GROUNDBOTH: SUPPORT–COUNTER GROUNDING FOR VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounded video question answering (VideoQA) asks a model to support its answer with localized temporal evidence. However, a correct answer with accurately grounded evidence does not by itself show that the model understood the video. When a wrong option describes an event that also occurs in the video and resembles the correct one, a model can choose correctly without ever telling the two apart. We extend grounded video question answering to support–counter grounded VideoQA: a model grounds the selected answer and, for each rejected option, grounds the competing event it refers to and explains why that event does not answer the question, making each rejection inspectable rather than taken on trust. To study this task, we build SC-GROUND by refining NExT-GQA, ReXTime, and CG-Bench so that every distractor is anchored to a real event elsewhere in the video, with support and counter annotations. Used as a diagnostic, it reveals that even strong proprietary models ground the selected answer well yet largely miss the events behind the wrong options; grounding the answer alone is far from enough. We therefore curate a support–counter training set and introduce GROUNDBOTH, a 2B model that emits the answer, its support, and per-option counter evidence in a single pass, trained by role-wise on-policy distillation (ROLE-WISE OPD) and support–counter GRPO (SC-GRPO) that rewards both evidence roles. Despite its size, GroundBoth rivals 7B models on the original benchmarks, raising grounded answering by 18.0, 15.5, and 2.7 points over its 2B backbone, and leads all evaluated open-source single-call models on SC-Ground, lifting joint grounding nearly eightfold on NExT-GQA-SC (1.48 to 11.67).

1 INTRODUCTION

Video–language models such as InternVideo2 (Wang et al., 2024a), VideoLLaMA 2 (Cheng et al., 2024), and Qwen-VL (Bai et al., 2025b;a) have advanced video question answering and long-video understanding. Yet a correct answer alone does not establish which visual events justify it. Grounded video question answering therefore couples answer prediction with temporal evidence localization, as studied in NExT-GQA (Xiao et al., 2024), ReXTime (Chen et al., 2024b), and CG-Bench (Chen et al., 2024a). This makes the evidence for the selected answer inspectable.

Existing methods improve grounding through temporal modeling, evidence verification, and reinforcement learning. Grounded-VideoLLM (Wang et al., 2025a) learns fine-grained temporal representations, while VideoMind (Liu et al., 2026) coordinates grounding, verification, and answering. TVG-R1 (Chen et al., 2025) improves temporal localization through a grounding-specific cold start and reinforcement learning. VideoChat-R1 (Li et al., 2025) jointly trains grounding and answering with correctness and temporal-overlap rewards. LOVER (Chen et al., 2026) further improves answer-support evidence spotting through an evidence-duration curriculum and an IoP-based reward. Their grounding targets nevertheless center on the queried event or selected answer, without requiring option-conditioned locations for rejected alternatives.

This omission matters even when both the selected answer and its support are correct. When a wrong option describes a similar event elsewhere in the video, support-only grounding does not show whether the model distinguished the two. Figure 1 illustrates this gap. Asked how to prepare the roof for putting cement, TVG-R1 selects pressure washing (B) and localizes it. Air-blowing (A) and membrane spraying (C) also appear in the video, but both happen after cementing. Even when

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: **Grounding the selected answer does not guarantee grounded rejection of plausible alternatives.** TVG-R1 selects pressure washing (B) and localizes its support, but even when asked for counter evidence, it largely misses air blowing (A) and only partly locates membrane spraying (C). GROUNDBOTH localizes B before cementing and both competing events afterward, explaining why A and C do not answer the question.

explicitly asked for counter evidence, TVG-R1 fails to locate the air-blowing and only partly locates the membrane spraying. Its localized support does not establish why the alternatives are wrong. We refer to the segments associated with such competing options as *counter evidence*. Localizing these competing events makes each rejection inspectable rather than taken on trust. A grounded response should therefore identify both the selected answer's support and the events associated with rejected options.

