MovieCORE: COgnitive REasoning in Movies

Anonymous ACL submission

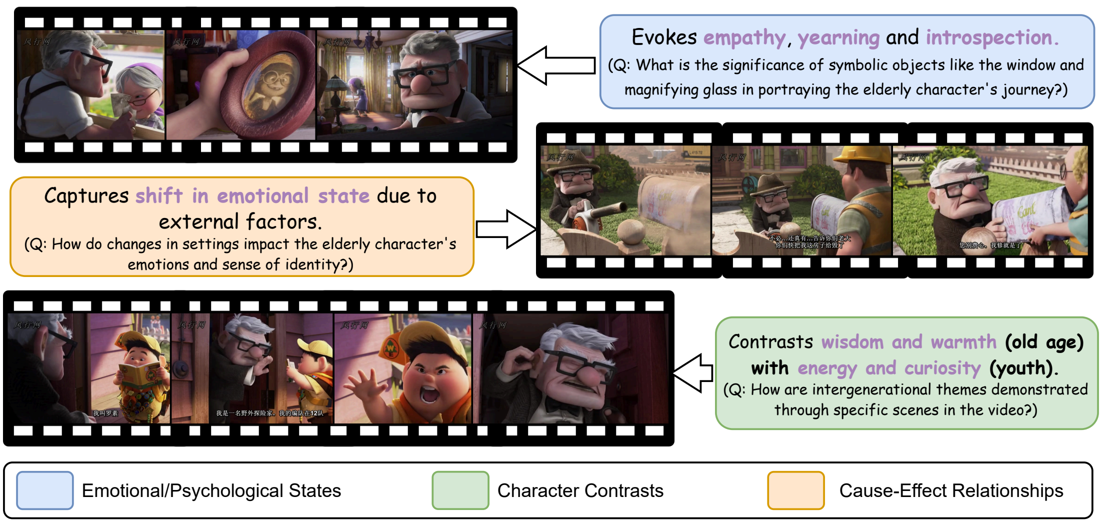


Figure 1: **Beyond Shallow Video Understanding:** The proposed benchmark, MovieCORE, challenges vision-language models (VLMs) to understand the subtle interplay between emotions *(Top, Middle)*, character dynamics and causality *(Middle, Bottom)*, and psychological complexity *(Top, Middle)*. From empathy to introspection, from wisdom to curiosity MovieCORE tests VLMs' ability to comprehend the deeper elements of movies.

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 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. To address the limitations of existing video-language models (VLMs), we introduce an agentic enhancement module, Agentic Choice Enhancement (ACE), which improves model reasoning capabilities post-training by 25%. 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, existing movie-based VQA datasets have yet to explore this aspect of film comprehension.

Recent movie-based VQA datasets (Wu and Krahenbuhl, 2021; Song et al., 2024; Rawal et al., 2024) 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 (Mangalam et al., 2023) attempts to delve beyond the obvious, its more profound questions often remain general.

We propose MovieCORE, a novel VQA dataset designed to engage System-2 thinking—the slow, deliberate, and logical cognitive processes—while maintaining strict relevance to specific video content. Unlike existing datasets, MovieCORE embraces the inherent subjectivity of "why" and "why not" questions as a feature rather than a limitation, creating both meaningful challenges and research opportunities. To generate comprehensive and faithful question-answer pairs, we develop an agentic brainstorming approach that leverages multiple large language models (LLMs) as interactive thought agents that engage in continuous discussions to refine QA pairs. We validate the quality of the QAs through rigorous human review of a representative subset. Additionally, we employ quantitative cognitive metrics to measure our dataset's depth and syntactic complexity relative to existing benchmarks. Our evaluation of current VQA models on MovieCOREreveals critical insights about their performance on these challenging cognitive tasks. To address identified limitations and improve existing VLMs' deeper cognitive reasoning capabilities, we introduce Agentic Choice Enhancement(ACE), which demonstrates relative performance improvements of up to 25% compared to baseline approaches.

Our key contributions are the following:

- We introduce MovieCORE, a VQA dataset focused on thought-provoking questions and answers specific to movie content.
- We develop an agentic brainstorming approach using multiple LLMs as agents to generate and refine high-quality QA pairs.
- We implement a set of cognitive tests to evaluate the depth, thought-provocation, and complexity of VQA datasets.
- We design a comprehensive evaluation scheme to assess the accuracy, comprehensiveness, depth, and coherence of answers from existing Video Language Models (VLMs).
- We evaluate several VLMs on our dataset in both zero-shot and fully-supervised settings, offering insights into their performance on deeper cognitive tasks.
- We propose a post-training "agentic selection" plugin to improve existing VLMs and show a relative improvement of up to 25% compared to the baseline.

2 Related Work

Movie-Based Question-Answering Datasets. Recent video understanding benchmarks are often based on movie scenes because films offer a rich blend of multimodal content, combining visual, linguistic, and temporal elements within complex narratives. Early efforts like MovieQA (Tapaswi et al., 2016) explores entire movie understanding but were limited by questions heavily relying on dialogue. TVQA (Lei et al., 2018) requires reasoning over multiple events in short TV series clips, integrating visuals and subtitles. LVU (Wu and Krahenbuhl, 2021) addresses scaling video comprehension to extended sequences, necessitating models to process long temporal contexts. MAD (Soldan et al., 2022) and its extension (Han et al., 2023) focus on scene-level descriptions through audio and visuals but were mainly used for scene annotation tasks with limited narrative comprehension. MoVQA (Zhang et al., 2023) introduces multi-level questions, challenging models in temporal perception, causal reasoning, and narrative synthesis. CinePile (Rawal et al., 2024) automates large-scale question generation across varied scenes and question type and MovieChat-1k (Song et al., 2024) focuses on basic understanding of cinematic contexts.

Video Question-Answering Reasoning. Text-based reasoning datasets like DROP (Dua et al., 2019) and GSM8K (Cobbe et al., 2021) handle discrete reasoning tasks, including counting and arithmetic, but are limited to textual inputs and do not address the complexities involved in integrating visual reasoning. Egocentric datasets, such as EpicKitchens (Damen et al., 2018), Ego4D (Grauman et al., 2022), and EgoSchema (Mangalam

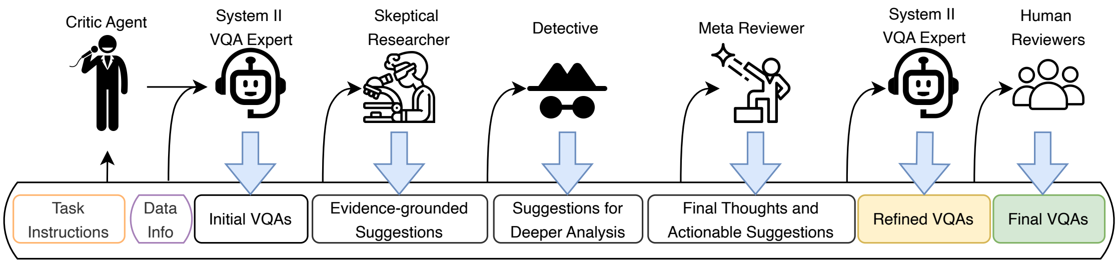


Figure 2: The *Critic Agent*, acting as the master of ceremonies (MC), orchestrates interactions among specialized agents using video context and task instructions. It sequentially engages the *System II VQA Expert*, *Skeptical Researcher*, *Detective*, and *Meta Reviewer*, accumulating insights at each stage. Upon receiving final recommendations from the *Meta Reviewer*, the MC relays them to the *System II VQA Expert* for VQA refinement. Subsequently, a subset of these refined VQAs undergoes evaluation by human experts for final validation.

et al., 2023), challenge models to interpret subjective interactions and continuous activities from a first-person perspective, requiring both perceptual understanding and intention reasoning. Perception Test (Patraucean et al., 2024) broadens perceptual reasoning to varied video contexts, assessing high-level reasoning abilities. Multi-task and complex video benchmarks, such as MVBench (Li et al., 2024), Video-MME (Fu et al., 2024), and MLVU (Zhou et al., 2024), integrate multiple reasoning challenges, requiring predictive reasoning, memory recall, and cross-modal inference over long video sequences. While these datasets have advanced various aspects of video understanding, they predominantly rely on surface-level comprehension of video content. Our work introduces the first dataset specifically designed to evaluate System-2 reasoning in the video domain, requiring models to engage in slow, deliberate, and analytical thinking processes aiming to mirror human approaches to complex movie understanding.

