See More, Think Deeper: Query-Expanded Visual Evidence and Answer-Clue Guided Reflection for Long Video Understanding

Anonymous ACL submission

Abstract

Recent advances in Video Large Language Models (Video-LLMs) have enabled performance on long-video understanding tasks. However, existing methods still face two key limitations: evidence acquisition often relies on a single search intent, and answer generation lacks an effective visual feedback mechanism. To address these limitations, we propose **CoVER**, a Comprehensive Visual Evidence and Reflection framework for long-video understanding. CoVER enables Video-LLMs to **See More** by dynamically gathering query-expanded visual evidence, and **Think Deeper** by verifying draft answers with effective answer-specific visual feedback. Together, these mechanisms shift long-video understanding from answer-centric generation to evidence-centric and visually verifiable reasoning. Experimental results show that CoVER-7B substantially outperforms models with the same parameter scale and even surpasses state-of-the-art closed-source models on certain metrics.

1 Introduction

Recent advances in Video Large Language Models (Video-LLMs) have greatly improved video question answering, event understanding, and long-video reasoning (Fu et al., 2025b; Lin et al., 2024; Ouyang et al., 2025; Li et al., 2024d). These models can now interpret complex video content and respond in natural language (Liu et al.). However, long-video understanding is not merely about seeing more frames. It requires identifying sparse but crucial evidence across extended temporal contexts. Key evidence may appear only in a few segments, local regions, or brief moments. Thus, models must not only capture global context, but also locate fine-grained visual evidence during reasoning. Recent studies have therefore shifted from coarse-grained video understanding to fine-grained temporal or spatio-temporal grounding. For example, VideoChat-R1.5 uses Iterative Perception to progressively focus on high-confidence spatio-temporal regions (Yan et al., 2026), while LongVT first skims the video globally and then inspects relevant clips with a video-cropping tool (Yang et al., 2025).

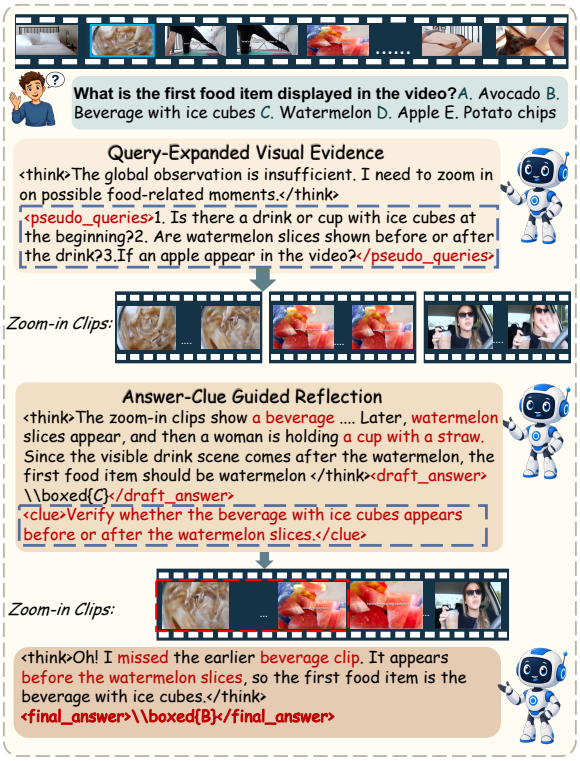

Figure 1: **Illustration of CoVER reasoning process.** CoVER retrieves task-relevant zoom-in evidence with pseudo-queries, then uses an answer-specific clue to verify and refine the draft answer with visual evidence.

Although these methods improve evidence localization, existing VideoQA (Zhang et al., 2025a; Liu et al., 2025a) still face a fundamental challenge: a correct answer may rely on incorrect or incomplete evidence. Prior work on visually grounded VideoQA shows that high answer accuracy does not necessarily imply accurate evidence grounding (Ouyang et al., 2025; Wang et al., 2026). VideoZeroBench (Meng et al., 2026) further high-

lights this gap, showing that model performance drops sharply when both answer correctness and spatio-temporal evidence correctness are required. This suggests that current video reasoning often relies on linguistic priors, salient partial cues, or incomplete context, rather than complete and verifiable visual evidence.

One reason is incomplete evidence acquisition. Existing methods (Wang et al., 2025b; Fu et al., 2025b) often retrieve evidence using only the original question or a single search intent. This may be sufficient for simple questions, but complex video reasoning usually requires multiple cues, such as object attributes, human actions, state changes, spatial relations, and temporal order. A single coarse-grained query may capture only salient cues while missing other critical details. To address this, we introduce query-expanded evidence acquisition. Given the original question, we generate diverse evidence-seeking queries that decompose the reasoning demand into observable and retrievable visual cues. In this paper, we refer to these model-generated evidence-seeking queries as pseudo-queries. This helps the model collect more comprehensive evidence before reasoning.

Another limitation is the lack of visual verification after answer generation. Existing Video-LLMs typically reason over available visual or textual context and directly produce an answer (Wu et al., 2026b; Yang et al., 2026). Once the answer is generated, the process stops. The model rarely checks whether the answer is supported by visual evidence, nor does it revise the answer when new evidence is found. To address this, we introduce answer-clue-guided visual reflection. After producing an initial answer, the model derives verifiable visual clues from the answer and retrieves related evidence from the video. The visual feedback is then used to re-evaluate and revise the answer when necessary.

Based on these insights, we propose **CoVER**, a **Co**mprehensive **V**isual **E**vidence and **R**eflection framework for long-video understanding. CoVER is equipped with a text-aware video evidence tool that retrieves relevant video segments and provides magnified visual information for reasoning. As shown in Fig. 1, it contains two modules: (1) query-expanded visual evidence acquisition, which generates diverse evidence-seeking queries to help the model **See More**; (2) answer-clue-guided visual reflection, which retrieves evidence based on clues derived from the initial answer and uses visual feedback to **Think Deeper**. Together, these modules shift long-video understanding from answer-centric generation to evidence-centric and visually verifiable reasoning. Experiments show that CoVER improves long-video understanding performance. Compared with Qwen2.5-VL-7B, CoVER-7B achieves gains of +3.9% on MLVU and +4.6% on LVBench.

Our contributions are summarized as follows:

- We are the first to introduce visual-feedback reflection into Video-LLM reasoning, transforming open-loop inference into a closed-loop paradigm.
- We propose CoVER, a framework that combines query-expanded visual evidence acquisition with answer-clue guided reflection for evidence-centric reasoning.
- Experiments on long-video benchmarks show that CoVER brings substantial gains over the baseline.

2 Related Work

Long Video Understanding. Large Video Language Models (LVLMs) have advanced and demonstrated strong capability across video understanding tasks (Bai et al., 2025a; Zhang et al.; Li et al., 2025c, 2024a; Lin et al., 2024). Existing approaches commonly adopt uniform frame sampling (Maaz et al., 2024), where frames are sampled at fixed intervals and concatenated as model input. However, in long-video settings, this strategy leads to a rapid increase in visual tokens. To improve efficiency, prior studies explore token compression, keyframe selection, long-context modeling, and agent-based multi-stage processing (Tang et al., 2025; Wang et al., 2025a; Tang et al., 2026; Wang et al., 2026). Unlike these approaches, CoVER uses query expansion to adaptively retrieve task-relevant evidence and verifies draft predictions with answer-specific visual clues, enabling evidence-grounded long-video reasoning.

Chain-of-Thought for Multimodal Reasoning. Recent reasoning models show that long Chain-of-Thought (CoT) can improve complex reasoning by decomposing problems into intermediate steps (Li et al., 2025a; Wei et al., 2023; Wu et al., 2026a). For video understanding, prior studies further incorporate visual information into reasoning, enabling operations such as zoom-in, grounding, and evidence selection (Zhang et al., 2025c; Jiang et al., 2025; Jian et al., 2025). Recent methods, including VITAL (Zhang et al., 2025a), LOVE-R1 (Fu

et al., 2025b), REVISOR (Li et al., 2025b), and Conan (Ouyang et al., 2025), explore dynamic video reasoning and multimodal reflection with tool use or multi-scale visual evidence. Other evidence-centric approaches study search-guided grounding, semantic-visual evidence consensus, and explicit spatio-temporal evidence for long-video answers (Wu et al., 2026b; Sheng et al., 2025; Wei et al., 2026). Unlike these methods, CoVER unifies question-side evidence acquisition and answer-side verification in a closed-loop process, using pseudo-queries to expand visual evidence and answer-specific clues to refine draft predictions.

