REAL: Reinforced Evidence-Aware Long Video Understanding

Yuan Xie[1] **Tianshui Chen**[2,3*] **Deyu Zhou**[1] **Lionel Ni**[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Guangdong University of Technology
[3]X-Era AI Lab

Abstract

Despite strong benchmark performance, multimodal large language models (MLLMs) remain prone to hallucinations, as they are optimized for linguistic plausibility rather than faithful answering based on observation. This issue is particularly severe in long videos, where critical cues are temporally sparse and often missed by uniform frame sampling. Yet existing methods either operate on fixed pre-sampled inputs, implicitly assuming that all necessary evidence has already been captured, or augment reasoning with retrieval without requiring answers to be grounded in visual evidence. We propose **R**einforced **E**vidence-**A**ware **L**earning (**REAL**), a framework that reformulates long-video question answering as an iterative process of evidence gathering and verification, enabling faithful answering with explicit evidence support, thereby reducing hallucinations. At each reasoning step, the model assesses whether the current observation is sufficient to answer the question, and either searches for missing visual evidence from the video or produces a final answer with explicitly cited evidence. To internalize this reasoning policy, we design an evidence-aware reward for reinforcement learning post-training that jointly requires answer correctness and evidence faithfulness. The latter is evaluated by a pretrained cross-modal consistency verifier that matches the generated evidence descriptions against the visual content of the source frames, eliminating the need for manual annotations. Built on Qwen2.5-VL-7B, REAL achieves state-of-the-art results on four long-video benchmarks (Video-MME, MLVU, LongVideoBench, and EgoSchema) and substantially suppresses hallucinations on the dedicated video-hallucination benchmarks VideoHallucer and ELV-Halluc, with only 96 input frames surpassing the 768-frame base model in both accuracy and hallucination resistance.

1 Introduction

Multimodal Large Language Models have achieved remarkable progress in video understanding tasks [2, 47, 16, 42], driven by the alignment of powerful language reasoning with rich visual perception. Yet they remain prone to *hallucinations*, generating seemingly plausible responses that lack factual grounding in the actual video content [37, 17, 3, 21, 22]. This issue is particularly acute in long video understanding, where decisive evidence is often sparsely distributed over temporal windows of minutes to hours [10, 46, 38]. For a 50-minute video, a query might be, "What color are the gloves when cleaning parts?" Even with 256 sampled input frames, the target object "the gloves" may be entirely absent, leaving the model to infer an answer without factual support from the video.

The challenge arises from two complementary limitations. Standard approaches rely on uniform frame sampling to stay within computational budgets, which can miss critical moments in long,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

content-rich videos. Without the relevant evidence in the input, even a strong reasoning model cannot produce a grounded answer. Moreover, even when relevant frames are available, MLLMs may rely on language priors or dataset biases to generate plausible answers without faithfully grounding them in the visual input [24, 3]. Together, these issues limit both the accuracy and reliability of long-video reasoning: even a correct answer may not reflect faithful visual reasoning, but instead result from an unsupported inference.

Existing efforts to mitigate hallucinations in video MLLMs [17, 37, 24] have made progress but remain fundamentally limited. Most assume that the relevant evidence is already contained in the sampled input frames and focus on improving how the model interprets these fixed inputs, through better decoding strategies, contrastive objectives, or hallucination-aware prompting. These approaches cannot address hallucinations that arise from missing evidence in the first place. In parallel, recent agentic video understanding methods [7, 43, 45, 4] enable models to iteratively retrieve additional visual context. While these methods improve evidence coverage, they primarily treat retrieval as *inference-time context augmentation*: the final answer is not explicitly constrained to faithfully use the retrieved evidence and may still produce fabricated claims. This motivates a shift from retrieval as context augmentation toward a *closed-loop* process that integrates evidence retrieval, verification, and answer generation into the learning process.

We argue that mitigating hallucinations in long-video understanding requires a different paradigm: one that (1) detects when the currently available evidence is insufficient, (2) actively retrieves additional visual evidence on demand, and (3) requires the model to explicitly cite and reason over that evidence before answering. The cited evidence should be verified by the system against the source video, with the verification feedback returned to the model. Crucially, these constraints should not be imposed only at test time; they should be integrated into the training objective so that evidence-grounded reasoning becomes an internalized behavior.

To this end, we propose REAL, a framework for trustworthy long-video understanding that reformulates long-video question answering as an iterative process of evidence seeking and verified reasoning. REAL operates in a SEARCH-ANSWER loop: at each step, the model either generates search cues to retrieve additional evidence frames or produces a final answer with explicit evidence citations. A cross-modal consistency verifier then checks whether the cited evidence matches the source video frames during training, and only answers that are both correct and supported by verified evidence receive high reward. REAL is trained with pure RL post-training under Reinforcement Learning from AI Feedback (RLAIF) using an evidence-aware reward function. Instead of relying on human-annotated grounding supervision, it uses a pretrained cross-modal consistency verifier to provide online feedback on whether the retrieved visual evidence supports the generated answer. The reward jointly considers final-answer correctness and evidence consistency, encouraging the policy to retrieve relevant evidence and faithfully ground its final answer in that evidence.

Our contributions are summarized as follows:

- We propose REAL, a novel framework for trustworthy long-video understanding that transforms long video QA into an iterative evidence-seeking and verified-reasoning process, with a SEARCH-ANSWER decision loop and cross-modal evidence verification.
- We design an evidence-aware reward for pure RL post-training that jointly optimizes answer accuracy and evidence faithfulness. By using a pretrained cross-modal verifier to provide AI feedback, our method internalizes evidence-grounded reasoning into the learning objective without requiring manual grounding annotations.
- Extensive experiments show that REAL achieves state-of-the-art results on four long video understanding benchmarks and significantly suppresses hallucinations on dedicated video-hallucination benchmarks. Notably, with only 96 input frames, it surpasses the 768-frame base model in both task accuracy and hallucination resistance.

2 Related Work

2.1 Long Video Understanding

Recent MLLMs have substantially advanced video understanding, and many have been extended to long-video settings through compression- or memory-based mechanisms, such as LLaMA-

VID [19], MovieChat [34], and Chat-UniVi [14], or expanded token capacity, such as LongVA [42], LongVILA [5], and VideoXL [32]. Yet most perform reasoning in a single pass over fixed inputs, making them vulnerable when decisive evidence is missed during sampling or buried in long, cluttered contexts that fail to attract sufficient attention. A complementary line of work introduces agentic or retrieval-based methods to address this limitation, including VideoAgent [7], DVD [43], DrVideo [27], LVAgent [4], and related frameworks that equip language models with tools to query additional information. While these methods improve evidence coverage over long temporal horizons, they generally treat retrieval as input augmentation, without enforcing that the reasoning process and final answer remain grounded in the retrieved evidence. Our method instead makes retrieval an integral part of reasoning and internalizes faithful evidence use into the learning objective.

2.2 Hallucination Suppression in MLLMs

Hallucination remains a persistent challenge in multimodal reasoning, and recent work has begun to study it more systematically in video MLLMs [37, 17, 3, 24, 22]. Existing mitigation efforts mainly improve robustness under a fixed visual input through better representations [26], prompting or decoding strategies [37], and alignment-based objectives [22]. However, these methods largely assume that the decisive evidence is already present in the sampled frames, and therefore cannot address hallucinations caused by missing evidence in long videos. In contrast, our method extends hallucination mitigation to the input side: when the model detects insufficient evidence, it actively acquires missing visual cues and answers only after sufficient evidence has been gathered, forming a closed-loop process that more effectively suppresses hallucinations in long videos.

2.3 Reinforcement Learning for MLLMs

Recent reasoning-oriented post-training has been strongly shaped by the success of DeepSeek-R1 [12], which showed that large-scale RL can be highly effective when rewards are computed automatically from task outcomes. This paradigm has since been extended to multimodal settings [23, 30, 9, 41], where it has likewise improved reasoning on visual tasks. However, most existing verified-reward pipelines rely solely on outcome-level supervision. While such signals are suitable for final-task correctness, they are insufficient in our setting, where the model must not only answer correctly but also ground its answer in dynamically retrieved visual evidence. Rather than relying on human-annotated grounding data or purely rule-based rewards, REAL adopts RLAIF with a frozen cross-modal verifier that automatically evaluates whether the cited evidence supports the answer. This online AI feedback is then used as the RL reward, enabling scalable post-training that jointly optimizes answer correctness and evidence faithfulness without additional grounding annotations.

3 Method

REAL reformulates long-video question answering as a closed-loop process of evidence gathering, evidence-grounded reasoning, and evidence verification. We first formalize the task, with Figure 1 providing an overview of the framework. We then describe the key modules of the iterative reasoning process, followed by the evidence-aware RL post-training objective that internalizes this behavior.

3.1 Iterative Evidence-Grounded Reasoning

Given a long video V and a natural-language question q, we formulate long-video understanding as generating a correct answer $\hat{y}$ along with verifiable visual evidence $\mathcal{E} = \{(t_i, d_i)\}_{i=1}^{K}$, where each t_i is a cited timestamp and d_i is the corresponding generated frame description. We first uniformly sample a sparse set of N_0 frames from V to form the initial global observation $\mathcal{O}_0 = \{f_1, f_2, \ldots, f_{N_0}\}$. Together with the question q, these frames serve as input to the first reasoning step. At each subsequent step $t \geq 1$, the MLLM receives the question q, the current observation $\mathcal{O}_t$, and a short system feedback message m_{t-1} from the previous step (e.g., a reminder to change the search cues when retrieval is empty), and outputs one of the following two structured actions.

