EG-VQA: BENCHMARKING VERIFIABLE VIDEO QUESTION ANSWERING WITH GROUNDED TEMPORAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in Video Large Language Models (Video-LLMs) have yielded promising performance on Video Question Answering (VideoQA). Nevertheless, existing benchmarks are predominantly evaluated through answer correctness, while the faithfulness of predicted evidence supporting those answers remains insufficiently evaluated. This disconnect between answer generation and evidence verification motivates the construction of the **Evidence-Grounded Video Question Answering Benchmark (EG-VQA)**, a large-scale open-ended benchmark in which each QA pair is annotated with temporally localized textual evidence, requiring models to jointly produce answers and verifiable evidence. EG-VQA comprises 2,067 videos and 11,838 QA pairs with fine-grained evidence annotations. To evaluate predicted evidence, **Evidence-Grounded F1 (EG-F1)** is introduced as a metric that jointly measures temporal alignment and semantic consistency between predicted and ground-truth evidence. Experiments reveal a substantial discrepancy between answer correctness and evidence faithfulness: even strong proprietary models can achieve high answer accuracy while exhibiting gaps in temporal-semantic grounding. To investigate evidence-aware learning under EG-VQA, we develop **EG-Reasoner**, an evidence-aware model trained with explicit evidence supervision. EG-Reasoner achieves strong evidence grounding performance among open-source models while maintaining competitive answer generation results compared with proprietary systems. These findings highlight the importance of explicitly modeling evidence grounding and suggest that evidence-aware supervision provides an effective direction toward more reliable and interpretable VideoQA systems.

1 INTRODUCTION

Video Question Answering (VideoQA) serves as a critical testbed for evaluating a model's capacity to comprehend dynamic visual content and reason over temporal events (Zhong et al., 2022; Lin & Liu, 2026; Jiang et al., 2025). Despite recent advances in large multimodal models, a fundamental limitation persists: current benchmarks often fail to distinguish between genuine video understanding and the exploitation of superficial correlations (Ko et al., 2023).

Many widely used VideoQA datasets are constructed in a multiple-choice format, wherein models select answers from predefined options (Tapaswi et al., 2016; Xiao et al., 2021; Wu & Yu, 2024; Fu et al., 2025). Although this paradigm simplifies evaluation, it inadvertently introduces option bias, allowing models to rely on linguistic shortcuts rather than grounding predictions in visual evidence (Liu et al., 2023b; Chen et al., 2025; Liu et al., 2023a), as shown in Figure 1 (a). Furthermore, fine-grained evidence annotations are typically absent from these benchmarks, rendering it difficult to verify whether relevant video segments support a given prediction. Consequently, models frequently produce black-box answers devoid of interpretable reasoning (Liu et al., 2022). Even in open-ended settings, evaluation is conventionally restricted to answer correctness, while attention to correct temporal regions is overlooked (Xiao et al., 2024; Lei et al., 2020). These limitations underscore a critical gap: correct answers do not necessarily imply that the underlying reasoning is faithful to the video content. For VideoQA systems to be reliable, both the answer and its provenance within the video should be explicitly verifiable (Xiao et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

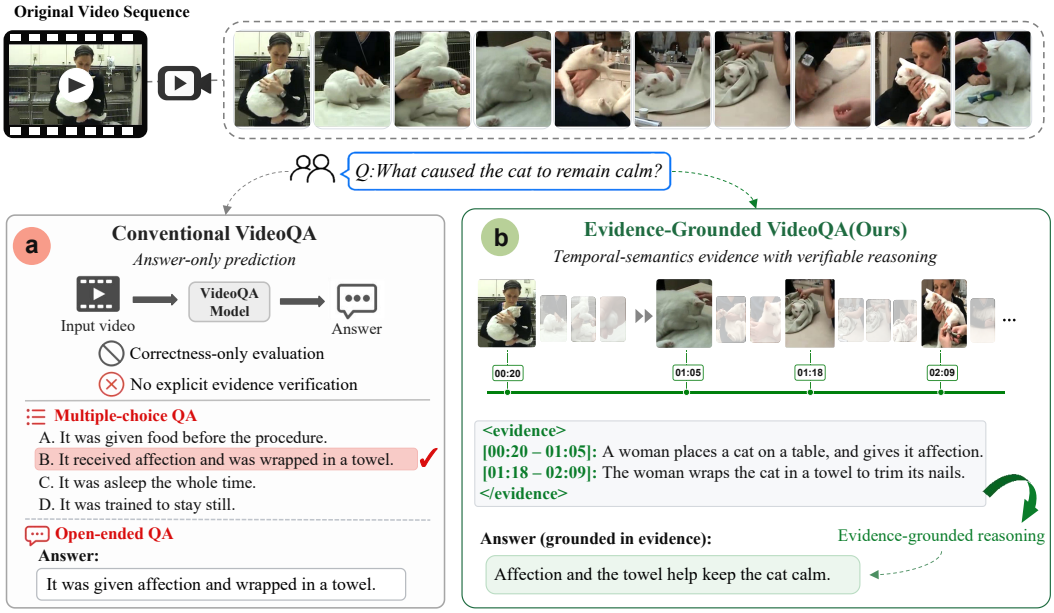

Figure 1: Overview of evidence-grounded VideoQA evaluation in EG-VQA. (a) Conventional VideoQA evaluation primarily focuses on answer prediction, without explicitly verifying whether the prediction is supported by relevant video evidence. (b) EG-VQA requires models to generate answers together with temporally localized textual evidence, enabling joint evaluation of answer correctness and evidence faithfulness.

Motivated by this challenge, we introduce the **Evidence-Grounded Video Question Answering Benchmark (EG-VQA)**, a large-scale open-ended VideoQA benchmark designed to jointly evaluate answer correctness and evidence faithfulness, i.e., whether the predicted evidence temporally and semantically supports the answer. Unlike conventional answer-only evaluation, EG-VQA requires models to generate both an answer and its supporting temporally localized evidence, enabling direct verification of whether predictions are grounded in the observed video content. As illustrated in Figure 1(b), each evidence unit is represented as a temporal interval paired with a textual event description, allowing evaluation of not only where relevant information appears but also whether the described evidence semantically supports the answer. The resulting dataset contains 2,067 videos and 11,838 QA pairs spanning four reasoning categories: descriptive, temporal, causal, and counterfactual.

Extensive evaluation of state-of-the-art multimodal models on EG-VQA reveals a substantial gap between answer correctness and evidence faithfulness. While strong models can often generate correct answers, their predicted evidence may still fail to accurately localize and describe the supporting video events. This discrepancy indicates that answer accuracy alone is insufficient for evaluating reliable VideoQA systems and motivates the study of evidence-aware learning strategies under the proposed benchmark.

To further study evidence-grounded VideoQA under EG-VQA, we develop **EG-Reasoner** as an evidence-aware reference model trained with explicit evidence supervision. EG-Reasoner generates answers together with temporally localized evidence, providing a model-level investigation of how evidence-aware supervision affects VideoQA performance. In addition, we introduce **Evidence-Grounded F1 (EG-F1)**, a metric tailored to the temporal-semantic evidence formulation of EG-VQA, which jointly evaluates temporal alignment and semantic consistency between predicted and reference evidence through optimal bipartite matching. Experiments show that evidence-aware optimization substantially improves evidence grounding performance and achieves competitive results among open-source Video-LLMs. Our contributions are summarized as follows:

- We introduce **EG-VQA**, a large-scale open-ended VideoQA benchmark that associates each QA pair with temporally localized textual evidence. By representing evidence as temporal intervals paired with event descriptions, EG-VQA enables joint evaluation of answer correctness and evidence faithfulness.

- We introduce **Evidence-Grounded F1 (EG-F1)**, a metric designed for temporal-semantic evidence evaluation, which jointly measures temporal alignment and semantic consistency between predicted and reference evidence through optimal bipartite matching.
- We develop **EG-Reasoner** as an evidence-aware reference model to investigate grounding-oriented learning under EG-VQA. Through explicit evidence supervision, EG-Reasoner generates answers jointly with temporally localized evidence and demonstrates the effectiveness of evidence-aware optimization.

2 Related Work

2.1 Video Question Answering Benchmarks

Video Question Answering (VideoQA) benchmarks have evolved from short-video understanding toward increasingly complex temporal and causal reasoning. Early datasets such as TGIF-QA (Jang et al., 2017), MovieQA (Tapaswi et al., 2016), and MSRVTT-QA (Xu et al., 2017) established VideoQA settings over GIFs, movies, and short web videos, while later benchmarks including ActivityNet-QA (Yu et al., 2019), TVQA (Lei et al., 2018), NExT-QA (Xiao et al., 2021), STAR (Wu & Yu, 2024), and Video-MME (Fu et al., 2025) expanded evaluation to longer videos and more diverse temporal and compositional reasoning. Many widely used benchmarks adopt multiple-choice protocols, which facilitate automatic evaluation but may introduce option-related shortcuts and do not by themselves establish whether an answer is supported by the relevant video content (Chen et al., 2025; Liu et al., 2023b; Wei et al., 2023; Liu et al., 2022).

To explicitly verify visual support for predicted answers, grounded VideoQA benchmarks augment answer supervision with evidence localization. TVQA+ (Lei et al., 2020) introduces spatio-temporal grounding, while NExT-GQA (Xiao et al., 2024) associates QA pairs with temporal labels for answer-supporting moments. MultiHop-EgoQA (Chen et al., 2024c) further requires models to answer questions while localizing multiple relevant temporal intervals in long-form egocentric videos, and TOGA (Gupta et al., 2025) studies temporally grounded open-ended VideoQA. HERBench (Ben Ami et al., 2026) focuses on the complementary challenge of integrating multiple temporally dispersed visual cues. Building on this direction, EG-VQA represents each supporting evidence unit as a temporal-semantic pair, consisting of a temporal span and a standalone textual description of the corresponding event. This requires models to predict not only where the supporting evidence occurs but also whether the evidence description semantically supports the answer, enabling direct verification of both temporal alignment and semantic fidelity.

2.2 Video-LLMs and Evidence-Grounded Reasoning

Recent Video Large Language Models (Video-LLMs) and multimodal large language models have substantially advanced video understanding through improved visual perception, temporal modeling, and language generation. Representative model families such as Qwen-VL (Bai et al., 2025b;a), InternVL (Zhu et al., 2025; Wang et al., 2025), and VideoChat (Li et al., 2023) demonstrate strong capabilities across video question answering, dialogue, and long-video understanding. More recent reasoning-oriented methods further employ post-training to improve temporal and spatio-temporal reasoning. Time-R1 (Wang et al., 2026) applies reinforcement-learning-based post-training for temporal video grounding, while VideoChat-R1 (Li et al., 2025) uses reinforcement fine-tuning with temporally aware rewards to improve spatio-temporal perception and reasoning.

These methods primarily optimize answer generation, temporal reasoning, or general video understanding objectives. EG-Reasoner instead focuses on generating explicit temporal-semantic supporting evidence together with the final answer, with evidence supervision defined over the same interval-description representation used by EG-VQA. This allows training to directly account for both the temporal and semantic quality of predicted supporting evidence.

3 Evidence-Grounded Video Question Answering Benchmark

This section describes the construction pipeline of the Evidence-Grounded Video Question Answering (EG-VQA) Benchmark, including the data curation, reasoning-oriented QA generation with prompt refinement, and two-stage quality control, as shown in Figure 2. More details discussed in A.