RL-based Reasoning in VLMs. Recent progress in reinforcement learning (RL) and preference optimization has become an effective paradigm for improving reasoning in Vision-Language Models (VLMs) (Guo et al., 2025; Yu et al., 2025; Rafailov et al., 2023). This trend has also been extended to video understanding, where existing efforts can be broadly grouped into three directions: optimizing temporal sampling for long-video understanding, such as TSPO (Tang et al., 2026); improving temporally grounded and consistent reasoning, as in Video-R2 (Maaz et al., 2025); and enhancing dynamic multi-step video reasoning, as explored by LOVE-R1 (Fu et al., 2025b) and Video-Thinker (Wang et al., 2025c). Unlike these methods, CoVER focuses on evidence completeness and answer verification: it uses query expansion to retrieve task-relevant visual evidence and answer-specific clues to verify and refine draft predictions.

3 Method

3.1 Overview

We propose CoVER, a **Co**mprehensive **V**isual **E**vidence and **R**eflection framework for long-video understanding. CoVER enables the model to acquire comprehensive visual evidence before answer generation and further reflect on draft answers with answer-specific visual clues. Instead of relying only on a low-resolution global observation, CoVER combines query-expanded visual evidence acquisition with answer-clue guided visual reflection, allowing the model to reason over both holistic video context and fine-grained visual evidence.

As illustrated in Fig. 2, the model first builds a low-resolution global observation of the input video V to capture holistic context with reduced computation. Based on the question q and the global observation, it decides whether to invoke the visual zoom-in tool for fine-grained evidence acquisition. When additional evidence is needed, the model generates pseudo-queries $P = \{p_m\}_{m=1}^{M}$ from the question and global observation. These pseudo-queries act as evidence-seeking queries that describe potentially missing visual facts. Together with the original question, they guide the zoom-in tool to retrieve evidence clips from the full video. The retrieved query-expanded evidence complements the global observation and supports draft answer generation. After producing a draft answer, the model derives an answer-specific visual clue and invokes the zoom-in tool again to retrieve verification evidence. It then checks whether the draft answer is visually supported and either preserves or revises it. Thus, CoVER forms a closed-loop reasoning process that acquires missing evidence, verifies draft predictions, and refines answers through explicit visual feedback. Overall, CoVER consists of two components: *Query-Expanded Visual Evidence (QVE)* for draft answer generation, and *Answer-Clue-Guided Reflection (AGR)* for answer verification and refinement.

3.2 CueZoom: Cue-Aware Visual Evidence Zoom-In Tool

To support evidence acquisition and verification in CoVER, we introduce **CueZoom**, a cue-aware visual evidence zoom-in tool. CueZoom retrieves visual evidence clips according to textual cues, including the original question, pseudo-queries, and answer-specific clues. Given textual cues $\mathcal{U} = \{u_j\}_{j=1}^{M}$ and sampled video frames $\{f_i\}_{i=1}^{N}$, CueZoom computes the relevance score s_i of each frame f_i by averaging its similarity to all cues:

$$s_i = \frac{1}{M}\sum_{j=1}^{M} \mathrm{sim}\big(\phi_t(u_j), \phi_v(f_i)\big), \qquad (1)$$

where $\phi_t(\cdot)$ and $\phi_v(\cdot)$ are the text and visual encoders, respectively, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Since frame-level relevance scores can be noisy and may contain isolated responses, we apply Laplacian Temporal Smoothing (LTS) to obtain more stable visual evidence:

$$\tilde{s}_i = \sum_{j=-r}^{r} \kappa_j s_{i+j}, \quad \kappa_j \propto \exp(-|j|/h), \qquad (2)$$

where $\tilde{s}_i$ denotes the smoothed relevance score of frame f_i, κ_j is the smoothing weight for its neighboring offset j, r is the smoothing radius, and h

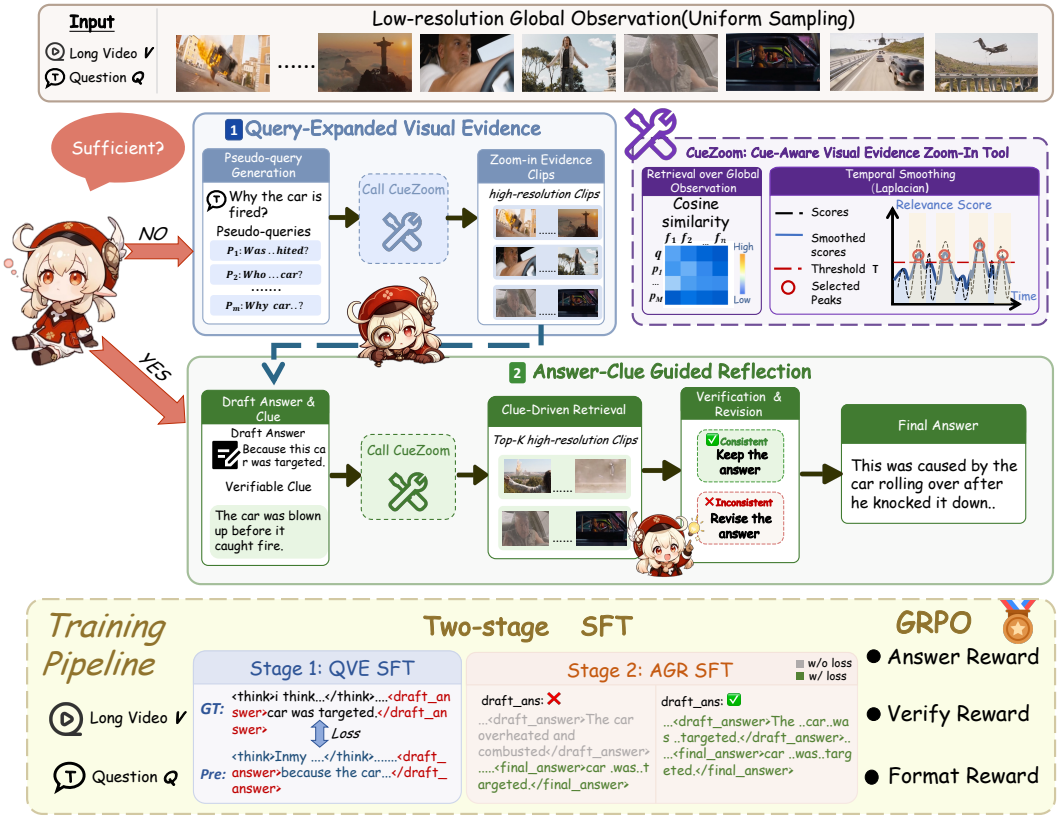

Figure 2: **The overview of the proposed CoVER framework.** CoVER first uses pseudo-queries to retrieve task-relevant zoom-in clips for evidence-aware draft generation, and then uses an answer-specific clue to retrieve verification evidence for final-answer refinement. The model is trained with staged supervised fine-tuning and evidence-guided reinforcement learning using answer, verification, and format rewards.

controls the decay rate of the smoothing kernel. Local normalization is applied near video boundaries. We then select frames whose smoothed scores exceed an adaptive threshold $\tau = \mu(\tilde{s}) + \alpha\sigma(\tilde{s})$, where $\mu(\tilde{s})$ and $\sigma(\tilde{s})$ are the mean and standard deviation of the smoothed relevance sequence, and α controls the selectivity of evidence selection. Finally, neighboring selected frames are merged into coherent visual evidence clips.

