VCA: Video Curious Agent for Long Video Understanding

Supplementary Material

1. Prompt Design

In this section, we elaborate the prompt design of our VCA framework. We include the prompt on how the reward model (R) generates relevance scores in Fig. 1 and Fig. 2, and the prompt used for the exploration agent (π) in Fig. 3, respectively.

As illustrated in in Fig. 1, for the first round, the reward model starts evaluating the relevance of each segment to the given question. For the subsequent rounds of exploration, since the reward model can only access the sampled frames from the selected local segment, we include reward history (H_r) in previous rounds to ensure consistency and alignment in terms of relevance score scale. The detailed template is shown in Fig. 2.

As illustrated in in Fig. 3, the Exploration Agent determines the next action based on these relevance scores. If the available information is sufficient, the agent proceeds to provide an answer. Otherwise, it continues to iteratively refine its exploration. Specifically, for the exploration agent, all frames stored in the memory buffer are presented together with the dialog history that is formatted in a conversational style.

2. Implementation Details

2.1. Dataset

Dataset	Avg. duration	Data size
EgoSchema [?]	180s	500
LVBench [?]	4,100s	1,549
MMBench-Video [?]	165s	1,998
VideoMME-long [?]	2,466s	900

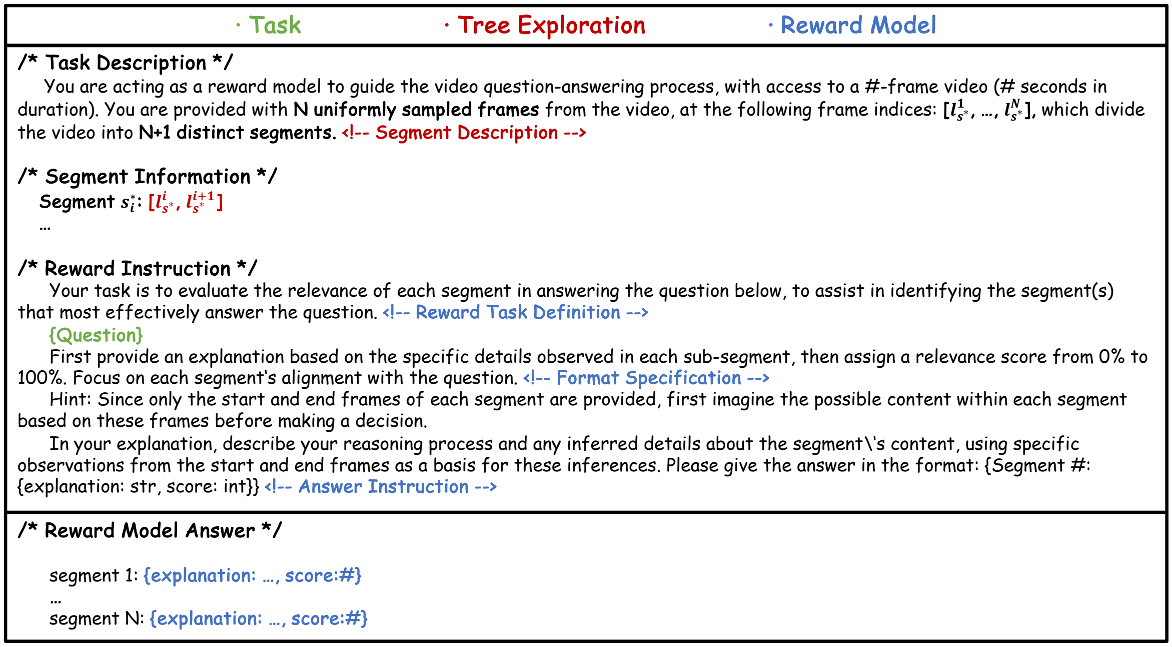
Table 1. **Dataset Statistics**. Overview of the data statistics across all four benchmarks: the minute-level EgoSchema and MMBench-Video, and the hour-level LVBench and VideoMME. For VideoMME, the statistics correspond to the *long* subset.

In this section, we elaborate the details and statistics of the benchmark datasets. Besides the two datasets mentioned in Sec. **??**, we also include the two recently released datasets, **MMBench-Video** [?] and **VideoMME** [?] into our evaluation. The statistics of the datasets are provided in Tab. 1. The official implementation of VideoMME utilizes GPT-4 for evaluation. To ensure consistency, we also use GPT-4o as the evaluator in this work.