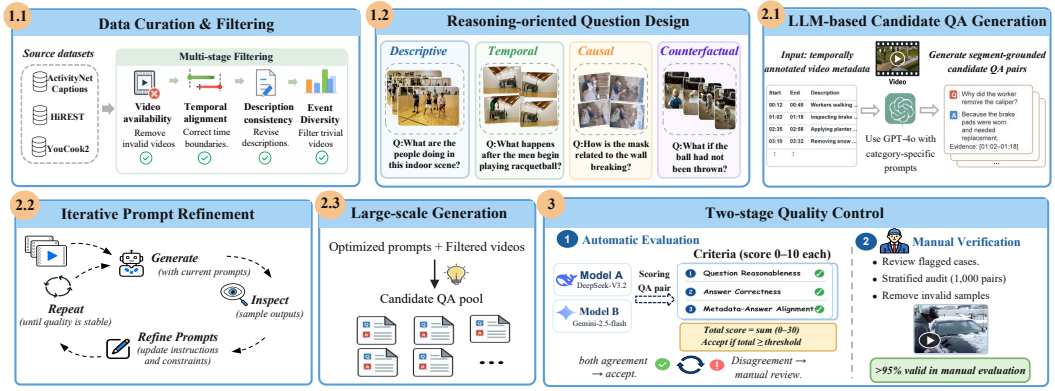

Figure 2: Overview of the EG-VQA dataset construction pipeline. The pipeline includes data curation and filtering, reasoning-oriented QA generation with prompt refinement, and two-stage quality control.

3.1 Data Curation and Question Design

The foundation of EG-VQA is built from three existing video sources: ActivityNet Captions (Caba Heilbron et al., 2015), HiREST (Zala et al., 2023), and YouCook2 (Zhou et al., 2018), which provide temporally segmented event annotations across diverse video domains, including open-domain activities, instructional procedures, and long-form event-centric videos. Specifically, we use the official validation splits of ActivityNet Captions and YouCook2, and the validation and test splits of HiREST as the source pool. These source partitions are used to obtain videos and temporal annotations, while EG-VQA subsequently constructs its own train–test partition at the video level. The raw annotations are standardized into a unified format in which each video is represented as an ordered sequence of timestamped segments paired with concise textual event descriptions. Because the original segment boundaries and descriptions occasionally deviate from the actual video content, we remove corrupted or inaccessible clips, correct misaligned temporal boundaries, revise inconsistent descriptions, and exclude videos with low event diversity that would otherwise yield trivial questions. After curation, 2,067 videos remain, with the corrected temporal descriptions serving as the evidence backbone. Detailed filtering criteria and source-wise statistics are provided in Appendix A.1 and A.4.

EG-VQA covers four reasoning categories. Descriptive (Desc.) questions focus on directly observable actions, objects, and states; Temporal (Temp.) questions require understanding event order and procedural sequences; Causal (Caus.) questions involve reasoning about event dependencies and consequences; and Counterfactual (Cntrf.) questions require reasoning over hypothetical alterations of observed events. Together, these categories span direct perception, temporal reasoning, causal inference, and counterfactual reasoning grounded in specific video events.

3.2 Dataset Construction and Quality Control

QA pairs are generated through a human-in-the-loop pipeline that iterates between automated generation and manual refinement. Given the temporally annotated video metadata, GPT-4o is prompted to generate candidate questions, answers, and supporting evidence for each reasoning category. The initial prompts are applied to a held-out subset and manually inspected for common failure modes, including video-independent questions, weak grounding to the metadata, and answers that omit necessary temporal context. The prompts are iteratively refined until the generated samples consistently satisfy the intended quality requirements, after which they are deployed across the curated video pool. Prompt templates and refinement details are provided in Appendix A.2.

We then apply a two-stage quality-control protocol to all candidate QA pairs. First, DeepSeek-V3.2 and Gemini-2.5-Flash independently evaluate each sample for question reasonableness, answer correctness, and evidence alignment using a structured scoring rubric. Samples receiving consistent accept/reject decisions are processed accordingly, while model disagreements are sent to human annotators for adjudication. After this full-dataset screening, we conduct an additional stratified random audit of 1,000 QA pairs, balanced across data sources, video durations, and question types. More than 95% of the audited samples are verified as valid. Minor, correctable issues are manually

revised and retained, whereas substantively invalid samples are removed. The complete validation protocol is provided in Appendix A.3.

Finally, we construct the train–test partition using a stratified video-level split, controlling for source dataset, video duration, and evidence density to maintain comparable distributions across the two partitions. All QA pairs and evidence annotations derived from the same video are assigned to the same partition, ensuring that no video appears in both the training and test sets. The resulting dataset contains 1,731 training videos with 8,949 QA pairs and 336 test videos with 2,889 QA pairs. Each question is associated with an average of 2.27 supporting evidence segments. Detailed source-wise and per-split statistics are reported in Appendix A.4.

3.3 Evaluation via Evidence-Grounded F1

In EG-VQA, evidence quality depends on both accurate temporal localization and semantic consistency with the supporting video events. Existing evaluation protocols often assess temporal overlap and textual similarity separately, allowing a prediction to perform well on one dimension while failing on the other. EG-F1 therefore evaluates predicted evidence as temporal-semantic units, requiring matched evidence to align with the reference in both time and meaning.

Formally, let the ground-truth evidence set be $\mathcal{G} = \{g_i\}_{i=1}^{|\mathcal{G}|}$ and the predicted set be $\mathcal{P} = \{p_j\}_{j=1}^{|\mathcal{P}|}$, where each item $g_i = (t_i, d_i)$ and $p_j = (t_j, d_j)$ consists of a temporal segment and a textual description. The metric proceeds by computing pairwise temporal overlap IoU_{ij} and semantic similarity Sim_{ij} between all pairs, retaining only those edges that satisfy thresholds on both dimensions, and then finding an optimal one-to-one matching via the Hungarian algorithm (Kuhn, 1955). Precision and recall are derived from the number of matched pairs M, yielding

$$\mathrm{EG{-}F1} = \frac{2 \cdot (M/|\mathcal{P}|) \cdot (M/|\mathcal{G}|)}{(M/|\mathcal{P}|) + (M/|\mathcal{G}|)}. \tag{1}$$

The complete algorithmic formulation, including threshold settings and the soft variant used during training, is deferred to Appendix B.

4 EG-Reasoner: An Evidence-Aware Baseline

To investigate how the evidence supervision provided by EG-VQA can be incorporated into model post-training, we develop EG-Reasoner as an evidence-aware baseline. EG-Reasoner generates structured responses containing supporting evidence, an intermediate reasoning trace, and a final answer. We adopt standard Group Relative Policy Optimization (GRPO) Shao et al. (2024); Guo et al. (2025) for post-training, while the task-specific design lies in a composite reward that evaluates output format, answer correctness, and temporal-semantic evidence quality. Figure 3 provides an overview of the training framework.

4.1 Structured Generation Formulation

Given an input pair (V, q), the model generates a structured response

$$y = (\hat{\mathcal{E}}, \hat{\mathcal{T}}, \hat{\mathcal{A}}), \tag{2}$$

where $\hat{\mathcal{E}} = \{(\hat{t}_k, \hat{d}_k)\}$ is a set of predicted evidence segments with temporal boundaries and textual descriptions, $\hat{\mathcal{T}}$ is a reasoning trajectory that chains the evidence into an explanatory narrative, and $\hat{\mathcal{A}}$ is a concise final answer. This three-block design, shown in the candidate responses of Figure 3, separates the predicted supporting evidence from the final answer, enabling the grounding quality of the response to be evaluated directly rather than inferred solely from answer correctness.

4.2 Composite Reward Design

Each sampled response is evaluated using a composite reward:

$$R(y) = \lambda_f R_{\mathrm{format}}(y) + \lambda_e R_{\mathrm{evi}}(y) + \lambda_a R_{\mathrm{ans}}(y). \tag{3}$$

where R_{format}, R_{ans}, and R_{evi} measure structure compliance, answer correctness, and evidence quality, respectively. The corresponding weights are provided in Appendix C.3.

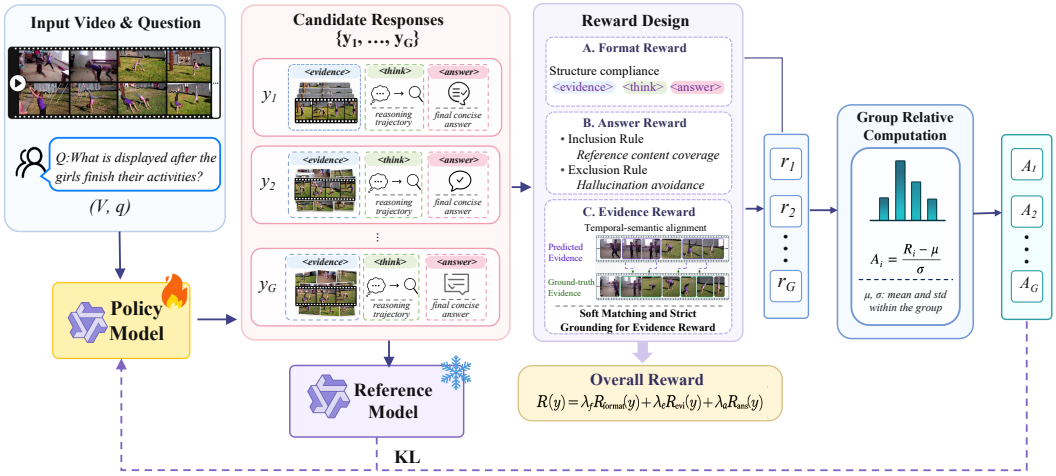

Figure 3: Overview of the EG-Reasoner training framework. Given a video-question pair, the model generates structured responses containing evidence, intermediate reasoning, and answer, and is optimized with GRPO using a composite reward for format compliance, answer correctness, and evidence faithfulness.

Format reward. The format reward encourages responses to follow the predefined output schema containing the <evidence>, <think>, and <answer> blocks. It assigns a binary score according to whether the required structure is correctly produced.

Answer reward. The answer reward evaluates whether the generated answer covers the key content of the reference answer while avoiding hallucinated or unsupported statements. A rule-guided LLM evaluator follows two complementary principles. The inclusion principle rewards coverage of relevant reference content, while the exclusion principle penalizes information that contradicts or cannot be supported by the video metadata. Partially correct answers receive intermediate credit to provide a denser optimization signal. Detailed scoring rules are provided in Appendix C.1.

Evidence reward. Directly using the thresholded EG-F1 as a training reward produces sparse supervision, since early predictions rarely satisfy both the temporal and semantic thresholds required for a valid evidence match. We therefore introduce a soft matching variant that assigns partial credit according to the product of temporal overlap and semantic similarity between predicted and reference evidence. Optimal bipartite matching is performed without hard thresholding to obtain a dense evidence score.

Soft matching alone, however, may reward loosely aligned evidence. We therefore combine the soft score with the original thresholded EG-F1:

$$R_{\text{evi}} = R_{\text{soft}} + R_{\text{EG-F1}}. \tag{4}$$

The soft component provides graded feedback for partially aligned evidence, while the hard component preserves the stricter temporal-semantic criterion used by EG-F1. The complete formulation of the soft variant is provided in Appendix B.2.

4.3 Optimization with GRPO

A structured response may admit multiple semantically valid realizations. Correct answers can be expressed through different phrasings, while supporting events may admit different but valid textual descriptions. A single-reference supervised objective observes only one annotated realization and optimizes token-level likelihood against that sequence, rather than directly optimizing task-level answer correctness and set-level evidence quality.

Given the composite reward defined above, we optimize EG-Reasoner using standard GRPO. For each video-question pair, the policy samples a group of candidate responses, each of which is evaluated by the composite reward. Their relative rewards are normalized within the group to estimate advantages, and the policy is updated to increase the likelihood of higher-quality responses. A KL regularization term constrains the policy from deviating excessively from the reference policy. The full GRPO objective, including group-relative advantage computation, clipping, and KL regularization, is

Table 1: Overall performance comparison on EG-VQA.