We formulate support–counter grounded VideoQA and construct SC-GROUND for its evaluation. Starting from the NExT-GQA test set, ReXTime validation set, and CG-Bench mini, we retain the original videos and questions and preserve the correct-answer semantics. We replace distractors with plausible alternatives anchored to other events in the same video and pair each with a counter interval and rejection rationale. The three refined sets, NEXT-GQA-SC, REXTIME-SC, and CG-BENCH-SC, carry both support and counter annotations. Our evaluation jointly measures answer correctness, support localization, and option-matched counter localization. It reveals a gap between support and counter grounding even among strong proprietary models.

For training, we build a support–counter training set from multiple-choice video questions in the Video-R1 (Feng et al., 2025) training collection. We augment the questions with support timestamps, option-conditioned counter timestamps where a corresponding event is present, event descriptions, and rejection analyses, providing supervision for both answer support and alternative rejection.

Our model, GROUNDBOTH, generates these structured outputs in a single pass. It learns through ROLE-WISE OPD, which combines annotation-based supervision for structure and timestamps with on-policy teacher guidance for descriptions and reasoning, followed by SC-GRPO, which jointly rewards answering and both evidence roles. Our contributions are threefold.

- We formulate support–counter grounded VideoQA, requiring localization of answer support and option-conditioned competing events, and build SC-GROUND to evaluate it.
- We curate support–counter training data and propose GROUNDBOTH, which pairs a role-wise cold start (ROLE-WISE OPD) with reinforcement learning that rewards the answer and both evidence roles (SC-GRPO).
- Our 2B model improves grounded-answering scores over Qwen3-VL-2B by 18.0, 15.5, and 2.7 points on NExT-GQA, ReXTime, and CG-Bench, respectively. Its results on these source benchmarks are comparable to those of the listed 7B models. On SC-GROUND, it leads the evaluated open-source single-call models and raises the joint grounding score on NEXT-GQA-SC nearly eightfold over its backbone, from 1.48 to 11.67.

2 Related Work

Video understanding and temporal representations. Video foundation models and multimodal language models increasingly combine visual recognition with language-based reasoning. InternVideo2 (Wang et al., 2024a) scales multimodal video pretraining, VideoLLaMA 2 (Cheng et al., 2024) improves spatial–temporal modeling, and the Qwen-VL family (Bai et al., 2025b;a) supports increasingly long video contexts and temporal understanding. More specialized models make time explicit: TimeChat (Ren et al., 2024) binds frames to timestamps, VTimeLLM (Huang et al., 2023) uses boundary-aware instruction tuning, and LITA (Huang et al., 2024) introduces time tokens and temporally detailed visual representations. Grounded-VideoLLM (Wang et al., 2025a) further develops fine-grained temporal grounding, while TimeLens (Zhang et al., 2026) revisits it with re-annotated benchmarks and training recipes. These advances address how video events are represented and localized. Our task additionally specifies which localized events support an answer and which correspond to its competitors.

Grounded answering and evidence retrieval. NExT-GQA (Xiao et al., 2024) formalizes the distinction between answering correctly and answering with localized visual support. Question guidance and answer calibration (Xu et al., 2024) improve evidence selection within this setting. ReXTime (Chen et al., 2024b) and CG-Bench (Chen et al., 2024a) emphasize cross-time reasoning and long-video clues, respectively. VideoZeroBench (Meng et al., 2026) verifies spatial and temporal evidence, EG-VQA (Huang et al., 2026) evaluates answer-supporting intervals, and E-VQA (Wang et al., 2026b) adds tracked object masks. VideoChat-TPO (Yan et al., 2025) aligns multimodal models with explicit vision tasks. VideoMind (Liu et al., 2026) uses a Chain-of-LoRA agent to plan, ground, verify, and answer, whereas Divid (Tang et al., 2026) disentangles spatial and temporal modeling inside the decoder. Retrieval, verification, and better temporal representations improve evidence for a queried event or answer. Our distinction is in the required output: support and rejected-option anchors are jointly annotated and evaluated. A grounder can be queried separately for each option, as in our multi-call baselines, but this adaptation differs from learning their roles together.