3 MovieCORE Creation and Curation

To address the challenge of obtaining question-answer pairs that delve into deeper levels of movie understanding, we propose an agentic annotation workflow. This approach leverages the deliberative capabilities of multiple 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.

3.1 Video Context Extraction

The videos for our dataset are sourced from MovieChat-1k (Song et al., 2024), 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, 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 captions, we observe inaccuracies and imbalanced descriptions. Therefore, we exclude these captions, focusing instead on the existing QA pairs and movie metadata.

To provide video context, we utilize MiniCPM-v2.6 (Yao et al., 2024), 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 S1 in the supplementary material) 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 agents.

3.2 Agentic Annotation Workflow

Our workflow, illustrated in Figure 2, employs a multi-agent system orchestrated by a Critic Agent acting as the master of ceremonies (MC). Using the Agentic AI framework autogen (Wu et al., 2024), we deploy instances of GPT4-o for the VQA Expert and Meta Reviewer roles (as these positions demand superior reasoning capabilities), 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 3.1 and sends them to the System II VQA Expert who generates questions that engage System-2 thinking. These initial QA pairs are then scrutinized by the Skeptical Researcher,

Single-Pass Annotation:

Q: How does the interaction between the two main characters (the one in red and black costume and the one in black with cat-like ears) evolve throughout the video, and what might this suggest about their relationship?
A: The interaction begins with a casual encounter and evolves into a clear partnership as they traverse the city together. This progression in their relationship suggests a deepening bond based on mutual goals or shared challenges, hinting at themes of teamwork and trust.

Agentic Annotation:

Q: Can you provide **specific scenes** that demonstrate the **evolution** and **motivations** of the main characters in **their relationship**?
A: Yes, there is a scene where **the character in the red and black costume helps the one in black with cat-like ears from falling, showing a bond of trust forming**. Another scene depicts **them working seamlessly together to navigate a tricky rooftop, reflecting their growing partnership**. **Their motivations seem rooted in overcoming a common threat**, as **further evidenced by the scene where they both face the glowing purple heart-shaped entity.**

Figure 3: **Comparison of single-pass and agentic annotation**. The agentic method (bottom) elicits specific scene details, concrete examples, and detailed story elements, demonstrating the enhanced granularity achieved through multi-agent refinement. Text in blue indicates new, specific details absent in the single-pass version. The single-pass annotation (top), on the other hand, while also attempting to ask deeper questions, remains at a more abstract level.

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 and biases. 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 QAs. 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 by harnessing collective intelligence and mitigating potential biases of any single agent[1].

To ensure the quality and reliability of our dataset, we implement a rigorous human verification process. Seven graduate students were recruited to assess a subset of 30 videos, 30 captions and 150 QA pairs. The final human validation ensures that the resulting VQAs meet the highest standards of quality and depth. We provide more details on the human validation in Appendix II.3 of the Supplementary material.

[1]Wondering why we chose these specific agents? Please see Appendix II.4 and II.5

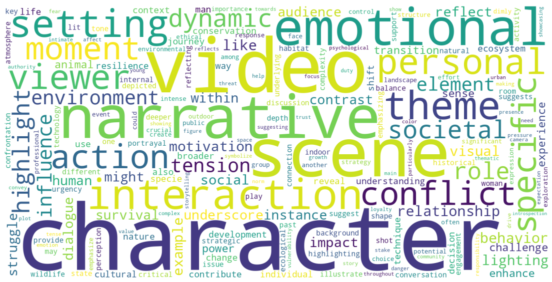


Figure 4: Wordcloud illustrating key themes and concepts of MovieCORE with terms such as "emotional", "character" and "influence" very prominent.

3.3 Agentic versus Single-Pass Annotation

To illustrate the effectiveness of the proposed Agentic Annotation workflow, we compare the quality of the VQAs generated by the System II VQA Expert in the initial round (single-pass) and those produced through our workflow after the agent has gathered feedback and enhancement ideas from other experts (agentic annotation). As shown in Figure 3, the agentic annotation approach demonstrates clear advantages over single-pass annotation. While the single-pass annotation provides a general, abstract description of character relationships, the agentic annotation generates questions that ask for and answers that deliver specific, concrete details about key scenes that support the relationship develop-

Dataset	Parse Tree Depth			F–K Grade Score			BT Level	HO-QA (%)
	Q	A	Avg	Q	A	Avg		
MovieChat-1k (Song et al., 2024)	3.58	1.31	2.45	3.19	-0.39	1.4	1.8	0.0
ActivityNetQA (Yu et al., 2019)	4.24	0.27	2.26	2.69	0.98	1.84	1.9	0.2
MVBench (Li et al., 2024)	3.96	1.71	2.84	4.74	1.47	3.11	2.2	3.4
EgoSchema (Mangalam et al., 2023)	**6.56**	<u>4.38</u>	<u>5.47</u>	<u>10.52</u>	<u>6.08</u>	<u>8.30</u>	<u>3.1</u>	<u>33.1</u>
MovieCORE	<u>5.38</u>	**6.39**	**5.88**	**12.98**	**15.07**	**14.03**	**4.9**	**99.2**

Table 1: **Syntactic Complexity and Cognitive Demand Comparison:** Parse tree depth, Flesch-Kincaid (F-K) grade scores, average Bloom's Taxonomy (BT) level, and percentage of higher-order questions and answers (HO-QA) across various VQA datasets. Q and A represent questions and answers respectively. Best results are in **bold**, second-best are <u>underlined</u>.

ment of the characters - including the falling scene, rooftop navigation, and confrontation with the purple heart-shaped entity. The agentic process elicits richer context and more granular evidence, making the annotations more specific and faithful to the movie content. It also makes the dataset much more valuable for training and evaluating AI systems' understanding of narrative progression and character dynamics. This suggests that using multiple AI agents as thought partners leads to more detailed and substantive annotations compared to traditional single-pass methods used by other auto-annotated datasets such as (Rawal et al., 2024) and (Mangalam et al., 2023). More comparisons between agentic and single-pass annotation can be found in Supplementary (Appendix II.4).

3.4 Dataset Description

MovieCORE is a video question-answering (VQA) dataset designed to probe deeper cognitive understanding of movie content. The dataset comprises 986 videos paired with 4,930 corresponding questions and answers and 986 captions. Following the splits of the original MovieChat-1k dataset (Song et al., 2024), we split MovieCORE into 4080 QAs for training (816 videos) and 850 for testing (170 videos). The primary application of MovieCORE lies in training and evaluating VQA models' capabilities in deeper cognitive tasks. The questions are specifically designed to assess models' abilities to comprehend complex narrative elements, character motivations, and subtle contextual cues – skills that are crucial for achieving human-like understanding of cinematic content. A wordcloud of MovieCORE is shown in Figure 4 suggesting complex themes regarding character dynamics, emotional resonance, and societal implications through terms like "tension," "psychological," "cultural," and "emotional." Also, the prominence of analytical terms such as "underscore","depth," and "critical," suggests questions that probe deeper interpretations and thematic elements rather than just literal plot descriptions.

4 Experiments

4.1 Linguistic and Cognitive Complexity

To evaluate the effectiveness of MovieCORE 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 MovieCORE against other VQA datasets.