3.3 Query-Expanded Visual Evidence

Given a long video V and a question q, the low-resolution global observation V_{global} provides overall semantic context, but may miss sparse or fine-grained evidence. To improve evidence coverage before answer generation, we introduce *Query-Expanded Visual Evidence (QVE)*. Specifically, the model generates a set of pseudo-queries $P = \{p_d\}_{d=1}^{D}$ from q and V_{global}, where $D \leq 3$ in our implementation. These pseudo-queries act as evidence-seeking queries that describe question-relevant visual facts to be further verified. We combine the original question and pseudo-queries into expanded textual cues: $\mathcal{U}_{\text{qe}} = \{q\} \cup P$. CueZoom uses these cues to locate and retrieve relevant video clips from the full video. The retrieved clips $\mathcal{C}_{\text{qe}}$ serve as high-resolution query-expanded visual evidence for draft answer generation.

3.4 Answer-Clue Guided Reflection

Although query-expanded visual evidence improves draft answer generation, the draft answer may still be insufficiently verified by visual evidence. To address this, we introduce *Answer-Clue-Guided Reflection (AGR)*. Given a draft answer $\hat{a}$, the model derives an answer-specific clue c, which describes the visual condition expected to hold if $\hat{a}$ is correct. We use c as the verification cue: $\mathcal{U}_{\text{ver}} = \{c\}$. CueZoom then retrieves answer-relevant evidence clips based on this cue. Unlike pseudo-query-based retrieval, which expands question-relevant evidence coverage, clue-guided retrieval focuses on evidence that directly supports or contradicts the draft answer. This enables the model to reflect on $\hat{a}$ using explicitly retrieved visual evidence rather than treating it as final.

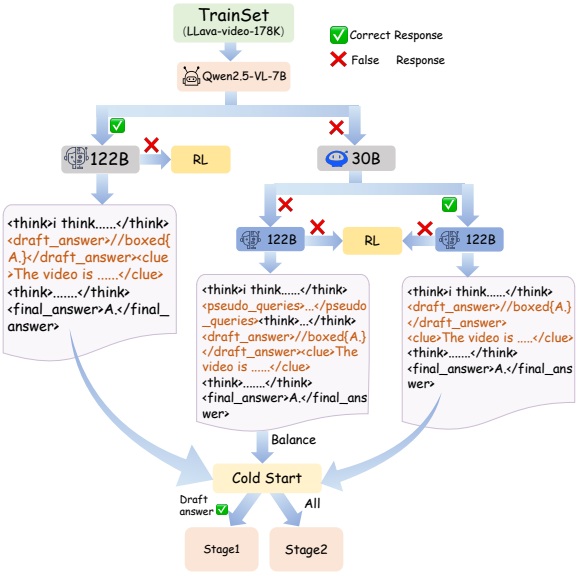

Figure 3: **Training data construction pipeline.** Samples are partitioned by difficulty and annotated into structured trajectories. Correct draft-answer trajectories are used for Stage-1 QVE SFT, while all annotated trajectories support Stage-2 AGR SFT; incorrectly answered samples are further used for RL optimization.

3.5 Data Construction and Two-Stage SFT

Data Construction. To provide a cold start for RL, we construct 11K high-quality SFT samples from LLaVA-Video-178K (Zhang et al., 2024c) through the automated pipeline shown in Fig. 3. The pipeline consists of three steps: (1) Data Partitioning. We first evaluate each sample with Qwen2.5-VL-7B. Samples correctly answered by Qwen2.5-VL-7B are treated as easier cases, while failed samples are further processed by Qwen3-VL-30B-A3B to identify tool-dependent or harder cases. This partition allows the dataset to cover both direct-answer and evidence-seeking reasoning patterns. (2) CoT Annotation Generation. Qwen3.5-122B-A10B is used to generate structured reasoning traces, including draft answers, pseudo-queries when additional evidence is needed, visual clues, retrieved evidence, and final verified answers. These traces provide supervision for both pseudo-query guided evidence acquisition and clue-guided answer verification. (3) Data Balancing and Cleaning. We balance different trace types, upsample rare answer-revision cases where retrieved evidence corrects an incorrect draft answer, and apply rule-based filtering to remove malformed or inconsistent samples. The detailed distribution of the dataset is show in Fig. 4.

Stage-1: QVE SFT. This stage aims to train the model to generate reliable draft answers from available visual evidence. Cold-start trajectories with correct draft answers are used as Stage-1 data. Given the global observation and question, the model learns an adaptive tool-use policy: it invokes the tool with pseudo-queries to retrieve additional visual evidence when needed, and bypasses the tool to directly produce a draft answer when the available evidence is sufficient. In both cases, the model also generates an answer-specific visual clue for subsequent verification. This stage initializes evidence-aware draft prediction while avoiding supervision from erroneous intermediate answers.

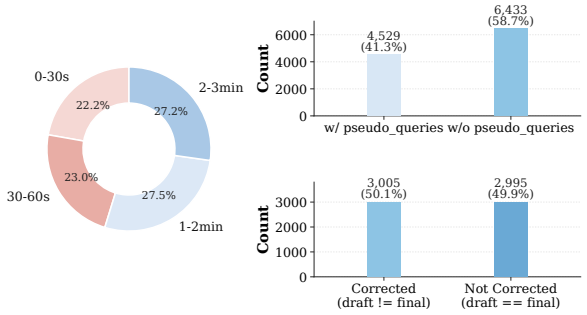

Figure 4: **Statistics of the constructed training data.** The dataset covers diverse video durations, both pseudo-query and direct-answer trajectories, and balanced corrected/no-corrected draft-answer cases.

Stage-2: AGR SFT. This stage trains the model to verify draft answers with answer-clue guided evidence. Given the question, global observation, draft answer, answer clue, and verification clips, the model learns to assess whether the draft answer is visually supported and decide whether the final answer should be revised. To avoid optimizing toward unreliable intermediate drafts, we mask the loss on incorrect draft answers while keeping supervision for verification reasoning and final-answer generation:

$$\mathcal{L}_{\text{AGR}} = -\sum_{t=1}^{T} m_t \log p_\theta(y_t \mid x, y_{<t}),$$
$$m_t = \begin{cases} 0, & \neg z_{\hat{a}} \wedge (t < t_{\text{da}}), \\ 1, & \text{otherwise}, \end{cases} \tag{3}$$

where T is the sequence length, y_t is the target token at position t, x is the input context, and $y_{<t}$ denotes preceding target tokens. The binary mask m_t controls whether token y_t contributes to the loss. $z_{\hat{a}}$ indicates whether the draft answer $\hat{a}$ is correct, and t_{da} denotes the position of the closing `</draft_answer>` tag. When the draft answer is

incorrect, tokens before this tag are masked out.

3.6 GRPO with Evidence-Guided Rewards

We adopt Group Relative Policy Optimization (GRPO) (Guo et al., 2025) to encourage dynamic and autonomous tool use during reinforcement learning. Given a generated reasoning trajectory τ, GRPO jointly optimizes final-answer correctness, answer-level verification, and format consistency through a composite reward:

$$R(\tau) = \lambda_{\text{ans}} R_{\text{ans}} + \lambda_{\text{ver}} R_{\text{ver}} + \lambda_{\text{fmt}} R_{\text{fmt}}, \quad (4)$$

where λ_{ans}, λ_{ver}, and λ_{fmt} are reward weights. We set them to 0.6, 0.25, and 0.15, respectively.

Answer Reward (Yu et al., 2026). The answer reward R_{ans} serves as the primary task-level signal for encouraging correct final predictions by comparing the final answer with the ground-truth.

Verification Reward. The verification reward measures whether the transition from the draft answer to the final answer is consistent with evidence-grounded verification. Given the extracted draft answer d, final answer f, and label answer y:

$$R_{\text{ver}} = \begin{cases} 1, & f = y, \\ -0.5, & d = y \text{ and } f \neq y, \\ -0.5, & d = \emptyset, \\ -1, & f = \emptyset, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

This reward assigns a positive score when the final answer is correct, regardless of whether the model preserves a correct draft or revises an incorrect one. It penalizes cases where a correct draft is revised into an incorrect final answer, thereby discouraging unsupported answer changes. Missing draft or final answers are also penalized to encourage complete verification trajectories.

