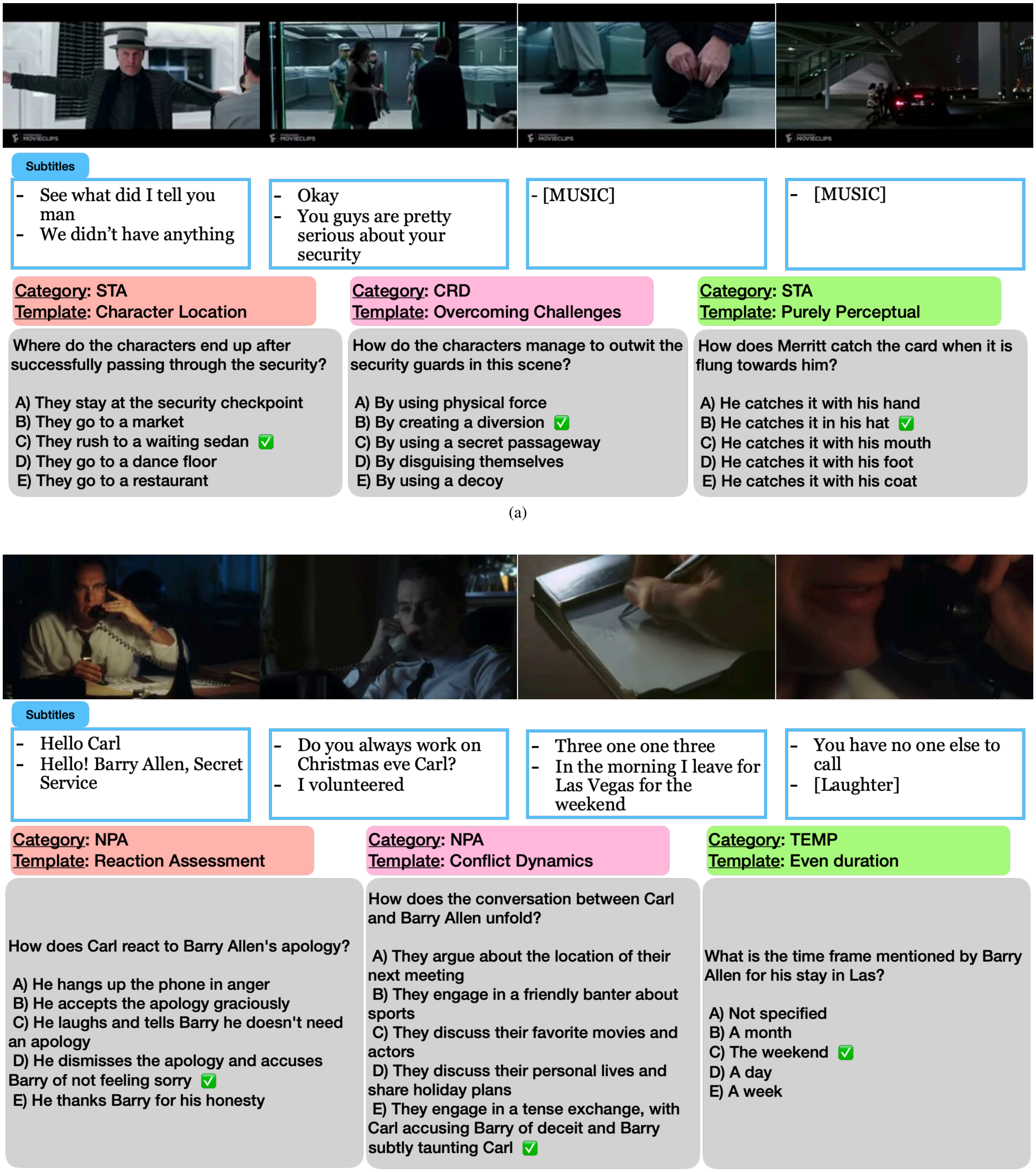
From a tie to a shirt

From a shirt to a tie

From a suit to a tie

From a tie to a suit

Figure 8. (*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.