We acknowledge the existence of several well-established long-term video question answering benchmarks [**?** **?**], such as MovieChat [**?**] and MovieQA [**?**]. However, since the videos in these benchmarks are often sourced from publicly available movies, there is a potential

System Prompt You are a helpful assistant. Please answer the following question.
EgoSchema Taking into account all the actions performed by c, what can you deduce about the primary objective and focus within the video content? Option 0: C is cooking. Option 1: C is doing laundry. Option 2: C is cleaning the kitchen. Option 3: C is cleaning dishes. Option 4: C is cleaning the bathroom. Please directly answer the option number.
LVBench What year appears in the opening caption of the video? (A) 1636 (B) 1366 (C) 1363 (D) 1633 Please select the best answer from the options provided and directly provide the letter representing your choice without giving any explanation.
MMBench-Video What is the name of the player who scored the first goal in the video? Please directly reply with your response.
VideoMME What is the video mainly about? A. Planes invented by the Wright Brothers. B. The structural difference between the planes created by Whitehead and planes created by the Wright Brothers. C. Who invented the first plane. D. How Whitehead and the Wright Brothers cooperated to invent the first motorized flight. Please select the best answer from the options provided and directly provide the letter representing your choice without giving any explanation.

Table 2. **Prompt Template Examples**. Examples of the prompt template used in our framework. Each dataset is represented by a sample instance, marked in red, paired with its corresponding instruction template for clarity.


· Task
· Tree Exploration
· Reward Model

/* Task Description */

You are acting as a reward model to guide the video question-answering process, with access to a #-frame video (# seconds in duration). You are provided with N **uniformly sampled frames** from the video, at the following frame indices: [$l_{s^*}^1$, ..., $l_{s^*}^N$], which divide the video into N+1 **distinct segments.** <!-- Segment Description -->

/* Segment Information */

Segment s_i^*: [$l_{s^*}^i$, $l_{s^*}^{i+1}$]

...

/* Reward Instruction */

Your task is to evaluate the relevance of each segment in answering the question below, to assist in identifying the segment(s) that most effectively answer the question. <!-- Reward Task Definition -->

{Question}

First provide an explanation based on the specific details observed in each sub-segment, then assign a relevance score from 0% to 100%. Focus on each segment's alignment with the question. <!-- Format Specification -->

Hint: Since only the start and end frames of each segment are provided, first imagine the possible content within each segment based on these frames before making a decision.

In your explanation, describe your reasoning process and any inferred details about the segment\'s content, using specific observations from the start and end frames as a basis for these inferences. Please give the answer in the format: {Segment #: {explanation: str, score: int}} <!-- Answer Instruction -->

/* Reward Model Answer */

segment 1: {explanation: ..., score:#}

...

segment N: {explanation: ..., score:#}

Figure 1. **Reward Model (R) Prompt Template for the First Round**. The prompt template begins by defining the reward scoring task and presenting the segment candidates. Given the target query, the reward model is instructed to first provide an explanation for its reasoning and then assign a relevance score to each candidate. <!-- --> represents comments or explanations about the given prompt.

Method	Frames	ER	EU	KIR	TG	Rea	Sum	Avg.
Ours	20.0	**43.7**	**40.7**	37.8	**38.0**	**46.2**	27.3	**41.3**
- w/o Reward	22.2	37.5	35.0	**43.6**	27.2	44.0	**40.0**	39.0
- w/o Tree Search	39.9	34.2	34.7	40.7	29.4	41.2	29.1	36.2

Table 3. **Detailed Ablation Study Results on LVBench**. Both the tree-search exploration and reward model contribute significantly to improvements in performance and efficiency. The memory buffer size of our agent is set to **16 frames**. We report the average number of observed frames and mark the best performance in **bold**.

risk that they may have been included in the training data of recent closed-source VLMs such as GPT-4o. To avoid data leakage and ensure fair evaluation, we chose not to use these benchmarks.

For further clarity, we provide detailed prompts of a randomly chosen example for each dataset in Tab. 2. For each dataset, we follow the prompt template provided in their paper or official implementation.