Format Reward (Zhang et al., 2025b). The format reward R_{fmt} encourages the model to follow the predefined response format.

4 Experiments

4.1 Experimental Setup

Implementations. We instantiate our framework with Qwen2.5-VL-7B (Bai et al., 2025b) as the base model. For supervised fine-tuning, we adopt two stages: Stage-1 is trained for 3 epochs to establish evidence-aware reasoning, and Stage-2 for 1 epoch to enhance final-answer verification and correction. For reinforcement learning, we further train with GRPO (Guo et al., 2025) on 4K high-quality reasoning samples using the proposed evidence-guided rewards. All experiments are conducted on 8 NVIDIA A800 GPUs, with more details reported in the Appendix.

Benchmarks for Evaluation. We conduct a comprehensive experimental analysis to assess the video reasoning capability of our method. For a broad evaluation, we adopt representative video understanding benchmarks, including MLVU (Zhou et al., 2025), Video-MME (Fu et al., 2025a), LongVideoBench (Wu et al., 2024), and LVBench (Wang et al., 2025d). MLVU and Video-MME evaluate general video understanding and reasoning. LongVideoBench and LVBench focus on long-video comprehension, evidence aggregation, and temporal-semantic reasoning.

4.2 Main Results

Table 1 provides a comprehensive comparison of CoVER with representative video MLLMs across four benchmarks, including open-source, agentic, and reasoning-oriented models such as Video-MTR (Xie et al., 2025) and Conan (Ouyang et al., 2025). For a fair comparison, Qwen2.5-VL-7B is evaluated with its default inference setting and a comparable visual-token budget using 128 uniformly sampled frames at approximately 448×448 resolution. CoVER consistently outperforms this baseline, achieving gains of +3.9% on MLVU, +0.7% on VideoMME, +1.1% on VideoMME-Long, +2.3% on LongVideoBench, and +4.6% on LVBench. CoVER also outperforms Video-MTR and Conan on their reported metrics. These results demonstrate the effectiveness of query-expanded visual evidence and answer-clue guided reflection for evidence-aware video reasoning.

CoVER maintains strong general video understanding on MLVU and VideoMME and achieves larger gains on LongVideoBench and LVBench, where sparse evidence requires visual verification. Its competitive performance against 7B-scale reasoning models shows that explicit evidence acquisition and clue-guided verification improve long-video reasoning without increasing model scale.

4.3 Ablation Study

Ablation Study on Training Stages. Table 2 evaluates the contribution of each training stage. CoVER achieves the best performance across all benchmarks, reaching 70.3 on MLVU, 63.8 on

Table 1: **Comparison with SOTA video-LLMs on long-video benchmarks.** We compare CoVER with proprietary, open-source, agentic, and reasoning-oriented models on MLVU, VideoMME, LongVideoBench, and LVBench. * denotes our reproduced result. Δ reports the performance improvement of CoVER over baseline. 768L+32H denotes 768 low-resolution and 32 high-resolution frames.

Models	Size	#Frames	Context	MLVU	VideoMME		LongVideoBench	LVBench
				3∼120 min	Overall	Long	0∼60 min	4101 sec
Proprietary Models								
GPT-4V	–	1fps	–	–	60.7	56.9	–	–
GPT-4o	–	1fps	–	66.2	77.2	72.1	66.7	34.7
Open-Source Video MLLMs								
Video-LLaVA (Lin et al., 2024)	7B	8	–	47.3	40.4	38.1	39.1	–
LLaMA-VID (Li et al., 2024d)	7B	1fps	–	33.2	–	–	–	23.9
ShareGPT4Video (Chen et al., 2024)	8B	16	–	46.4	43.6	37.9	39.7	–
LLaVA-NeXT-Video (Zhang et al., 2024b)	7B	32	–	–	46.5	–	43.5	–
VideoLLaMA2 (Cheng et al., 2024)	7B	32	–	48.5	46.6	43.8	–	–
LongVA (Zhang et al., 2024a)	7B	128	–	56.3	54.3	47.6	–	–
VideoChat2 (Li et al., 2024b)	7B	16	–	47.9	54.6	39.2	–	–
LLaVA-OneVision (Li et al., 2024a)	7B	32	–	64.7	58.2	46.7	–	–
Vamba (Ren et al., 2025)	10B	1024	–	65.9	57.8	–	55.9	42.1
VideoChat-T (Zeng et al., 2025)	7B	12	–	–	46.3	41.9	–	–
Video-XL (Shu et al., 2025)	7B	256	–	64.9	55.5	–	50.7	–
Video-XL-Pro (Liu et al., 2025b)	7B	240	–	70.6	60.0	–	56.7	–
LongVILA (Chen et al., 2025)	7B	256	–	–	60.1	53.0	57.1	–
LongVU (Shen et al., 2024)	7B	1fps	–	65.4	60.6	59.5	–	–
Hour-LLaVA (Lin et al., 2026)	7B	1fps	–	–	63.6	55.0	60.4	45.6
LongVITA-128K (Shen et al., 2025)	14B	256	–	–	66.4	58.8	60.9	–
VILAMP (Cheng et al., 2025)	7B	1fps	–	72.6	67.5	57.8	61.2	45.2
VideoChat-Flash (Li et al., 2024c)	7B	512	–	74.7	65.3	55.4	–	48.2
Open-Source Agent Video MLLMs								
VideoMind (Bhatnagar et al., 2026)	7B	–	–	64.4	58.2	49.2	56.3	40.8
Video-RAG (Luo et al., 2026)	7B	–	–	72.4	62.1	59.8	58.7	–
Open-Source Reasoning Video MLLMs								
Video-MTR (Xie et al., 2025)	7B	32	4k	48.4	59.0	51.0	–	–
Video-R1 (Feng et al., 2026)	7B	32	4k	–	59.3	–	–	–
VITAL (Zhang et al., 2025a)	7B	1024	32k	–	64.1	54.0	–	–
LongVILA-R1 (Chen et al., 2026)	7B	512	132k	–	65.1	55.2	58.0	–
Conan (Ouyang et al., 2025)	7B	32	–	63.4	60.5	–	56.6	39.2
Ours								
Qwen2.5-VL* (Bai et al., 2025b)	7B	128	16k	66.4	63.1	53.6	56.0	42.0
CoVER	**7B**	768L + 32H	**16k**	**70.3**	**63.8**	**54.7**	**58.3**	**46.6**
Δ	–	–	–	+3.9	+0.7	+1.1	+2.3	+4.6

Table 2: **Ablation on training stages.** LongVB denotes LongVideoBench.

Model	MLVU	VideoMME	LongVB	LVBench
CoVER	**70.3**	**63.8**	**58.3**	**46.6**
w/o Stage-1 SFT	67.5	62.9	57.0	45.8
w/o Stage-2 SFT	66.9	63.0	56.2	42.2
w/o RL	63.3	61.5	56.4	42.6
w/o SFT + RL	43.4	50.3	40.8	31.0

VideoMME, 58.3 on LongVideoBench, and 46.6 on LVBench. Removing either Stage-1 SFT or Stage-2 SFT degrades performance, showing that both stages are important for evidence-grounded reasoning. In particular, removing Stage-2 SFT causes a 4.4% drop on LVBench, indicating the importance of supervised verification for long-video understanding. Removing RL leads to a larger drop than removing either SFT stage on MLVU and LVBench, highlighting its role in further optimizing evidence-grounded reasoning after supervised training. The variant without SFT or RL performs worst, showing all training stages are important.

Table 3: **Ablation study on CoVER components.** LongVB denotes LongVideoBench.

Model	MLVU	VideoMME	LongVB	LVBench
w/o QVE	66.4	63.1	56.0	44.0
w/o AGR	67.4	63.3	57.1	45.3
CoVER	**70.3**	**63.8**	**58.3**	**46.6**

Table 4: **CueZoom hyperparameter sensitivity on LVBench.** Each hyperparameter is varied independently, with the default setting ($r = 4, h = 1, \alpha = 0.5$) highlighted in gray.