SEARCH When the current evidence is insufficient for a reliable answer, the model generates a set of search cues $\mathcal{C}_t = \{c_1, c_2, \ldots\}$. Each cue can be either a semantic description of the missing visual information (e.g., "a person picking up a red cup", "the black scoreboard") or a timestamp indicating

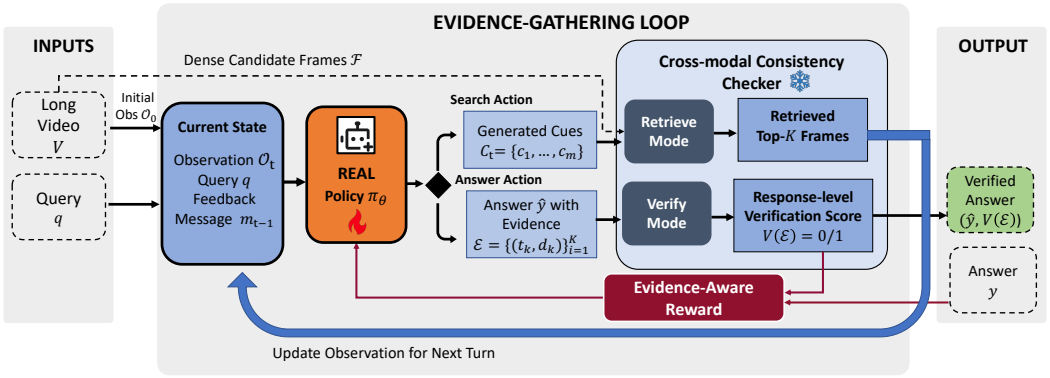


Figure 1: Overview of REAL. Given a long video and a question, the model starts with a sparse global sampling and iteratively decides between Search and Answer. Upon answering, a cross-modal verifier checks whether the cited evidence is consistent with the video. During training, an evidence-aware reward jointly optimizes answer correctness and evidence faithfulness.

a relevant moment. These cues are then used by the system for evidence gathering, which retrieves the most relevant frames from the video to form the next observation set $\mathcal{O}_{t+1}$.

ANSWER Once the accumulated evidence is deemed sufficient, the model outputs the final answer y together with a set of evidence citations $\mathcal{E} = \{(t_i, d_i)\}_{i=1}^{K}$, where each citation specifies a frame timestamp t_i and a textual description d_i of the corresponding frame content that supports the answer.

Evidence Gathering. Given the generated search cues $\mathcal{C}_t$, the system gathers additional evidence from a candidate pool of densely sampled frames $\mathcal{F} = \{f_1^c, f_2^c, \ldots, f_M^c\}$ extracted from the whole video. For each search cue $c_j \in \mathcal{C}_t$, we compute its cross-modal similarity to every candidate frame using a pretrained vision-language encoder ϕ:

$$s(c_j, f_i^c) = \mathrm{sim}\big(\phi_{\text{text}}(c_j), \phi_{\text{vis}}(f_i^c)\big), \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. When multiple cues are generated, candidate frames are retrieved independently for each cue, merged with deduplication, and scored by their maximum similarity across cues. Timestamp cues, when present, are directly resolved to their nearest candidate frames and inserted before semantic retrieval results, which are then deduplicated against them. The top-K candidates under the frame budget are finally selected to form the next observation:

$$\mathcal{O}_{t+1} = \mathrm{TopK}(\mathcal{R}_t, K), \tag{2}$$

where $\mathcal{R}_t$ denotes the merged candidate set. The updated observation $\mathcal{O}_{t+1}$ is passed to the next reasoning step.

Evidence Verification. During training, we use a cross-modal consistency check on cited evidence as a stricter faithfulness signal for RL. For each ANSWER action, every cited description is verified against its referenced frame. Only previously observed frames are admissible as evidence; citations to unobserved timestamps are treated as unsupported. For each evidence citation (t_i, d_i), we compute

$$v_i = \mathrm{sim}\big(\phi_{\text{text}}(d_i),\ \phi_{\text{vis}}(f_{t_i})\big), \tag{3}$$

where f_{t_i} is the frame at timestamp t_i. A citation is considered *verified* if $v_i > \tau$, with τ a consistency threshold. The overall consistency score for an answer is

$$V(\mathcal{E}) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}[v_i > \tau], \tag{4}$$

This verification is applied only during training and removed at inference time for two reasons. First, once optimized with the faithfulness reward, the MLLM already learns to produce evidence descriptions that are highly consistent with the cited frames, making an explicit online check unnecessary. Second, the verifier is itself an encoder-based feedback model and thus imperfect; enforcing it at inference time could inadvertently reject valid answers and unnecessarily constrain final accuracy.

In the default setting, the same frozen pretrained encoder is used for both evidence gathering and evidence verification, but in different roles, with no gradients propagated through it during policy optimization. This shared-encoder design is a practical choice rather than a requirement of REAL; the two modules can also use different frozen encoders. During SEARCH, it is used in a recall-oriented manner to retrieve and rank candidate frames under the frame budget. During ANSWER, it is used in a precision-oriented manner: each cited evidence pair (t_i, d_i) must exceed a consistency threshold τ. This design preserves broad evidence gathering during intermediate reasoning while enforcing strict grounding for the final answer. Implementation details are provided in Appendix A.1.

3.2 Evidence-Aware Policy Optimization

REAL is trained with pure RL post-training, without supervised evidence-grounding annotations. The optimization objective encourages an evidence-driven reasoning policy, with evidence faithfulness assessed by a frozen cross-modal verifier, enabling reward construction from AI feedback rather than human labels. Formally, the MLLM defines a policy π_θ over the structured SEARCH/ANSWER actions. A valid answer must be produced within at most $T_{\max}$ steps. Invalid or failed intermediate actions are returned via system feedback to guide subsequent decisions. Only the final ANSWER step of a valid trajectory contributes to the terminal reward; invalid trajectories receive zero reward.

Evidence-Aware Reward Function. We use a terminal reward that couples answer correctness with evidence faithfulness. Consider a rollout whose final step is an ANSWER action producing a predicted answer $\hat{y}$ and evidence $\mathcal{E} = \{(t_i, d_i)\}_{i=1}^{K}$. Answer correctness is measured by exact match against the ground-truth answer y, while evidence faithfulness is measured by the response-level verification score in Eq. 4. To reflect the precision-oriented role of final verification, we adopt a strict *All-Pass Rule*: a response passes evidence verification only if all cited evidence satisfies the consistency check, i.e., $V(\mathcal{E}) = 1$. The binary main reward is therefore

$$R(\hat{y}, \mathcal{E}) = \begin{cases} 1, & \text{if } R_{\text{acc}}(\hat{y}, y) > 0 \text{ and } V(\mathcal{E}) = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This strict design suppresses hallucinated correctness, where the model reaches the correct answer while citing unsupported evidence. In practice, the final reward is dominated by this answer-accuracy-based main reward (weight 0.9), together with an auxiliary format reward (weight 0.1) that encourages valid structured outputs.

Optimization and Curriculum. We optimize the policy with Group Relative Policy Optimization (GRPO) [12]. For each training question, we sample a group of G rollouts from the current policy π_θ^{old} and assign each rollout a terminal reward according to the weighted objective above. Each rollout follows the full SEARCH-ANSWER pipeline. Notably, SEARCH actions receive no explicit step-wise reward; their utility is reflected only through their contribution to the final verified answer. The advantage of rollout i is computed relative to the group:

$$\hat{A}_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \tag{6}$$

where μ_R and σ_R are the mean and standard deviation of rewards within the group. The policy is updated by maximizing

$$\mathcal{L}(\theta) = \mathbb{E}\left[\sum_{i=1}^{G} \hat{A}_i \cdot \min\left(\rho_i, \text{clip}(\rho_i, 1 \pm \epsilon_c)\right)\right] - \beta \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{7}$$

where ρ_i denotes the likelihood ratio for rollout i, ϵ_c is the clipping range, and the KL term regularizes the policy against a reference policy π_{ref}. With terminal-only reward, intermediate decisions are credited through their contribution to a successful verified answer, while GRPO's group-relative advantages help identify effective trajectories.