Model	Strict Accuracy					Relaxed Accuracy					Evidence Grounding		Joint
	Desc.	Temp.	Caus.	Cntrf.	Overall	Desc.	Temp.	Caus.	Cntrf.	Overall	Avg. F1	Avg. EG-F1	GAS
Human Results*	52.08	57.69	51.72	65.38	56.67	68.75	73.08	65.52	76.92	70.95	73.76	56.34	33.81
Commercial Large Multimodal Models													
GPT-4o Hurst et al. (2024)	18.90	23.08	37.20	35.10	28.83	40.10	38.94	48.00	53.50	45.24	15.58	3.33	2.29
Gemini-2.5-Flash Comanici et al. (2025)	14.58	20.45	33.93	28.89	24.75	37.50	36.36	43.75	46.67	41.15	17.03	4.99	2.74
GPT-5.6 Luna	27.08	21.00	39.58	40.40	32.12	43.23	35.00	47.92	57.58	45.95	32.79	10.56	6.81
Qwen3.7 Plus	24.00	18.45	30.61	23.47	24.13	42.00	33.50	45.41	41.84	40.63	20.73	5.89	2.58
Open-sourced Video-LLMs													
Qwen2.5-VL-7B Bai et al. (2025b)	9.91	7.73	22.36	16.80	14.31	24.96	20.07	36.21	30.30	27.96	13.86	3.33	1.80
InternVL-3-8B Zhu et al. (2025)	9.71	5.24	18.11	13.64	11.71	27.91	18.75	30.45	28.72	26.36	9.48	2.02	1.01
Time-R1-7B Wang et al. (2026)	10.52	7.74	22.73	21.15	15.66	27.52	20.49	39.24	34.24	30.44	24.89	6.87	3.23
VideoChat-R1-7B Li et al. (2025)	13.24	9.99	24.10	19.81	16.87	30.29	22.90	41.37	34.15	32.21	27.00	7.65	3.95
VideoChat-R1-Thinking-7B Li et al. (2025)	9.91	8.40	24.50	25.65	17.29	26.75	20.97	40.09	38.74	31.76	21.56	5.76	2.86
VideoChat-R1.5-7B Yan et al. (2026)	10.65	8.12	22.92	19.02	15.29	28.46	22.04	39.01	32.01	30.42	17.29	5.08	2.41
TW-GRPO Dang et al. (2025)	12.37	7.48	25.80	24.21	17.58	27.79	19.56	39.97	36.70	31.14	16.17	3.77	2.04
EG-Reasoner (Ours)	**14.20**	**10.92**	**33.16**	**48.16**	**26.88**	**34.12**	**27.10**	**48.46**	**60.50**	**42.71**	**37.89**	**12.28**	**7.14**

Note. Avg. F1 and Avg. EG-F1 denote averages across the threshold settings specified in Appendix D.3.

provided in Appendix C.2. Additional implementation details and hyperparameters are reported in Appendix C.3.

5 Experiments

5.1 Experimental Setup

Models. We evaluate a diverse set of Video-LLMs spanning proprietary multimodal models, general-purpose open-source models, and reasoning-oriented video models. The evaluated systems cover representative model families with different scales and training paradigms. Detailed descriptions of all baselines are provided in Appendix D.1.

Evaluation Metrics. We evaluate model performance in terms of answer correctness and evidence grounding. For open-ended answer evaluation, we use Gemini-2.5-Pro as an LLM judge and report strict and relaxed accuracy. On a 200-sample human audit, the judge achieves 96.5% exact agreement with human evaluation (Cohen's $\kappa = 0.944$). For evidence grounding, we report event-level F1 under multiple temporal IoU thresholds together with EG-F1. We additionally report the Grounded Answer Score (GAS), which measures whether answer correctness and evidence grounding are achieved on the same samples. Detailed evaluation protocols are provided in Appendix D.2 to D.4.

Implementation details. We adopt Qwen2.5-VL-7B-Instruct as the base model of EG-Reasoner and fine-tune it on the EG-VQA training split. During both training and evaluation, the model is required to generate structured outputs following the unified format defined in Section 4. Detailed training hyperparameters and evaluation configurations are provided in Appendix C.3 and Appendix D.5.

Human study. To estimate human performance on EG-VQA, we randomly sample 200 test QA pairs covering all four reasoning categories and ask three human annotators to provide open-ended answers together with supporting evidence under the same task setting as the evaluated models. Their responses are evaluated using the same answer and evidence metrics, and we report the average performance across annotators.

5.2 Main Results

Table 1 summarizes the main results on EG-VQA. GPT-5.6 Luna achieves the highest overall answer accuracy among the evaluated models, with 32.12% strict and 45.95% relaxed accuracy, while remaining below human performance (56.67% and 70.95%). Among open-source models, EG-Reasoner performs best overall, reaching 26.88% strict and 42.71% relaxed accuracy, improving over Qwen2.5-VL-7B by 12.57 and 14.75 percentage points, respectively. Its performance is particularly strong on counterfactual questions, where it achieves 48.16% strict and 60.50% relaxed accuracy.

EG-Reasoner shows a stronger advantage in evidence grounding, achieving 37.89 Avg. F1 and 12.28 Avg. EG-F1, outperforming GPT-5.6 Luna (32.79 and 10.56) and VideoChat-R1-7B (27.00 and 7.65). It also obtains the highest model GAS of 7.14, indicating stronger sample-level co-occurrence of correct answers and grounded evidence. Human performance remains higher overall, indicating substantial room for improvement in model performance. Detailed threshold-specific grounding

results are provided in Appendix D.6, and additional analyses of video dependency and the answer correctness–evidence grounding relationship are provided in Appendix D.7.

5.3 Ablation Studies

We perform controlled ablations to analyze the contribution of different training components in EG-Reasoner. Results are summarized in Table 2.

Table 2: Ablation study of evidence-aware training on EG-VQA.

Model	Strict Acc.↑	Relaxed Acc.↑	Avg. F1↑	Avg. EG-F1↑
Qwen2.5-VL-7B	14.31	27.96	13.86	3.33
+ SFT	13.12	24.47	20.29	4.85
+ RL w/o Evidence Reward	22.21	36.17	15.81	4.21
+ RL w/o Soft Evidence Reward	24.00	37.96	15.33	3.62
+ RL w/ Evidence Reward (EG-Reasoner)	**26.88**	**42.71**	**37.89**	**12.28**
VideoChat-R1-7B	16.87	32.21	27.00	7.65
+ RL w/o Evidence Reward	25.42	40.26	22.07	6.51

Effect of supervised fine-tuning. SFT improves evidence grounding quality, increasing Avg. F1 from 13.86 to 20.29 and Avg. EG-F1 from 3.33 to 4.85, while slightly reducing answer accuracy. This suggests that SFT improves structured evidence prediction, but token-level optimization on reference responses does not fully align with the temporal-semantic grounding objective.

Effect of reinforcement learning without evidence rewards. RL without evidence rewards improves answer generation, increasing relaxed accuracy from 27.96% to 36.17%. However, evidence grounding improves only marginally (Avg. EG-F1: 3.33 $\rightarrow$ 4.21), indicating that optimizing response-level objectives alone is insufficient for precise evidence grounding.

Effect of evidence-aware rewards. Introducing evidence-aware rewards leads to substantial improvements in evidence grounding compared with RL without evidence rewards, demonstrating that explicitly optimizing evidence quality better aligns training with the evaluation objective. Removing the soft evidence component further degrades grounding performance, highlighting the importance of continuous temporal-semantic supervision. Moreover, VideoChat-R1-7B trained with the same reinforcement learning procedure but without evidence rewards still underperforms, suggesting that access to EG-VQA training data alone is insufficient and that evidence-aware supervision is essential for grounded video reasoning.

5.4 Zero-Shot Generalization Beyond EG-VQA

To evaluate the transferability of evidence-aware training beyond the original benchmark, we conduct zero-shot evaluation on two existing video reasoning benchmarks, NExT-GQA and ReXTime. No additional training or task-specific adaptation is performed on these benchmarks. For ReXTime, we exclude all videos that appear in the EG-VQA training split, along with their associated QA pairs, before evaluation. Results are summarized in Table 3.

Table 3: Zero-shot transfer results on NExT-GQA and ReXTime.

	NExT-GQA (Xiao et al., 2024)			**ReXTime** (Chen et al., 2024b)		
Method	**Acc**	**mIoU**	**Acc@GQA**	**Acc**	**mIoU**	**Acc@GQA**
Qwen2.5-VL-7B Bai et al. (2025b)	72.7	16.6	11.1	60.4	14.6	9.6
Time-R1-7B Wang et al. (2026)	78.4	28.6	19.4	64.3	25.4	17.9
VideoChat-R1-7B Li et al. (2025)	75.6	30.8	20.7	63.6	29.1	21.4
TW-GRPO Dang et al. (2025)	74.0	17.4	10.5	60.8	16.0	11.4
VideoChat-R1.5 Yan et al. (2026)	77.9	15.4	9.1	**65.7**	12.2	6.3
EG-Reasoner (Ours)	**78.7**	**32.6**	**23.3**	64.8	**34.4**	**25.5**

EG-Reasoner demonstrates strong transferability across both benchmarks. On NExT-GQA, it achieves the best overall performance, leading both answer accuracy and temporal grounding metrics among all evaluated models. On ReXTime, while VideoChat-R1.5 obtains slightly higher answer accuracy, EG-Reasoner achieves the strongest temporal localization performance, obtaining the highest mIoU and Acc@GQA. These results indicate that evidence-aware training improves general temporal reasoning and grounding capabilities beyond the original EG-VQA setting.

5.5 Case Study

Figure 4 presents qualitative comparisons between the base model Qwen2.5-VL and EG-Reasoner in evidence-grounded reasoning. In Figure 4(a), the baseline model predicts the correct answer but fails to localize the relevant temporal segment, instead relying on loosely related context. In contrast, EG-Reasoner accurately identifies the key evidence region associated with the chalk-drawing action. Figure 4(b) further demonstrates the advantage of evidence-grounded reasoning in complex procedural understanding. The baseline model produces hallucinated objects and incorrect reasoning steps, whereas EG-Reasoner decomposes the task into fine-grained temporal evidence segments and generates a coherent reasoning chain aligned with the video content, leading to the correct answer. Additional qualitative examples and failure cases are provided in Appendix E.

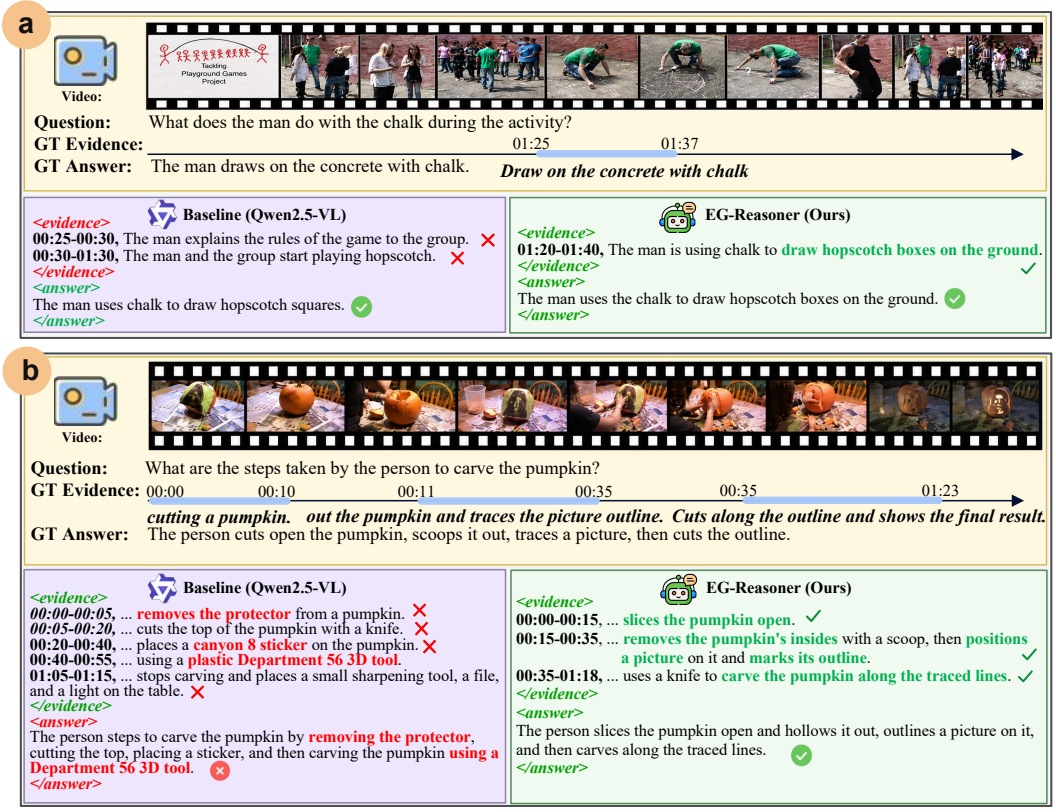

Figure 4: Qualitative comparison between the baseline model and EG-Reasoner. Hallucinated evidence is highlighted in red, and correctly grounded evidence is marked in green.