Reinforcement learning and on-policy distillation for grounding. GRPO (Shao et al., 2024) post-trains language models with verifiable rewards. Video-R1 (Feng et al., 2025) applies it to video reasoning, Time-R1 (Wang et al., 2025b) and TVG-R1 (Chen et al., 2025) add timestamp-aware overlap rewards, data curation, and grounding-specific cold starts, VideoChat-R1 (Li et al., 2025) combines answer and overlap rewards for grounded QA, and LOVER (Chen et al., 2026) adapts the reward and curriculum to answer-support evidence. On-policy distillation (Agarwal et al., 2024) instead supervises the student's own trajectories with dense token-level teacher signals, and recent work brings it to video. Video-OPD (Li et al., 2026) replaces sparse grounding rewards with reverse-KL supervision from a teacher, and Clue-OPSD (Wang et al., 2026a) distills a long-video student from a self-teacher that sees only the annotated clue intervals. All of these methods reward or distill the localization of evidence for the queried event or answer. SC-GRPO adds option-conditioned counter rewards to GRPO. Role-wise OPD uses annotations for structure and timestamps, and teacher distillation for semantic content and option decisions.

Negative context and competing evidence. Negative video content has been used as a nuisance to suppress or as a context to search. Invariant Grounding for VideoQA (Li et al., 2022) intervenes on question-irrelevant scenes to reduce spurious correlations, and HawkEye's InternVid-G (Wang et al., 2024b) defines a negative span as the retrieval context in which the positive span must be found, excluding highly similar events. Our counter evidence is different in kind. It localizes a video event described by a rejected option, even though that event does not satisfy the question's specific condition. Counter grounding therefore assigns a temporal target to a rejected option rather than mining background negatives.

3 METHODS

3.1 SUPPORT–COUNTER GROUNDED VIDEOQA

Given a video V, a question q, and answer options $\mathcal{A}$, our task requires a model to predict an answer and localize evidence for both its selection and the rejection of competing options. We denote the input by $x = (V, q, \mathcal{A})$ and the correct answer by a^*. The annotations comprise support intervals $\mathcal{S}$ for a^* and option-conditioned counter intervals $\mathcal{C}_a$ for each alternative $a \neq a^*$. The model predicts

$$\hat{y} = \big(\hat{a}, \hat{\mathcal{S}}, \{(a, \hat{\mathcal{C}}_a)\}_{a \in \mathcal{A} \setminus \{\hat{a}\}}\big). \tag{1}$$

Support intervals contain evidence that establishes a^* as the answer to q. Counter intervals localize the competing event associated with a rejected option. Such an event may occur in the video yet fail to answer the question because of its timing, the entity involved, or another contextual condition. The distinction between the two roles therefore depends on the event's relation to the question, rather than on whether the event occurs at all. For an alternative annotated as having no corresponding event, we set $\mathcal{C}_a = \varnothing$.

The structured response (Figure 2, left) supplements these predictions with event descriptions and analyses that explain how the localized evidence supports or rules out each option. An absent-event alternative receives an analysis without a temporal interval. Appendix F shows two examples of support–counter grounding.

3.2 BENCHMARK REFINEMENT AND TRAINING DATASET

Constructing SC-GROUND. We refine the test split of NExT-GQA (Xiao et al., 2024), the validation split of ReXTime (Chen et al., 2024b), and the mini split of CG-Bench (Chen et al., 2024a) to provide augmented competing alternatives of wrong option paired with temporal annotations. Specifically, we retain the videos, questions, and correct-answer semantics, but replace the distractors with plausible alternatives anchored to other events in the same video. GPT-5.6 first generates a timestamped event record for each video. From this record, it proposes candidate distractors, each paired with a counter interval and a rejection rationale. Gemini-3.1-Pro checks the candidates against the record, testing whether their events align with the proposed intervals but do not answer the question in the full-record context. After distractor selection and option restyling, a final check retains questions for which the correct answer remains unique. The resulting subsets form SC-GROUND; the full construction procedure is given in Appendix A. Because both distractors and temporal annotations are refined, we evaluate the source benchmarks separately under their original protocols.