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 and recursively compute their depth as follows. Let $d(t)$ be the depth of a token t in the tree. For a token with children $C(t)$, the depth is defined as:

$$d(t) = \begin{cases} 0 & \text{if } C(t) = \emptyset \\ 1 + \max_{c \in C(t)} d(c) & \text{if } C(t) \neq \emptyset \end{cases} \quad (1)$$

where $d(t) = 0$ if t is a leaf node (no children), $d(t) = 1 + \max_{c \in C(t)} d(c)$ if t has children $C(t)$, with $\max_{c \in C(t)} d(c)$ representing the maximum depth of the children of t. For a sentence with multiple tokens, the depth of the parse tree D rooted at the token r (root of the sentence) is $D = d(r)$.

The depth of these trees are then 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 compared to the other VQA datasets.

The Flesch-Kincaid (F-K) 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 below

$$\text{F-K Grade Score} = 0.39\left(\frac{W}{S}\right) + 11.8\left(\frac{Y}{W}\right) - 15.59 \tag{2}$$

where W is the total number of words in the text, S, total number of sentences and Y the total number of syllables.

While our goal is not to make the content unnecessarily difficult, a moderately high Flesch-Kincaid score indicates that the QAs require a more advanced level of comprehension and thinking. As shown in Table 1, MovieCORE substantially outperforms other datasets with an average grade score of 14.03, with its closest competitor – EgoSchema (Mangalam et al., 2023) – standing at 8.3.

Bloom's Taxonomy is a hierarchical model used to classify educational learning objectives into levels of complexity and specificity (Mcdaniel, 1970). 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 (BT Level) of 4.9, indicating that our questions and answers predominantly engage higher-order cognitive skills, significantly surpassing the other datasets. Additionally, we report the percentage of higher-order questions and answers (HO-QA), representing 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.

Algorithm 1 ACE: Agentic Choice Enhancement

1: **Input**: Video V, Question Q, Beam width $k = 5$
2: **Output**: Best response R^*
3: $C \leftarrow$ VLM.generate$(V, Q, \text{beam_width} = k)$
4: $S \leftarrow$ Llama-3.2.score(C) ▷ Score candidates
5: $R^* \leftarrow \arg\max_{c \in C} S(c)$ ▷ Select best response
6: **return** R^*

5 ACE: Agentic Choice Enhancement

We propose ACE, a straightforward yet effective approach to improving existing video language model (VLM) outputs through post-generation refinement. Our approach, detailed in Algorithm 1, uses an existing VLM and leverages beam search with a width of 5 to generate diverse candidate responses, which are then re-ranked using the compact 1B-parameter Llama-3.2 (MetaAI, 2024) language model. We hypothesize that, when engaging in a task requiring deeper deliberation, it is advisable to have a second pair of eyes to refine one's thinking. The lightweight nature of Llama-3.2 (1B) ensures that this enhancement remains computationally efficient while significantly improving the quality of generated responses. We prompt the model without specific evaluation guidelines, allowing it to leverage its inherent understanding of "answer quality". Table 2 show that this "agentic selection" approach paired with HERMES (Faure et al., 2024) (HERMES + ACE) registers an absolute gain of 0.48 compared to the baseline VLM, which translates to roughly a 16 percent improvement in answer quality. It also improves InstructBLIP (Dai et al., 2023) by 25% (2.63→3.29) and MA-LMM (He et al., 2024) by 20% (2.79→3.35). These results suggest that existing VLMs have untapped potential that can be realized through a simple post-generation "second pair of eyes" strategy, offering a practical path to training-free improvement.

Table 3 shows similar performance across beam widths (3, 5 and 7) for HERMES, suggesting ACE's effectiveness stems from the agentic selection mechanism itself rather than hyperparameter choices. These results validate our framework's fundamental premise: lightweight post-generation refinement can unlock significant untapped potential in existing VLMs.

Model	Accuracy	Comprehensiveness	Depth	Evidence	Coherence	Avg.
Proprietary Models						
Gemini-1.5-pro	3.91	3.81	3.90	3.87	3.79	3.86
GPT-4o (08-06)	**4.18**	**4.00**	**3.98**	**3.96**	**3.96**	**4.02**
Zero-Shot Results						
InstructBlip (Dai et al., 2023)	1.03	0.43	0.85	0.33	0.40	0.61
MA-LMM (He et al., 2024)	1.14	0.63	0.93	0.57	0.67	0.79
HERMES (Faure et al., 2024)	1.77	1.21	1.41	1.28	0.37	1.41
LongVU (Shen et al., 2024)	2.95	2.01	1.94	2.06	2.12	2.22
InternVL2 (IntenVL, 2024)	**3.80**	**3.42**	**3.10**	**3.37**	**3.51**	**3.44**
Fully-Supervised Results						
InstructBlip (Dai et al., 2023)	3.25	2.43	2.47	2.61	2.38	2.63
MA-LMM (He et al., 2024)	3.42	2.54	2.66	2.81	2.50	2.79
HERMES (Faure et al., 2024)	**3.52**	**2.72**	**2.83**	**2.98**	**2.62**	**2.93**
Fully-Supervised Results + ACE (Ours)						
InstructBlip (Dai et al., 2023)	3.71	3.15	3.02	3.30	3.25	3.29 (+0.66)
MA-LMM (He et al., 2024)	3.76	3.24	3.09	**3.39**	3.30	3.35 (+0.56)
HERMES (Faure et al., 2024)	**3.81**	**3.30**	**3.12**	3.38	**3.42**	**3.41** (+0.48)

Table 2: **Performance Comparison of Video Question-Answering Models.** We evaluate various open-source and proprietary Vision-Language Models (VLMs) on five criteria: Accuracy, Comprehensiveness, Depth, Evidence, and Coherence. Zero-shot results, in particular, highlight significant limitations in multi-step inference and evidence gathering, indicating these models often fail to piece together context from complex sequences.

w/ ACE	Acc.	Com.	Dep.	Evi.	Coh.	Avg.
Beam=3	3.81	3.40	3.19	3.42	3.43	3.45
Beam=5	3.81	3.30	3.12	3.38	3.42	3.41
Beam=7	3.79	3.29	3.08	3.36	3.35	3.37

Table 3: ACE improves performance across all evaluation dimensions regardless of the beam size.

6 Quantitative Evaluation

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 (Maaz et al., 2023), attempt to solve this issue by considering synonyms or paraphrases as valid matches. This works for VQAs 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 insufficient, 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. For all of these metrics, we prompt GPT-4o-mini (OpenAI, 2024) to assign a score between 0 to 5 to each.

Table 2 presents a comprehensive evaluation of model performance across our five assessment criteria. Several key insights emerge from these results: (1) Proprietary models significantly outperform their open-source counterparts. This performance gap indicates that large-scale proprietary training data likely contains more diverse reasoning tasks than those available in public datasets. (2) In the zero-shot setting, most open-source models struggle considerably with complex reasoning, except InterVL-2. The particularly low scores in Depth and Evidence metrics highlight these models’ difficulty in formulating multi-step inferences and grounding their responses in specific visual content. (3) Fine-tuning on MovieCORE yields substantial improvements for all models, with HERMES showing the strongest performance. However, even with full supervision, these models still underperform compared to proprietary alternatives, suggesting architectural limitations in handling complex reasoning tasks. (4) Our proposed ACE post-generation strategy delivers consistent and substan-