6 Conclusion

This paper introduces EG-VQA, a benchmark that evaluates VideoQA models through joint assessment of answer correctness and temporal-semantic evidence faithfulness. Additionally, the Evidence-Grounded F1 (EG-F1) metric is introduced to jointly measure temporal alignment and semantic consistency between predicted and reference evidence. Experiments show that Video-LLMs, including strong proprietary models, may achieve high answer correctness while still exhibiting gaps in temporal-semantic evidence grounding, highlighting the discrepancy between answer correctness and evidence faithfulness. To further investigate evidence-aware learning under EG-VQA, we develop EG-Reasoner as an evidence-aware reference model trained with explicit evidence supervision. EG-Reasoner achieves the strongest evidence grounding performance among open-source models while maintaining competitive answer generation performance. Overall, our findings suggest that evidence-aware supervision provides an effective direction for improving grounded video understanding, and we hope EG-VQA can serve as a useful testbed for future research on evidence-aware multimodal reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for both research-related tasks and manuscript preparation. For dataset construction, GPT-4o was used to generate candidate question–answer pairs and supporting evidence. DeepSeek-V3.2 and Gemini-2.5-Flash were used as independent evaluators during automatic quality control, with model disagreements further reviewed by human annotators. Gemini-2.5-Pro was used as an automatic judge for open-ended answer evaluation, and its agreement with human annotations was separately validated as described in the paper.

Generative AI tools were also used to assist in drafting portions of the manuscript and to improve the readability and language quality of the paper. All AI-assisted text was reviewed, revised, and verified by the authors.

The authors reviewed all AI-assisted research artifacts and manuscript content and take responsibility for the final content of this work, including all text, claims, experimental results, and conclusions.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Herbench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4505–4514, 2026.

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

Guo Chen, Yicheng Liu, Yi-Fei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *ArXiv*, abs/2412.12075, 2024a. URL https://api.semanticscholar.org/CorpusID:274788907.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang F Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673, 2024b.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *ArXiv*, abs/2408.14469, 2024c. URL https://api.semanticscholar.org/CorpusID:271957561.

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24087–24096, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv preprint arXiv:2505.24718*, 2025.

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 IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

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.

Ayush Gupta, Anirban Roy, Rama Chellappa, Nathaniel D. Bastian, Alvaro Velasquez, and Susmit Jha. Toga: Temporally grounded open-ended video qa with weak supervision. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23593–23603, 2025. URL https://api.semanticscholar.org/CorpusID:279305993.

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.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9091–9101, 2025.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4300–4316, 2023.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8211–8225, 2020.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding. *Science China Information Sciences*, 68, 2023. URL https://api.semanticscholar.org/CorpusID:258588306.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Liang Lin and Yang Liu. Multimodal large models: A new paradigm of artificial intelligence. *Cham: Springer*, 2026.

Y. Liu, Weixing Chen, Guanbin Li, and Liang Lin. Causalvlr: A toolbox and benchmark for visual-linguistic causal reasoning. *ArXiv*, abs/2306.17462, 2023a. URL https://api.semanticscholar.org/CorpusID:259308889.

Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. Causal reasoning meets visual representation learning: A prospective study. *Machine Intelligence Research*, 19(6):485–511, 2022.

Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641, 2023b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026.

Yushen Wei, Yang Liu, Hong Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 377–386, 2023.

Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. *ArXiv*, abs/2405.09711, 2024. URL https://api.semanticscholar.org/CorpusID:244907062.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781, 2021. URL https://api.semanticscholar.org/CorpusID:234763093.

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*, pp. 13204–13214, 2024.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

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

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

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 6439–6455, 2022.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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 Dataset Construction Details

A.1 Filtering protocol

To ensure the reliability and consistency of the benchmark, we apply a structured filtering protocol to the collected videos and their associated annotations. This process is designed to correct annotation errors, remove low-quality samples, and retain videos that support meaningful reasoning.

Unified Metadata Format. The source datasets used in EG-VQA adopt different annotation schemas and temporal segmentation formats. To facilitate unified processing and evidence-grounded question generation, all videos are first converted into a standardized metadata representation consisting of video-level metadata, temporally ordered timestamps, and corresponding event descriptions.

Each video is represented as a structured sequence of timestamp–description pairs. An example is shown below:


```
"TEOPA76qOKA": {
  "duration": 88,
  "title": "How to Make a Lemon Battery",
  "timestamps": [
    [30.0, 34.0],
    [34.0, 53.0],
    [53.0, 63.0],
    [63.0, 76.0],
    [76.0, 85.0]
  ],
  "descriptions": [
    "Roll the lemons",
    "Put copper wire and paper clips",
    "Connect alligator clips",
    "Touch the unconnected ends",
    "Connect the positive to LED"
  ]
}
```


This unified representation serves as the basis for subsequent filtering, question generation, evidence annotation, and quality control procedures.

Data Validation. We first remove corrupted, incomplete, or inaccessible video samples. Videos that cannot be decoded, contain missing segments, or fail to load correctly are excluded from further processing.

Temporal Alignment Refinement. Although the source datasets provide temporally segmented annotations, the original segment boundaries and descriptions may not always accurately reflect the underlying video content. To address this, human annotators are instructed to review each video and its corresponding segments, and to refine temporal boundaries when necessary. Segments that exhibit significant misalignment with the actual visual events are corrected or removed.

Description Consistency Check. We further examine the consistency between segment descriptions and the observed video content. Descriptions that are ambiguous, incomplete, or inconsistent with the visual evidence are revised or discarded. Samples with severely noisy or unreliable metadata are excluded entirely.

Informational Density Filtering. To ensure that the resulting dataset supports non-trivial reasoning, we remove videos with low informational density. Specifically, videos that contain only a single event, minimal temporal variation, or insufficient interaction between entities are filtered out, as they tend to yield trivial or purely descriptive questions.

Human Verification. All refinement steps are conducted by trained human annotators following a standardized review protocol. Annotators are instructed to prioritize both temporal accuracy and semantic consistency when validating segment annotations.

After applying the above filtering criteria, we retain a total of 2,067 high-quality videos, which serve as the foundation for constructing the EG-VQA benchmark.

A.2 Prompt Design and Iterative Refinement

We construct question–answer (QA) pairs using a human-in-the-loop pipeline that combines large language model generation with iterative prompt refinement. The goal is to produce high-quality, reasoning-oriented questions that are grounded in temporally annotated video content.

Initial Prompt Design. Given the temporally segmented video metadata, we design type-specific prompts for each reasoning category (descriptive, temporal, causal, and counterfactual). Each prompt explicitly encourages the generation of open-ended questions that require reasoning over one or multiple temporal segments, rather than relying on superficial patterns or static descriptions.

Failure Mode Identification. Direct application of large language models often results in suboptimal QA pairs. To systematically identify common issues, we apply the initial prompts to a held-out subset of videos and manually inspect the outputs. The most frequent failure modes include:

- **Video-independent questions:** Questions that can be answered without referring to the video content.
- **Weak grounding:** Questions or answers that are only loosely related to the temporal metadata.
- **Incomplete reasoning:** Answers that omit necessary temporal or contextual information.

Iterative Refinement. Based on these observations, we iteratively refine the prompts by introducing stricter constraints on grounding, temporal specificity, and reasoning depth. This process follows a generate–inspect–refine loop, where prompts are repeatedly adjusted and re-evaluated until the generated QA pairs consistently satisfy our quality criteria. During this process, we explicitly filter templated or weakly grounded questions to reduce potential shortcut patterns that could be exploited without genuine video understanding.

Final Deployment. Once stabilized, the refined prompts are applied to the full set of filtered videos to generate candidate QA pairs at scale. This iterative design ensures that the resulting dataset emphasizes non-trivial, reasoning-intensive, and evidence-grounded questions.

A.3 Quality Control Protocol

To ensure the reliability of generated QA pairs, we employ a two-stage quality control protocol that combines structured automatic evaluation with human verification.

Stage 1: Structured Multi-model Evaluation. Each candidate QA pair is independently evaluated by two large language models from distinct architectural families (DeepSeek-V3.2 and Gemini-2.5-Flash). We adopt a structured scoring rubric consisting of three criteria:

- **Question Reasonableness:** whether the question is well-posed and requires video understanding.
- **Answer Correctness:** whether the answer is consistent with the video content.
- **Evidence Alignment:** whether the answer is properly grounded in the relevant temporal segments.

Each dimension is scored on a scale of 0–10, and the total score is obtained by summing the three dimensions. A sample is considered acceptable if the total score is greater than or equal to 25.

Agreement-based Filtering. We adopt a dual-model agreement strategy to improve robustness. When both models agree on the decision (accept or reject), the result is directly adopted. Otherwise, the sample is flagged for manual review.

Stage 2: Human Adjudication. All samples for which the two evaluators disagree are reviewed by human annotators using the same criteria of question reasonableness, answer correctness, and evidence alignment. Annotators determine whether a sample should be retained, corrected, or removed.

Post-hoc Validation. After the full-dataset screening and adjudication process, we conduct an additional stratified random audit of 1,000 QA pairs. The audit is stratified across the three source datasets, video-duration ranges, and four question types to provide broad coverage of the resulting benchmark. For each audited sample, human annotators inspect the original video together with the corresponding metadata, generated question, reference answer, and supporting evidence. They verify whether the question is answerable from the observed video content, whether the answer is correctly supported by the video, and whether the annotated evidence is temporally and semantically aligned with the answer. Of the 1,000 audited samples, 932 (93.2%) are judged valid without modification, while 24 (2.4%) contain only minor correctable issues and are retained after manual revision.

Invalid-sample Handling. The remaining 44 samples (4.4%) are removed due to substantive validity issues, including questions that are not answerable from the video, answers that are inconsistent with the observed content, supporting evidence that fails to justify the answer. Overall, 956 of the 1,000 audited samples (95.6%) are retained after post-hoc validation and correction.

A.4 Dataset Statistics and Benchmark Comparison

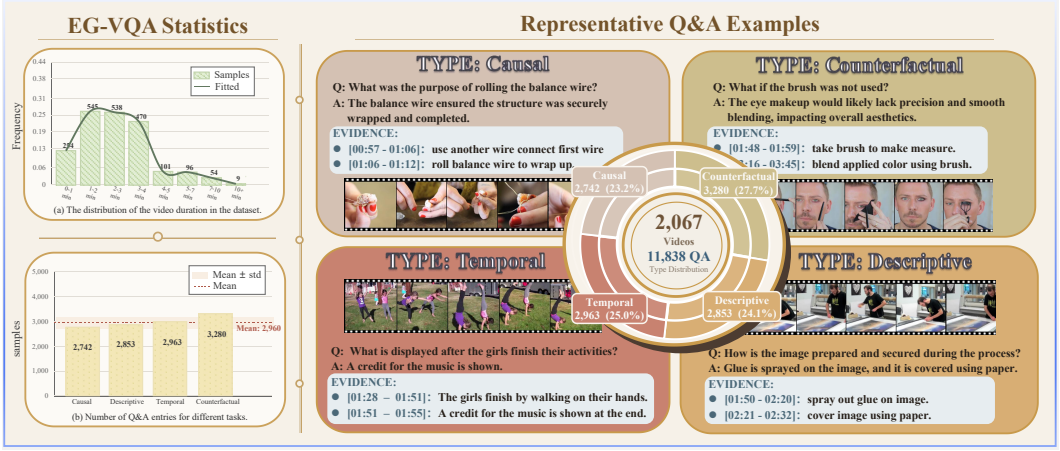

Figure 5: Overview of EG-VQA statistics and representative QA examples.