Param.	r			h			α		
Val.	0	**4**	8	0.5	**1**	2	0.25	**0.5**	1
Score	43.8	**46.6**	45.7	46.0	**46.6**	46.0	45.6	**46.6**	45.0

Ablation Study on CoVER Components. Table 3 ablates CoVER's two core components: Query-Expanded Visual Evidence (QVE) and Answer-Clue Guided Reflection (AGR). Removing QVE or AGR consistently degrades performance

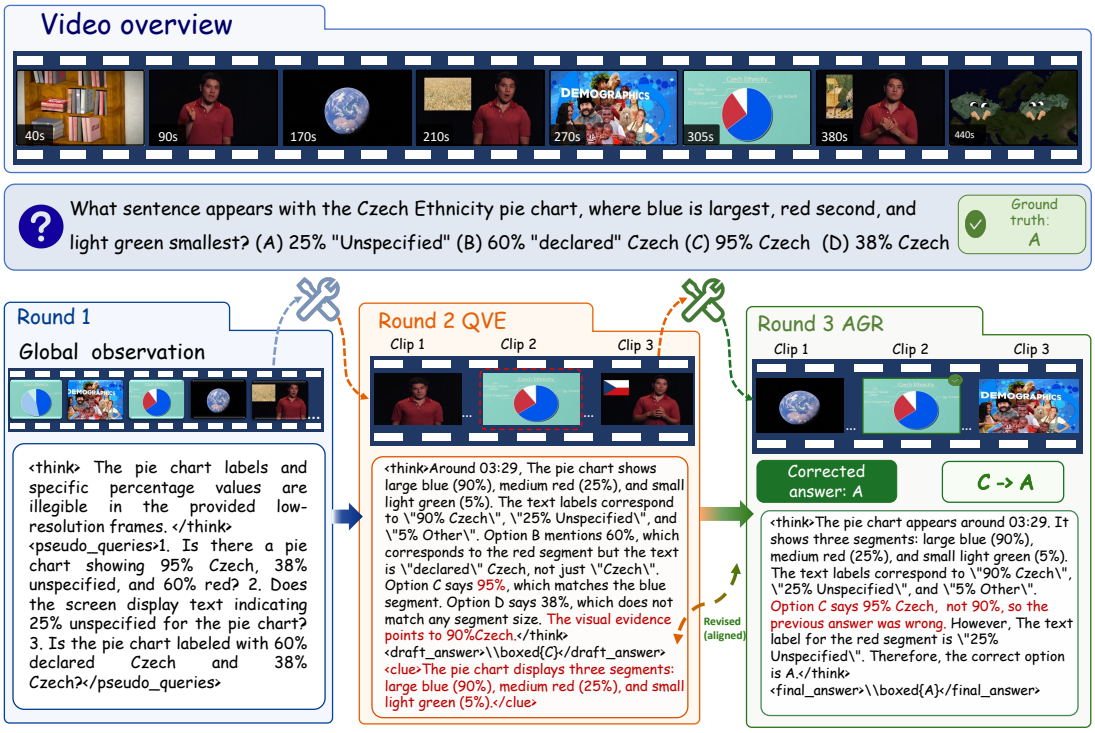

Figure 5: **Case study of CoVER's reasoning process.** This example illustrates CoVER's two-step evidence process: pseudo-query based zoom-in for fine-grained evidence retrieval, followed by answer-specific verification to revise an unsupported draft into the correct answer.

across all benchmarks, showing that both components contribute to CoVER's final performance. Compared with w/o QVE, CoVER improves by +3.9 on MLVU, +0.7 on VideoMME, +2.3 on LongVideoBench, and +2.6 on LVBench. Compared with w/o AGR, CoVER further improves by +2.9, +0.5, +1.2, and +1.3, respectively.

4.4 Hyperparameter Sensitivity.

We analyze the sensitivity of CueZoom Tool's three hyperparameters: the smoothing radius r, Laplacian bandwidth h, and adaptive threshold coefficient α. As shown in Table 4, the default setting ($r = 4$, $h = 1$, $\alpha = 0.5$) achieves the best performance. Removing temporal smoothing ($r = 0$) clearly hurts performance, while overly large radius values also lead to degradation. This indicates that moderate smoothing helps suppress noisy peaks without blurring fine-grained evidence. Results for h and α further highlight the importance of proper temporal aggregation and thresholding.

4.5 Case Study

Fig. 5 shows a representative case of CoVER's evidence-grounded reasoning. When the low-resolution global observation is insufficient, pseudo-queries help retrieve fine-grained zoom-in evidence that improves evidence coverage. The answer-specific clue then guides verification toward evidence directly related to the draft answer, enabling CoVER to identify unsupported predictions and revise them. This case highlights the complementary roles of QVE and AGR in finding missing evidence and verifying answer grounding. Additional qualitative results, including more case studies and representative failure cases, are provided in the appendix.

5 Conclusion

We present **CoVER**, a visual evidence and reflection framework for long-video understanding. CoVER improves Video-LLM reasoning from two aspects: it gathers more complete visual evidence through query-expanded evidence acquisition, and it verifies answers through answer-clue guided visual reflection. By integrating these two modules, CoVER shifts long-video understanding from open-loop answer generation to evidence-centric and visually verifiable reasoning. Experiments show that CoVER improves over the baseline, demonstrating the importance of comprehensive evidence acquisition and visual-feedback reflection for reliable long-video reasoning.

6 Limitations

Although CoVER improves evidence-grounded long-video reasoning, it still depends on the quality of generated evidence-seeking queries and answer-specific clues. Ambiguous or biased queries may retrieve incomplete or misleading evidence. In addition, answer-clue-guided reflection is conditioned on the initial answer, so incorrect initial predictions may lead to biased verification clues. The zoom-in based evidence retrieval process also introduces extra computation compared with direct uniform sampling. Finally, CoVER may be less effective for questions that require holistic video understanding rather than localized visual evidence. Future work can improve query robustness, tool invocation, and retrieval efficiency.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shubhang Bhatnagar, Renxiong Wang, Kapil Krishnakumar, Adel Ahmadyan, Zhaojiang Lin, Lambert Mathias, Xin Luna Dong, Babak Damavandi, Narendra Ahuja, and Seungwhan Moon. 2026. Videomind: Thinking in steps for long video understanding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 406–416.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, and 1 others. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2026. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2025. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pages 18227–18246.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. 2025. Scaling video-language models to 10k frames via hierarchical differential distillation. *arXiv preprint arXiv:2504.02438*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. 2025b. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. 2025. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9262–9281.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jiachun Li, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025a. Towards better chain-of-thought: A reflection on effectiveness and faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10747–10765.

Jiaze Li, Hao Yin, Wenhui Tan, Jingyang Chen, Boshen Xu, Yuxun Qu, Yijing Chen, Jianzhong Ju, Zhenbo Luo, and Jian Luan. 2025b. Revisor: Beyond textual reflection, towards multimodal introspective reasoning in long-form video understanding. *arXiv preprint arXiv:2511.13026*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025c. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102.

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

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, and 1 others. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024d. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and 1 others. 2026. Unleashing hour-scale video training for long video-language understanding. *Advances in Neural Information Processing Systems*, 38:17523–17552.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. 2025a. Longvideoagent: Multi-agent reasoning with long videos. *arXiv preprint arXiv:2512.20618*.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025b. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and 1 others. 2026. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033.

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. 2025. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. 2026. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. 2025. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21197–21208.

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*.

Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Yi-Fan Zhang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Shaohui Lin, and 1 others. 2025. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. *arXiv preprint arXiv:2502.05177*.

Yuan Sheng, Yanbin Hao, Chenxu Li, Shuo Wang, and Xiangnan He. 2025. Sevices: Unifying semantic-visual evidence consensus for long video understanding. *arXiv preprint arXiv:2510.20622*.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. 2026. Tspo: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9368–9376.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128.

Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. 2025a. Dynamic-vlm: Simple dynamic visual token compression for videollm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20812–20823.

Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, and Jiaheng Liu. 2025b. Vr-thinker: Boosting video reward models through thinking-with-image reasoning. *arXiv preprint arXiv:2510.10518*.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. 2025c. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025d. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. 2026. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. *arXiv preprint arXiv:2603.04977*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zhaoyang Wei, Guoliang Wang, Guohua Gao, Yanchao Hao, Mingda Li, Wenchao Ding, Xi Chen, Shizhu He, and Xuehui Yu. 2026. Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10584–10592.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Zhiheng Wu, Tong Wang, Shuning Wang, Naiming Liu, and Yumeng Zhang. 2026a. See further, think deeper: Advancing vlm's reasoning ability with low-level visual cues and reflection. *arXiv preprint arXiv:2604.24339*.

Zhixuan Wu, Quanxing Zha, Teng Wang, Genbao Xu, Wenyuan Gu, Wei Rao, Nan Ma, Bo Cheng, and Soujanya Poria. 2026b. Chain-of-glimpse: Search-guided progressive object-grounded reasoning for video understanding. *arXiv preprint arXiv:2604.14692*.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. 2025. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2026. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184.

Ruoliu Yang, Chu Wu, Caifeng Shan, Ran He, and Chaoyou Fu. 2026. Videodetective: Clue hunting via both extrinsic query and intrinsic relevance for long video understanding. *arXiv preprint arXiv:2603.22285*.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and 1 others. 2025. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong

Shi, Zhengrong Yue, Yi Wang, and 1 others. 2025. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pages 38057–38081.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. 2025a. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, and 1 others. 2025b. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*.

Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. 2025c. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5267–5276.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. *URL https://arxiv.org/abs/2410.02713*, 17.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701.

A Implementation Details and Hyperparameters

During training, we adopt a three-stage pipeline consisting of two supervised fine-tuning stages followed by evidence-guided reinforcement learning. For SFT, we use a learning rate of 5×10^{-6}, a warm-up ratio of 0.05, BF16 precision, gradient checkpointing, and a per-device batch size of 1 with gradient accumulation steps of 4 on 8 GPUs. Stage-1 SFT trains the model for 3 epochs to generate draft answers from the global observation and CLIP-retrieved (Radford et al., 2021) zoom-in clips, while Stage-2 SFT continues training for 1 epoch with answer-clue guided reflection clips for final-answer supervision.

For video processing, CoVER uses a token-controlled fast-slow visual input setting. Specifically, we sample 768 low-resolution global frames to preserve long-range temporal context and retrieve 32 high-resolution evidence frames for fine-grained visual verification. The reproduced Qwen2.5-VL-7B baseline follows its default inference setting and uniformly samples 128 frames at approximately 448×448 resolution, which gives its best observed performance under a comparable visual-token budget. Both settings are controlled to use a comparable visual-token budget of about 16K, ensuring that the comparison mainly reflects evidence selection and verification rather than a larger visual-token budget.

For reinforcement learning, we further optimize the model for 1 epoch with evidence-guided rewards, using a learning rate of 1×10^{-4}, a warm-up ratio of 0.1, and 8 generations per question. We apply LoRA with rank 32, alpha 64, and dropout 0.05. Detailed hyperparameters are reported in Table 5 and Table 6.

Video Processing and Evidence Retrieval. For evidence retrieval, we first uniformly sample N candidate frames from each input video and encode them with the CLIP ViT-L/14 (Radford et al., 2021) visual encoder. Textual cues, including the original question, pseudo-queries, and answer-specific clues, are encoded with the corresponding CLIP text encoder. We compute the cosine similarity between each textual cue and all sampled frames, average the scores over cues, and apply Laplacian Temporal Smoothing to suppress isolated noisy peaks. Frames whose smoothed scores exceed the adaptive threshold are selected, and temporally adjacent selected frames are merged into evidence clips.

Table 5: Hyperparameters used during two-stage supervised fine-tuning.

Hyper-parameters	Value
Stage-1 epochs	3
Stage-2 epochs	1
Batch size	8
Gradient accumulation	4
Learning rate	5×10^{-6}
Warmup ratio	0.05
Number of clips	3
GPUs	8
Optimizer	AdamW

Table 6: Hyperparameters used during GRPO training.

Hyper-parameters	Value
Epochs	1
Batch size	8
Gradient accumulation	4
Rollout	8
Learning rate	1×10^{-4}
Warmup ratio	0.1
Number of clips	3
LoRA rank	32
LoRA alpha	64
GPUs	8
Optimizer	AdamW

B Reward Dynamics

As shown in Fig. 6, all reward components improve steadily during evidence-guided reinforcement learning. The total reward rises rapidly at the beginning and stabilizes after around 2K steps, while the answer and verification rewards continue to increase, indicating improved answer accuracy and evidence-grounded verification. The format reward quickly saturates, suggesting that the model learns the predefined response format early in training. Overall, these trends show stable optimization of both reasoning quality and trajectory consistency.

C Generalization on Short-Video Benchmarks.

We further evaluate CoVER on short-video benchmarks to examine whether the gains on long-video reasoning compromise general short-video understanding. Specifically, we report results on

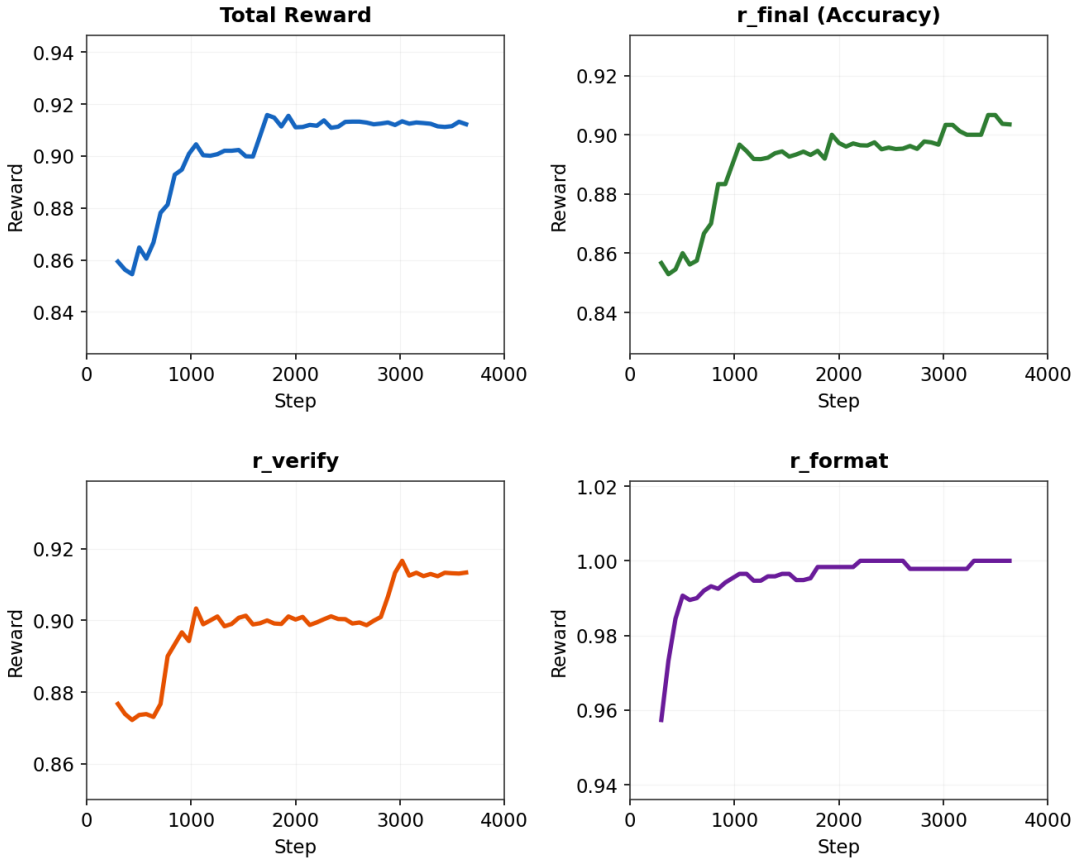

Figure 6: Reward dynamics during GRPO training. The total reward is composed of r_{final}, r_{verify}, and r_{format}, which denote the answer reward, verification reward, and format reward, respectively.

