MovieCORE: COgnitive REasoning in Movies

Gueter Josmy Faure[1] **Min-Hung Chen**[2] **Jia-Fong Yeh**[1] **Ying Cheng**[4]
Hung-Ting Su[1] **Shang-Hong Lai**[4] **Winston H. Hsu**[1,3]*
[1]National Taiwan University [2]NVIDIA [3]Mobile Drive Technology
[4]National Tsing Hua University

Abstract

This paper introduces MovieCORE, a novel video question answering (VQA) dataset designed to probe deeper cognitive understanding of movie content. Unlike existing datasets that focus on surface-level comprehension, MovieCORE emphasizes thought-provoking questions that engage System 2 thinking while remaining specific to the video material. We present an innovative agentic brainstorming approach, utilizing multiple large language models (LLMs) as thought agents to generate and refine high-quality question-answer pairs. To evaluate dataset quality, we develop a set of cognitive tests assessing depth, thought-provocation potential, and syntactic complexity. We also propose a comprehensive evaluation scheme for assessing VQA model performance on deeper cognitive tasks. Our work contributes to advancing movie understanding in AI systems and provides valuable insights into the capabilities and limitations of current VQA models when faced with more challenging, nuanced questions about cinematic content. We will make our agentic annotation system, the dataset and its metadata publicly available.

1 Introduction

Movie audiences consciously or subconsciously absorb information about actors' states of mind, body language, and expressions to infer their moods and empathize with their situations. Most people would agree that such inferences are crucial to truly understanding a movie. Despite the significance of this deeper level of understanding in cinematic experiences, existing movie-based video question answering (VQA) datasets have yet to fully explore this aspect of film comprehension.

Recent movie-based VQA datasets [14, 13, 12] primarily focus on surface-level understanding, neglecting the challenge of comprehending movies at a deeper cognitive level. They predominantly address the "what" by posing questions such as "What is the relationship between the actors?" or "What time does the video take place?", and largely overlook the "how," "why," and "why not" questions crucial for achieving a profound understanding of movies. While EgoSchema [11] attempted to delve beyond the obvious, its more profound questions often remain general and could apply to virtually any video.

This paper presents MovieCORE, a VQA dataset that emphasizes questions and answers with the potential to engage our System 2 thinking – the slow, deliberate, and logical cognitive processes – while remaining specific to the video in question. We acknowledge the inherent subjectivity in responses to "why" or "why not" questions, but argue that this subjective element is precisely what makes the task both challenging and significant. To address the challenge of obtaining comprehensive and faithful question-answer pairs, we implement an agentic brainstorming approach. This method utilizes multiple large language models (LLMs) as thought agents to achieve superior question-answer (QA) refinements through continuous discussions. We then rigorously review a subset of the generated QA pairs to ensure high quality. Furthermore, we employ a set of cognitive tests to assess

*Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

our dataset's depth and syntactic complexity. Lastly, we evaluate existing VQA models on our dataset to gauge their performance on these more challenging cognitive tasks.

Our key contributions can be summarized as follows:

- We introduce MovieCORE, a VQA dataset focused on deep, thought-provoking questions and answers specific to movie content.
- We develop an agentic brainstorming approach using multiple LLMs as thought agents to generate and refine high-quality QA pairs.
- We implement a set of cognitive tests to evaluate the depth, thought-provocation, and syntactic complexity of VQA datasets.
- We design a comprehensive evaluation scheme to assess the accuracy, comprehensiveness, depth, and coherence of answers from existing VQA models.
- We evaluate several VQA models on our dataset in both zero-shot and fully-supervised settings, offering insights into their performance on deeper cognitive tasks.

2 Dataset Creation and Curation

To address the challenge of obtaining question-answer pairs that delve into deeper levels of movie understanding, we introduce an agentic annotation workflow. This approach leverages the deliberative capabilities of multiple Large Language Models (LLMs) acting as specialized agents, each contributing unique perspectives to the annotation process. We start with video context extraction to make sure our text-only annotation agents have enough information about the video.