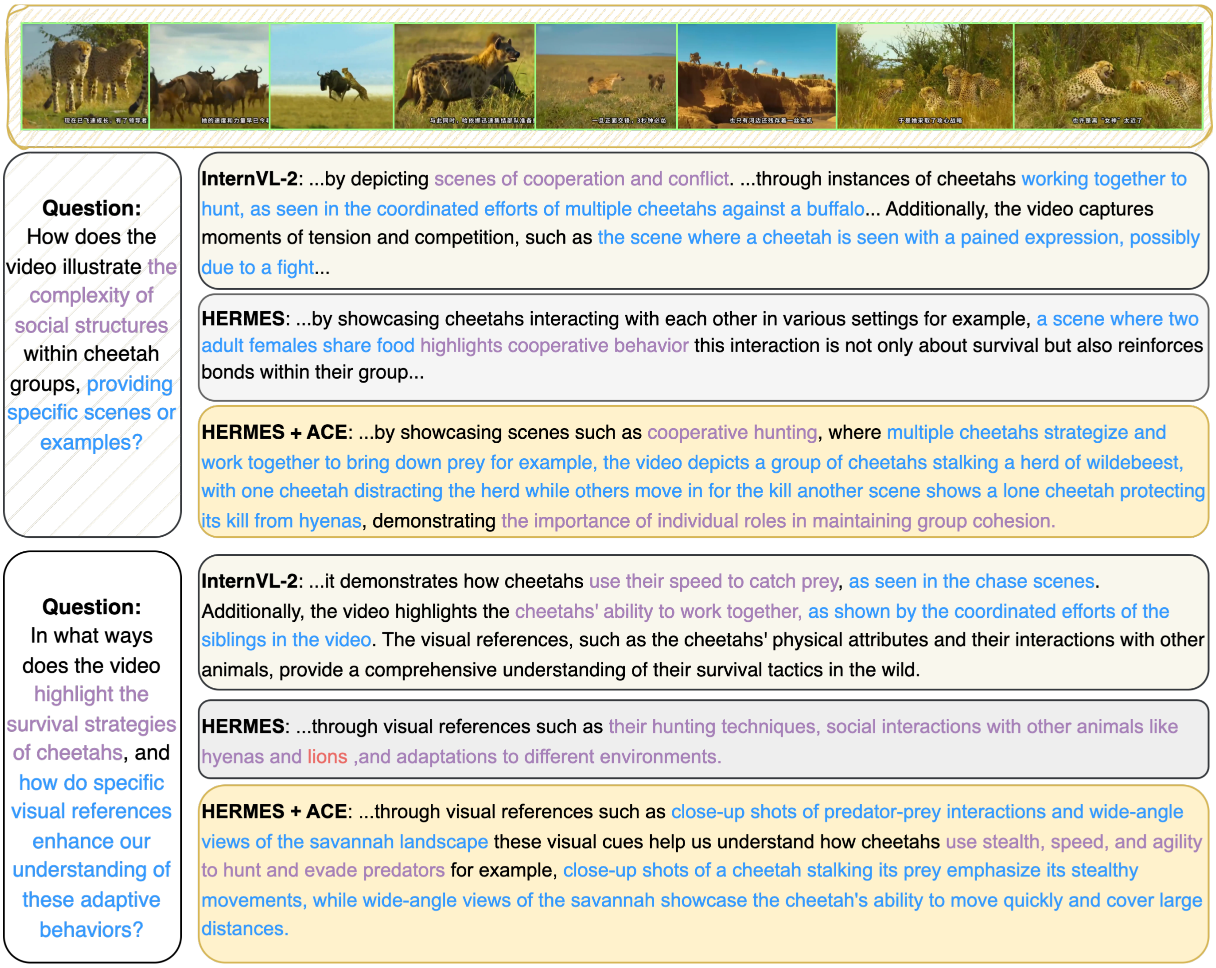


Figure 5: **Qualitative Comparison of Model Responses**. This figure contrasts responses from InternVL-2 (zero-shot), HERMES (fully-supervised), and HERMES+ACE on two questions about cheetah behaviors. Purple text highlights conceptual understanding while blue text indicates specific visual evidence and contextual details. Note how ACE enhances responses with more precise scene descriptions and behavioral insights.

tial improvements across models and metrics.

7 Qualitative Results

Figure 5 provides a qualitative comparison between different models' responses to questions that require understanding of complex animal behaviors. The figure illustrates how different approaches handle the same queries about cheetah social structures and survival strategies. InternVL-2, a strong zero-shot model, provides basic observations but lacks sufficient depth and details. HERMES, a fully-supervised model, also struggles with the details and performs worse than InternVL. HERMES+ACE, demonstrates enhanced response quality by incorporating specific visual evidence and richer contextual details. As highlighted in the responses, ACE significantly improves the model's ability to reference specific scenes and provide concrete examples to support its assertions.

8 Conclusion

We introduce MovieCORE, a novel VQA dataset that fills a critical gap in existing movie-based VQA datasets by emphasizing questions designed to engage System-2 thinking. Our agentic workflow, which leverages brainstorming agents, enables the generation and refinement of high-quality QA 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 dataset. To tackle the challenges posed by MovieCORE, we propose ACE, a lightweight inference-time agentic answer selection plug-in which yields up to 25% relative improvement in answer quality compared to baseline methods, providing insights for future works on this topic.

9 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.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736.

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

Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Winston H. Hsu, and Shang-Hong Lai. 2024. Hermes: temporal-coherent long-form understanding with episodes and semantics. *Preprint*, arXiv:2408.17443.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. 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.

InternVL. 2024. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

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

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

Rhett Mcdaniel. 1970. Bloom's taxonomy. `https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/`.

MetaAI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models — ai.meta.com. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. [Accessed 03-11-2024].

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. [Accessed 01-11-2024].

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2024. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. 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.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. 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.

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

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

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. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

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

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

The Supplementary material is organized as follows:

I Reproducibility Statement

The dataset will be made public as soon as this paper is accepted (or rejected) for publication, as well as the evaluation scheme with clear examples. We will also release the annotation agents used for generating and refining question-answer pairs, including the code and configurations for the large language models (LLMs) employed in the agentic brainstorming process. Additionally, we provide detailed instructions for data preprocessing, agent configuration, and evaluation protocols, enabling reproduction of both the dataset generation process and the evaluation scheme. Our annotation system is scalable and has the potential to inspire other researchers to create massive video benchmarks.

II More Details on MovieCORE

II.1 Extracting "Video Info"

To generate meaningful interpretations of video content, we employ a structured question framework designed to probe various aspects of the video's narrative, emotional tone, and intended purpose. This framework consists of eight prompts, each targeting specific dimensions of video understanding. The prompts and a continuation of the sample answers they elicit are listed in Figure S1 and roughly contains the following:

1. **Step-by-step explanation:** Encourages a chronological breakdown of events in the video.
2. **Main subject or focus:** Identifies the central theme or entity in the video.
3. **Overall mood or atmosphere:** Captures the emotional tone conveyed by the video.
4. **Significant events or actions:** Highlights key actions and turning points within the narrative.
5. **Main characters or entities:** Focuses on the individuals or groups driving the video's story.
6. **Settings and locations:** Explores the physical or contextual backdrop of the video.
7. **Genre or category:** Classifies the video into a relevant category or type.
8. **Intended audience:** Identifies the target demographic for the video.

II.2 Agentic Annotation Details

Figure S2 depicts the system messages for the different agents involved in the task of creating system-2 thinking VQAs from system-1 VQAs. The agents and their respective roles are:

System-2 Video Question Answering Assistant Responsible for generating up to five system-2 thinking VQA pairs from the given system-1 VQAs. The focus is on creating questions and answers that encourage deeper analysis, critical thinking, and meaningful reflection, while ensuring the insights are grounded in the actual video content.

Critic Agent Evaluates the system-2 VQAs created by the System-2 Video Question Answering Assistant and passes them to various Expert Agents for detailed analysis. The Critic Agent then compiles the constructive feedback from the experts and returns it to the System-2 Video Question Answering Assistant, emphasizing the importance of aligning the VQAs with the actual video context.