The terminal reward in Eq. 5 is highly sparse at the start of training. To stabilize optimization, we adopt a two-stage curriculum. Stage I trains the model in a single-step setting with only ANSWER outputs and valid evidence citations, allowing the answer-with-evidence branch to become stable

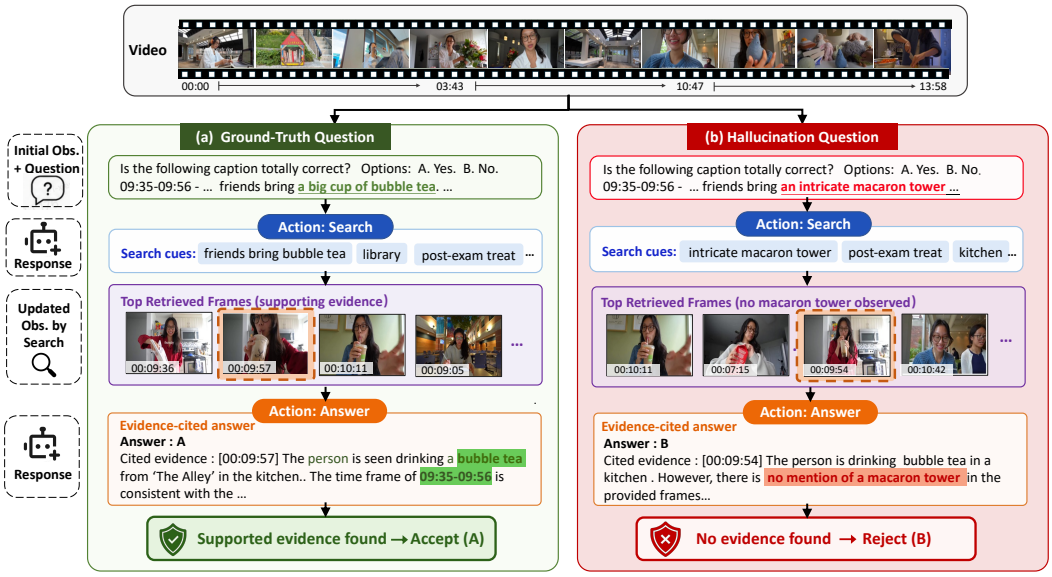


Figure 2: Case study on ELV-Halluc.REAL answers the ground-truth question correctly and rejects the hallucinated counterpart when no supporting evidence is retrieved. Dashed boxes mark the frames cited in the final answer.

before introducing search. Stage II enables evidence gathering and trains the full multi-step SEARCH-ANSWER policy. During the early phase of this stage, we add a temporary bonus to SEARCH actions when the tool-use rate is below a preset threshold, encouraging the model to discover nontrivial search behavior. Once search usage exceeds the threshold, this bonus is removed and training proceeds using only the terminal reward.

Emergent Search Strategy. Under this training scheme, search behavior is learned without explicit supervision on search timing or any SFT demonstrations; instead, it emerges from the terminal reward and GRPO-based credit assignment. When the current evidence is insufficient, answering immediately is unlikely to succeed, since either the answer will be incorrect or the cited evidence will fail verification. The policy is thus encouraged to issue SEARCH actions when additional evidence is needed, and to ANSWER only when it can produce a verifiably grounded response.

4 Experiments

4.1 Setup

Model and Rollout Settings. REAL is built on Qwen2.5-VL-7B [2] and uses LongCLIP-L [40] as the default frozen encoder for evidence retrieval and verification. We additionally evaluate different encoder configurations using DINOv3 [33] and ASP-CLIP [4] as alternative choices. The total input frame budget is 96 frames, with a maximum of $T_{\max} = 3$ iterations in the SEARCH-ANSWER loop. This budget is allocated across the rollout: 72 frames for the initial observation and up to 24 additional frames from at most two search rounds. The candidate pool consists of up to 256 frames uniformly subsampled from the full video after 1-FPS sampling.

Training Details. We perform RL post-training in the `verl` [31] framework. For each prompt, we sample $G = 15$ rollouts and train for 3 epochs, using a constant learning rate of 1×10^{-6} and a KL regularization coefficient of 0.001. Training runs on two 8-H800 nodes: 1 GPU is reserved for the consistency-checking model, while the remaining 15 GPUs are used for policy training, yielding a global batch size of 60. The training set is constructed from NExT-GQA [39], QVHighlights [15], and LongVideoReason [5], retaining samples with duration $\geq$ 30s for the first two datasets and $\geq$ 600s for LongVideoReason, for a total of 13K examples.

Table 1: Main results on four standard long-video benchmarks and two hallucination benchmarks. For Qwen2.5-VL-7B, 768f denotes the official setting, while 96f denotes our re-evaluation under the matched 96-frame budget. Best and second-best per category are **bolded** and underlined, respectively. * denotes results from our re-implementation/evaluation under different input settings.

Model	Size	Standard Long-Video Benchmarks				Hallucination Benchmarks	
		Video-MME	MLVU	LongVideoBench	EgoSchema	VideoHallucer	ELV-Halluc
		Overall (w/o sub.)	Test	Val	Subset	Overall	Overall.
Proprietary Models							
GPT-4o [13]	-	71.9	**54.9**	**66.7**	**72.2**	**53.3**	**8.7**
Gemini-1.5-Pro [35]	-	**75.0**	-	64.0	71.1	37.8	-
Qwen2.5-VL-7B(768f) [2]	7B	65.1	**51.4***	56.0	65.0	-	-
Open-Source Baselines							
Video-LLaVA [20]	7B	39.9	30.7	39.1	36.8	17.8	-
VideoChat2 [18]	7B	39.5	30.1	39.3	-	7.8	-
LLaVA-OV-7B [16]	7B	58.2	-	56.3	60.1	48.5	9.9
LLaVA-Video-7B [44]	7B	63.3	-	58.2	57.3	-	3.6
Video-CCAM [8]	9B	50.3	42.9	43.1	-	-	-
LongVA [40]	7B	52.6	41.1	47.8	-	-	-
LongVILA [5]	7B	60.1	-	57.1	67.7	-	-
Video-XL [32]	7B	55.5	45.6	50.7	-	-	-
VideoAgent [36]	-	56.0	-	-	60.2	-	-
Video-R1 [9]	7B	61.4	47.6	-	51.8	-	-
PaMi-VDPO [6]	7B	58.5	-	-	-	53.8	-
Qwen2.5-VL-7B(96f)* [2]	7B	57.1	46.6	54.7	63.2	43.1	17.3
REAL(Ours)	7B	**63.6**	**53.9**	**58.6**	**67.8**	**54.8**	**19.8**

Benchmarks. We evaluate long-video understanding on several challenging benchmarks, including Video-MME [10], MLVU [46], LongVideoBench [38], and EgoSchema [29]. For hallucination analysis, we additionally use VideoHallucer [17] and ELV-Halluc [25] for a more fine-grained evaluation of different hallucination types.

4.2 Main Results.

Table 1 reports the main accuracy results on four standard long-video benchmarks and two hallucination benchmarks. For the latter, accuracy is defined by a stricter pair-based metric; formal definition and detailed analysis are deferred to Sec. 4.3. We compare REAL against top closed-source models, strong open-source baselines of comparable scale, and hallucination-oriented methods such as PaMi-VDPO [6], with particular emphasis on fair comparisons against its backbone, Qwen2.5-VL-7B.

REAL consistently improves the matched 96-frame backbone baseline across all six benchmarks, by +6.5% on Video-MME, +7.3% on MLVU, +3.9% on LongVideoBench, +4.6% on EgoSchema, +11.7% on VideoHallucer, and +2.5% on ELV-Halluc. Since both are evaluated under the same visual budget, these gains are attributable not to scaling visual input, but to more effective evidence acquisition and evidence-grounded reasoning. On the four standard long-video benchmarks, REAL achieves the strongest overall performance among open-source models of comparable scale while remaining competitive with proprietary systems. Notably, despite using only 96 input frames, it surpasses the official 768-frame Qwen2.5-VL-7B baseline on MLVU (53.9% vs. 51.4%), LongVideoBench (58.6% vs. 56.0%), and EgoSchema (67.8% vs. 65.0%), while remaining close on Video-MME (63.6% vs. 65.1%). This highlights the frame efficiency of REAL: actively seeking missing evidence is more effective than relying on a larger but fixed single-pass input. The gains also extend to hallucination robustness. Measured by the primary metrics in the last two columns, REAL substantially outperforms its 96-frame backbone on both VideoHallucer and ELV-Halluc, but also surpasses strong proprietary models such as GPT-4o and Gemini-1.5-Pro.

4.3 Hallucination Suppression.

This section provides a detailed analysis of hallucination mitigation on two dedicated benchmarks. VideoHallucer evaluates intrinsic and extrinsic hallucinations, while ELV-Halluc targets semantic aggregation hallucinations. Both benchmarks employ adversarial paired binary questions: one assesses video-grounded recognition, and the other tests rejection of plausible but unsupported claims. Their primary metrics, *Overall Acc.* and *Avg Acc.*, count a sample as correct only if both questions in the pair are answered correctly, making them substantially stricter than standard accuracy. Unless otherwise noted, baseline results are drawn from the original benchmark or method papers under

Table 2: Hallucination evaluation. Arrows indicate preferred direction. * denotes results from our own evaluation.