2.1 Video Context Extraction

The videos for our dataset are sourced from MovieChat-1k [13], a collection of 1,000 movie clips averaging 10 minutes each. We use 986 of these clips, as 14 were either unavailable or lacked necessary annotations. MovieChat-1k, being a VQA dataset, already provides high-level information for each video, such as temporal setting (e.g., ancient or modern) and metadata like the movie's genre. Although some videos in the original dataset include dense captions, we observe inaccuracies and imbalanced scene descriptions. As a result, we exclude these captions, focusing instead on the existing QA pairs and movie metadata.

To provide video context, we utilize MiniCPM-v2.6 [17], an open-source model with visual capabilities comparable to GPT-4V. We prompt it with a carefully curated set of eight questions (shown in Figure 2) designed to extract a multi-dimensional understanding of the video. These questions address narrative structure, thematic focus, emotional tone, key events, character dynamics, genre, and target audience. The extracted information serves as *Data Info* priors for our annotation agents.

2.2 Agentic Annotation Workflow

Our workflow, illustrated in Figure 1, employs a multi-agent system orchestrated by a central Critic Agent acting as the master of ceremonies (MC). Using the Agentic AI framework autogen [15], we deploy instances of GPT4-o for the VQA Expert and Meta Reviewer roles, with GPT4-o-mini powering the other expert agents. The process begins as the Critic Agent receives task instructions and video context (*Data Info*) extracted as described in Section 2.1. The System II VQA Expert initiates the process by generating deep, thought-provoking questions that engage system 2 thinking.

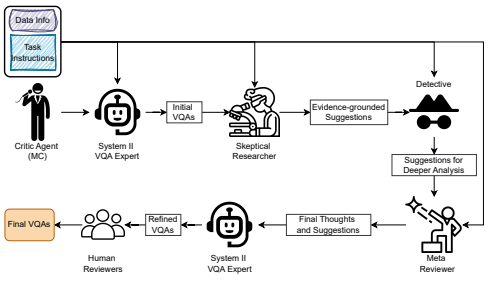

Figure 1: The *Critic Agent*, acting as the master of ceremonies (MC), orchestrates interactions among specialized agents using video context and task instructions.

Table 1: **Syntactic Complexity and Cognitive Demand Comparison:** Parse Tree Depth (PTD), Flesch-Kincaid Grade (FKG), average Bloom's Taxonomy Level (BTL), and percentage of higher-order QAs (HO-QA) across various VQA datasets. Best results are in **bold**, second-best are underlined.

Dataset	PTD	FKG	BTL	HO-QA
MovieChat-1k [13]	2.45	1.4	1.8	0.0
ActivityNetQA [2]	2.26	1.84	1.9	0.2
MVBench [9]	2.84	3.11	2.2	3.4
EgoSchema [11]	5.47	8.30	3.1	33.1
MovieCORE (Ours)	**5.88**	**14.03**	**4.9**	**99.2**

Table 2: **Performance Comparison of Video Question-Answering Models.** We evaluate various models on five criteria: Accuracy (Acc), Comprehensiveness (Comp), Depth, Evidence (Evid), and Coherence (Coh).

Model	Acc	Comp	Depth	Evid	Coh	Avg
Zero-shot Results						
MA-LMM [7]	1.06	0.65	0.90	0.50	0.64	0.75
HERMES [6]	1.40	0.84	1.00	0.76	0.97	0.99
MovieChat [13]	2.90	2.29	2.14	2.30	2.23	2.37
Goldfish [1]	**3.43**	**3.15**	**3.20**	**3.00**	**3.20**	**3.20**
Fine-tuned Results						
InstructBLIP [4]	**4.13**	**3.85**	1.18	2.03	**3.60**	2.96
MA-LMM [7]	3.78	3.20	3.09	3.36	3.32	3.35
HERMES [6]	3.80	3.29	**3.16**	**3.39**	3.36	**3.40**