Skeptical Researcher Reviews the questions and answers in the context of the video, analyzing the context and evaluating the system-2 VQAs for their contextual relevance and accuracy. The Skeptical Researcher challenges the assumptions behind the QAs and encourages further evidence-based exploration, providing concise and relevant suggestions.

Detective Given the video information and the system-2 VQAs, the Detective identifies additional questions that could uncover underlying causes, motivations, or potential biases. The suggestions should be concise, realistic, and directly relevant to the video's actual content.

Meta Reviewer Aggregates the feedback and suggestions from all reviewers (Skeptical Researcher, Detective) and provides final insights and suggestions to refine and improve the system-2 VQAs. The Meta Reviewer ensures the feedback

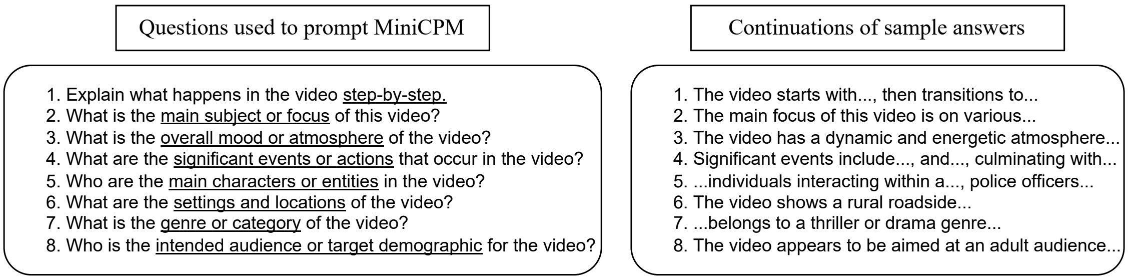


Figure S1: **Extracting Detailed Context from Videos:** We input each video to MiniCPM-v2.6, prompting it with a series of carefully crafted questions (left). The model's responses (right) provide rich, multi-faceted details about the video, including narrative flow, character information, setting, mood, 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.

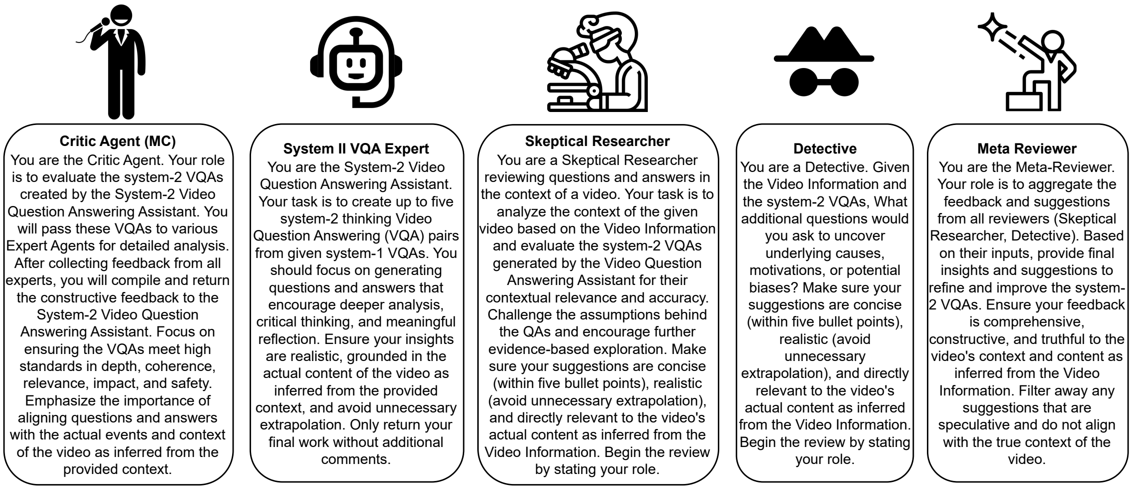


Figure S2: System Messages for the Annotation Agents

is comprehensive, constructive, and truthful to the video's context and content, filtering out any speculative suggestions.

II.3 Human Verification

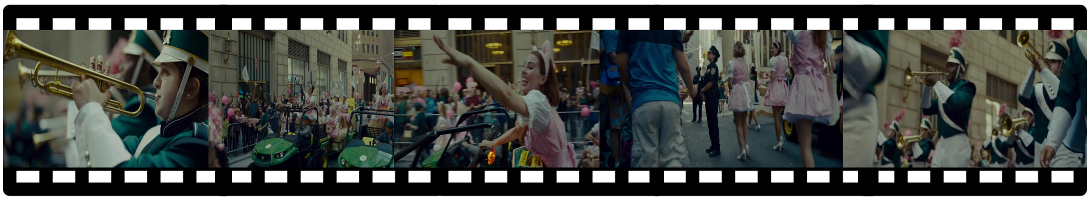
Verification Rules To ensure the quality and reliability of our dataset, we implemented a rigorous human verification process. Seven qualified evaluators, each holding at least a Bachelor's degree, were recruited to assess a subset of 30 videos and 150 QA pairs. The verification was conducted through a standardized evaluation form (Figure S4) that assessed four key dimensions:

- **Relevance (1-5):** Evaluates how directly the question/answer relates to the video content
- **Clarity (1-5):** Measures the linguistic clarity and absence of ambiguity
- **Depth (1-5):** Assesses the level of cognitive analysis required
- **Answerability (1-5):** Determines whether the question can be answered solely from the video content

As for the captions, we assessed accuracy, clarity and depth.

Evaluators were instructed to watch each video in its entirety and carefully consider the scenes, characters, actions, and dialogues before rating the associated QA pairs. To maintain objectivity, evaluators were required to focus solely on the video content when reviewing the QA pairs and encouraged to replay videos when necessary. The evaluation process also included assessing the accuracy and clarity of video captions to ensure comprehensive content accessibility.

Verification Result The human verification process (the rules and interface are illustrated in Figure S4) yields consistently high scores across all evaluated dimensions, as shown in Table S1. Ques-



Question: What cultural or historical references are present in the parade scenes, and what significance do they add to the narrative?

Answer: The parade scenes likely feature cultural or historical motifs through costumes, floats, and performances that could reflect specific traditions, societal events, or historical periods. These elements enrich the narrative by contextualizing the celebratory aspects within a broader cultural framework, offering viewers deeper insights into the world of the characters. Understanding these references can also underscore themes of heritage, identity, and the collective human experience, enhancing the video's relevance and emotional impact.

Figure S3: A parade scene from MovieCORE featuring various cultural and historical elements. This particular QA receives low answerability and relevance scores from one of our reviewers but was still kept following thorough review by a human meta-reviewer.

Metric	Captions	Questions	Answers
Accuracy	3.9	–	–
Clarity	4.0	4.3	4.3
Depth	4.1	4.5	4.2
Relevance	–	4.0	3.8
Answerability	–	3.8	4.1