As shown in Figure 5, the EG-VQA benchmark contains 2,067 videos and 11,838 QA pairs. The benchmark is constructed from the official validation splits of ActivityNet Captions and YouCook2, and the validation and test splits of HiREST. After source-specific curation, we perform a stratified video-level split by controlling the distributions of source dataset, video duration, and evidence density. All QA pairs and evidence annotations derived from the same video are assigned to the same partition, ensuring that no video appears in both the training and test sets. Detailed source-wise statistics are summarized in Table 4, while per-split question statistics are provided in Table 5.

Reasoning-oriented Question Distribution. EG-VQA contains four categories of reasoning tasks: descriptive, temporal, causal, and counterfactual. These categories progressively increase the reasoning difficulty, ranging from direct perception of observable content to hypothetical reasoning over altered events. As shown in Figure 5, the dataset maintains a relatively balanced distribution across different reasoning types, ensuring diverse evaluation of video understanding capability.

Table 4: Source composition and video-level split statistics of EG-VQA.

Source Dataset	Source Split	#Videos	Train	Test
ActivityNet Captions	Validation	1,500	1,277	223
YouCook2	Validation	209	168	41
HiREST	Val. + Test	358	286	72
Total	–	2,067	1,731	336

Table 5: Question and evidence statistics of EG-VQA.

	Train		Test	
Question Type	#Pairs	Avg. Evid.	#Pairs	Avg. Evid.
Descriptive	2,196	2.12	657	2.24
Temporal	2,212	2.66	751	2.89
Causal	1,994	2.23	748	2.26
Counterfactual	2,547	1.96	733	2.04
Total	8,949	2.24	2,889	2.36

Evidence-grounded Reasoning Characteristics. EG-VQA explicitly associates each question with one or more temporally localized evidence units. Each evidence unit consists of a temporal span and a textual description of the corresponding event, enabling models to identify both where supporting evidence occurs and what semantic content it conveys. On average, each question is paired with 2.27 evidence segments, reflecting the multi-step nature of video reasoning.

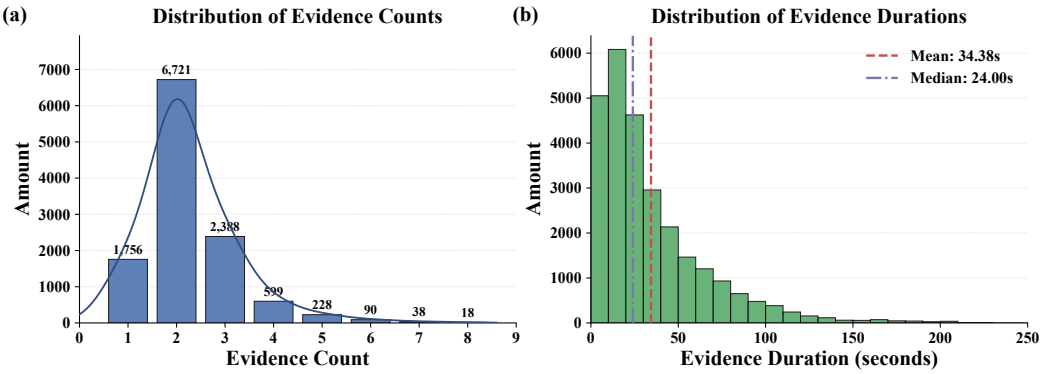

Figure 6: Evidence statistics of EG-VQA. (a) Distribution of evidence counts per question. (b) Distribution of evidence duration.

Figure 6(a) shows the distribution of evidence counts per question. Most questions require multiple evidence segments rather than a single localized event, indicating that successful reasoning often depends on integrating information across temporally separated video events.

In addition, the temporal spans of evidence vary substantially across samples. Figure 6(b) illustrates the distribution of evidence duration. The benchmark contains both short fine-grained actions and long procedural activities, requiring models to perform reasoning at multiple temporal scales.

The average video duration is 177.6 seconds, while the average evidence segment length is 34.4 seconds. This large gap further increases the difficulty of evidence grounding, since models must identify relevant moments within long video contexts.

Benchmark Comparison. Table 6 compares EG-VQA with representative VideoQA benchmarks that provide explicit evidence-related supervision. Existing grounded benchmarks have made important progress in associating answers with relevant temporal evidence. However, they typically represent evidence primarily through temporal localization annotations. In contrast, EG-VQA explicitly represents supporting evidence as temporal-semantic units, where each evidence segment is paired with a textual description of the corresponding event.

Table 6: Comparison between evidence-grounded VideoQA benchmarks and EG-VQA.

Benchmark	Format	Temporal	Multiple	Textual Evidence	Semantic Matching
TVQA+ (Lei et al., 2020)	MC	✓	×	×	×
NExT-GQA (Xiao et al., 2024)	MC	✓	✓	×	×
MultiHop-EgoQA (Chen et al., 2024c)	OE	✓	✓	×	×
CG-Bench (Chen et al., 2024a)	MC	✓	✓	×	×
HERBench (Ben Ami et al., 2026)	MC	×	✓	×	×
EG-VQA	**OE**	✓	✓	✓	✓

Note. Here, "Temporal" denotes temporal evidence localization, "Multiple" denotes multi-segment evidence annotation, "Textual Evidence" denotes standalone textual descriptions for evidence segments, and "Semantic Matching" denotes semantic evaluation of evidence descriptions.

Therefore, EG-VQA enables evaluation beyond whether a model identifies relevant moments, by additionally assessing whether the generated evidence descriptions are semantically consistent with the supporting video events. This design supports research on open-ended VideoQA with explicitly verifiable temporal-semantic evidence.

B EG-F1: Complete Formulation

This section expands the EG-F1 metric introduced in Section 3.3.

B.1 Computation Steps

Step 1: Pairwise similarity. For a ground-truth item $g_i = (t_i, d_i)$ and a prediction $p_j = (t_j, d_j)$, we compute both temporal overlap and semantic similarity:

$$\mathrm{IoU}_{ij} = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}, \tag{5}$$

$$\mathrm{Sim}_{ij} = \cos\big(\phi(d_i), \phi(d_j)\big), \tag{6}$$

where $\phi(\cdot)$ denotes a sentence embedding function used to encode textual evidence descriptions.

Step 2: Edge construction. Valid candidate edges are defined by

$$\mathcal{E} = \{(i,j) \mid \mathrm{IoU}_{ij} \geq \alpha \wedge \mathrm{Sim}_{ij} \geq \beta\}, \tag{7}$$

where α and β denote the temporal IoU threshold and semantic similarity threshold, respectively. Different threshold settings are used in our experiments to evaluate grounding performance under varying levels of strictness.

Step 3: Optimal matching. A bipartite graph with edge weights

$$s_{ij} = \begin{cases} \mathrm{IoU}_{ij} \cdot \mathrm{Sim}_{ij}, & (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

is constructed between $\mathcal{G}$ and $\mathcal{P}$. The Hungarian algorithm is applied to obtain a maximum-weight one-to-one matching $\mathcal{M}$. Let $M = |\{(i,j) \in \mathcal{M} \mid s_{ij} > 0\}|$.

Step 4: Metric computation. Precision, recall, and F1 are computed as

$$P = \frac{M}{|\mathcal{P}|}, \quad R = \frac{M}{|\mathcal{G}|} \tag{9}$$

$$\mathrm{EG{-}F1} = \frac{2PR}{P+R}. \tag{10}$$

B.2 SOFT VARIANT FOR TRAINING.

To mitigate reward sparsity during reinforcement learning, a soft variant is employed by removing the hard threshold in Step 2. The pairwise weight becomes $s_{ij}^{\text{soft}} = \text{IoU}_{ij} \cdot \text{Sim}_{ij}$ for all (i, j), and soft precision/recall are defined as

$$P_{\text{soft}} = \frac{\sum_{(i,j)\in\mathcal{M}} s_{ij}^{\text{soft}}}{|\mathcal{P}|}, \quad R_{\text{soft}} = \frac{\sum_{(i,j)\in\mathcal{M}} s_{ij}^{\text{soft}}}{|\mathcal{G}|}. \tag{11}$$

The soft variant is then

$$\text{EG-F1}_{\text{soft}} = \frac{2P_{\text{soft}}R_{\text{soft}}}{P_{\text{soft}} + R_{\text{soft}}} \tag{12}$$

, which is combined with the hard EG-F1 as described in the main text.

B.3 COMPLEXITY.

Computing all pairwise similarities costs $O(|\mathcal{G}||\mathcal{P}|)$; the Hungarian algorithm adds $O(\max(|\mathcal{G}|, |\mathcal{P}|)^3)$. In practice, $|\mathcal{G}|$ and $|\mathcal{P}|$ are small (average 2.27), making the metric efficient to compute at both training and inference time.

C EG-REASONER: SUPPLEMENTARY DETAILS

C.1 ANSWER REWARD DETAILS

The answer reward evaluates whether the generated answer is semantically correct and consistent with the reference answer while avoiding hallucinated or unsupported content.

Evaluation Principles. We adopt a rule-guided LLM evaluator that assesses each generated answer based on two complementary principles:

- **Inclusion principle:** the generated answer should cover as many key elements of the reference answer as possible.
- **Exclusion principle:** the answer should not contain information that contradicts or cannot be inferred from the video metadata.

Scoring Scheme. To reduce reward sparsity, especially during early training, we employ a relaxed scoring scheme that assigns partial credit:

$$R_{\text{ans}}(y) = \begin{cases} 1, & \text{fully correct,} \\ 0.5, & \text{partially correct,} \\ 0, & \text{incorrect.} \end{cases} \tag{13}$$

This design provides denser supervision signals compared to binary evaluation, enabling more stable optimization.

Prompt-based Evaluation. The evaluation is performed using a prompt-based LLM judge, which compares the generated answer against the reference answer and the associated video metadata. The prompt explicitly instructs the evaluator to follow the inclusion and exclusion principles and to output a discrete score according to the predefined rubric.

C.2 GRPO OBJECTIVE DETAILS

We adopt Group Relative Policy Optimization (GRPO) to optimize the structured generation policy.

Given an input (V, q), the policy π_θ samples a group of candidate responses $\{y_1, \ldots, y_G\}$. Each candidate is assigned a reward R_i, and the relative advantage is computed as:

$$A_i = \frac{R_i - \mu}{\sigma}, \tag{14}$$

where μ and σ denote the mean and standard deviation of rewards within the sampled group.

Let $r_i = \frac{\pi_\theta(y_i|V,q)}{\pi_{\theta_{\text{old}}}(y_i|V,q)}$ denote the importance sampling ratio. The clipped GRPO objective is:

$$\mathcal{L}(\theta) = \mathbb{E}_{(V,q),\{y_i\}\sim\pi_{\theta_{\text{old}}}} \left[\sum_{i=1}^{G} \min(r_i A_i,\ \text{clip}(r_i, 1-\epsilon, 1+\epsilon) A_i) - \beta_{\text{KL}}\, \text{KL}(\pi_\theta \| \pi_{\text{ref}}) \right], \quad (15)$$

where $\epsilon = 0.2$ is the clipping parameter and $\beta_{\text{KL}} = 0.01$ controls the strength of KL regularization.