VideoMME-Short and MVBench (Li et al., 2024b), which assess short-duration video question answering and temporal/event-level understanding, respectively. As shown in Table 7, CoVER improves VideoMME-Short from 73.2% to 73.9%, yielding a gain of 0.7%, and also improves MVBench from 65.1% to 66.3%. These results suggest that CoVER preserves, and even slightly enhances, short-video understanding while primarily strengthening evidence-aware reasoning in long-video scenarios.

Table 7: Generalization on short-video benchmarks.

Model	VideoMME-Short	MVBench
Qwen2.5-VL*	73.2	65.1
CoVER	73.9	66.3
Δ	+0.7	+1.2

D Prompt Templates

We design three prompt templates that are shared by both automated data construction and inference. Prompt 1 is used for the initial full-video observation, where the model either directly produces a draft answer with a visual clue or generates pseudo-queries when the evidence is insufficient. Prompt 2 is applied after pseudo-query based retrieval, requiring the model to reason over both the low-resolution full video and high-resolution zoom-in clips to generate a draft answer and a verifiable clue. Prompt 3 uses answer-clue guided reflection clips to check whether the draft answer is visually supported and outputs the final answer. Together, these prompts guide the model to acquire evidence, form a draft prediction, and verify it before producing the final response.

E Qualitative Analysis

We present three visual cases to illustrate different behaviors of CoVER. The first case shows correct-answer preservation, where the draft answer is already supported by the visual evidence, and answer-clue guided reflection keeps it as the final prediction. The second case shows draft answer generation without query-expanded visual evidence: when the global observation is sufficient,

CoVER directly produces a draft answer and an answer-specific clue, then uses answer-clue guided reflection to confirm the final prediction. The third case presents a failure mode, where the retrieved evidence is insufficient or misleading, causing the model to produce or preserve an incorrect final answer. Together, these cases illustrate both the effectiveness and limitations of CoVER in evidence acquisition, answer verification, and visually grounded refinement.

Prompt 1: Initial Global Observation

System Prompt:
You are a helpful video analyst. The red numbers overlaid on each frame represent the timestamp in seconds. You can refer to them for temporal grounding.
Video details:
- Duration: {duration}s | FPS: {fps} | Total frames: {total_frames}
- Question: {question}
Watch the full video carefully.
- If the visual evidence is sufficient, provide:
(1) your reasoning in <think>...</think>
(2) a draft answer in <draft_answer>\boxed{X}</draft_answer> where X is EXACTLY one letter from the given options (A/B/C/D/E).
Do NOT write anything other than a single letter inside \boxed{}.
(3) a single discriminative visual clue in <clue>...</clue>
- If the evidence is insufficient, provide:
(1) your reasoning in <think>...</think>
(2) Pseudo-queries in <pseudo_queries>...</pseudo_queries>.
Pseudo-queries are used to retrieve the missing visual evidence needed to answer the question. Each query will be encoded by a retrieval model (CLIP) and its text-frame similarity scores will be computed across all video frames. The frames with the highest similarity scores will be selected as zoom-in clips and shown to you in the next turn for further reasoning.
Each query must target ONE specific missing visual fact, such as object identity, human action, object state, or temporal relation. Write them as direct English questions. Do NOT be vague.

Prompt 2: Query-Expanded Visual Evidence

You are given two streams of visual input above:
- **Full video (low resolution)**: the complete video at reduced resolution for global context.
- **Zoom-in clips (high resolution)**: the most relevant segments retrieved using your pseudo-queries, selected by computing text-frame similarity between each query and all video frames.
Based on these clips, reason over the visual evidence in <think>...</think>.
In your reasoning, use the full video for temporal context and the zoom-in clips for fine-grained details. Also state what observable facts should be visible in the video if your answer is correct, i.e., form a verifiable hypothesis.
Then provide your candidate answer in <draft_answer>\boxed{X}</draft_answer>, where X is EXACTLY one letter from the given options (A/B/C/D/E).
Do NOT write anything other than a single letter inside \boxed{}. Then extract one short, visually checkable clue in <clue>...</clue> that distinguishes this answer from other options.
You MUST output <draft_answer>\boxed{X}</draft_answer>, even if uncertain, pick the most likely option. You MUST also output <clue>...</clue> after <draft_answer>.
Do NOT output <pseudo_queries> in this turn. You have already retrieved the zoom-in clips; now you must commit to an answer.
Question: {question}

Prompt 3: Answer-Clue Guided Reflection

You are given two streams of visual input above:
- **Full video (low resolution)**: the complete video at reduced resolution for global context.
- **Verification clips (high resolution)**: segments retrieved by computing CLIP similarity between your clue and all video frames, then selecting the highest-scoring frames.
In <think>, use the full video for temporal context and the verification clips for fine-grained checking. Verify whether the visual evidence in the clips supports your clue and candidate answer.
- If the clips provide clear and sufficient evidence that your candidate answer is wrong, revise to the correct answer.
- If the evidence is ambiguous, unclear, or does not strongly contradict your candidate answer, keep your original candidate answer unchanged.
Then output the final answer in <final_answer>\boxed{X}</final_answer>, where X is EXACTLY one letter from the given options (A/B/C/D/E). Do NOT write anything other than a single letter inside \boxed{}.
You MUST output <final_answer>\boxed{X}</final_answer> – no exceptions, no refusals. When in doubt, trust your original candidate answer.

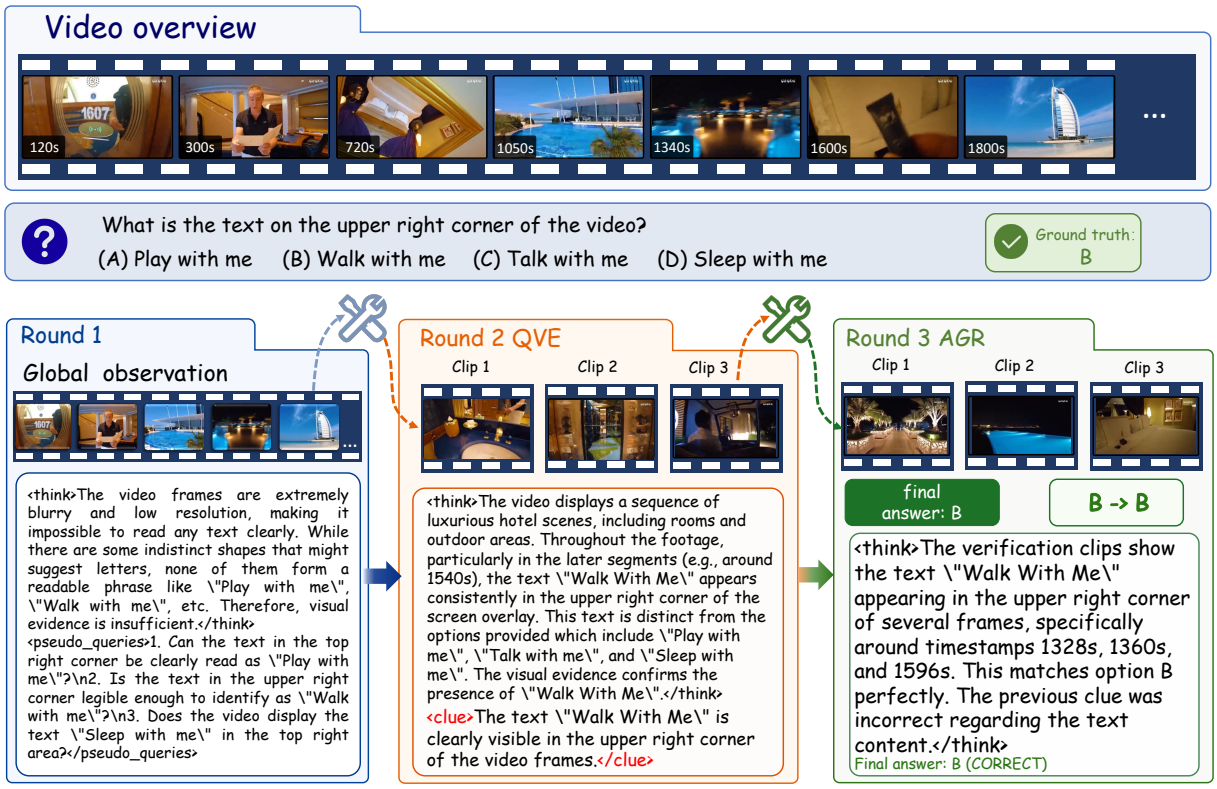

Figure 7: **Verification of a correct draft answer.** CoVER retrieves zoom-in evidence with pseudo-queries and uses answer-clue guided reflection to preserve the correct draft answer.