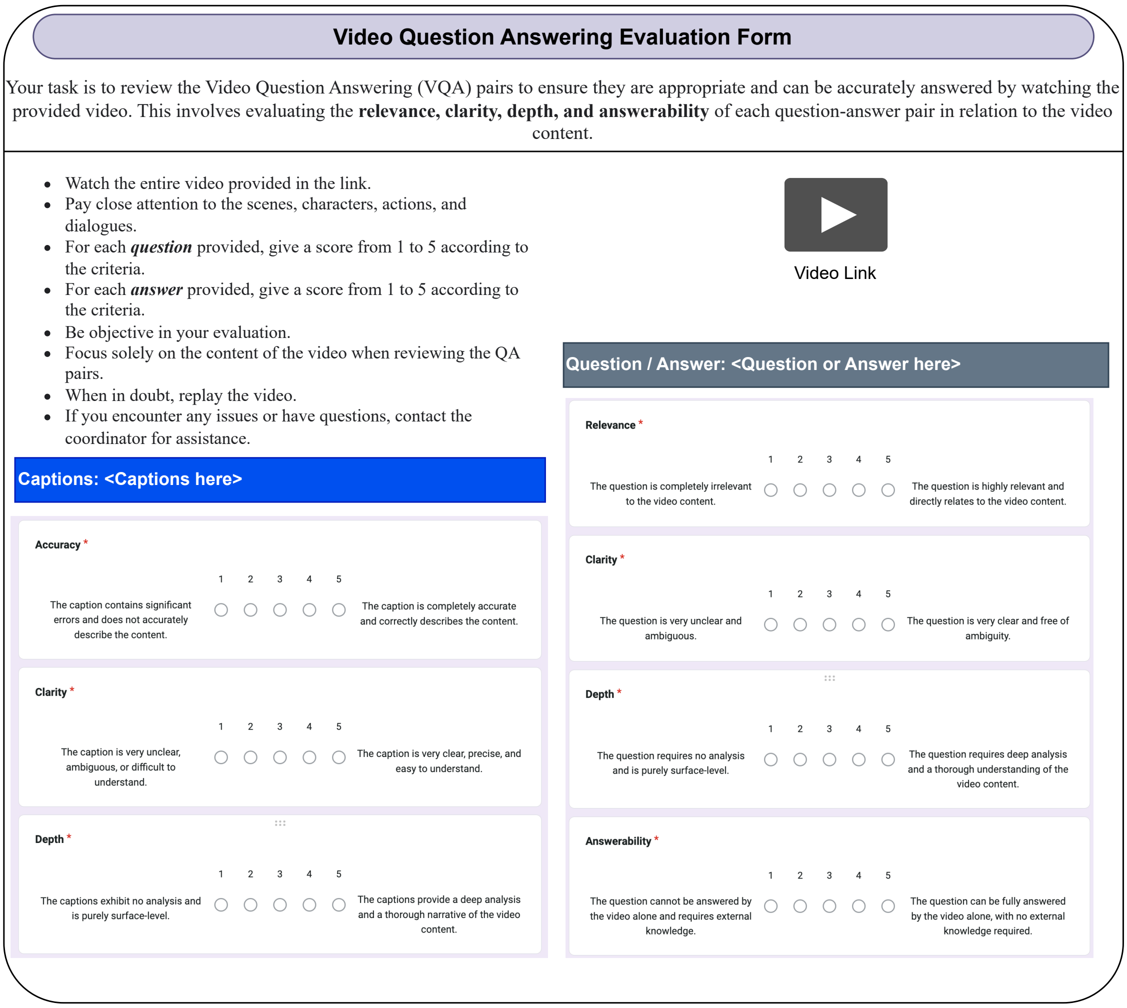
Table S1: Human verification scores across different dimensions for captions, questions, and answers. Scores range from 1 to 5, with 5 being the highest quality. Dashes (–) indicate metrics not applicable to that content type. The scores, being above 3.8 indicate strong quality across all evaluated dimensions.

tions and answers received notably high scores in clarity (4.3) and depth (4.5 and 4.2 respectively), validating our dataset's emphasis on deep cognitive understanding. The captions also demonstrate strong quality with scores above 3.8 across applicable metrics. While answerability scores were slightly lower (3.8 for questions), they remain well above acceptable thresholds, confirming that the questions can be reasonably answered from the video content alone.

The sample QA pair for the video depicted in Figure S3 received low scores of 2 each for Answerability and Relevance from the human evaluators. However, our human meta-reviewer has determined that the question and answer offer meaningful insights and contextual relevance (underlined in the figure).

II.4 Agentic versus Single-Pass Annotation

As shown in Figure S5, the single-pass annotation provides a general interpretation of the themes suggested by the presence of the hippopotamus, focusing on human-animal conflict and critiques of captivity. In contrast, the agentic annotation delves deeper by exploring how the hippopotamus functions as a symbol throughout the video, detailing its evolution from a chaotic force to a representation of innocence and victimhood. This nuanced analysis offers specific, concrete details about the symbolic transformation, enhancing the understanding of the narrative's thematic complexity. In the other example shown in Figure S6, the single-pass annotation mentions general visual and narrative elements like close-ups and quick scene transitions to build suspense. The agentic annotation specifies how visual techniques such as dramatic lighting, shadow play, and strategic camera angles enhance the emotional weight and suspense of key scenes. By providing detailed examples—like capturing a character's raw emotion through close-ups or creating an ominous atmosphere with dim lighting—the agentic approach offers a more granular and faithful depiction of the cinematic techniques used. These comparisons further illustrate that the agentic annotation process elicits richer context and more detailed evidence, reinforcing the idea that using multiple AI agents as thought partners leads to more substan-



Video Question Answering Evaluation Form

Your task is to review the Video Question Answering (VQA) pairs to ensure they are appropriate and can be accurately answered by watching the provided video. This involves evaluating the **relevance, clarity, depth, and answerability** of each question-answer pair in relation to the video content.

- Watch the entire video provided in the link.
- Pay close attention to the scenes, characters, actions, and dialogues.
- For each ***question*** provided, give a score from 1 to 5 according to the criteria.
- For each ***answer*** provided, give a score from 1 to 5 according to the criteria.
- Be objective in your evaluation.
- Focus solely on the content of the video when reviewing the QA pairs.
- When in doubt, replay the video.
- If you encounter any issues or have questions, contact the coordinator for assistance.

Video Link

Captions: <Captions here>

Accuracy *
1 2 3 4 5
The caption contains significant errors and does not accurately describe the content. — The caption is completely accurate and correctly describes the content.

Clarity *
1 2 3 4 5
The caption is very unclear, ambiguous, or difficult to understand. — The caption is very clear, precise, and easy to understand.

Depth *
1 2 3 4 5
The captions exhibit no analysis and is purely surface-level. — The captions provide a deep analysis and a thorough narrative of the video content.

Question / Answer: <Question or Answer here>

Relevance *
1 2 3 4 5
The question is completely irrelevant to the video content. — The question is highly relevant and directly relates to the video content.

Clarity *
1 2 3 4 5
The question is very unclear and ambiguous. — The question is very clear and free of ambiguity.

Depth *
1 2 3 4 5
The question requires no analysis and is purely surface-level. — The question requires deep analysis and a thorough understanding of the video content.

Answerability *
1 2 3 4 5
The question cannot be answered by the video alone and requires external knowledge. — The question can be fully answered by the video alone, with no external knowledge required.

Figure S4: **Video Question Answering Evaluation Form used in our human verification process.** The form assesses four critical dimensions (relevance, clarity, depth, and answerability) on a 5-point scale. Each dimension is clearly defined with anchored endpoints to ensure consistent evaluation. The form includes sections for both question/answer assessment and caption verification to ensure comprehensive content quality. Evaluators use this standardized form to systematically review each QA pair while referring to the corresponding video content.

Single-Pass Annotation:

Q: What underlying themes can be inferred from the presence of the large hippopotamus and the reactions of the people around it?
A: The presence of the large hippopotamus in an indoor setting, combined with the chaos and panic among the people, suggests themes of human-animal conflict and perhaps a critique of captivity or mistreatment of animals. The transition to a more focused narrative on animal rights, indicated by discussions around a tablet displaying disturbing images, highlights a deeper thematic concern with ethical treatment and possibly animal liberation.

Agentic Annotation:

Q: How does the hippopotamus function as a symbol throughout the video, and **how does its portrayal evolve**?
A: Initially, the hippopotamus is portrayed as a **chaotic and disruptive force**, symbolizing the **intrusion of wild nature into human spaces**. **As the narrative unfolds and the characters engage with the animal rights theme**, the hippopotamus evolves into a symbol of **innocence and victimhood**, representing the **broader plight of animals in captivity**. This shift is emphasized through changes in cinematography, with **early chaotic scenes giving way to more empathetic** and focused portrayals, reflecting the growing recognition of the animal's plight.