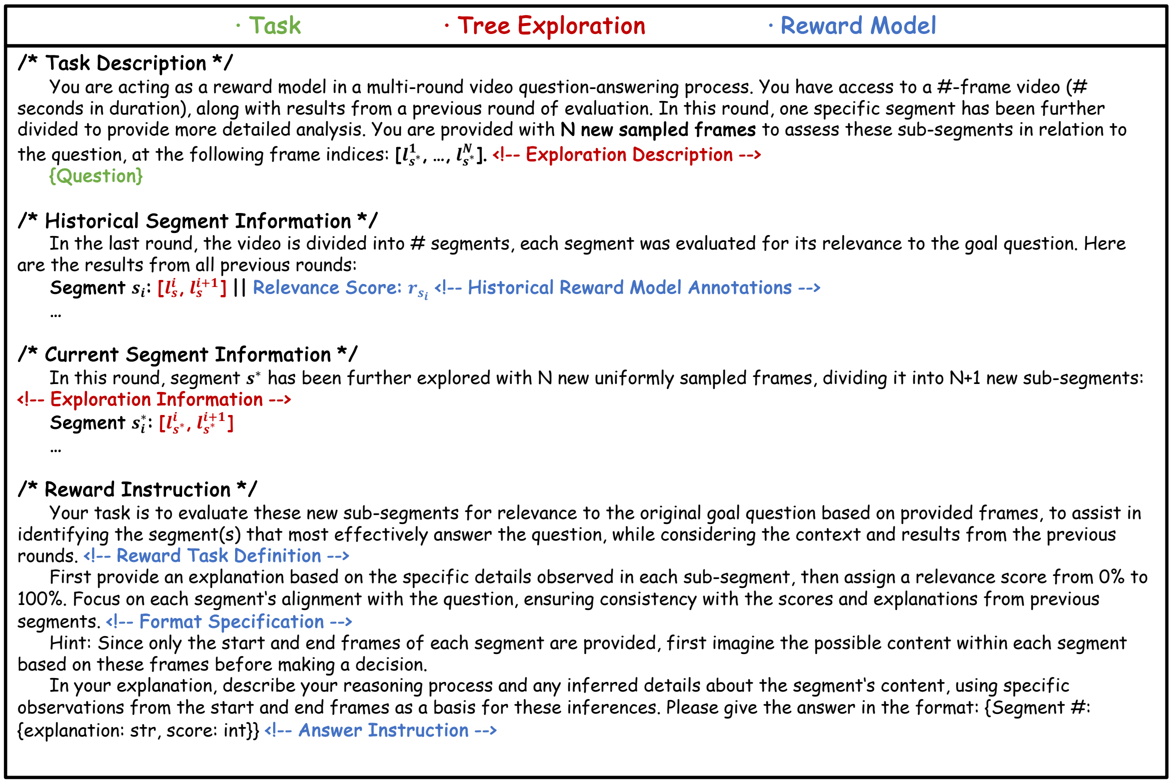
2.2. Baselines

In this section, we elaborate the implementation details of the baselines. For VideoAgent [**?**] and VideoTree [**?**], we begin by adhering to their official implementations[1][2]. However, we notice that both methods utilize GPT-4 as the reasoning agent, which is suboptimal compared to GPT-4o, the model we employ as the exploration agent. Therefore, to ensure a fair comparison, we re-implement both methods using GPT-4o as the reasoning agent across all benchmarks. Furthermore, as the original implementations do not provide extracted captions for the videos, we followed their methodology by extracting captions at 0.2 FPS using the current state-of-the-art VLM, Qwen2-VL [**?**], due to its powerful video understanding capabilities. **However, using GPT-4o as the captioner is computationally prohibitive, as a single 30-minute video would require over 4M tokens, making it infeasible in practice.** For their visual encoders, we leverage EVA-CLIP-8B [**?**] with the same settings as [**? ?**].

For a fair comparison with VideoAgent, we experimented with varying the number of initial frames of VideoAgent to optimize its performance, accounting for changes in both the agent and caption models. The top two results on EgoSchema are presented in Tab. **??**, aligned with the optimal parameter reported in the original paper.

[1] https://github.com/wxh1996/VideoAgent

[2] https://github.com/Ziyang412/VideoTree



· Task · Tree Exploration · Reward Model

/* Task Description */

You are acting as a reward model in a multi-round video question-answering process. You have access to a #-frame video (# seconds in duration), along with results from a previous round of evaluation. In this round, one specific segment has been further divided to provide more detailed analysis. You are provided with **N new sampled frames** to assess these sub-segments in relation to the question, at the following frame indices: [$l_{s^*}^1$, ..., $l_{s^*}^N$]. <!-- Exploration Description -->

{Question}

/* Historical Segment Information */

In the last round, the video is divided into # segments, each segment was evaluated for its relevance to the goal question. Here are the results from all previous rounds:

Segment s_i: [l_s^i, l_s^{i+1}] || Relevance Score: r_{s_i} <!-- Historical Reward Model Annotations -->

...

/* Current Segment Information */

In this round, segment s^* has been further explored with N new uniformly sampled frames, dividing it into N+1 new sub-segments: <!-- Exploration Information -->

Segment s_i^*: [$l_{s^*}^i$, $l_{s^*}^{i+1}$]

...

/* Reward Instruction */

Your task is to evaluate these new sub-segments for relevance to the original goal question based on provided frames, to assist in identifying the segment(s) that most effectively answer the question, while considering the context and results from the previous rounds. <!-- Reward Task Definition -->

First provide an explanation based on the specific details observed in each sub-segment, then assign a relevance score from 0% to 100%. Focus on each segment's alignment with the question, ensuring consistency with the scores and explanations from previous segments. <!-- Format Specification -->

Hint: Since only the start and end frames of each segment are provided, first imagine the possible content within each segment based on these frames before making a decision.