KL Regularization. The KL term discourages excessive deviation from the reference policy during post-training. It is evaluated over the generated tokens and added to the policy objective as a regularization term.

GRPO uses group-relative rewards to estimate advantages and therefore does not require a separately trained value model. We otherwise follow the standard GRPO optimization procedure.

C.3 Training Configuration

The training configuration of EG-Reasoner is summarized in Table 7. We initialize the policy from Qwen2.5-VL-7B-Instruct, freeze the visual backbone, and optimize only the language-model parameters. Unless otherwise specified, all EG-Reasoner variants use the same training configuration.

Table 7: Training configuration of EG-Reasoner.

Hyperparameter	**Value**
Base model	Qwen2.5-VL-7B-Instruct
Trainable parameters	Language model only
Frozen parameters	Visual backbone
Epochs	1
Training frames	32
Frame sampling strategy	Uniform sampling
Per-device batch size	4
Learning rate	1e-6
Training steps	558
Group size G	8
λ_f / λ_e / λ_a	0.1 / 0.6 / 0.3
ϵ (clip)	0.2
β_{KL}	0.01
Sentence Embeddings	all-MiniLM-L6-v2
Number of GPUs	4 $\times$ NVIDIA A800
Training time	$\sim$50 hours
Output format	Unified structured format in Section 4

D Experimental Details

D.1 Baseline Details

We evaluate representative Video-LLMs covering proprietary systems, general-purpose open-source models, and recent reasoning-enhanced variants.

GPT-4o (Hurst et al., 2024) is a proprietary multimodal large language model capable of processing visual and textual inputs in a unified framework. We use it as a representative commercial baseline.

Gemini-2.5-Flash (Comanici et al., 2025) is a proprietary multimodal model from the Gemini family, designed to balance efficiency and reasoning capability. We include it to evaluate commercial multimodal systems.

GPT-5.6 Luna is a proprietary multimodal model from OpenAI, designed for general-purpose multimodal understanding and reasoning. We include it to assess the capability of recent commercial models.

Qwen3.7 Plus is a proprietary multimodal model from the Qwen family, developed for multimodal perception and reasoning. We include it as a representative commercial model.

Qwen2.5-VL-7B (Bai et al., 2025b) is an open-source multimodal model built upon the Qwen2.5 backbone. It supports image and video inputs and is used as a representative general-purpose open-source baseline.

InternVL-3-8B (Zhu et al., 2025) is an open-source multimodal model from the InternVL family, designed for multimodal perception and reasoning. It is included as a general multimodal understanding baseline.

Time-R1-7B (Wang et al., 2026) is a reasoning-oriented video model designed to improve temporal reasoning and event-level understanding. We include it to evaluate temporal reasoning capability.

TW-GRPO (Dang et al., 2025) is a GRPO-based reasoning model that introduces token-wise weighting and soft reward mechanisms to provide more fine-grained supervision. It is included as a reinforcement learning-based reasoning baseline.

VideoChat-R1-7B (Li et al., 2025) is a video reasoning model trained with reinforcement learning to enhance spatio-temporal perception and reasoning. It serves as a strong open-source video reasoning baseline.

VideoChat-R1-Thinking-7B (Li et al., 2025) extends the VideoChat-R1 framework by incorporating explicit thinking-style reasoning, and is included as a reasoning-oriented video baseline.

VideoChat-R1.5-7B (Yan et al., 2026) is an improved version of VideoChat-R1 that emphasizes test-time perceptual refinement. We include it as a recent video reasoning baseline.

D.2 Answer Evaluation Protocol

We evaluate answer correctness using an LLM-based protocol, as EG-VQA is formulated as open-ended question answering where semantically equivalent answers may have different surface forms, making traditional string-based metrics insufficient.

Given a test set of N samples, let x_i denote the model response and y_i the reference answer. The strict accuracy is defined as:

$$\mathrm{Acc}_{\mathrm{strict}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Match}_{\mathrm{full}}(x_i, y_i), \tag{16}$$

where $\mathrm{Match}_{\mathrm{full}}(\cdot)$ returns 1 if the response fully matches the reference answer in meaning.

To account for partially correct answers, we define relaxed accuracy as:

$$\mathrm{Acc}_{\mathrm{relaxed}} = \frac{1}{N}\sum_{i=1}^{N} \left(\mathrm{Match}_{\mathrm{full}}(x_i, y_i) + \gamma \cdot \mathrm{Match}_{\mathrm{partial}}(x_i, y_i)\right), \tag{17}$$

where $\mathrm{Match}_{\mathrm{partial}}(\cdot)$ indicates partial correctness and γ controls the credit assigned to partially correct answers. In our experiments, $\gamma = 0.5$.

We use Gemini-2.5-Pro as the evaluator. For each prediction, the evaluator receives the question, reference answer, model response, video title, and temporally annotated video metadata, and assigns one of three labels: Fully Correct, Partially Correct, or Incorrect. The same protocol is applied to all evaluated models.

To validate the reliability of the evaluator, we collect responses generated by Qwen2.5-VL-7B on the EG-VQA evaluation set. Annotator 1 manually inspected these real model outputs and selected a balanced validation subset of 200 responses, including 100 Fully Correct, 50 Partially Correct, and 50 Incorrect cases. The selected samples are balanced across the three source datasets and four question types. Annotator 2, blind to Annotator 1's selections, independently evaluated the same responses using the 1/0.5/0 scoring rubric. Four responses selected as Fully Correct by Annotator 1 were reclassified as Partially Correct by Annotator 2, resulting in 96 Fully Correct, 54 Partially Correct, and 50 Incorrect human judgments. Gemini-2.5-Pro independently evaluated the same inputs under the same protocol.

Table 8: Confusion matrix between human and Gemini-2.5-Pro judgments.

Human \ Gemini-2.5-Pro	Incorrect	Partial	Full	Total
Incorrect	50	0	0	50
Partial	0	47	7	54
Full	0	0	96	96
Total	50	47	103	200

Gemini-2.5-Pro achieves 96.5% exact agreement with human judgments (193/200), with a Cohen's κ of 0.944. All seven disagreements occur between Partially Correct and Fully Correct judgments, with none crossing the Incorrect-versus-Correct boundary. In particular, all 50 responses judged Incorrect by human annotators are also identified as Incorrect by Gemini-2.5-Pro, indicating that the remaining discrepancies are confined to the boundary of answer completeness rather than substantive correctness.

D.3 Evidence Evaluation Settings

We evaluate evidence grounding using both event-level temporal F1 and the proposed EG-F1 metric.

For temporal localization, we report F1 scores under multiple IoU thresholds $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ to evaluate localization performance at different granularities.

For EG-F1, we evaluate multiple threshold combinations (α, β), where α denotes the temporal IoU threshold and β denotes the semantic similarity threshold. Specifically, we report results under $(0.3, 0.5)$, $(0.3, 0.75)$, and $(0.5, 0.75)$, corresponding to increasingly strict grounding criteria.

D.4 Grounded Answer Score

Answer correctness and evidence grounding measure complementary aspects of evidence-grounded VideoQA, but reporting them independently does not indicate whether both are achieved on the same samples. We therefore additionally report the Grounded Answer Score (GAS), an auxiliary sample-level metric that couples answer correctness with evidence quality.

For sample i, let $a_i \in \{0, 0.5, 1\}$ denote the relaxed answer score corresponding to Incorrect, Partially Correct, and Fully Correct predictions, respectively. Let $s_i^{\text{soft}} \in [0, 1]$ denote the Soft EG-F1 score of the predicted evidence, computed using the soft temporal-semantic matching formulation in Appendix B.2. GAS is defined as:

$$\text{GAS} = \frac{100}{N} \sum_{i=1}^{N} a_i \, s_i^{\text{soft}}. \tag{18}$$

A high GAS therefore requires a model to produce both a correct answer and well-grounded evidence for the same sample. Incorrect answers contribute zero regardless of evidence quality, while partially correct answers receive half credit. We use Soft EG-F1 as the evidence component to retain continuous temporal-semantic evidence quality without introducing an additional hard threshold. GAS is computed by coupling answer correctness and evidence quality at the sample level before dataset-level aggregation, and therefore differs from multiplying their independently averaged scores.

D.5 Evaluation Configuration

The key evaluation settings for our experiments on the EG-VQA benchmark are summarized in Table 9.

Table 9: Evaluation configuration used for EG-VQA.

Evaluation setting	**Value**
Benchmark	EG-VQA
Task format	Open-ended QA
Evaluation frames	64
Frame sampling strategy	Uniform sampling
Maximum visual resolution budget	$256 \times 28 \times 28$
Output format	Unified structured answer-and-evidence format
Answer metric	Strict / Relaxed accuracy
Evidence metric	EG-F1 and event-level F1

D.6 Full Evidence Grounding Results

Table 10 reports the complete evidence grounding results under the evaluation settings defined in Appendix D.3. These results complement the evidence-grounding summary reported in Table 1 with threshold-specific breakdowns.

Table 10: Threshold-specific evidence grounding results on EG-VQA.

Model	**Event-level F1**				**EG-F1**		
	@0.1	**@0.3**	**@0.5**	**@0.7**	**(0.3, 0.5)**	**(0.3, 0.75)**	**(0.5, 0.75)**
Human Results*	93.29	89.48	68.75	43.52	84.61	47.67	36.74
Commercial Large Multimodal Models							
GPT-4o Hurst et al. (2024)	35.85	17.85	5.67	2.95	8.49	1.18	0.33
Gemini-2.5-Flash Comanici et al. (2025)	28.87	19.81	12.56	6.88	12.64	1.48	0.86
GPT-5.6 Luna	57.13	41.88	22.39	9.78	25.60	4.09	1.99
Qwen3.7 Plus	36.64	25.18	14.63	6.49	15.57	1.37	0.73
Open-sourced Video-LLMs							
Qwen2.5-VL-7B Bai et al. (2025b)	29.36	15.83	7.70	2.55	8.18	1.15	0.66
InternVL-3-8B Zhu et al. (2025)	20.89	11.02	4.68	1.34	5.11	0.60	0.35
Time-R1-7B Wang et al. (2026)	46.46	32.03	15.64	5.41	16.46	2.75	1.40
VideoChat-R1-7B Li et al. (2025)	48.24	34.50	18.43	6.84	18.71	2.70	1.56
VideoChat-R1-Thinking-7B Li et al. (2025)	40.90	26.90	13.69	4.77	14.51	1.82	0.94
VideoChat-R1.5-7B Yan et al. (2026)	33.26	21.44	10.88	3.58	11.86	2.17	1.22
TW-GRPO Dang et al. (2025)	33.30	19.48	8.97	2.94	9.65	1.07	0.58
EG-Reasoner (Ours)	**65.97**	**48.02**	**27.27**	**10.30**	**28.79**	**4.95**	**3.12**

Performance generally decreases as the evaluation criteria become more stringent, highlighting the difficulty of precise evidence grounding. EG-Reasoner achieves the strongest model performance across all reported event-level F1 and EG-F1 settings, including under the strictest temporal thresholds, reaching 27.27 F1@0.5 and 10.30 F1@0.7.

D.7 Benchmark Validity and Diagnosis

We conduct two complementary diagnostic analyses to examine whether EG-VQA genuinely requires the corresponding visual content and whether answer correctness faithfully reflects evidence-grounding quality.

Dependence on visual input. We evaluate Qwen2.5-VL-7B and EG-Reasoner under three input conditions. In the *Full Video* setting, the model receives the original video together with the question.

Table 11: Video dependency analysis under different input conditions.

Model	Input Condition	Answer Correctness		Evidence Grounding	
		Relaxed Acc. ↑	Drop	Avg. EG-F1 ↑	Drop
Qwen2.5-VL-7B	Full Video	27.96	–	3.33	–
	Question-only	4.36	23.60	0.06	3.27
	Mismatched Video	1.97	25.99	0.02	3.31
EG-Reasoner	Full Video	42.71	–	12.28	–
	Question-only	9.61	33.10	0.22	12.06
	Mismatched Video	6.77	35.94	0.31	11.97

Note: Drop denotes the absolute decrease relative to the Full Video setting.