Figure S5: **Additional Comparison of single-pass and agentic annotation**. The agentic method (bottom) delves into specific scene details, such as the hippopotamus's evolution from a chaotic force to a symbol of innocence, and highlights changes in cinematography that reflect this transformation. The single-pass annotation (top) provides a general interpretation of themes like human-animal conflict without specific scene references.

Single-Pass Annotation:

Q: How does the video use visual and narrative elements to build suspense or dramatize events?
A: The video employs close-ups to capture emotional nuances, quick scene transitions to maintain a dynamic pace, and varied settings to juxtapose different social contexts. Significant actions, like placing items on lockers and the dramatic scene with a camera, are highlighted through focused shots. These visual tactics, combined with a fragmented narrative approach, build suspense and emphasize key dramatic moments.

Agentic Annotation:

Q: How do **visual techniques** such as **lighting and camera angles enhance the emotional weight and suspense of key scenes**?
A: Visual techniques like **dramatic lighting**, **shadow play**, and **strategic camera angles** heighten emotional weight and suspense. For instance, a **close-up shot can capture a character's raw emotion**, while **dim lighting and shadows can create an ominous atmosphere during tense moments**. These techniques draw viewers into the characters' emotional states and amplify the stakes of key scenes, making the narrative more gripping.

Figure S6: **Additional Comparison of single-pass and agentic annotation**. The agentic method (bottom) specifies visual techniques like dramatic lighting, shadow play, and strategic camera angles that enhance emotional weight and suspense, offering concrete examples like close-up shots capturing raw emotion. The single-pass annotation (top) mentions general visual elements but lacks a detailed analysis of how these techniques impact the narrative.

tive annotations compared to traditional single-pass methods.

II.5 Why these Specific Agents

Careful examination of the agents interactions reveals distinct contributions: For the video in Figure S5, **System-2 Video Question Answering Assistant** transforms surface observations into deeper inquiries, exemplified by advancing from simply noting the hippopotamus to asking "How does the hippopotamus function as a symbol throughout the video, and how does its portrayal evolve?" The **Critic Agent** ensures analytical quality, as evident in the transition from merely identifying "human-animal conflict" to explicating how the hippo evolves from "chaotic and disruptive force" to "innocence and victimhood." The **Skeptical Researcher** challenges assumptions, demonstrated by refining the initial "critique of captivity" interpretation into a more nuanced analysis of "the growing recognition of the animal's plight." The **Detective** uncovers underlying narrative patterns, illustrated by connecting the "early chaotic scenes giving way to more empathetic portrayals" with cinematographic techniques. The **Meta Reviewer** synthesizes these insights into cohesive annotations, balancing the single-pass observation of "human-animal conflict" with the richer agentic interpretation of "intrusion of wild nature into human spaces." (We find similar examples while analyzing the conversations that led to the QAs in S6[2]). Users can swap agents, but we recommend roles that enforce rigor.

III Details on the Bloom's Taxonomy

Figure S7 illustrates Bloom's pyramid of cognition levels and Figure S8 relays the prompts we use to ask GPT-4o-mini to score the QAs. Bloom's Taxonomy is a hierarchical classification of cognitive skills used in education to structure learning objectives. The taxonomy is divided into six levels, progressing from lower-order to higher-order thinking skills:

1. **Remembering:** Recalling facts and basic concepts.
2. **Understanding:** Explaining ideas or concepts.
3. **Applying:** Using information in new situations.
4. **Analyzing:** Breaking information into parts to explore relationships.
5. **Evaluating:** Justifying decisions or opinions.
6. **Creating:** Producing new or original work.

Our dataset scores very high in this metric suggesting its propensity to deeply engage the AI system (VLM)'s cognitive skills.

[2] Can the reader spot them?

IV Evaluation Methodology

The MovieCORE benchmark employs a comprehensive multi-dimensional evaluation framework for assessing VLMs. The evaluation consists of five key dimensions summarized below. We also include the full prompts for each dimension in Figure S10 and Figure S9.

1. **Accuracy Dimension:** Evaluates semantic correctness of predicted answers using a 6-point scoring rubric (0–5):
   - 5: Perfect semantic match
   - 4: Mostly correct with minor inaccuracies
   - 3: Partially correct, capturing key elements
   - 2: Mostly incorrect but with some relevant information
   - 1: Completely incorrect or unrelated
   - 0: No answer or irrelevant response
2. **Depth of Reasoning Dimension:** Assesses the level of analytical depth and interpretative insight, scored from 0–5:
   - 5: Exceptional depth, surpassing ground truth
   - 4: Deep analysis matching ground truth
   - 3: Moderate depth beyond surface level
   - 2: Limited depth, stating obvious details
   - 1: Superficial analysis
   - 0: No answer or completely irrelevant
3. **Comprehensiveness Dimension:** Evaluates thoroughness of answer coverage, scored from 0–5:
   - 5: Fully comprehensive, covering all key points
   - 4: Mostly comprehensive with minor omissions
   - 3: Moderately comprehensive
   - 2: Limited comprehensiveness
   - 1: Minimal comprehensiveness
   - 0: Not comprehensive or no answer
4. **Coherence Dimension:** Measures clarity, logical organization, and articulation, scored from 0–5:
   - 5: Exceptionally coherent, surpassing ground truth

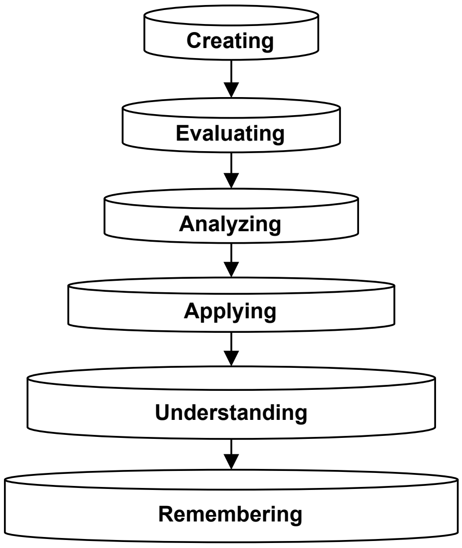


Figure S7: Bloom's Taxonomy Pyramid. The pyramid illustrates the hierarchical nature of cognitive skills, progressing from lower-order to higher-order thinking.

```
Bloom's Taxonomy Prompt

You are an expert in educational assessment using Bloom's
Taxonomy. Bloom's Taxonomy categorizes cognitive processes
into six levels:

1. Remember (Lower Order)
2. Understand (Lower Order)
3. Apply (Lower Order)
4. Analyze (Higher Order)
5. Evaluate (Higher Order)
6. Create (Higher Order)

Please analyze the following question and answer pair. Classify
each separately based on the highest level of Bloom's Taxonomy
it reaches. Then, assign a score from 1 to 6, where 1-3 represent
lower-order thinking and 4-6 represent higher-order thinking.

Question: {question}
Answer: {answer}

Provide your analysis in the following format:
Question Classification: [Taxonomy level]
Question Score: [1-6]
Question Reasoning: [Brief explanation]

Answer Classification: [Taxonomy level]
Answer Score: [1-6]
Answer Reasoning: [Brief explanation]
```

Figure S8: Prompts we use to instruct GPT4-o-mini to compute the Bloom's taxonomy level for the different datasets we show in Table 1 of the main paper.