These initial QA pairs are then scrutinized by the Skeptical Researcher, who evaluates their contextual relevance and accuracy, often challenging the VQA Expert to provide more concrete evidence. The Detective agent follows, suggesting additional questions to uncover underlying motivations, biases, and character development. The Meta Reviewer synthesizes these insights, proposing enhancements to the initial VQAs. The Critic Agent then consolidates this feedback for the VQA Expert to refine the questions and answers. The process concludes with human expert evaluation of a subset of the refined VQAs, assessing their clarity, depth, relevance, and answerability. This agentic annotation workflow mimics collaborative human expert discussions, leveraging diverse LLM perspectives to produce nuanced, critical, and comprehensive VQAs that probe deeper aspects of movie understanding. By employing multiple specialized agents in a structured, iterative process, we harness collective intelligence and mitigate potential biases of any single LLM. The final human validation ensures that the resulting VQAs meet the highest standards of quality and depth, representing a significant advancement over traditional LLM-assisted annotation methods.

3 Experiments

To evaluate the effectiveness of our dataset in engaging System 2 thinking and promoting deeper cognitive processing, we conduct a series of tests designed to assess the complexity, readability, and cognitive demand of our questions and answers. These tests include well-established metrics such as parse tree depth, Flesch-Kincaid grade score, and Bloom's taxonomy classification. Each provides unique insights into different aspects of our dataset's ability to stimulate higher-order thinking. Table 1 presents a comparative analysis of our MovieCOREagainst other VQA data sets on these metrics.

3.1 Parse Tree Depth (PTD)

Parse tree depth measures the syntactic complexity of sentences by analyzing their hierarchical structure. We utilize this metric to assess the structural intricacy of our questions and answers. We employ the spaCy library to generate parse trees for each question and answer in our dataset. The depth of these trees is then calculated and averaged across the dataset. A greater parse tree depth often correlates with more complex sentence structures, which typically require more cognitive resources to process. By measuring this, we aim to quantify the linguistic sophistication of our VQAs as compared to existing datasets', hypothesizing that questions and answers with higher parse tree depths are more likely to engage System 2 thinking. Table 1 shows that MovieCORE has the highest average parse tree depth (PTD) compared to the other VQA datasets.

3.2 Flesch-Kincaid Grade Score (FKG)

The Flesch-Kincaid Grade Score is a readability measure that indicates the U.S. grade level needed to understand a text. We calculate this score for both questions and answers in our dataset using the standard Flesch-Kincaid formula, which considers factors such as word length and sentence length.

$$\text{FK Grade Score} = 0.39 \left(\frac{\text{Total Words}}{\text{Total Sentences}} \right) + 11.8 \left(\frac{\text{Total Syllables}}{\text{Total Words}} \right) - 15.59 \quad (1)$$

While our goal is not to make the content unnecessarily difficult, a moderately high Flesch-Kincaid score indicates that the VQAs require a more advanced level of comprehension and thinking. MovieCORE substantially outperforms other datasets with an average grade score of 14.03, suggesting that its QAs demand college-level reasoning.

3.3 Bloom's Taxonomy Level (BTL)

Bloom's Taxonomy is a hierarchical model used to classify educational learning objectives into levels of complexity and specificity. We prompt GPT-4o-mini with a comprehensive breakdown of the Bloom's Taxonomy and ask it to classify each question and answer into one of six cognitive levels: Remember (1), Understand (2), Apply (3), Analyze (4), Evaluate (5), and Create (6). Such classification helps us assess the cognitive demand of the QAs. Questions falling into higher levels of Bloom's Taxonomy (Analyze, Evaluate, Create) require deeper analysis and critical thinking skills susceptible to trigger System 2 thinking. MovieCORE achieves the highest average Bloom Taxonomy Level (BTL) of 4.9, indicating that our questions and answers predominantly engage higher-order cognitive skills, significantly surpassing the other data sets. Additionally, we report the percentage of higher-order questions and answers (HO-QA), representig the proportion of both questions and answers that fall into the upper levels of Bloom's Taxonomy (levels 4-6). MovieCORE excels in this metric with 99.2% of its questions and answers classified as higher-order, far surpassing the next best dataset.