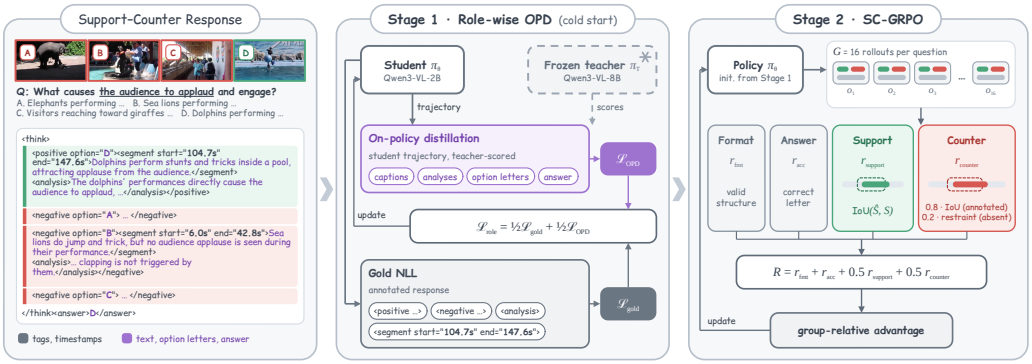
Training dataset. For training, we retain the original multiple-choice questions and options from the Video-R1 (Feng et al., 2025) training dataset. We annotate support intervals for correct answers and counter intervals only for distractors that match visible events. This yields 7,648 cold-start examples and a separate set of 2,400 examples for reinforcement learning. Since Video-R1 includes NExT-QA and ActivityNet videos, which also supply NExT-GQA and ReXTime, we carefully exclude training examples with potential video overlap with the NExT-GQA test or ReXTime validation split. Unlike SC-GROUND, which uses temporally anchored distractors, some original Video-R1 distractors describe events absent from the video. A concrete detail is shown in (Appendix A).

3.3 GROUNDBOTH: ROLE-WISE COLD START AND SUPPORT–COUNTER GRPO

GROUNDBOTH uses a two-stage training procedure (Figure 2). ROLE-WISE OPD initializes the policy by assigning different learning objectives to the structural and semantic roles of its response. SC-GRPO subsequently optimizes answer correctness and the localization of both evidence roles.

3.3.1 ROLE-WISE ON-POLICY DISTILLATION

A structured grounding response combines fields with different supervision requirements. Tags specify its structure, and timestamps have explicit annotation targets. Event descriptions and analyses, however, admit multiple valid formulations. We therefore apply supervised fine-tuning


Support-Counter Response
Q: What causes the audience to applaud and engage?
A. Elephants performing ... B. Sea lions performing ...
C. Visitors reaching toward giraffes ... D. Dolphins performing ...
<think>
<positive option="D">Dolphins perform stunts and tricks inside a pool, attracting applause from the audience.
<analysis>The dolphins' performances directly cause the audience to applaud, ...</analysis></positive>
<negative option="A"> ... </negative>
<negative option="B">Sea lions do jump and trick, but no audience applause is seen during their performance.
<analysis>... clapping is not triggered by them.</analysis></negative>
<negative option="C"> ... </negative>
</think><answer>D</answer>
tags, timestamps
text, option letters, answer
Stage 1 · Role-wise OPD (cold start)
Student π_θ
Qwen3-VL-2B
Frozen teacher π_T
Qwen3-VL-8B
trajectory
scores
On-policy distillation
student trajectory, teacher-scored
captions
analyses
option letters
answer
$\mathcal{L}_{OPD}$
update
$\mathcal{L}_{role} = ½\mathcal{L}_{gold} + ½\mathcal{L}_{OPD}$
Gold NLL
annotated response
<positive ...>
<negative ...>
<analysis>

$\mathcal{L}_{gold}$
Stage 2 · SC-GRPO
Policy π_θ
init. from Stage 1
G = 16 rollouts per question
o_1
o_2
o_3
...
o_n
Format
r_{fmt}
valid structure
Answer
r_{acc}
correct letter
Support
$r_{support}$
IoU($\hat{S}$, S)
Counter
$r_{counter}$
0.8 · IoU (annotated)
0.2 · restraint (absent)
$R = r_{fmt} + r_{acc} + 0.5\, r_{support} + 0.5\, r_{counter}$
update
group-relative advantage

Figure 2: Overview of GROUNDBOTH. ROLE-WISE OPD combines supervised learning of tags and timestamps with on-policy distillation of descriptions, analyses, and option predictions from a frozen teacher. SC-GRPO then optimizes answer correctness and support–counter grounding with structured rewards.