(a) VideoHallucer

Model	Basic Acc.↑	Halluc. Acc.↑	Overall Acc.↑
GPT-4o	75.1	74.2	53.3
Gemini-1.5-Pro	83.6	42.3	37.8
VideoChat2 [18]	29.7	25.8	7.8
Video-LLaVA [20]	95.1	20.3	17.8
MiniGPT4-Video [1]	79.4	28.6	22.3
LLaVA-OV-7B [16]	84.1	60.8	48.5
PaMi-VDPO [6]	82.8	69.9	53.8
Qwen2.5-VL-7B(256f)*	94.1	45.8	40.9
Qwen2.5-VL-7B(96f)*	92.8	48.6	43.1
+REAL (Ours)	80.6	**68.5**	**54.8**

(b) ELV-Halluc

Model	Avg Acc.↑	Avg Diff.↓	SAH Ratio ↓
GPT-4o	8.7	0.9	1.0
Video-chatgpt [28]	2.0	0.2	0.1
LLaVA-OV-7B [16]	9.9	2.8	3.0
LLaVA-Video-7B [44]	3.6	-0.6	-0.6
InternVL3-8B [47]	16.3	5.9	6.8
Qwen2.5-VL-7B(256f)	18.1	7.6	8.8
Qwen2.5-VL-7B(96f)*	17.3	7.6	9.6
+REAL (Ours)	**19.8**	**5.2**	**6.3**

default inference settings; detailed metric definitions and inference configurations are provided in Appendix B.1 and Appendix B.2, respectively. For Qwen2.5-VL-7B, we report both 96-frame and 256-frame settings, where 96 frames match the final REAL budget for a fair comparison, and 256 frames follow ELV-Halluc as a stronger baseline. Under the 32K context limit with 128×28×28 per-frame resolution, 256 frames represent the maximum supported input. As shown in Table 2, REAL achieves the highest primary metric on both benchmarks, substantially improving over its backbone and even surpassing strong proprietary models including GPT-4o and Gemini-1.5-Pro. Figure 2 provides a representative example. For the ground-truth query, REAL retrieves supporting evidence and answers correctly. For the hallucinated counterpart, which introduces an unsupported detail (an "intricate macaron tower"), REAL finds no supporting frame and rejects the claim. The example illustrates that REAL grounds both positive and negative judgments in retrieved evidence.

On **VideoHallucer**, the gap between *Basic Acc.* and *Hallucinated Acc.* reveals whether a model can reject plausible but unsupported claims rather than merely answer the original question correctly. Models that are overly prone to accept plausible textual claims can score highly on basic questions. For example, Video-LLaVA achieves 95.1% *Basic Acc.* but only 20.3% *Hallucinated Acc.*, resulting in a low *Overall Acc.* of 17.8% (Table 2a). This shows that *Overall Acc.* is a more faithful indicator of hallucination suppression than basic-question accuracy alone. Built on Qwen2.5-VL-7B, REAL improves *Hallucinated Acc.* to 68.5% and achieves the best *Overall Acc.* of 54.8%, substantially outperforming open-source models and even surpassing GPT-4o.

ELV-Halluc targets semantic aggregation hallucinations, where semantics from different events in the same video are incorrectly conflated. Besides the primary metric *Avg Acc.*, it reports *Avg Diff.* and *SAH Ratio* to quantify the gap between in-video semantic aggregation errors and ordinary out-of-video hallucinations. The benchmark is particularly challenging because it involves long videos with multiple confusable events, and increasing visual input alone does not reliably resolve this issue. By requiring the final answer to be grounded in the correct temporal evidence, REAL improves *Avg Acc.* from 17.3% to 19.8%, while reducing *Avg Diff.* from 7.6% to 5.2% and *SAH Ratio* from 9.6% to 6.3% relative to the 96-frame Qwen2.5-VL-7B baseline. It also outperforms the stronger 256-frame baseline on all three metrics, yielding the best overall trade-off on ELV-Halluc.

Consistency-Grounding Gap. Evidence consistency does not necessarily guarantee sufficient support for the final answer. To examine this distinction, we audit 1,000 REAL outputs using two independent VLM judges, GPT-4o and Gemini 3 Flash [11]. Both judges report smaller consistency–support gaps for correct than incorrect predictions (11.8 vs. 22.3 and 12.6 vs. 32.5 percentage points, respectively). These results suggest that evidence consistency is an informative but incomplete signal for answer-level grounding. Detailed results are provided in Appendix B.3.

4.4 Ablation Studies

Effect of Reward and Evidence Gathering. Table 3 reports a controlled ablation of REAL's two core components: Evidence Gathering (EG) and Evidence-Aware Reward (EAR). All variants share the same backbone, training data, and 96-frame budget, differing only in component activation. Vanilla RL, without EG or EAR, yields modest gains, particularly on hallucination suppression. Adding EG improves long-video understanding but minimally affects hallucination, showing that additional evidence alone is insufficient for faithful reasoning. In contrast, EAR consistently boosts performance on both Video-MME and VideoHallucer, showing explicitly rewarding evidence faith-

Table 3: Ablation on EG and EAR for long-video understanding and hallucination suppression.

Variant	EG	EAR	Video-MME		VideoHallucer	
			Long	Overall	Halluc.	Overall
Qwen2.5-VL-7B	✗	✗	47.6	57.1	48.6	43.1
+ Vanilla RL	✗	✗	51.6	60.7	49.3	44.2
+ EG	✓	✗	52.3	61.3	49.0	45.7
+ EAR	✗	✓	53.7	63.0	58.6	51.0
REAL (Ours)	✓	✓	**54.9**	**63.6**	**68.5**	**54.8**

Figure 3: Training dynamics under different cross-modal encoders.

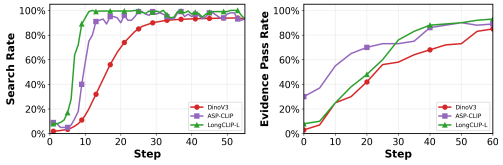


Table 4: Final performance across encoder variants on Video-MME and ELV-Halluc.

Visual Encoder	Video-MME		ELV-Halluc		
	Long	Overall	Avg Acc.↑	Avg Diff.↓	SAH Ratio↓
DINOv3	53.9	64.0	19.3	6.9	8.2
ASP-CLIP	54.2	65.4	19.0	8.4	7.9
LongCLIP-L	54.9	63.6	19.8	5.2	6.3

fulness, substantially reduces hallucinated answers even without search. The full REAL framework combining EG and EAR attains the highest accuracy and hallucination robustness, confirming their complementarity: EG enhances evidence acquisition, while EAR ensures faithful utilization.

Encoder Design Sensitivity. We investigate whether REAL depends critically on using the same encoder for evidence retrieval and verification. *(a) Encoder Decoupling.* We first perform a more direct decoupling experiment by fully training REAL with different frozen encoders for retrieval and verification. With LongCLIP for retrieval and DINOv3 for verification, REAL achieves 61.81% on Video-MME and 57.60% on VideoHallucer. Reversing the configuration, with DINOv3 for retrieval and LongCLIP for verification, yields 63.50% and 55.60%, respectively. These results show that retrieval and verification remain effective with different encoders, indicating that encoder sharing is not a requirement of REAL. *(b) Shared Encoder Variants.* We further evaluate the original shared-encoder design with different cross-modal encoders. Figure 3 shows consistent training dynamics across encoder variants: the search rate increases rapidly before stabilizing, while the evidence consistency pass rate improves steadily. Table 4 further confirms strong performance across all three encoders. Together, these results suggest that REAL is robust to the choice of encoder, while encoder sharing remains a practical design choice rather than a strict requirement.

Scalability with Frame Budget. Figure 4 shows the performance of REAL under 32-, 64-, and 96-frame budgets on ELV-Halluc. Two trends are clear. First, under every matched budget, REAL consistently outperforms the backbone, validating the effectiveness of evidence-driven reasoning . Second, REAL scales favorably with additional frames: accuracy improves steadily from 32 to 96 frames, while SAH Ratio remains consistently lower than the backbone and further decreases at 96 frames. These results indicate that REAL not only uses limited frame budgets more effectively, but also retains strong headroom as the visual budget increases.

Search Behavior Analysis. On VideoHallucer, Table 5 reveals several behaviors that highlight the advantage of REAL. Hallucinated questions trigger more intensive search: the probability of requiring two search rounds rises from 9.7% to 27.5%, and the average number of search rounds increases from 1.1 to 1.3. This suggests that the model can detect insufficient evidence and allocate additional retrieval effort to harder cases. Hallucinated questions also induce more complex but less efficient retrieval: the mean number of cues rises from 2.9 to 3.5, while the average number of frames per search drops from 11 to 8. This likely reflects the presence of misleading or unsupported elements, which makes cue-conditioned retrieval less likely to match genuine evidence. Although hallucinated questions remain more difficult overall (68.5% vs. 80.6%), these results indicate that REAL learns an adaptive evidence-driven search policy for robust long-video reasoning and hallucination mitigation.

5 Conclusion and Limitations

We present REAL, a framework for trustworthy long-video understanding built on an evidence-driven decision process. By unifying evidence gathering, evidence-grounded reasoning, and verification

Figure 4: Scaling with frame budget on ELV-Halluc.

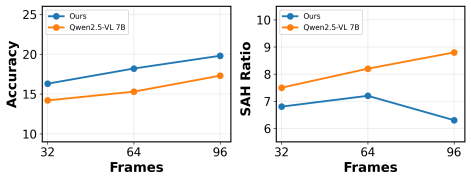


Table 5: Search behavior analysis on VideoHallucer. We report the distribution of search rounds and search statistics for basic and hallucinated questions.

Type	Search-round Distribution		Search Statistics		
	$P(n{=}1)$	$P(n{=}2)$	Mean Cues	Avg Frames / Search	Acc.
Basic	90.3%	9.7%	2.9	11	80.6%
Hallucinated	72.5%	27.5%	3.5	8	68.5%

within a closed-loop policy, and internalizing this behavior through RL, REAL consistently improves performance on both standard long-video and dedicated hallucination benchmarks.