4 Evaluation Metrics and Assessment of Existing VQA models

VQA datasets usually use top-1 accuracy as metrics, but a valid match has to be a perfect match. For instance, there can be one strict answer to the question "Does sea appear in the video?", which is "Yes" or "No". However, in the age of LLMs and especially for zero-shot evaluation settings, we might get answers such as "it does" or "no sea appears in the video". In such cases the accuracy would be 0. Recently LLM-assisted evaluation schemes such as the one introduced by [10], attempt to solve this issue by considering synonyms or paraphrases as valid matches. This works for shallow QA datasets where there is a perfect answer, and would not work in our case, especially since accuracy for a system 2 answer is not binary but exists in a spectrum. Furthermore, we posit that accuracy alone is not sufficient, therefore we design four other LLM-assisted metrics: *depth* to assess the depth of reasoning in the answers, *comprehensiveness* to assess how fully the answer covers all key points and relevant details, *coherence and clarity*, and *evidence* to evaluate the quality and relevance of the evidence provided in the answers. For all of these metrics, we assign a score of 0 to 5 for each answer and report the average.

Table 2 presents a comparative performance evaluation of video question-answering models across five key criteria: Accuracy (Acc), Comprehensiveness (Comp), Depth, Evidence (Evid), and Coherence (Coh). Results are provided for both zero-shot and fine-tuned scenarios, highlighting the best scores in each category. The zero-shot results highlight a significant underperformance across all metrics, with only one of the models achieving satisfactory scores, reinforcing the challenge posed by deeper cognitive tasks in the absence of fine-tuning. These findings underscore the need for targeted datasets, like MovieCORE, that focus on deeper, System 2 reasoning to enhance model performance on complex video content.

5 Conclusion

We introduce MovieCORE, a novel VQA dataset that fills a critical gap in existing movie-based VQA datasets by emphasizing thought-provoking questions designed to engage System 2 thinking. Our agentic workflow, which leverages brainstorming agents with feedback loops, enables the generation and refinement of high-quality question-answer pairs. To measure the cognitive depth of VQA datasets, we devise a set of tests that demonstrate the superiority of MovieCORE over existing datasets. Additionally, we propose a comprehensive evaluation framework to assess the performance of VQA models on this challenging dataset.

Acknowledgement This work was supported in part by National Science and Technology Council, Taiwan, under Grant NSTC 112-2634-F-002-006.

References

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

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

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[5] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

[6] Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Winston H. Hsu, and Shang-Hong Lai. Hermes: temporal-coherent long-form understanding with episodes and semantics, 2024.

[7] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024.

[8] 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.

[9] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

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

[11] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[12] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

[13] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[14] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.

[15] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*, 2024.

[16] 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.

[17] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[18] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019.

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

A Appendix

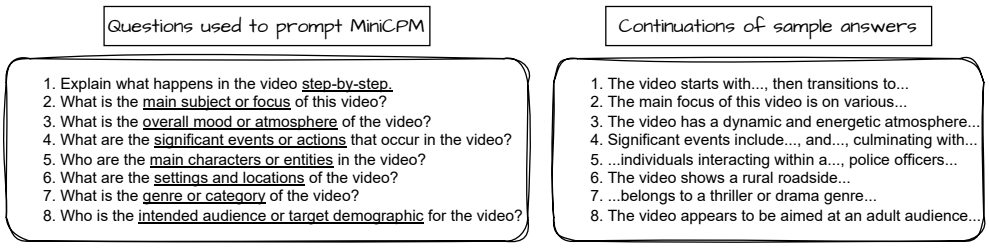

Figure 2: **Extracting Detailed Context from Videos:** We input each video to MiniCPM-v2.6, prompting it with a series of carefully crafted questions (left). The model's responses (right) provide rich, multi-faceted details about the video, including narrative flow, character information, setting, mood, and target audience. This extracted information serves as *Data Info* priors to inform our annotation agents, ensuring a comprehensive understanding of the video content before the VQA generation process.