In the *Question-only* setting, the video input is removed. In the *Mismatched Video* setting, the original video is replaced with an unrelated video while the question remains unchanged. As shown in Table 11, removing or mismatching the video leads to substantial degradation in both answer correctness and evidence grounding. Relaxed accuracy drops by 23.60–35.94 percentage points, while Avg. EG-F1 falls to 0.02–0.31 under the ablated input conditions. These results indicate that EG-VQA substantially depends on the corresponding visual content rather than being solvable from question-side cues alone.

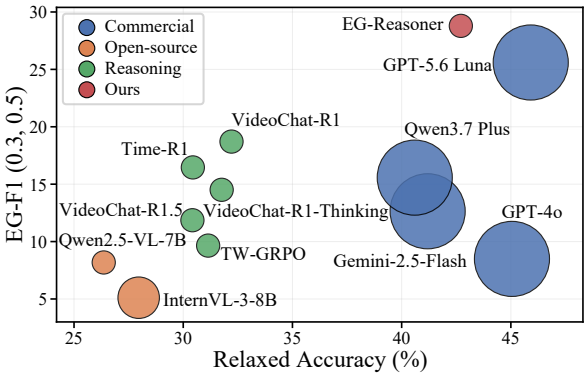

Figure 7: Relationship between answer correctness and evidence grounding across models. Marker size indicates model scale.

Answer correctness does not necessarily coincide with evidence grounding. Figure 7 compares relaxed answer accuracy with evidence grounding across the evaluated models. Models with similar answer accuracy can exhibit substantially different grounding quality, and stronger answer performance does not consistently correspond to stronger evidence grounding. Although larger models generally achieve competitive answer accuracy, increased model scale does not necessarily translate into proportionally better grounding performance. EG-Reasoner exhibits a more balanced profile across the two dimensions despite its smaller scale. Together, this discrepancy highlights the limitations of answer-only evaluation and further motivates evidence-grounded evaluation beyond answer accuracy.

E ADDITIONAL CASE STUDIES

We provide additional qualitative examples to further analyze the behavior of EG-Reasoner on evidence-grounded video reasoning.

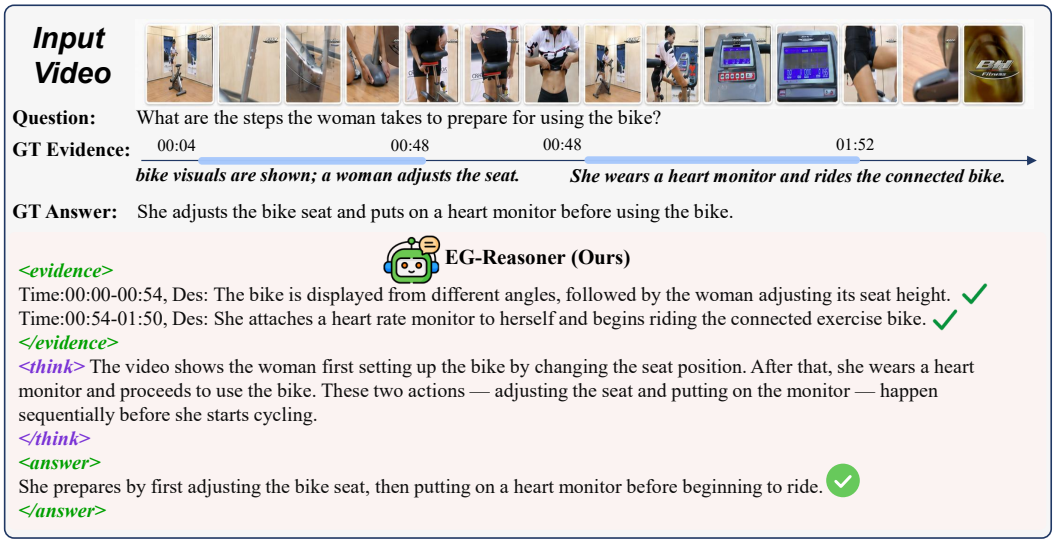

Figure 8: Qualitative example of successful evidence-grounded reasoning. EG-Reasoner correctly identifies both the preparation steps and the corresponding temporal evidence segments.

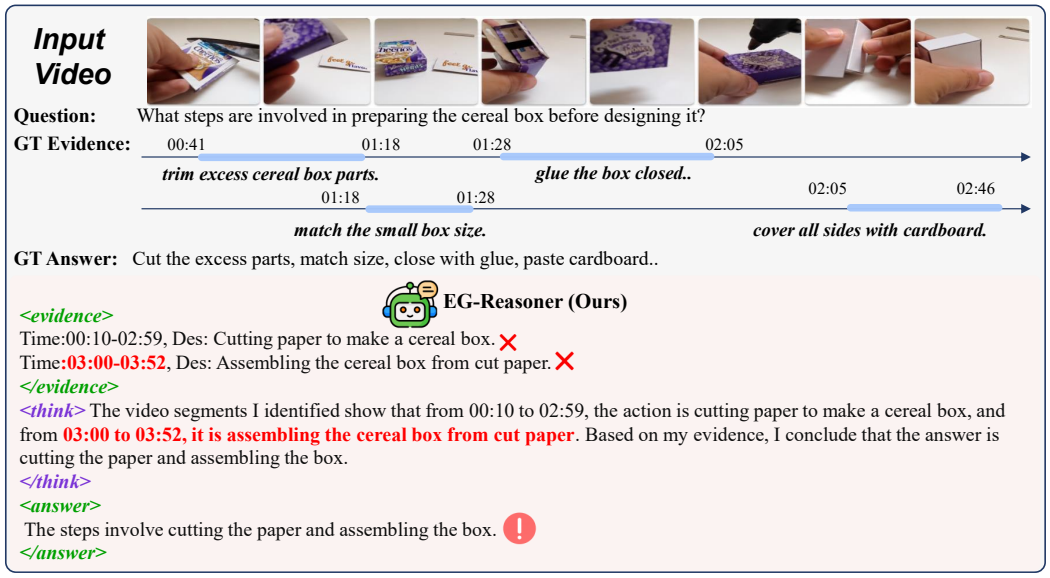

Figure 9: Qualitative example of incomplete procedural grounding. EG-Reasoner captures the overall procedure but misses several fine-grained intermediate steps required for detailed reasoning.

F LIMITATIONS

Despite the proposed benchmark and evaluation framework, several limitations remain:

- Although EG-VQA is constructed with newly generated question-answer pairs and avoids using the original training splits of the source datasets, we cannot completely exclude potential exposure of the underlying videos in the pretraining corpora of existing Video-LLMs. This limitation is particularly difficult to eliminate for proprietary models whose training data are unavailable.
- Potential human annotation bias may exist due to variations among annotators during the benchmark construction and validation processes.
- A large portion of QA pairs in EG-VQA are generated with the assistance of large language models. Despite iterative prompt refinement, multi-stage screening, and human validation, the generated QA pairs may still inherit potential biases from large language models.
- Despite careful prompt design and human validation, model-based automatic evaluation may still face challenges where evaluation results may not always align with human assessment.

We will further investigate and address these limitations in future work.

G PROMPT TEMPLATES

Figures 10–16 present the complete set of prompt templates designed in this paper. Specifically, Fig. 10 illustrates the prompt template for descriptive question generation. Figs. 11, 12, and 13 show the prompts for temporal, causal, and counterfactual reasoning generation, respectively. Fig. 14 presents the prompt used for model cross-validation. Fig. 15 provides the training and inference prompt for all MLLMs. Fig. 16 gives the prompt for judge model evaluation.

You are an **expert video understanding and descriptive question designer**. Your task is to generate **high-quality descriptive open-ended questions** that require describing **what is observable, stated, or directly inferable** from timestamped segments of a video.

Input Format:

You will receive a JSON object containing:

- `title`: overall content focus of the video
- `timestamps`: a list of `[start, end]` time segments
- `descriptions`: a list of corresponding descriptions for each segment

Task Instructions:

1. Generate **1 to 3 open-ended questions**, adjusting the number based on the richness of video content available. Each question must be **descriptive** in nature, asking about **what** is observable, stated, or directly inferable from the descriptions. Focus on: events, actions, objects, people, states, locations, relationships, and summaries. If the video content does not support genuine descriptive questions, output `{"questions": []}` with no additional text.
2. Questions may require integrating information from **one or more timestamp segments**, depending on the descriptive content being tested.
3. Questions must be **self-contained** and clear. **Do not use any words that explicitly or implicitly reference the video's timestamps or segment order.**
4. Base your questions **strictly on the provided descriptions**. Do not introduce outside knowledge, assumptions, or speculative content (e.g., the person's identity, emotions, intentions, or behind-the-scenes details).
5. The question should be **concise, and low in information density**. Do **not** ask temporal reasoning questions that rely on order/sequence (e.g., "what happened after/before/first/next/during", "what is the sequence of"). Questions should **not** test temporal order.

Output Format:

Return a valid JSON array. The array can contain **multiple items** (1 to 3 as appropriate).
Each item follows the structure below; the evidence array can also contain multiple segment entries:

```
[
    {
        "question": "A clear, open-ended question string testing descriptive reasoning.",
```

```
        "reasoning": "A step-by-step explanation (2-4 sentences) showing how evidence supports the answer.",
        "answer": "A concise answer no more than 30 words.",
        "evidence": [
            {"timestamp": [start, end], "description": "string"}
        ]
    }
]
```

Constraints:

- The `timestamp` and `description` fields within each evidence item must be copied **verbatim** from the input, without modification, omission, or addition.
- Evidence may come from **one or more** segments; when multiple segments are used, list them in chronological order.

Now generate descriptive questions for the following video data: {metadata}

Figure 10: Prompt template for descriptive question generation.

You are an **expert video understanding and temporal reasoning question designer**. Your task is to generate **high-quality temporal reasoning open-ended questions** that require understanding **order/sequence, change detection, and process/procedure** across timestamped segments from a video.

Input Format:

You will receive a JSON object containing:

- `title`: overall content focus of the video
- `timestamps`: a list of `[start, end]` time segments
- `descriptions`: a list of corresponding descriptions for each segment

Task Instructions:

1. Generate **1 to 3 open-ended questions**, adjusting the number based on the richness of temporal relationships available. If the video content does not support genuine temporal reasoning, output `{"questions": []}` with no additional text. Each question must be **temporal reasoning** in nature, testing one of the following:
   - **Order/Sequence**: Understanding what happens before/after/during specific events (ask "what happens before/after", "what occurs first/last")
   - **Change Detection**: Recognizing how something changes between different points in time (ask "how does X change", "what difference appears")
   - **Process/Procedure**: Identifying the sequence of steps in a process or workflow (ask "what are the steps", "in what order")
2. Questions must be **self-contained** and clear. **Do not use any words that explicitly or implicitly reference the video's timestamps or segment order.**
3. Base your questions **strictly on the provided descriptions**. Do not introduce outside knowledge, assumptions, or speculative content (e.g., the person's identity, emotions, intentions, or behind-the-scenes details).
4. The question should be **concise, and low in information density**, shouldn't contain information that allows guessing the answer **without watching the video**.
5. **Ensure diversity:** Do not generate duplicate questions that ask the same thing in different wording. Each question should explore a different temporal relationship in the video.

Output Format:

Return a valid JSON array. The array can contain **multiple items** (1 to 3 as appropriate).