Despite its promising results, our work still has several limitations. First, although we verify the robustness of REAL across multiple encoders, the overall capability of the framework remains bounded by the quality of the verifier, which mainly captures coarse semantic alignment at the frame level and may be less reliable for fine-grained temporal relations or subtle discrepancies. Second, REAL currently grounds answers primarily in visual evidence. However, some long-video questions require grounding in additional modalities, such as narration, audio, or subtitles. Extending the framework to support multimodal grounding is therefore an important direction for future work.

Acknowledgements and Disclosures

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62676107 and 62632003.

References

[1] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[4] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents, 2025. URL `https://arxiv.org/abs/2503.10200`.

[5] Yukang Chen, Fuzhao Xue, Dacheng Li, et al. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations (ICLR)*, 2025.

[6] Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*, 2025.

[7] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024. doi: 10.48550/arXiv.2403.11481. URL `https://arxiv.org/abs/2403.11481`.

[8] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. URL `https://arxiv.org/abs/2503.21776`.

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[11] Google DeepMind. Gemini 3 flash, 2025. URL `https://deepmind.google/models/model-cards/gemini-3-flash/`.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[14] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[15] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[17] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13723–13733, 2025.

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

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

[20] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 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, 2024.

[21] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[22] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025. URL `https://arxiv.org/abs/2503.01785`.

[24] Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding, 2025. URL `https://arxiv.org/abs/2508.21496`.

[25] Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[26] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160, 2024.

[27] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18936–18946, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 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, 2024.

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

[30] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

[31] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

[32] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 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, 2025.

[33] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[34] Enxin Song, Tao Chen, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[36] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

[37] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

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

[39] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.

[40] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024.

[41] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025. URL `https://arxiv.org/abs/2503.12937`.

[42] Pan Zhang et al. Long context transfer from language to vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[43] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025. doi: 10.48550/arXiv.2505.18079. URL `https://arxiv.org/abs/2505.18079`.

[44] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025. URL `https://arxiv.org/abs/2410.02713`.

[45] Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. Llm-based agentic reasoning frameworks: A survey from methods to scenarios. *arXiv preprint arXiv:2508.17692*, 2025. doi: 10.48550/arXiv.2508.17692. URL `https://arxiv.org/abs/2508.17692`.

[46] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406, 2024.

[47] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Implementation Details

A.1 Search and Verification Hyperparameters

At each SEARCH step, we keep at most five search cues. Cues can be either semantic descriptions or timestamps. Timestamp cues are directly resolved to the nearest frame in the candidate pool, while semantic cues are used for cross-modal retrieval.

For evidence gathering, the candidate pool contains at most 256 frames per video. Candidate frames are encoded by the frozen cross-modal encoder, and retrieval scores are computed by the dot product between text and frame embeddings. When multiple semantic cues are provided, retrieval is performed independently for each cue, followed by merging and deduplication. If timestamp cues are present, their nearest frames are inserted first, and semantic retrieval results are deduplicated against them before final ranking.

For evidence verification, REAL computes a frame-text consistency score for each cited evidence pair (t_i, d_i), as defined in Eq. 4. In the default LongCLIP-based setup, the consistency threshold is fixed to $\tau = 65.0$. The LongCLIP threshold is calibrated once by manual inspection on a small set of evidence-frame pairs collected from early training rollouts. Concretely, we sample roughly 100 description-frame pairs, manually label whether each pair is visually consistent, and choose a threshold that best separates positive and negative cases. This threshold is then fixed throughout

training and evaluation. For alternative cross-modal encoders such as DINOv3 and ASP-CLIP, we calibrate the verification threshold by matching the empirical acceptance rate of the LongCLIP verifier on a held-out calibration set. Specifically, let N be the number of evidence-frame pairs accepted by LongCLIP under $\tau = 65.0$. We then score the same pairs with the target encoder, sort the scores in descending order, and choose the N-th largest score as its threshold. This yields a verifier threshold with the same empirical pass rate as LongCLIP, enabling a fairer comparison across encoder choices.

A.2 Structured Action Format and System Feedback

REAL uses a structured two-action interface in the multi-turn `Search-Answer` loop. At each step, the model must choose exactly one of the following actions in the structured format:

Search action. When the current observation is insufficient, the model invokes the search tool by emitting a `<search>` block:

```
<search>{"cues": ["cue1", "cue2", ...]}</search>
```

The content inside `<search>` must be a valid JSON object containing the key `"cues"`, whose value is a list of search cues. In practice, the system uses at most the first `max_search_cues` entries (default: 5). Search cues can be either (i) semantic phrases describing relevant objects, actions, events, or scenes, or (ii) timestamps in `HH:MM:SS` format, which are resolved to the nearest frame in the candidate pool.

Answer action. When sufficient evidence has been gathered, the model emits its final prediction inside an `<answer>` block:

```
<answer>{...}</answer>
```

The content inside `<answer>` must be exactly one JSON object with no extra text. This JSON object contains:

- `"final_answer"`: the final prediction (e.g., `"A"`, `"B"`, ... for multiple-choice tasks),
- `"evidence"`: a non-empty list of evidence items, where each item contains
  - `"timestamp"` in `HH:MM:SS` format, and
  - `"description"` as a textual description of the supporting frame.

Each turn permits exactly one action: either `<search>` or `<answer>`. Mixed outputs are disallowed. After receiving search results, the model either issues another search or gives the final answer.

System Feedback for Search Outcomes. When the model emits a valid `<search>` block, the system parses the cue list, retrieves the top matching frames, and returns them through a system feedback message for the next step. If the number of search cues exceeds the allowed limit, the feedback message briefly notes that only the first few cues will be used. Under normal successful retrieval, the feedback message contains the retrieved frames together with their timestamped frame tags. If the current step has reached the last allowed search round, the feedback message additionally instructs the model not to call the tool again and to provide a final answer in the next step. A search may also execute successfully but return no new frames; in this case, the feedback message provides a brief reminder to generate different search cues in the next step. By contrast, if the output does not contain a valid `<search>` block, the search is not executed, and the resulting failure is handled implicitly through trajectory termination and format-related rewards. In this way, successful retrieval, empty retrieval, and invalid search formatting are all reflected in the closed-loop rollout through system feedback or downstream reward signals.

A.3 Prompt Templates

We use a unified system prompt to define the multi-turn SEARCH-ANSWER workflow, the required output schema, and the tool-use constraints. The complete system prompt is shown below.

System Prompt for Multi-turn Search-and-Answer

```
You are a helpful assistant. You will receive sampled frames from a video and
    related questions. You have access to the search tool. Use the search tool
    to find more relevant frames when the current frames are insufficient to
    answer.

[Workflow]
1. Analyze the question and the frames provided.
2. If you have enough information to answer: place your final answer inside <
    answer> and </answer>.
3. If you need more frames to answer: use the search tool by outputting
   <search>{"cues": ["cue1", "cue2", ...]}</search>.
   The system will return the top matching frames for your next turn.

[Search Tool Format]
When you need to search for more relevant frames, output:
<search>{"cues": ["cue1", "cue2", "cue3"]}</search>

Cues are typically semantic phrases (descriptive text for similarity search).
Optionally, you may also include timestamps in HH:MM:SS format ( "00:12:35") to
     get the frame nearest to that time.

Guidelines for cues:
- Provide 3-5 cues. If you provide more, only the first 5 will be used.
- Semantic cues: include objects, actions, events, scene elements; be specific
    and concrete.
- Timestamp cues (optional): use HH:MM:SS when you know the time.
- Examples:
  "person holding a red umbrella"
  "car turning left at intersection"
  "woman opening a gift box"
  or mixed:
  ["race winner", "00:12:35", "cars crossing finish line"]

[Final Answer Format]
When you are ready to give your final answer, place it inside <answer> and </
    answer>. The content between <answer> and </answer> must contain exactly
    one JSON object with no extra text.

The JSON must include:
(1) "final_answer": required. For multiple-choice questions use one of "A", "B
    ", "C", etc.; for other tasks it can be any string.
(2) "evidence": required, a non-empty array. Each element must have "timestamp"
     (format HH:MM:SS) and "description" (string).

Example:
<answer>
{"final_answer": "B", "evidence": [{"timestamp": "00:01:15", "description": "
    the person is searching the room"}, {"timestamp": "00:01:28", "description
    ": "the person found the keys on the table"}]}
</answer>

[Rules]
- In each turn, either use <search> to search, OR give your <answer>. Do not do
     both in the same turn.
- After receiving search results, analyze the new frames and either search
    again (if still insufficient) or provide your final answer.
- Do not repeat the same search cues you have already used; provide new
    specific cues to search again.
```

A.4 Details of RL Optimization and Curriculum

To stabilize training under the sparse evidence-aware reward, we adopt a two-stage curriculum. In **Stage I**, we train a single-turn ANSWER-only model using the evidence-aware reward, without enabling SEARCH. This stage requires only 150–250 steps. The purpose of this stage is to first make the model reliably produce valid final answers together with admissible evidence citations under the strict faithfulness constraint, before requiring it to additionally learn when and how to search. Without this warm-up stage, learning evidence-grounded answering and multi-turn search behavior simultaneously is substantially less stable, since both the action format and the reward signal are sparse and highly constrained.