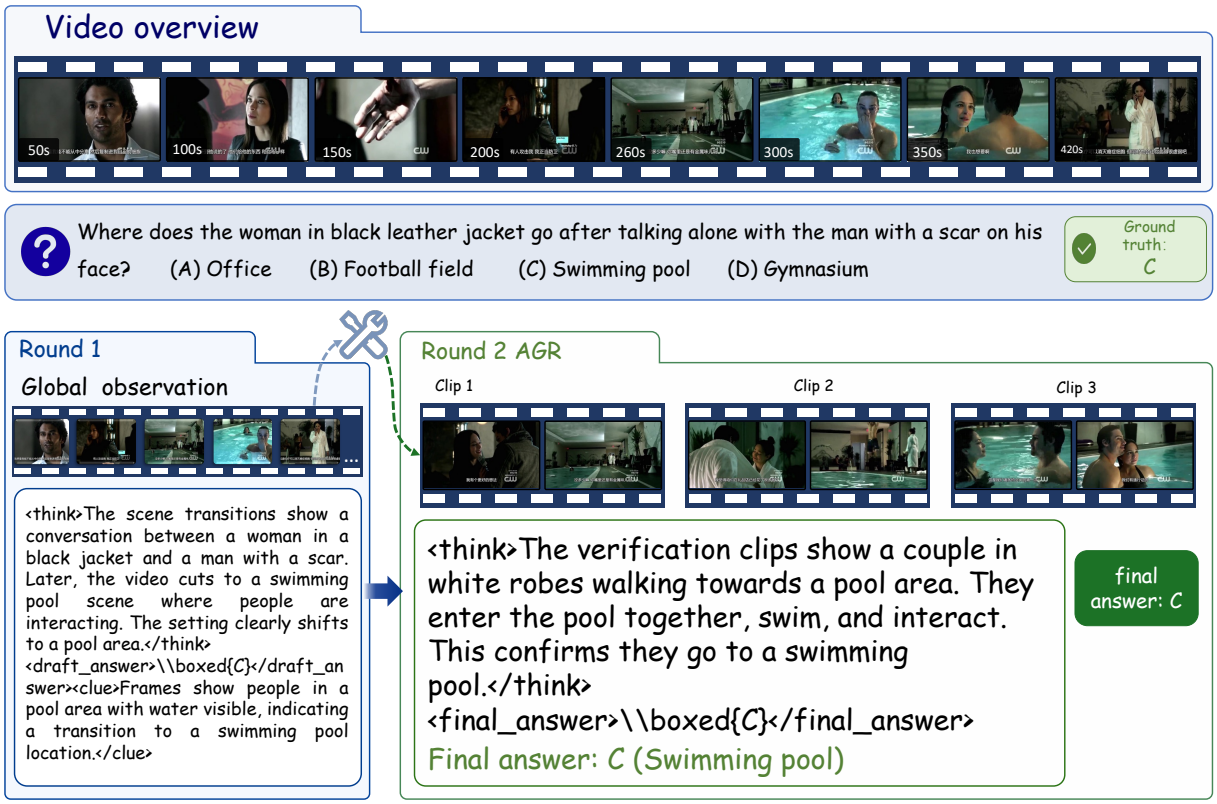

Figure 8: **Draft answer without QVE.** When the global observation is sufficient, CoVER directly produces a draft answer and uses answer-clue guided reflection to confirm the final prediction.

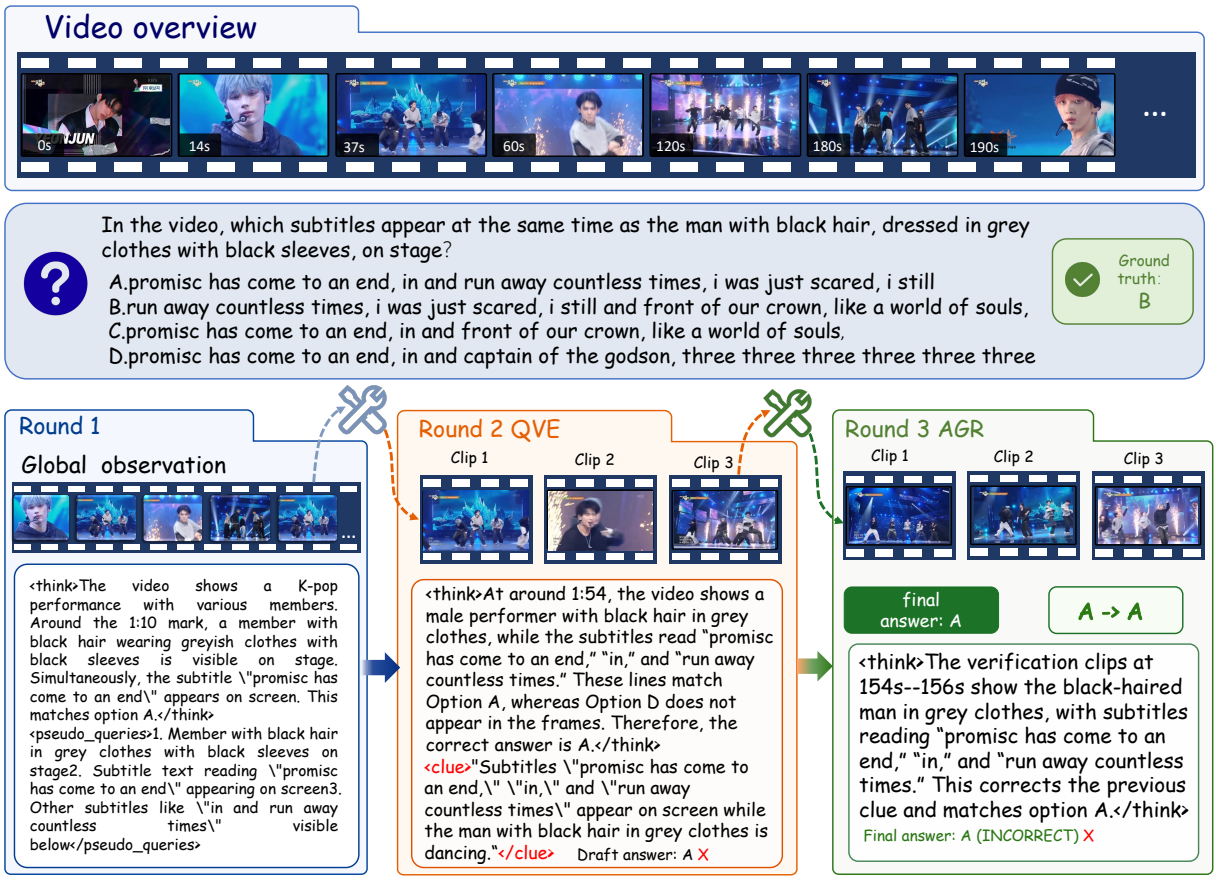

Figure 9: **Failure case of CoVER.** Although CoVER retrieves zoom-in evidence, the verification clips remain insufficient or misleading, causing the model to preserve an incorrect draft answer.