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 ✓

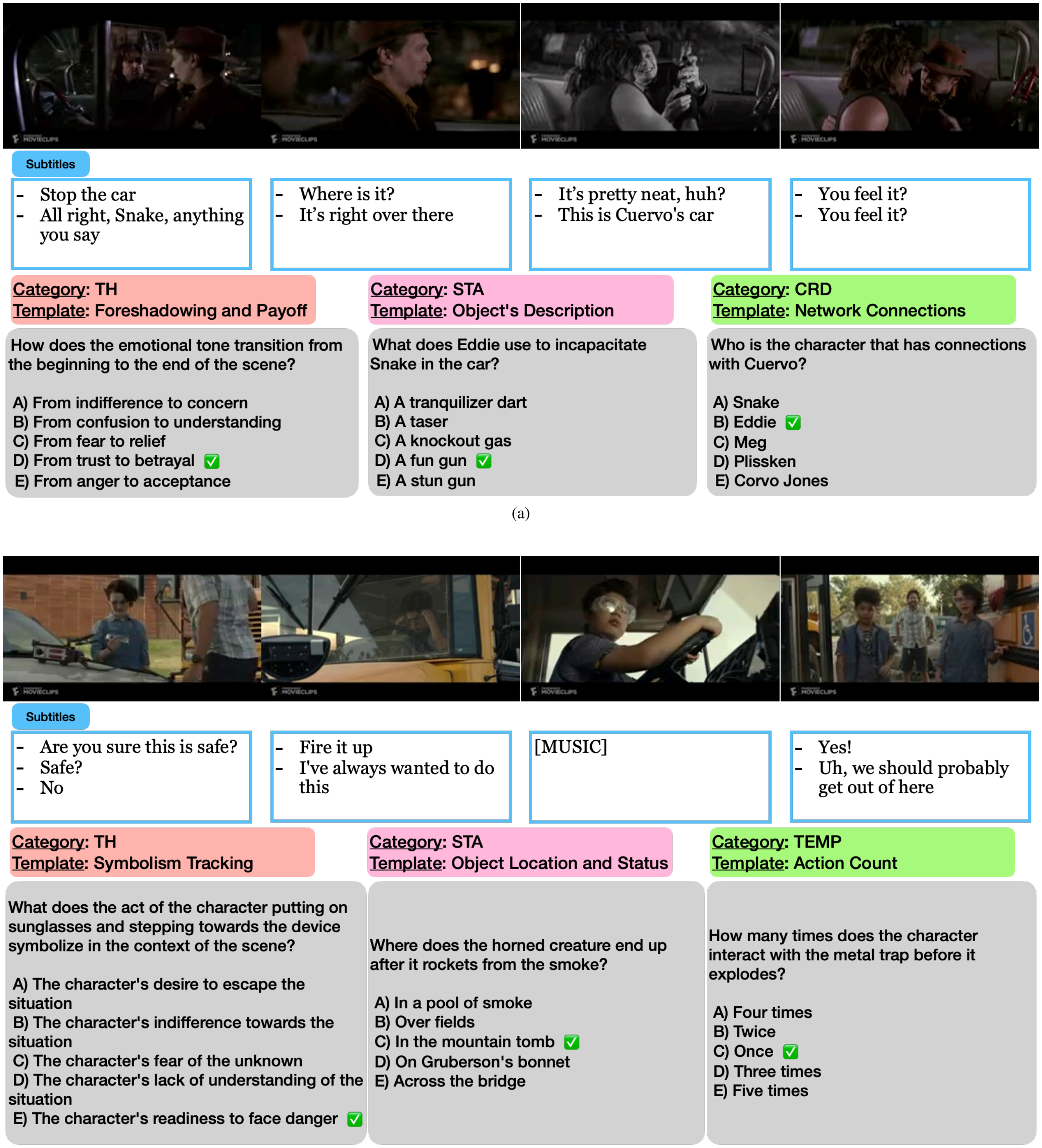
Category: TEMP
Template: Even duration

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 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) 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 Tab. 2 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