In **Stage II**, we initialize from the Stage-I model and continue RL training with the full multi-turn SEARCH-ANSWER loop enabled. At this stage, the objective is to teach the model to transition from single-turn evidence-grounded answering to active evidence gathering when the current observation is insufficient. To bootstrap tool use in the early phase, we apply a temporary search bonus with reward `tool_call_reward`=0.2 whenever the empirical tool-use rate is below `tool_use_rate_threshold`=0.3. Once the tool-use rate exceeds this threshold, the bootstrap bonus is removed. In addition, during the later phase of training, we apply a small over-search penalty of `late_phase_tool_penalty`=0.1 to discourage unnecessary search calls after the model has already acquired the basic search capability.

B Experimental Setup

B.1 Metric Definitions for Hallucination Benchmarks

VideoHallucer Metrics. VideoHallucer is a benchmark for hallucination detection in large video-language models. It adopts an adversarial binary VideoQA setting. The benchmark organizes hallucinations into two main categories, *intrinsic* and *extrinsic*, and further includes five subtypes: object-relation hallucination, temporal hallucination, semantic-detail hallucination, extrinsic factual hallucination, and extrinsic non-factual hallucination. In total, the benchmark contains 1,800 question-answer pairs over 948 videos, with 400 question-answer pairs for each subtype except semantic-detail hallucination, which also follows the same paired construction. VideoHallucer reports three metrics: *Basic Acc.*, *Hallucinated Acc.*, and *Overall Acc.*. Here, *Basic Acc.* is the accuracy on the original video-grounded questions, *Hallucinated Acc.* is the accuracy on their adversarially hallucinated counterparts, and *Overall Acc.* counts a sample as correct only when both questions in the pair are answered correctly. Formally, let $\mathcal{D}$ denote the set of paired samples, and let $b_i, h_i \in \{0, 1\}$ indicate whether the model answers the basic and hallucinated questions in pair i correctly. Then:

$$\text{Basic Acc.} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} b_i,$$

$$\text{Hallucinated Acc.} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} h_i,$$

$$\text{Overall Acc.} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbf{1}[\, b_i = 1 \wedge h_i = 1 \,].$$

These definitions follow the benchmark's adversarial paired evaluation protocol.

ELV-Halluc Metrics. ELV-Halluc is designed to evaluate *Semantic Aggregation Hallucination* (SAH), i.e., cases where a model captures local video semantics but misattributes them across different events within the same long video. To isolate this effect, the benchmark constructs two adversarial settings: *In-Video Hallucination*, where the hallucinated content is drawn from another event in the same video, and *Out-of-Video Hallucination*, where the hallucinated content does not appear anywhere in the video. The benchmark forms adversarial QA pairs by combining a ground-truth question with a hallucinated question, resulting in two pairs per triplet: $(\mathrm{GT}, \mathrm{InVideo})$ and $(\mathrm{GT}, \mathrm{OutVideo})$. A pair is counted as correct only if the model predicts `Yes` for the ground-truth question and `No` for the hallucinated question. In total, ELV-Halluc contains 4,800 binary QA pairs, which can be grouped into 3,200 adversarial QA pairs.

Table 6: VLM-based audit of evidence consistency and answer support. Gap denotes the difference between consistency and support in percentage points (pp).

Judge	Prediction	Cons. (%)	Supp. (%)	Gap (pp)↓
GPT-4o	Correct	83.5	71.7	11.8
	Incorrect	76.6	54.3	22.3
Gemini 3 Flash	Correct	79.1	66.5	12.6
	Incorrect	67.0	34.5	32.5

Let InAcc and OutAcc denote the accuracies on the in-video and out-of-video adversarial QA pairs, respectively. The benchmark reports:

$$\begin{aligned} \text{Avg Acc.} &= \frac{\mathrm{InAcc} + \mathrm{OutAcc}}{2}, \\ \text{Avg Diff.} &= \mathrm{OutAcc} - \mathrm{InAcc}, \\ \text{SAH Ratio} &= \frac{\mathrm{OutAcc} - \mathrm{InAcc}}{1 - \mathrm{InAcc}}. \end{aligned}$$

Here, *Avg Acc.* measures the average performance across the two hallucination settings. *Avg Diff.* is the raw accuracy gap between out-of-video and in-video adversarial pairs, and is used to quantify a model's sensitivity to semantic misalignment across events. *SAH Ratio* further normalizes this gap by the in-video error rate, measuring the proportion of hallucinations that can be attributed to semantic aggregation hallucination. Lower *Avg Diff.* and lower *SAH Ratio* therefore indicate stronger robustness to event-level semantic misattribution.

B.2 Inference Configurations

This section provides the detailed inference configurations used for evaluation on the hallucination benchmarks.

ELV-Halluc. Following the benchmark protocol, the Qwen2.5-VL series uses a maximum video resolution of $128 \times 28 \times 28$ and up to 256 frames; InternVL3 limits the input to 64 frames; GPT-4o uniformly samples 50 frames. All other settings follow the official defaults. In addition to the reported benchmark results, we evaluate Qwen2.5-VL-7B under a matched 96-frame setting for comparison with REAL.

VideoHallucer. All baseline settings (frame counts and generation hyperparameters) follow the original benchmark or method papers. Since VideoHallucer does not report results for the Qwen2.5-VL series, we additionally evaluate Qwen2.5-VL-7B under both 96-frame and 256-frame settings to support the controlled comparisons in the main text. The PaMi-VDPO result is taken from the original paper itself.

B.3 Additional Grounding Analysis

Consistency–Grounding Gap. Cross-modal evidence consistency does not necessarily imply sufficient support for the final answer. To further examine this distinction, we conduct an audit of 1,000 REAL outputs using two independent VLM judges, GPT-4o and Gemini 3 Flash [11]. Each judge evaluates evidence consistency and answer support based only on the question, cited frames, evidence descriptions, and predicted answer, without access to the full video.

As shown in Table 6, correct predictions achieve higher consistency and answer support (83.5%/71.7% for GPT-4o and 79.1%/66.5% for Gemini 3 Flash), with consistency–support gaps of 11.8 and 12.6 percentage points, respectively. In contrast, incorrect predictions exhibit substantially larger gaps of 22.3 and 32.5 points. These findings suggest that correct predictions are associated with stronger evidence support and smaller consistency–support gaps. Nevertheless, source consistency remains an informative but incomplete signal for answer-level grounding, particularly when temporal relationships beyond the cited frames are required.

Table 7: **Accuracy-efficiency trade-off of REAL.** Left: comparison with the 96-frame Qwen2.5-VL-7B backbone. Right: inference cost breakdown of REAL. The added cost is dominated by backbone-side MLLM reasoning, while retrieval contributes only a small fraction of the total latency.

Method	Peak GPU Mem.	Latency (ms)	Video-MME (%)	VideoHallucer (%)
Qwen2.5-VL-7B (96f)	25.2 GB	2614.7	57.1	43.1
REAL	**26.8 GB**	**6467.3**	**63.6**	**54.8**

Stage	Avg. Latency (ms)
LLM reasoning	5149.0
Candidate encoding	342.0
Similarity search	314.0
Other overhead	602.0

C Efficiency Analysis

C.1 Search Overhead Decomposition

Table 7 reports both the end-to-end trade-off and the stage-wise latency breakdown of REAL. Although REAL introduces auxiliary retrieval steps, the additional cost is incurred by lightweight frozen cross-modal encoding, similarity search, and extra reasoning turns, while the expensive backbone-side MLLM reasoning still operates under the same 96-frame input budget as the single-pass baseline. In practice, the added latency is dominated by backbone-side reasoning: out of 6408.0 ms total latency, 5149.0 ms comes from LLM reasoning, whereas candidate encoding and similarity search take only 342.0 ms and 314.0 ms, respectively. This is because retrieval is performed by a frozen CLIP-style encoder over a sparse 1-FPS candidate pool and operates only on embedding-level matching, rather than full MLLM inference over all candidate frames.

C.2 End-to-End Comparison with the Backbone

As shown in Table 7, REAL trades additional latency and a small memory increase for substantially higher accuracy under the same reported 96-frame budget. Relative to the 96-frame Qwen2.5-VL-7B backbone, REAL increases latency from 2614.7 ms to 6467.3 ms and peak GPU memory from 25.2 GB to 26.8 GB, but improves Video-MME from 57.1% to 63.6% and VideoHallucer from 43.1% to 54.8%. Therefore, the benefit of REAL is a stronger accuracy-efficiency trade-off: it spends a modest amount of lightweight retrieval overhead to substantially improve evidence selection and final task performance.

D Additional Qualitative Examples

We provide additional qualitative examples to further illustrate the strengths and limitations of REAL.

D.1 Success Cases

Figures 5 and 6 highlight a common strength of REAL: retrieving temporally sparse but decision-critical semantic details. In Figure 5, the model progressively refines the search from the broad cabinet-opening event to a highly specific object cue ("glass with a wooden stirrer and a straw inserted"), and finally retrieves the decisive frame. In Figure 6, REAL similarly localizes a rare entity-attribute cue, the swallow wearing a hat with a pink feather from a much longer video. Together, these cases show that REAL is particularly effective when the answer depends on localized and semantically precise evidence that cannot be reliably recovered from coarse scene context alone.