In your explanation, describe your reasoning process and any inferred details about the segment's content, using specific observations from the start and end frames as a basis for these inferences. Please give the answer in the format: {Segment #: {explanation: str, score: int}} <!-- Answer Instruction -->

Figure 2. **Reward Model (R) Prompt Template for the Subsequent Rounds**. This prompt template, used after the initial round, begins by outlining the exploration strategy and presenting the target query. Besides, to ensure consistency in relevance scores throughout the exploration trajectory, it also includes historical segment information along with the reward annotations previously assigned. `<!-- -->` represents comments or explanations about the given prompt.

For LVBench and Video-MME, we set the initial frame count to 15 to better accommodate the processing requirements of longer videos. For the clustering component of VideoTree, we provide the hyper-parameter settings as follows: max_breadth = 32, max_depth = 3, branch_width = 4, and rele_num_thresh = 4.

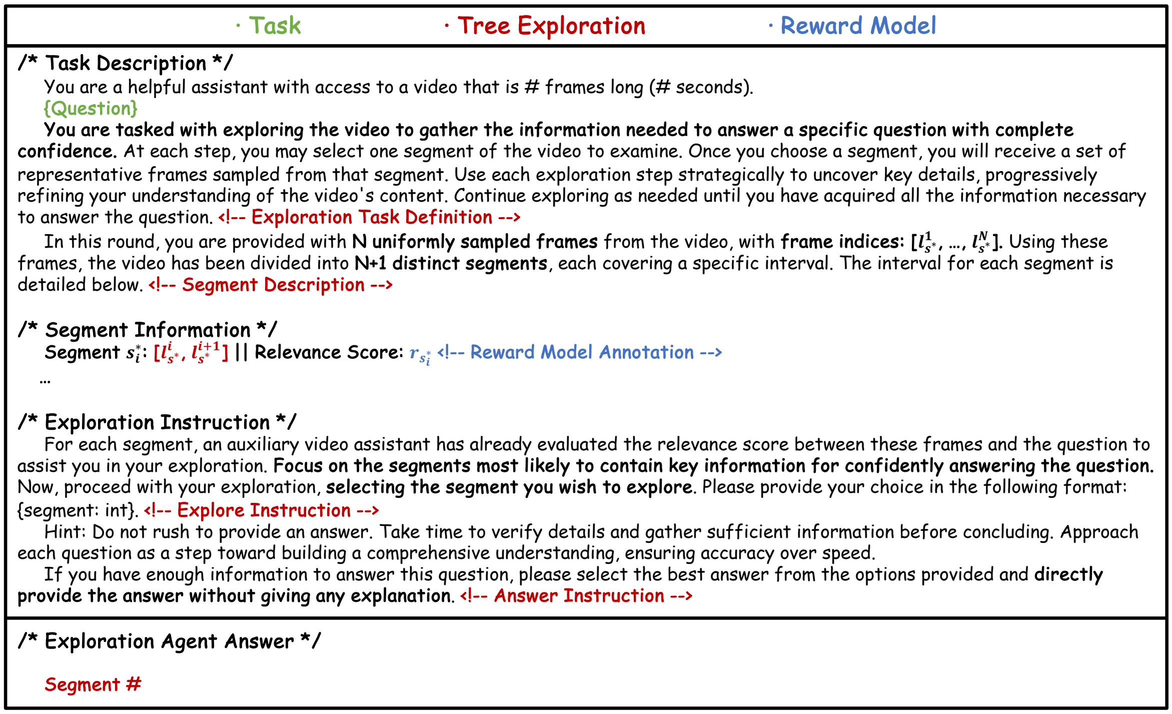
3. Experiments

3.1. Experimental Results

As discussed in Sec. **??**, we present the experimental results on VideoMME and MMBench-Video in Tab. 4 and Tab. 5, respectively. For VideoMME, we evaluate our framework on the long split, with videos averaging over 2,000 seconds in duration. Similarly, to ensure a fair comparison, we re-implement the baselines using GPT-4o as the base VLM. As shown in Tab. 4, our method achieves a significant improvement, outperforming VideoTree by 4.9% with less than 20% of observed frames, and outperforming VideoAgent by 12.1% with about 75% of observed frames. Similar trends are observed in the results for MMBench-Video in Tab. 5. We observe that both VideoAgent and VideoTree perform poorly on this benchmark. We conjecture that this is due to the long split's emphasis on detailed visual clues, while captioning-based methods inherently struggle with tasks requiring detailed visual comprehension.

Method	VideoAgent†	VideoTree†	Ours
Avg. Frames	24.6	98.0	**18.1**
Knowledge	52.2	**60.7**	56.9
Film & Television	42.5	52.5	**55.0**
Sports Competition	42.7	48.6	**59.3**
Artistic Performance	47.5	51.6	**65.8**
Life Record	44.7	49.5	**51.9**
Multilingual	36.6	40.0	**46.7**
Overall	46.4	53.1	**56.3**

Table 4. **Experimental Results on VideoMME Long Split**. Methods marked with † indicate implementations where **GPT-4o** serves as the main component. The memory buffer size of our agent is set to **16 frames**. We report the average number of observed frames and mark the best performance in **bold**.