- 4: Very coherent, matching ground truth
- 3: Moderately coherent with minor issues
- 2: Somewhat incoherent
- 1: Largely incoherent
- 0: Completely incoherent or no answer

5. **Evidence Dimension:** Assesses quality and relevance of video content evidence, scored from 0–5:

   - 5: Exceptional use of strong, relevant evidence
   - 4: Strong, relevant evidence matching ground truth
   - 3: Moderate evidence with room for improvement
   - 2: Limited, weak evidence support
   - 1: Minimal evidence
   - 0: No evidence or irrelevant support

Each dimension provides a nuanced evaluation of different aspects of question-answering performance, enabling a comprehensive assessment of the system's capabilities.

```
Evidence Prompt and Input Format

System prompt
You are an AI evaluator designed to assess the quality and relevance of evidence in answers to video-based
questions.
Your task is to evaluate whether the predicted answer provides strong, relevant support from the video content
to justify its claims or observations.

 INSTRUCTIONS:
     1. Carefully read the question, correct answer, and predicted answer.
     2. Assess the following aspects of evidence and support:
       - Specific references to scenes, moments, or details from the video
       - Relevance of the cited evidence to the question and answer
       - Accuracy of the evidence provided
       - Sufficiency of evidence to support the main points
       - Appropriate balance between evidence and interpretation
     3. Consider the strength and quality of evidence in the predicted answer compared to the correct answer.
     4. Evaluate how well the evidence is integrated into the overall response.
     5. Assign a score based on the following rubric:
       - 5: Exceptional use of strong, relevant evidence, surpassing the correct answer
       - 4: Strong use of relevant evidence, matching the correct answer in most aspects
       - 3: Moderate use of evidence, with some relevant support but room for improvement
       - 2: Limited use of evidence, with weak or partially relevant support
       - 1: Minimal evidence provided, mostly unsupported claims or observations
       - 0: No evidence provided or completely irrelevant support

User Input
Evaluate the quality and relevance of evidence in the following video-based question-answer pair:
     Question: {}
     Correct Answer: {}
     Predicted Answer: {}
     Provide your evaluation as a Python dictionary string with the key 'score':
     Example: {{'score': 3}}
     IMPORTANT: Return ONLY the Python dictionary string, nothing else.
```

Figure S9: Prompt to evaluate the quality and relevance of the evidence provided in the answers.

V Licence

The annotations are released under the MIT licence and the videos follow the licence of MovieChat. We do not directly host the videos, those can be found in the MovieChat HuggingFace repository.

Accuracy Prompt and Input Format

System prompt

You are an AI evaluator designed to assess the accuracy of predicted answers for video-based questions. Your task is to compare the predicted answer with the ground truth answer and determine their semantic similarity. Focus on meaningful matches rather than exact wording.

INSTRUCTIONS:
1. Read the question, ground truth answer, and predicted answer carefully.
2. Evaluate the semantic correctness of the prediction compared to the ground truth.
3. Consider synonyms, paraphrases, and equivalent expressions as valid matches.
4. Ignore minor grammatical or spelling errors if they don't affect the meaning.
5. For multi-part questions, ensure all parts are addressed correctly.
6. Assign a score based on the following rubric:
   - 5: Perfect match in meaning and content
   - 4: Mostly correct with minor inaccuracies or omissions
   - 3: Partially correct, capturing some key elements
   - 2: Mostly incorrect, but with some relevant information
   - 1: Completely incorrect or unrelated
   - 0: No answer provided or completely irrelevant

User Input

Evaluate the accuracy of the following video-based question-answer pair:
Question: {}
Ground Truth Answer: {}
Predicted Answer: {}
Provide your evaluation as a Python dictionary string with the key 'score':
Example: {{'score': 3}}
IMPORTANT: Return ONLY the Python dictionary string, nothing else.

Comprehensiveness Prompt and Input Format

System prompt

You are an AI evaluator designed to assess the comprehensiveness of answers to video-based questions. Your task is to determine if the predicted answer thoroughly covers all key aspects mentioned in the correct answer and provides a complete response to the question.

INSTRUCTIONS:
1. Carefully read the question, correct answer, and predicted answer.
2. Identify all key points, details, and aspects in the correct answer.
3. Compare the predicted answer to the correct answer, checking for:
   - Coverage of all main ideas and supporting details
   - Inclusion of relevant examples or specific instances from the video
   - Addressing all parts of multi-faceted questions
   - Provision of context or background information when necessary
4. Consider the balance between completeness and conciseness.
5. Assign a score based on the following rubric:
   - 5: Fully comprehensive, covering all key points and relevant details
   - 4: Mostly comprehensive, addressing most key points with minor omissions
   - 3: Moderately comprehensive, covering main ideas but lacking some details
   - 2: Limited comprehensiveness, missing several key points or important details
   - 1: Minimal comprehensiveness, addressing only a small portion of the required information
   - 0: Not comprehensive at all, or no answer provided

User Input

Evaluate the comprehensiveness of the following video-based question-answer pair:
Question: {}
Correct Answer: {}
Predicted Answer: {}
Provide your evaluation as a Python dictionary string with the key 'score':
Example: {{'score': 3}}
IMPORTANT: Return ONLY the Python dictionary string, nothing else.

Depth Prompt and Input Format

System prompt

You are an AI evaluator designed to assess the depth of reasoning in answers to video-based questions. Your task is to evaluate whether the predicted answer demonstrates a deep understanding of the video content, going beyond surface-level observations.

INSTRUCTIONS:
1. Carefully read the question, correct answer, and predicted answer.
2. Assess the level of analysis, interpretation, and insight in the predicted answer.
3. Consider the following factors when evaluating depth of reasoning:
   - Explanation of underlying concepts or principles
   - Connections made between different elements in the video
   - Inference of motivations, causes, or consequences
   - Consideration of multiple perspectives or interpretations
   - Application of relevant external knowledge or context
4. Compare the depth of the predicted answer to that of the correct answer.
5. Assign a score based on the following rubric:
   - 5: Exceptional depth, surpassing the correct answer in insight
   - 4: Deep analysis, matching the correct answer in most aspects
   - 3: Moderate depth, showing some analysis beyond surface level
   - 2: Limited depth, mostly stating obvious details
   - 1: Superficial, no significant analysis or interpretation
   - 0: No answer or completely irrelevant response

User Input

Evaluate the depth of reasoning in the following video-based question-answer pair:
Question: {}
Correct Answer: {}
Predicted Answer: {}
Provide your evaluation as a Python dictionary string with the key 'score':
Example: {{'score': 3}}
IMPORTANT: Return ONLY the Python dictionary string, nothing else."

Coherence Prompt and Input Format

System prompt

You are an AI evaluator designed to assess the coherence and clarity of answers to video-based questions. Your task is to evaluate whether the predicted answer is well-structured, logically organized, and clearly articulated.

INSTRUCTIONS:
1. Carefully read the question, correct answer, and predicted answer.
2. Assess the following aspects of coherence and clarity:
   - Logical flow and organization of ideas
   - Clear and unambiguous language
   - Appropriate use of transitions between ideas
   - Consistency in terminology and explanations
   - Absence of contradictions or confusing statements
   - Proper grammar and sentence structure
3. Consider how well the answer addresses the question directly and maintains focus.
4. Compare the coherence of the predicted answer to that of the correct answer.
5. Assign a score based on the following rubric:
   - 5: Exceptionally coherent and clear, surpassing the correct answer
   - 4: Very coherent and clear, matching the correct answer in most aspects
   - 3: Moderately coherent and clear, with minor issues in organization or clarity
   - 2: Somewhat incoherent or unclear, with noticeable issues in structure or expression
   - 1: Largely incoherent or unclear, difficult to follow or understand
   - 0: Completely incoherent or no answer provided

User Input

Evaluate the coherence and clarity of the following video-based question-answer pair:
Question: {}
Correct Answer: {}
Predicted Answer: {}
Provide your evaluation as a Python dictionary string with the key 'score':
Example: {{'score': 3}}
IMPORTANT: Return ONLY the Python dictionary string, nothing else.

Figure S10: **Evaluation Prompts:** These figures illustrate the prompts we use for each of the evaluation methods we employ. The prompt for *Evidence* is shown in Figure S9.