Figure 7 shows that the advantage of REAL is not limited to object- or attribute-level grounding. Rather than relying on a single matched frame, REAL issues search cues aligned with these concepts and retrieves evidence spanning multiple relevant stages of the video, enabling a correct sequence-level judgment. This case suggests that the search-and-answer loop can also support higher-level organization of video content when the answer depends on how multiple concepts unfold over time.

D.2 Failure Cases

Figure 8 reveals a limitation that goes beyond evidence localization. In this railway example, REAL successfully retrieves visually relevant evidence, including the map frame containing key textual and geographic elements. However, locating such evidence is still not sufficient: answering the question correctly requires a further reasoning step to determine whether the retrieved element is actually the

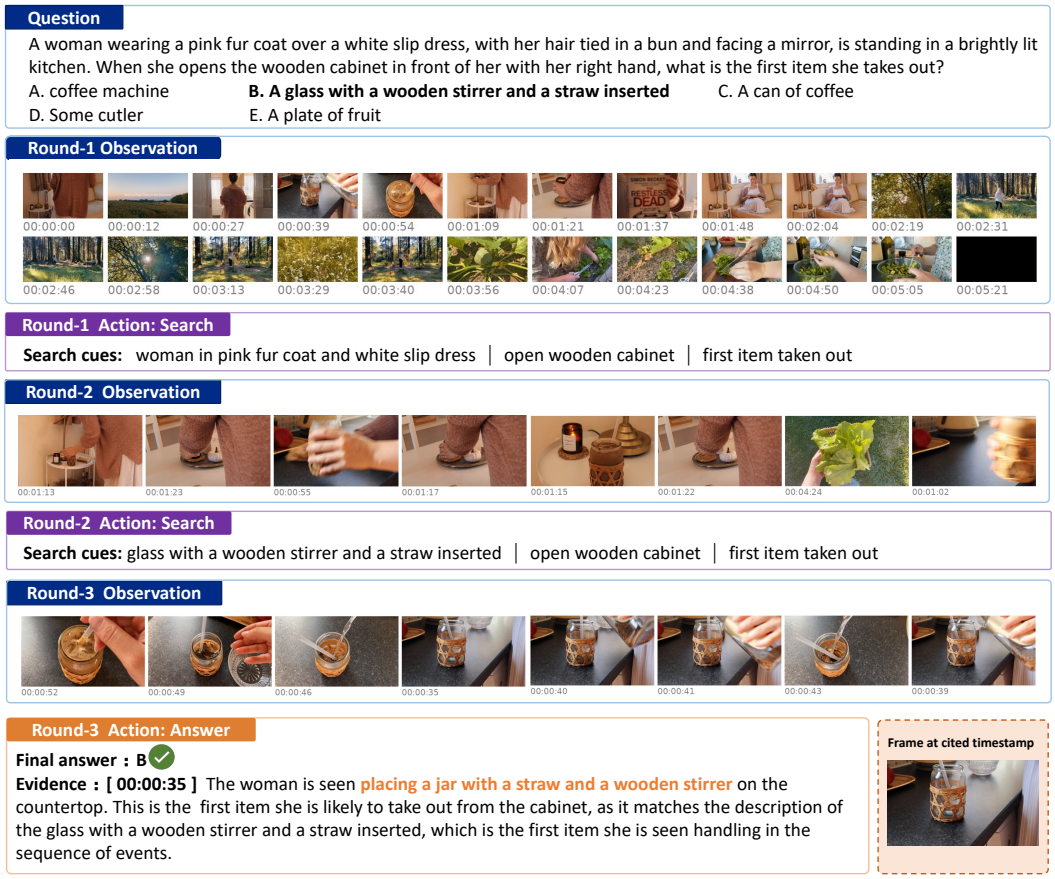


Figure 5: Iterative retrieval of fine-grained object evidence(success case).

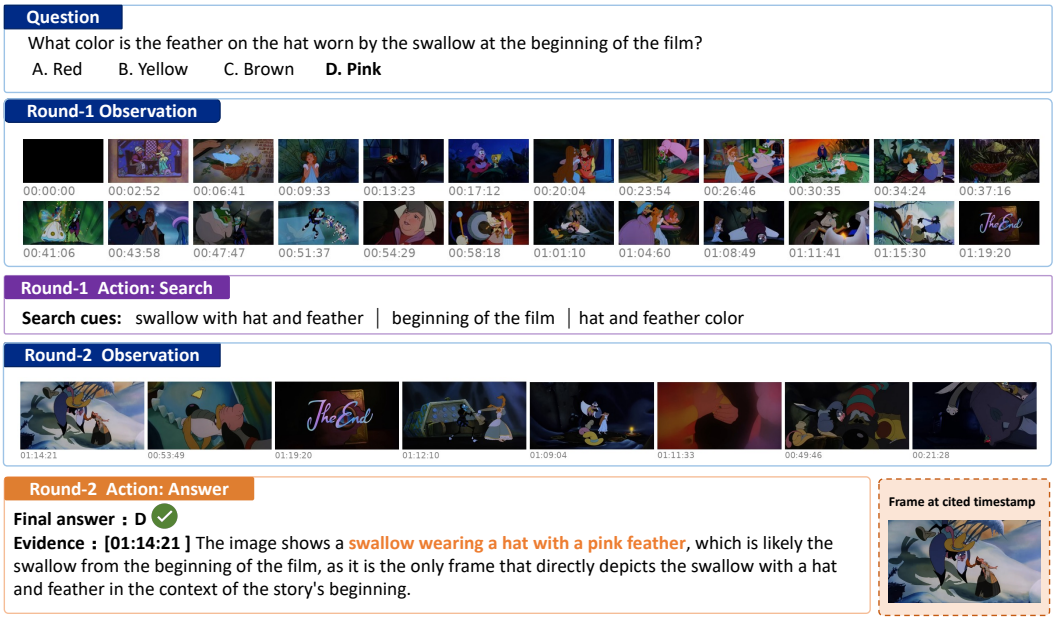


Figure 6: Retrieval of sparse entity-attribute evidence in long-form video(success case).

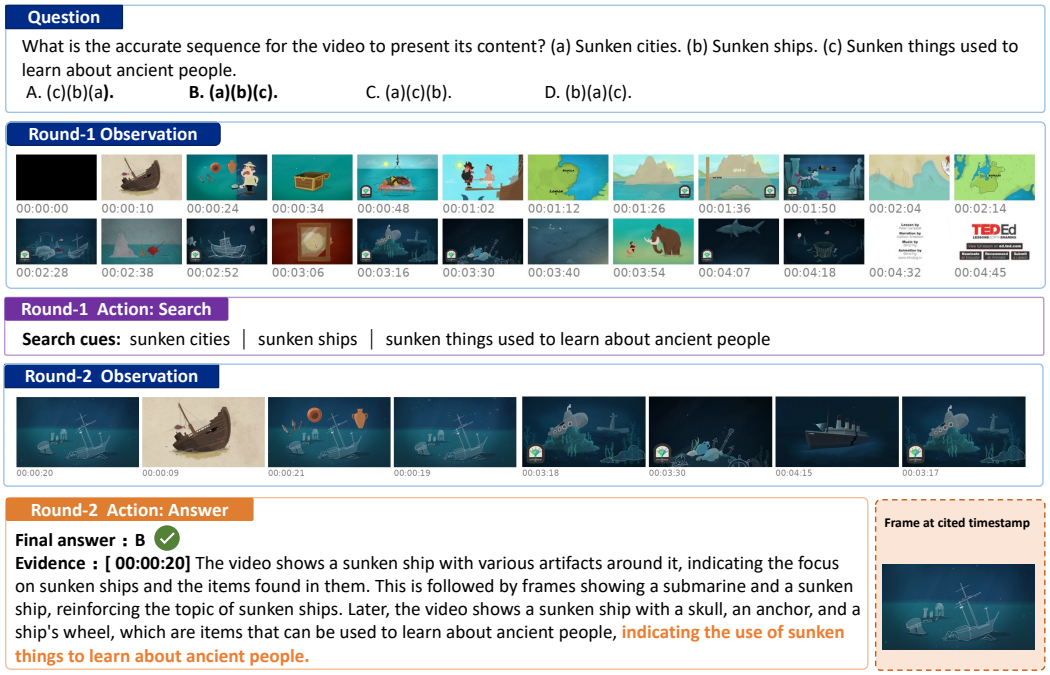


Figure 7: Sequence-level reasoning over multiple semantic segments(success case).