A.1 Related Works

Video Question-Answering Datasets. Several datasets have been developed to address various aspects of video question-answering. [8, 16, 18] introduced datasets focusing on general video content, asking questions about events and characters. In the domain of movies and dramatic scenes, [14, 13, 12] created datasets that require models to comprehend dialogues and scene interactions. [11] moved towards deeper reasoning with more complex questions. Recently, [9, 19] explored challenges in cross-modal and causal reasoning.

Video Question-Answering Reasoning. While text-based datasets like DROP [5] and GSM8K [3] have promoted complex reasoning, similar advancements in video datasets are still emerging with substantial room for improvement in this area. Our work builds upon these foundations by introducing a dataset that emphasizes System-2 reasoning, focusing on intricate logical reasoning and abstract thinking in the context of movie content.

A.2 Dataset Card: MovieCORE

A.2.1 Dataset Overview

MovieCORE is a video question-answering (VQA) dataset designed to probe deeper cognitive understanding of movie content. It consists of annotated videos paired with corresponding questions and answers that emphasize System 2 thinking.

A.2.2 Statistics

- **Total Videos:** 986
- **Total QA Pairs:** 4,930
- **Human-Verified:**
  - 30 videos
  - 150 QA pairs

A.2.3 Human Verification

To ensure the quality of a subset of the dataset, we employed 7 human verifiers with at least a Bachelor's degree to validate 30 videos, and 150 QA pairs.

Table 3: Sample Question-Answer Pairs from MovieCORE

Question	Answer
In what ways do the characters' key actions reveal their motivations, and how do these actions develop throughout the narrative?	The characters' key actions, such as the police officer's search efforts and the woman's protective measures, reveal their motivations rooted in duty and safety. The police officer's persistence and visible distress indicate a deeply personal or professional investment in resolving the situation. Meanwhile, the woman's continuous vigilance and actions to protect her home suggest her motivation to keep her loved ones safe. These motivations evolve as the narrative progresses, with each character becoming more desperate and determined in their respective quests.
What deeper meanings can be inferred from the juxtaposition of a barren landscape with a single leaf, and how do these visuals contribute to the narrative?	The barren landscape with a single leaf represents the profound themes of desolation and resilience. It contrasts the current state of environmental decay with the potential for regeneration, emphasizing hope amidst seemingly insurmountable challenges. This visual metaphor enriches the narrative by suggesting that even in the bleakest circumstances, there exists a spark of life and renewal.
How does the video illustrate the duality of natural beauty through serene and chaotic moments, and what are the broader themes about nature that emerge from this contrast?	The video begins with the peaceful image of baboons grooming by a tranquil riverbank at sunset, embodying serene natural beauty. This is juxtaposed with chaotic scenes like the black dog aggressively leaping through water in pursuit of prey and hyenas engaging in rough play. These contrasting moments reveal the dual nature of the wilderness, where tranquility coexists with survival-driven chaos. Broader themes that emerge include the resilience of wildlife, the balance between life and death, and the inherent beauty in both the peaceful and volatile aspects of nature. This duality enriches the viewer's appreciation for the complexity and resilience of natural ecosystems.

A.2.4 Applications

MovieCORE is tailored for evaluating VQA models on deeper cognitive tasks. The inclusion of challenging, thought-provoking questions makes it ideal for advancing research in movie understanding and System 2 reasoning in AI systems.

A.2.5 Availability

We will release the dataset, along with a customizable agentic annotation pipeline and an evaluation scheme, to facilitate further research.

A.3 Limitations

While MovieCORE offers a significant advancement in video question-answering (VQA) by targeting deeper cognitive understanding, it is not without limitations. Although we incorporate human verification for a subset of the dataset, only 30 videos, and 150 QA pairs were manually verified. While this enhances quality control for a portion of the data, the majority of the dataset relies on automated processes. Furthermore, the dataset's reliance on the MovieChat-1k dataset may limit its genre diversity and focus. Certain movie genres or narrative styles might dominate, potentially making the dataset less representative of all types of cinematic content. This could restrict the generalizability of the dataset to broader contexts or other forms of video media. Despite these limitations, MovieCORE represents a critical step towards enhancing the depth and cognitive challenge of VQA tasks, and we believe it will drive further research and refinement in this area.