· Task · Tree Exploration · Reward Model

/* Task Description */

You are a helpful assistant with access to a video that is # frames long (# seconds).

{Question}

You are tasked with exploring the video to gather the information needed to answer a specific question with complete confidence. At each step, you may select one segment of the video to examine. Once you choose a segment, you will receive a set of representative frames sampled from that segment. Use each exploration step strategically to uncover key details, progressively refining your understanding of the video's content. Continue exploring as needed until you have acquired all the information necessary to answer the question. <!-- Exploration Task Definition -->

In this round, you are provided with **N uniformly sampled frames** from the video, with **frame indices: [$l_{s^*}^1$, ..., $l_{s^*}^N$]**. Using these frames, the video has been divided into **N+1 distinct segments**, each covering a specific interval. The interval for each segment is detailed below. <!-- Segment Description -->

/* Segment Information */

Segment s_i^*: [$l_{s^*}^i$, $l_{s^*}^{i+1}$] || Relevance Score: $r_{s_i^*}$ <!-- Reward Model Annotation -->

...

/* Exploration Instruction */

For each segment, an auxiliary video assistant has already evaluated the relevance score between these frames and the question to assist you in your exploration. **Focus on the segments most likely to contain key information for confidently answering the question.** Now, proceed with your exploration, **selecting the segment you wish to explore**. Please provide your choice in the following format: {segment: int}. <!-- Explore Instruction -->

Hint: Do not rush to provide an answer. Take time to verify details and gather sufficient information before concluding. Approach each question as a step toward building a comprehensive understanding, ensuring accuracy over speed.

If you have enough information to answer this question, please select the best answer from the options provided and **directly provide the answer without giving any explanation.** <!-- Answer Instruction -->

/* Exploration Agent Answer */

Segment #

Figure 3. **Exploration Agent (π) Prompt Template**. The prompt template begins by introducing the target question, followed by a definition of the exploration task. Next, the segment candidates' information is provided, guiding the agent to either select a segment for further exploration or answer the question if sufficient information has been gathered. <!-- --> represents comments or explanations about the given prompt.

Method	GPT-4o	VideoAgent†	VideoTree†	Ours
Avg. Frames	8	7.8	27.1	7.4
Score	1.62	1.05	1.38	**1.68**

Table 5. **Experimental Results on MMBench-Video**. Methods marked with † indicate implementations where **GPT-4o** serves as the main component. The memory buffer size of our agent is set to **8 frames**. We report the average number of observed frames and mark the best performance in **bold**.

Method	Frames	Acc.	Frames	Acc.
	(LongVA)		(GPT-4o)	
VCA (ours)	13.7	33.4	20.0	41.3
- w/ GT Reward	27.2	**36.1**	24.0	**44.2**

Table 6. **Comparison of Accuracy w/ and w/o Ground Truth (*GT*) Reward Scores on LVBench**. Incorporating *GT* reward scores further improves performance on both LongVA and GPT-4o models. Memory buffer size is set to 16 frames.

3.2. Ablation Study

In this section, we present the detailed results of the ablation study discussed in Sec. ??. The results across different domains of LVBench are summarized in Tab. 3. Overall, the results demostrate that both the reward model and the tree-search exploration mechanism play significant roles in enhancing performance. Interestingly, we observe that the agent without the reward model or tree-search exploration performs better in the Key Information Retrieval (KIR) and Summarization (Sum) domains. This result is unsurprising, as the absence of tree-search exploration prompts the agent to explore nearly twice as many frames. Similarly, without the reward model, the agent lacks focused guidance and instead explores more broadly across the video.

Conversely, these components contribute significantly to performance improvements in tasks like Event Recognition (ER) and Temporal Grounding (TG), where precise and focused exploration is essential. Overall, the reward model and tree-search exploration, each addressing distinct aspects of the exploration process, work together to drive large performance gains across diverse tasks, aligning with our findings in Sec. ??.