to tags and timestamps, while using on-policy distillation to guide the student's semantic content on its own generated trajectories. Specifically, for an annotated response y^* and a student-generated response $\tilde{y}$, we assign token roles to two groups: $\mathcal{R}_{\text{gold}} = \{\text{tags}, \text{timestamps}\}$ and $\mathcal{R}_{\text{OPD}} = \{\text{captions}, \text{analyses}, \text{option letters}, \text{answer}\}$. Option letters and the final answer are distilled together with the analyses that determine these predictions. The frozen teacher conditions on the same video–question prompt and the student's generated prefix; its token probabilities supervise only the roles in $\mathcal{R}_{\text{OPD}}$.

For annotation-supervised roles, the student policy π_θ minimizes the negative log-likelihood on the corresponding tokens of y^*:

$$\mathcal{L}_{\text{gold}} = -\left\langle \log \pi_\theta(y_t^* \mid x, y_{<t}^*) \right\rangle_{\mathcal{R}_{\text{gold}}}. \tag{2}$$

Here, $\langle\cdot\rangle_{\mathcal{R}}$ averages token losses within each example and role, then across examples and active roles. This normalization balances the contributions of roles with different sequence lengths.

For distilled roles, we sample $\tilde{y}$ from the behavior policy π_b and optimize a sampled reverse-KL surrogate:

$$\rho_t = \frac{\pi_\theta(\tilde{y}_t \mid x, \tilde{y}_{<t})}{\pi_b(\tilde{y}_t \mid x, \tilde{y}_{<t})}, \quad \mathcal{L}_{\text{OPD}} = \left\langle \rho_t \, \text{sg}[\log \pi_\theta(\tilde{y}_t \mid x, \tilde{y}_{<t}) - \log \pi_T(\tilde{y}_t \mid x, \tilde{y}_{<t})] \right\rangle_{\mathcal{R}_{\text{OPD}}}, \tag{3}$$

where π_T is the frozen teacher, ρ_t is the token-level importance ratio, and sg denotes stop-gradient. The two branches are combined with equal weights:

$$\mathcal{L}_{\text{role}} = \tfrac{1}{2}\mathcal{L}_{\text{gold}} + \tfrac{1}{2}\mathcal{L}_{\text{OPD}}. \tag{4}$$

The role assignment, rather than a uniform mixture of losses over all tokens, anchors structure and temporal boundaries to the annotations while allowing teacher-guided learning of student-generated semantic content.

3.3.2 SUPPORT–COUNTER GRPO

Starting from the ROLE-WISE OPD policy, SC-GRPO jointly rewards answer correctness, support localization, and option-conditioned counter grounding. For each sampled response, the total reward is

$$R = r_{\text{fmt}} + r_{\text{acc}} + 0.5 r_{\text{support}} + 0.5 r_{\text{counter}}. \tag{5}$$

Answer and support rewards. The binary format reward r_{fmt} requires a valid response envelope and a support block whose option letter agrees with the final answer. The binary answer reward r_{acc} checks whether the final answer matches a^*; any support block present must agree with that answer. The support reward is the temporal set IoU between predicted and annotated intervals, $r_{\text{support}} = \text{IoU}(\hat{\mathcal{S}}, \mathcal{S})$. Set IoU measures the intersection duration divided by the union duration after merging overlapping intervals within each set.

Counter localization and restraint. The counter reward distinguishes two types of alternatives: those with annotated temporal evidence, $\mathcal{A}_{\text{loc}}$, and those annotated as having no corresponding event, $\mathcal{A}_{\text{abs}}$ (Section 3.2). For $a \in \mathcal{A}_{\text{loc}}$, we score the prediction by $\text{IoU}(\hat{\mathcal{C}}_a, \mathcal{C}_a)$, assigning 0 if no interval is predicted. For $a \in \mathcal{A}_{\text{abs}}$, the desired output is an explicit counter block without an interval. We use the following restraint reward:

$$r_{\text{restraint}}(a) = \begin{cases} 1, & \hat{\mathcal{C}}_a = \varnothing, \\ 0, & \hat{\mathcal{C}}_a = \perp \text{ or } \hat{\mathcal{C}}_a \approx \mathcal{S}, \\ \frac{1}{2}, & \text{otherwise}, \end{cases} \tag{6}$$