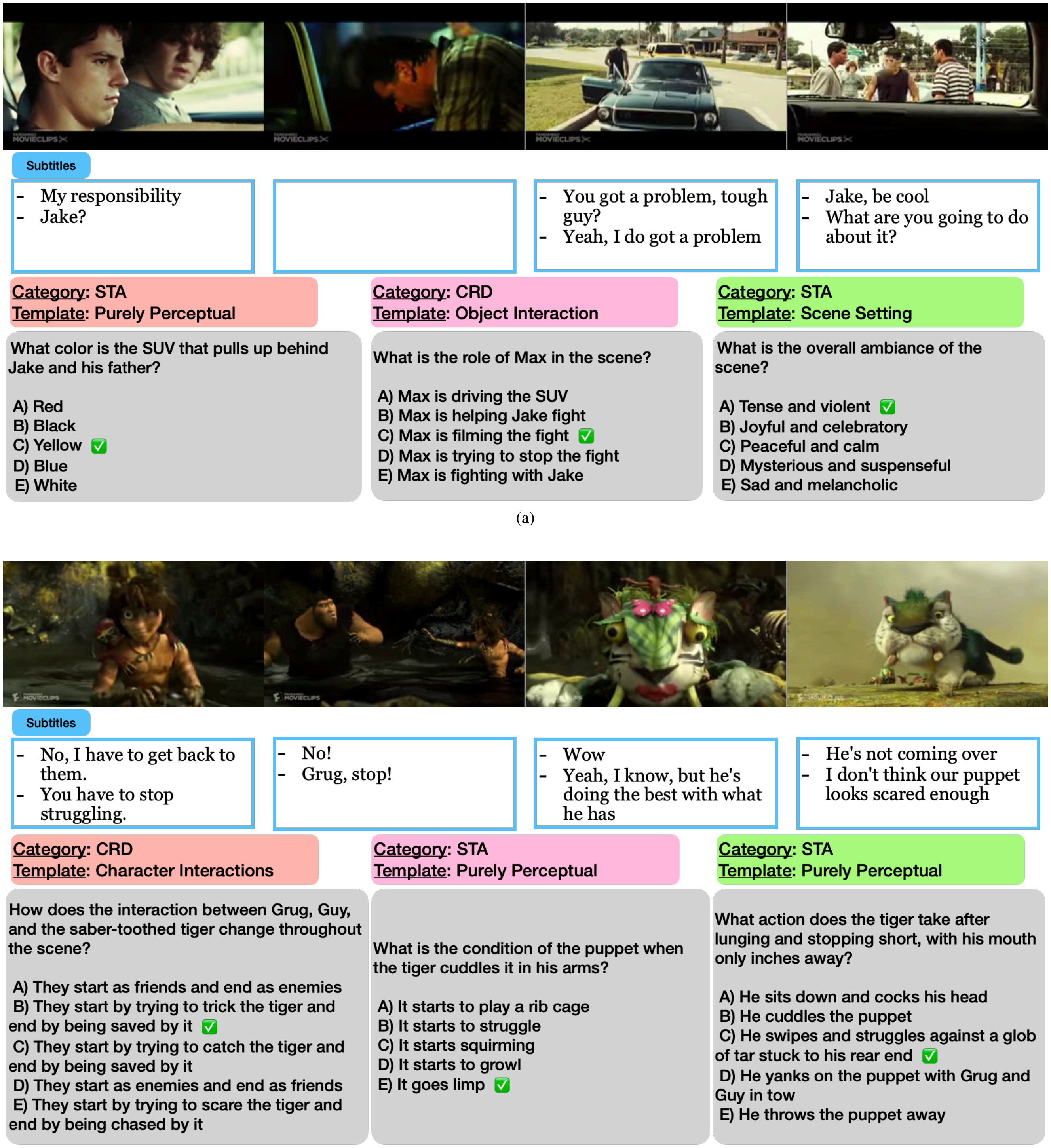
Category: TEMP
Template: Action Count

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 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)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 Tab. 2 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 ✓