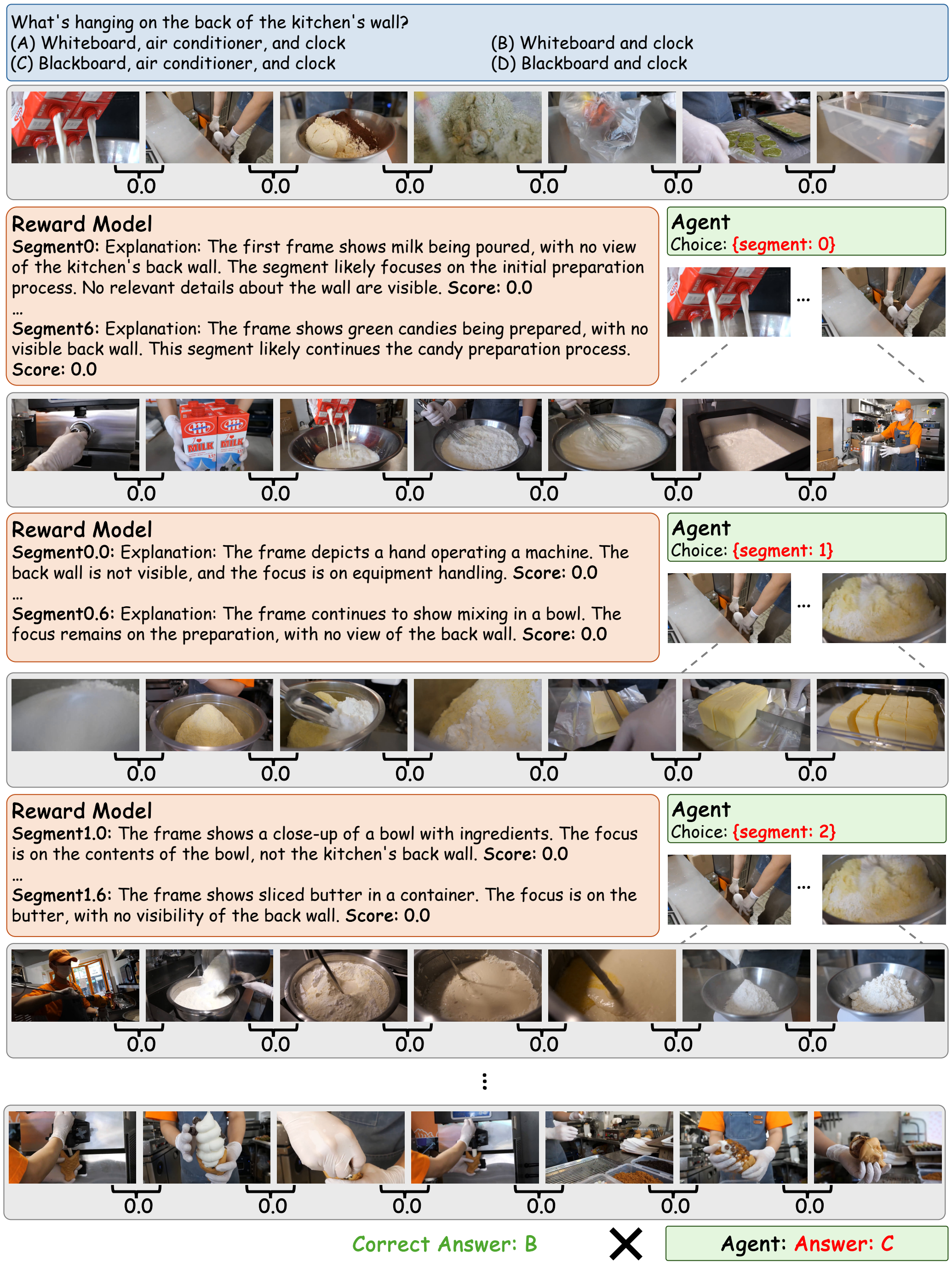
3.3. Can VCA Benefit from Better Reward?

Previous analysis demonstrates the effectiveness of the reward model, though a gap remains compared to ground truth guidance. To examine the upper bound of our agent's performance with optimal reward guidance, we substitute the reward scores with ground truth time references. The results, shown in Tab. 6, indicate that using ground truth scores yields a consistent 3% improvement on LVBench, regardless of whether GPT-4 or LongVA is used as the model. This finding highlights the great potential of our framework when paired with stronger, specialized reward models that can provide more accurate and informative guidance.

4. Case Study

As mentioned in Sec. **??**, in this section, we investigate the common failure cases of our framework, aiming to provide data points and insights for the future research.

4.1. Failure Mode: Inability to Detect Subtle Visual Details



What's hanging on the back of the kitchen's wall?
(A) Whiteboard, air conditioner, and clock
(B) Whiteboard and clock
(C) Blackboard, air conditioner, and clock
(D) Blackboard and clock

Reward Model
Segment0: Explanation: The first frame shows milk being poured, with no view of the kitchen's back wall. The segment likely focuses on the initial preparation process. No relevant details about the wall are visible. **Score:** 0.0
...
Segment6: Explanation: The frame shows green candies being prepared, with no visible back wall. This segment likely continues the candy preparation process. **Score:** 0.0

Agent
Choice: {segment: 0}

Reward Model
Segment0.0: Explanation: The frame depicts a hand operating a machine. The back wall is not visible, and the focus is on equipment handling. **Score:** 0.0
...
Segment0.6: Explanation: The frame continues to show mixing in a bowl. The focus remains on the preparation, with no view of the back wall. **Score:** 0.0