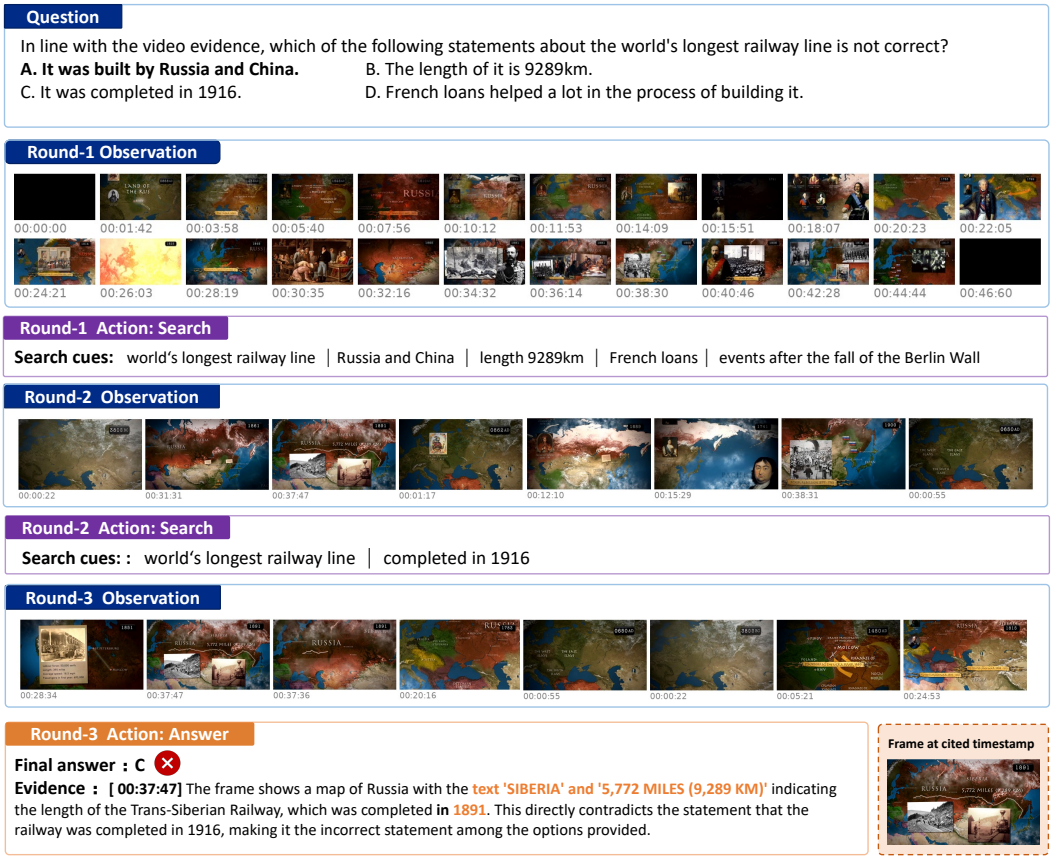


Figure 8: Relevant evidence is retrieved, but further reasoning remains insufficient(failure case).

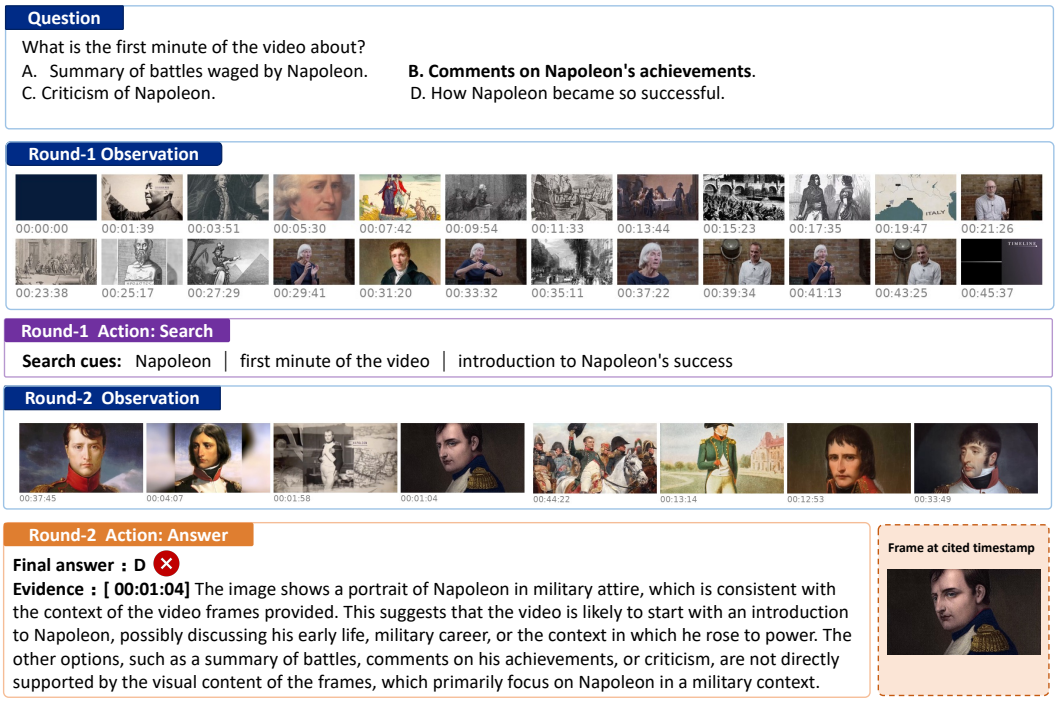


Figure 9: Visually relevant evidence is found, but the discourse-level focus is misidentified(failure case).

decisive clue for identifying the incorrect statement. This suggests that the current framework can still fail when the answer depends on more complex fact-level or logic-level reasoning built on top of the retrieved visual evidence.

Figure 9 highlights a related but distinct limitation. In this example, REAL retrieves a visually relevant portrait of Napoleon, but the question asks about the discourse-level topic of the first minute rather than simply the presence of Napoleon as an entity. The error therefore does not mainly come from missing visual evidence, but from insufficient higher-order reasoning over whether the retrieved evidence truly captures the semantic focus required by the question.

D.3 Case Analysis

Taken together, the successful cases show that REAL is especially strong when the answer depends on localized, sparse, or semantically specific evidence. In such cases, the search-and-answer loop helps the model move from coarse relevance to decision-critical grounding, including fine-grained object details, rare entity-attribute cues, and even multi-concept sequence organization.

By contrast, the failure cases suggest that the current bottleneck is increasingly not whether relevant visual evidence can be retrieved, but whether the model can correctly reason over that evidence once retrieved. More specifically, two challenges remain: (1) determining whether a retrieved visual element is truly the key clue required by the question, and (2) performing more abstract semantic or discourse-level reasoning beyond localized visual grounding. More broadly, some long-video questions, especially in benchmarks such as Video-MME, may also depend on information conveyed through narration, audio, or subtitles. Since the current version of REAL grounds answers only in visual evidence, extending the framework to broader multimodal grounding is a promising direction for future work.

E Additional Responsible Research Details

Existing Assets and Licenses. This work uses publicly available pretrained models, codebases, datasets, and benchmarks, including Qwen2.5-VL-7B, LongCLIP-L, DINOv3, ASP-CLIP, verl,

NExT-GQA, QVHighlights, LongVideoReason, Video-MME, MLVU, LongVideoBench, EgoSchema, VideoHallucer, and ELV-Halluc. We cite the original papers for these assets and follow their licenses and terms of use. We do not redistribute raw datasets or benchmark data; instead, we provide data preparation scripts and instructions for obtaining them from the original sources.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's main contributions, including the evidence-aware search-answer framework, the reinforcement learning objective, and evaluation on long-video understanding and hallucination benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper discusses limitations in Section 5, including the dependence on verifier quality and the current focus on visual-only grounding. Additional failure-case analysis in the appendix further discusses remaining challenges.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results such as theorems or formal guarantees. The equations are used to define the method, reward function, and optimization objective rather than to state theoretical claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: The paper discloses the main algorithmic components, reward design, search-answer procedure, model backbone, encoders, frame budgets, datasets, benchmarks, and training configuration. Additional implementation details, structured action formats, and prompt templates are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized code and model assets together with setup instructions, evaluation scripts, and commands to reproduce the main results reported in the paper. All the experiments use existing public datasets and benchmarks.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and evaluation setup, including the backbone model, visual encoders, frame budget, rollout configuration, datasets, hyperparameters, optimizer framework, and benchmark settings. Further details on search, verification, and prompting are provided in Appendix A and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports deterministic benchmark results under fixed evaluation protocols, but does not provide error bars or statistical significance tests. Multiple full RL training runs would be computationally expensive given the scale of the training setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the compute resources required to reproduce the main experiments , including the H800 GPU configuration, and estimated epochs for training. The paper also reports the main training allocation, with one GPU reserved for the verifier and the remaining GPUs used for policy training.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses existing datasets, pretrained models, and benchmark evaluations, and does not involve human-subject experiments or private data collection. The authors have reviewed and followed the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The current paper focuses on the method, experiments, and limitations, but does not include a dedicated discussion of positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk assets such as a new pretrained language model, image generator, or scraped dataset requiring special safeguards. Therefore, responsible-release safeguards for such assets are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers and/or repositories for all pretrained models, codebases, datasets, and benchmarks used in this work, and summarize our use of these assets and their license/term compliance in Appendix. We do not redistribute raw datasets or benchmark data, and the released code and model assets are provided under terms compatible with the underlying assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce new code and model assets for REAL and provide documentation alongside the supplemental release. The released assets are anonymized for submission and are based on existing datasets and pretrained models, without introducing newly collected human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. No participant instructions, screenshots, or compensation details are therefore applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research. Therefore, IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The use of LLMs is central to the proposed method: REAL is built on Qwen2.5-VL-7B as the policy model in the iterative search-answer framework. The paper describes this usage in the method and experimental setup sections.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.