Each item follows the structure below; the evidence array can also contain multiple segment entries:

```
[
    {
        "question": "A clear, open-ended question string testing temporal reasoning.",
        "reasoning": "A step-by-step explanation (2-4 sentences) showing how evidence supports the answer.",
        "answer": "A concise answer no more than 30 words.",
        "evidence": [
            {"timestamp": [start, end], "description": "string"}
        ]
    }
]
```

Constraints:

- The `timestamp` and `description` fields within each evidence item must be copied **verbatim** from the input, without modification, omission, or addition.
- Evidence may come from **one or more** segments; when multiple segments are used, list them in chronological order.

Now generate temporal reasoning questions for the following video data: {metadata}

Figure 11: Prompt template for temporal question generation.

You are an **expert video understanding and causal reasoning question designer**. Your task is to generate **high-quality causal reasoning open-ended questions** that require identifying and explaining cause-effect relationships based on factual evidence from timestamped segments of a video.

Input Format:

You will receive a JSON object containing:

- `title`: overall content focus of the video
- `timestamps`: a list of `[start, end]` time segments
- `descriptions`: a list of corresponding descriptions for each segment

Task Instructions:

1. Generate **1 to 3 open-ended questions**, adjusting the number based on the richness of causal connections available. Each question must be **causal** in nature, asking **"why"** or **"how"** something happened based on evidence from the video. If the video content does not support genuine multi-segment reasoning or lacks causal chains, output `{"questions": []}` with no additional text rather than forcing questionable ones.
2. Each question **must**:
   - Require integrating information from **at least two distinct timestamp segments** (non-adjacent segments are encouraged).
   - Be framed as either:
     - **Cause-seeking:** "Why did X happen?", "How did X happen?" or "What caused X?" (X is an outcome)
     - **Effect-seeking:** "What happened as a result of Y?" or "What effect did Y have?" (Y is an action/event)
   - Be **self-contained** and clear. **Do not use any words that explicitly or implicitly reference the video's timestamps or segment order.**
   - **Not** include the answer or obvious causal cues in the question itself. The question should name only the event/outcome of interest, without explaining its cause or consequence.
   - Require the answer to be logically deducible **only** from the causal evidence explicitly shown or described in the video segments.
3. Base your questions **strictly on the provided descriptions**. Do not introduce outside knowledge, assumptions, or speculative content (e.g., the person's identity, emotions, intentions, or behind-the-scenes details). Do not invent any fact, event, or detail not explicitly stated or logically implied.
4. The question should be **concise, abstract and low in information density**:
   - Use generic terms (e.g., "the person", "the individual", "the object") rather than specific labels (e.g., "the gymnast", "the chef") that reveal professional or situational context.
   - Do **not** mention unobserved entities or consequences (e.g., "audience", "judges", "reaction") in the question, as this leaks answer cues.
   - Should **not** contain specific action descriptions or visible states that would give away the answer (e.g., avoid phrases like "using the rope," "bending the knee," or "pulling the row"). Instead, phrase it in terms of the **outcome, role, or purpose** that requires viewing the video to understand. The question should be answerable **only** by synthesizing evidence from multiple segments.
   - **Critical:**
     - Do **not** state the cause when asking about an effect (e.g., avoid "Why did the tree look beautiful after adding lights?" — this gives away the cause).
     - Do **not** state the effect when asking about a cause (e.g., avoid "What did marinating the chicken cause to happen to the final dish?" — this gives away the effect domain).
     - Do **not** include any phrasing that situates the event within a larger sequence as a positional marker (e.g., "before serving", "during the routine", "after adding X", "when making Z"). The question should ask about the event itself.
     - Do **not** phrase the question as a full action that includes the surrounding context. Instead, use a minimal or noun-phrase structure. The question should be short and abstract, requiring the video to reveal what that element is used for and what consequence follows.
5. **Ensure diversity:** Do not generate duplicate questions that ask the same thing in different wording. Each question should explore a different causal relationship in the video.

Output Format:

Return a valid JSON array. The array can contain **multiple items** (1 to 3 as appropriate).

Each item follows the structure below; the evidence array can also contain multiple segment entries:

```
[
    {
        "question": "A clear, open-ended question string testing causal reasoning.",
```

```
        "reasoning": "A step-by-step explanation (2-4 sentences) showing how evidence supports the answer.",
        "answer": "A concise answer no more than 30 words.",
        "evidence": [
            {"timestamp": [start, end], "description": "string"}
        ]
    }
]
```

Constraints:

- The `timestamp` and `description` fields within each evidence item must be copied **verbatim** from the input.
- Evidence may come from **one or more** segments; when multiple segments are used, list them in chronological order.
- **Critical: The `question` field should be as short and abstract as possible. Avoid describing the surrounding scene, the broader task, or the role of the action/event. The video should be required to understand both the context and the causal chain.

Now generate causal reasoning questions for the following video data: {metadata}

Figure 12: Prompt template for causal question generation.

You are an **expert video understanding and counterfactual reasoning question designer**. Your task is to generate **high-quality counterfactual reasoning open-ended questions** that require reasoning about hypothetical scenarios ("what if") based on factual evidence from timestamped segments of a video.

Input Format:

You will receive a JSON object containing:

- `title`: overall content focus of the video
- `timestamps`: a list of `[start, end]` time segments
- `descriptions`: a list of corresponding descriptions for each segment

Task Instructions:

1. **First, evaluate the video content:** Determine if it contains clear causal relationships, dependencies, or alternative possibilities that support meaningful counterfactual reasoning.
   - **Unsuitable scenarios (output {"questions": []}):** Static scenes, purely descriptive monologues without actions, videos lacking sequential logic, or scenarios where altering a step leads to trivial outcomes.
   - If suitable, generate **1 to 3 open-ended questions**, adjusting the number based on the richness of causal connections available. Prefer quality over quantity.
2. **Ensure diversity:** Do not generate duplicate questions. Each question should test a distinct counterfactual scenario or explore a different causal relationship.
3. **Ensure reasoning complexity:** Prioritize questions that require synthesizing information from **at least two distinct timestamped segments** to establish a causal chain.
4. Each question must:
   - Begin with a counterfactual premise ("What if...", "If... not/had...")
   - Alter **one specific event or condition** from the actual video, while keeping all other events unchanged
   - Do **not** include in the question any description of the consequence, downstream step, or final context (e.g., do not say "before adding to the salad", "before serving", "during the routine", "when making the dish"). The question should name only the altered event itself, without situating it within the broader sequence or outcome. The causal relationship must be inferred from the video, not stated in the question.
   - Do **not** include in the question any description that situates the altered event within a larger action or process. The question should name only the core element being altered (e.g., an object, a substance, a tool, or a simple action), not the full action that surrounds it. Avoid phrasing like "decorated with lights" or "placed the star on the tree"; instead use structures like "lights were not used" or "the golden star was not used".
   - Require the answer to be logically deducible from: (a) the altered premise, and (b) the actual events/relationships shown in the video
5. Base your questions **strictly on the provided descriptions**. Do not invent any fact, event, or detail not explicitly stated or logically implied.

Output Format:

Return a valid JSON array. The array can contain **multiple items** (1 to 3 as appropriate).

Each item follows the structure below; the evidence array can also contain multiple segment entries:

```
[
    {
        "question": "A clear, open-ended question string testing counterfactual reasoning.",
        "reasoning": "A step-by-step explanation (2-4 sentences) showing how evidence supports the answer.",
        "answer": "A concise answer no more than 30 words.",
        "evidence": [
            {"timestamp": [start, end], "description": "string"}
        ]
    }
]
```

Constraints:

- The `timestamp` and `description` fields within each evidence item must be copied **verbatim** from the input.
- Evidence may come from **one or more** segments; when multiple segments are used, list them in chronological order.
- **Critical:** The `question` field should be as short and abstract as possible. Avoid describing the surrounding scene, the broader task, or the role of the element being altered. The video should be required to understand both the context and the causal chain.

Now generate counterfactual reasoning questions for the following video data: {metadata}

Figure 13: Prompt template for counterfactual question generation.

You are a strict but fair **video question-answer validator**. Your task is to verify whether a generated question and its answer are reasonable based on the provided video metadata.

Input Format:
- `metadata`: Contains video duration, timestamp segments, and corresponding descriptions for each segment
- `question`: The generated question to evaluate
- `question_type`: Type of question, which belongs to one of the following four types: causal, counterfactual, temporal, descriptive
- `generated_answer`: The generated answer to evaluate
- `generated_evidence`: The supporting temporal evidence generated for the QA pair

Validation Criteria:
1. **Question Reasonableness (0-10 points)**
   - Does the question require watching the video to answer? (Cannot be answered by common sense alone)
   - Is the question relevant to the video content described in metadata?
   - Is the question clear, specific, and answerable based on the metadata?
2. **Answer Correctness (0-10 points)**
   - Is the answer directly supported by the metadata? (No hallucination)
   - **REASONABLE INFERENCES ARE ENCOURAGED**: If the answer makes minimal, logical inferences that are directly implied by the metadata (e.g., "to remove moisture" when metadata says "wipe with dry cloth"), this is acceptable.
   - Does the answer DIRECTLY address the specific question being asked? Not just being relevant or true, but actually answering what was asked (Not off-topic)
   - Is the answer concise and accurate based on available information?
3. **Evidence-Answer Alignment (0-10 points)**
   - Can all claims in the answer be traced back to specific segments?
   - Are there any contradictions between answer and evidence?

Scoring Guide:
- Total Score $\geq$ 25: PASS
- Total Score $<$ 25: FAIL
- **Automatic FAIL only if these critical issues exist:**
  * Answer contains information that directly contradicts metadata
  * Answer invents major events/objects/actions not present in metadata at all
  * Question is completely unrelated to the video content
  * Question can be answered without any reference to the video (pure common sense)

Now validate the following:
Video metadata: {metadata}
Question: "{question}"
Question Type: {question_type}
Generated Answer: "{answer}"
Generated Evidence: "{evidence}"

Then, provide your final judgment in **valid JSON format only** (no additional text before or after):
{{
  "Score": <total_score>,
  "Pass": <true_or_false>
}}

Figure 14: Prompt template for model cross-validation.

You are a video understanding assistant. Given a video and a question, respond in the following format:
First output the seen relevant video segments within <evidence> </evidence> tags in chronological order. Then based on the observed video and evidence, analyze and reason through the question. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in <answer> </answer> tags.

Format: <evidence>Time:MM:SS-MM:SS, Des: description (one or more lines)</evidence>. <think> reasoning based on evidence </think>. <answer> concise direct answer </answer>.

IMPORTANT RULES:
- Output EXACTLY 3 sections: <evidence>, <think>, <answer>
- Evidence: Each line MUST follow: "Time:MM:SS-MM:SS, Des: ..."
- NO repeated descriptions or time segments
- Think: 2-4 sentences maximum, based ONLY on the evidence above
- Answer: 1 short sentence, no explanations

Now answer the following question:
Question: {question}

Figure 15: Prompt template for training and inference.

You are an expert specializing in evaluating whether a respondent's answer after watching a video matches the golden answer. We will provide the video's title, video segments' descriptions, question, golden answer, and the response to be judged below.

Video's Title:
{title}

Video Segments' Descriptions (in chronological order):
{metadata}

Question:
{question}

Golden Answer:
{golden_answer}

Response to be judged:
{model_answer}

Rules:
1. If the response to be judged contains ALL key information of the golden answer or expresses the same meaning using other sentences or synonyms, it is considered a match, and the output is 1.
2. If the response to be judged does NOT contain the key information from the golden answer, it is considered a mismatch, and the output is 0.
3. The response to be judged should NOT contain any content that is contradictory, conflicting, or unreasonable when inferred from the video content description. If such content exists, it is considered a mismatch, and the output is 0.
4. If the response to be judged contains MOST of the key information of the golden answer, and does NOT contain any information that is contradictory, conflicting, or unreasonable when inferred from the video content description, it is considered a partial match, and the output is 0.5.

Instructions:
Follow the format below and do not give any extra outputs:
Score: 0 (if the response does not match)
Score: 0.5 (if the response partially matches)
Score: 1 (if the response matches)

Figure 16: Prompt template for model evaluation.