Agent
Choice: {segment: 1}

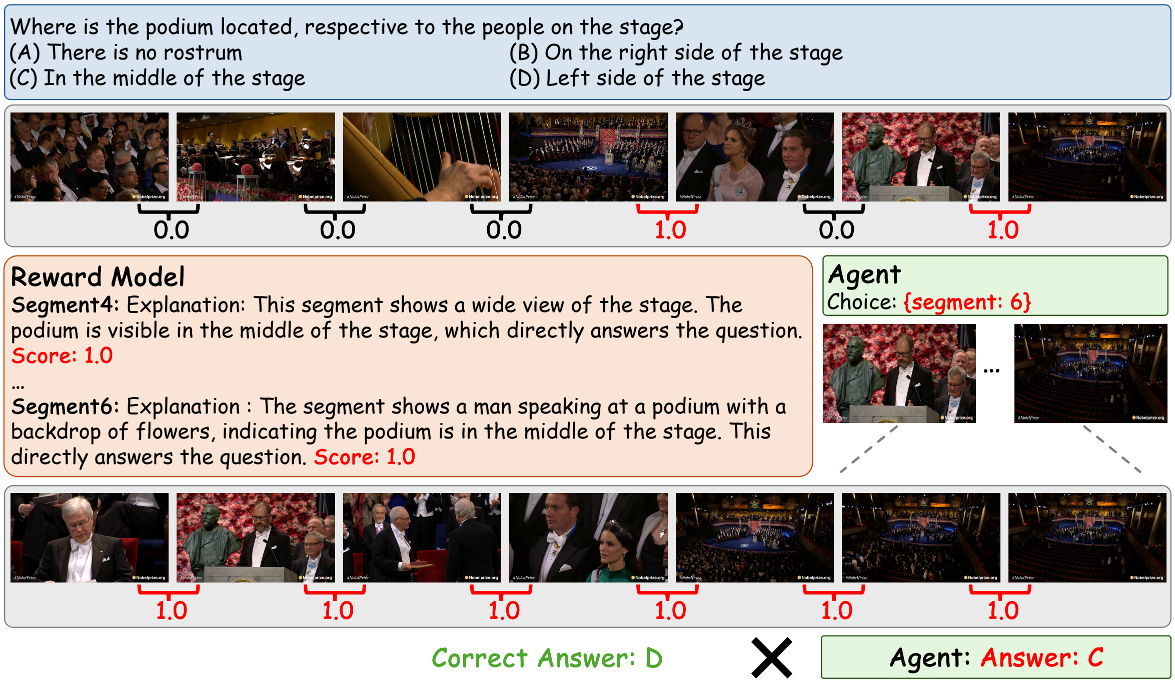
Reward Model
Segment1.0: The frame shows a close-up of a bowl with ingredients. The focus is on the contents of the bowl, not the kitchen's back wall. **Score:** 0.0
...
Segment1.6: The frame shows sliced butter in a container. The focus is on the butter, with no visibility of the back wall. **Score:** 0.0

Agent
Choice: {segment: 2}

Figure 4. **Failure Exploration Trajectory Example**. The agent struggles to locate crucial information for questions requiring visual details, even after prolonged exploration, ultimately resulting in a random generated answer.

The most common error occurs when the agent overlooks subtle clues and fails to extract crucial information, even after extensive exploration. A representative example is illustrated in Fig. 4. In an egocentric cooking video, the agent is tasked with identifying decorations on the back wall of the kitchen, details that appear only briefly during rapid camera movements, making the task challenging even for humans. After multiple rounds of exploring irrelevant segments, the agent fails to locate the key frames and ends up with terminating exploration, and provides a guessed answer. In such cases, the number of frames observed is typically three times the average. We attribute these errors to the inherently high difficulty of questions including subtle visual details, an extremely tricky task for humans as well.

4.2. Failure Mode: Guidance Errors from the Reward Model



Where is the podium located, respective to the people on the stage?

(A) There is no rostrum

(B) On the right side of the stage

(C) In the middle of the stage

(D) Left side of the stage

Reward Model

Segment4: Explanation: This segment shows a wide view of the stage. The podium is visible in the middle of the stage, which directly answers the question. Score: 1.0

...

Segment6: Explanation : The segment shows a man speaking at a podium with a backdrop of flowers, indicating the podium is in the middle of the stage. This directly answers the question. Score: 1.0

Agent

Choice: {segment: 6}

Agent: Answer: C

Figure 5. **Failure Exploration Trajectory Example**. The agent is misled by inaccurate relevance scores from the reward model, causing it to focus on incorrect segments and provide an inaccurate answer.