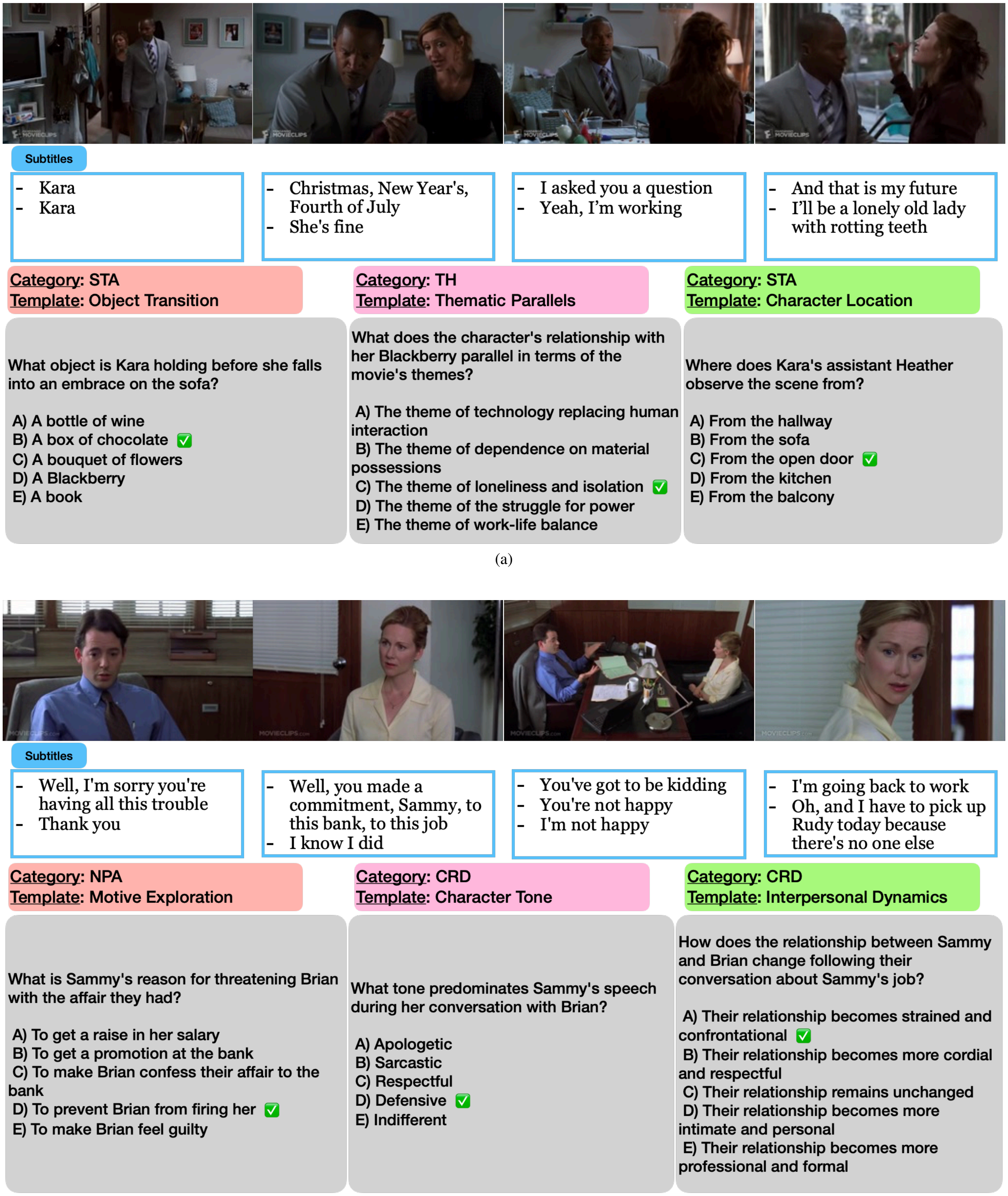
Category: STA
Template: Purely Perceptual

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 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) 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 Tab. 2 for other acronyms.



Subtitles

- 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)

Subtitles

- 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 12. **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 Tab. 2 for other acronyms.