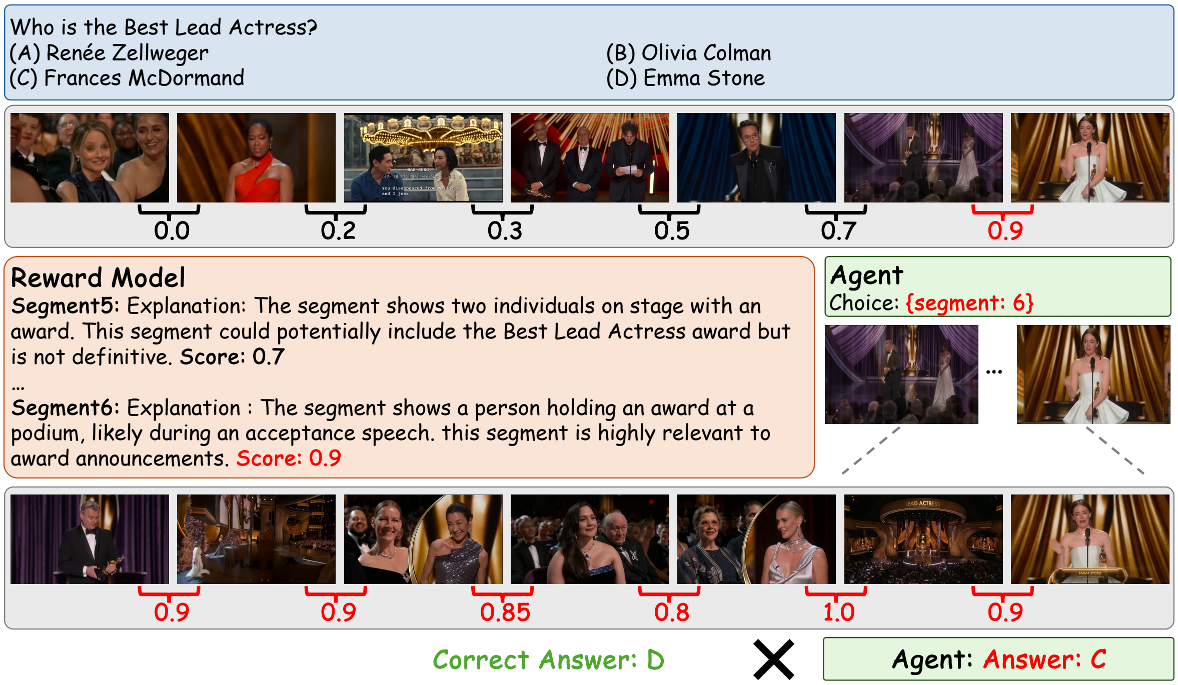
Another common failure mode occurs when the agent is misled by the reward model, as demonstrated in Fig. 5. For instance, when tasked with identifying the spatial relationship between the podium and the stage, the reward model incorrectly assigns 100% relevance scores to the 4th and 6th segments, while the correct information lies in the 5th segment. Consequently, the agent focuses on the wrong segments and provides an incorrect answer. These errors stem from the inherent limitations of the reward model, which fails to provide robust and accurate guidance for the exploration agent. This limitation is also discussed in Sec. **??**.

4.3. Failure Mode: Limited Multi-modal Reasoning Abilities

Another typical failure arises when the agent successfully identifies the crucial segments but fails to produce the correct answer due to limited spatial reasoning capabilities or inadequate alignment of multi-modal knowledge. For example, as shown in Fig. 6, the agent correctly identifies the frame indicating the winner of the Best Lead Actress, with the name clearly annotated in the subtitles. However, even when provided with ground truth visual information, the agent is unable to generate the correct response. We attribute these errors to the inherent limitations of the exploration agent, which suggests the potential for further improving our framework by integrating more advanced foundation models.

4.4. Discussion

The analysis of failure modes reveals the large potential of our framework. While detecting subtle visual details remains challenging even for human, the error stemming from guidance failures by the reward model and reasoning failures by the



Who is the Best Lead Actress?
(A) Renée Zellweger
(B) Olivia Colman
(C) Frances McDormand
(D) Emma Stone

Reward Model

Segment5: Explanation: The segment shows two individuals on stage with an award. This segment could potentially include the Best Lead Actress award but is not definitive. **Score:** 0.7

...

Segment6: Explanation : The segment shows a person holding an award at a podium, likely during an acceptance speech. this segment is highly relevant to award announcements. Score: 0.9

Agent
Choice: {segment: 6}

Figure 6. **Failure Exploration Trajectory Example**. The agent successfully identifies the ground truth visual information but fails to deliver the correct response.

exploration agent could be mitigated by incorporating more robust multi-modal foundation models. As these foundational models evolve, our framework can seamlessly integrate these improvements, indicating its strong potential for tackling more challenging tasks in long-video understanding.