Here, $\varnothing$ denotes an explicit block without an interval, whereas $\perp$ denotes an omitted block. We write $\hat{\mathcal{C}}_a \approx \mathcal{S}$ when any predicted counter interval overlaps annotated support by more than half of its own duration, treating this as support copying. The intermediate value $\frac{1}{2}$ imposes a softer penalty on non-copying but unsupported localization than on omission or support copying, while assigning the highest reward to explicit abstention from localization. The copying penalty applies only to $\mathcal{A}_{\text{abs}}$; annotated counter intervals may overlap support and are evaluated against their own temporal targets.

When both groups are nonempty, their scores are averaged separately and combined as

$$r_{\text{counter}} = (1-\lambda)\,\frac{\sum_{a\in\mathcal{A}_{\text{loc}}} \text{IoU}(\hat{\mathcal{C}}_a, \mathcal{C}_a)}{|\mathcal{A}_{\text{loc}}|} + \lambda\,\frac{\sum_{a\in\mathcal{A}_{\text{abs}}} r_{\text{restraint}}(a)}{|\mathcal{A}_{\text{abs}}|}, \tag{7}$$

where $\lambda \in [0,1]$ balances localization and restraint, with its value given in Section 4.1. Averaging the two groups separately makes their relative weighting independent of the number of alternatives in each. If only one group is present, its mean alone defines r_{counter}.

We optimize this reward within the GRPO (Shao et al., 2024) framework, following the advantage estimator of Dr. GRPO (Liu et al., 2025) and asymmetric clipping of DAPO (Yu et al., 2026); optimization details are provided in Appendix C. During training and inference, GROUNDBOTH generates the answer, support evidence, and option-conditioned counter evidence in a single response.

4 EXPERIMENTS

Our experiments address 3 questions. Section 4.2 first asks whether existing video models can ground both support and counter evidence, and how far GROUNDBOTH advances this ability on SC-GROUND. It then asks whether GROUNDBOTH retains its answering and support grounding on the original benchmarks, which annotate support evidence only. Section 4.3 asks which parts of the training account for the gains, by ablating the cold start and the reward composition and by examining how answering and grounding relate.

4.1 IMPLEMENTATION DETAILS

Training and decoding. We initialize the 2B policy from Qwen3-VL-2B-Instruct (Bai et al., 2025a) and use a frozen Qwen3-VL-8B-Instruct teacher. ROLE-WISE OPD is trained for 800 updates with a batch size of 8 and a learning rate of 5×10^{-7}, sampling 1 trajectory per prompt at temperature 0.7. SC-GRPO is then trained for 1 epoch with a batch size of 8, a learning rate of 10^{-6}, and $G = 16$ rollouts per question at temperature 1.0. Its reward weights are $(1, 1, 0.5, 0.5)$ for format, answer, support, and counter, respectively (Eq. 5); the counter-reward mixing coefficient is $\lambda = 0.2$ (Eq. 7). In the reduced-reward ablations, retained components keep these weights. During training and evaluation, each input video is uniformly sampled to 64 frames, with a maximum input resolution of $144 \times 32 \times 32$ pixels. The generation limit is 2,048 tokens during training and 8,192 tokens at evaluation, where the base model and GROUNDBOTH decode greedily. Additional optimizer and output-format details are given in Appendix C.

Benchmarks. Our primary evaluation uses the three refined benchmarks of SC-GROUND, namely NEXT-GQA-SC, REXTIME-SC, and CG-BENCH-SC (Section 3.2), on which every wrong option is anchored to a counter interval. To verify that GROUNDBOTH retains answering and support grounding, we also evaluate on the original NExT-GQA test, ReXTime validation, and CG-Bench mini splits.

Table 1: **Support–counter grounding on SC-GROUND.** In our multi-call adaptation, one answer call is followed by one grounding call per option ($M + 1$ calls for M options); single-call models return the answer and all evidence in one call. Shaded columns show the joint grounding score (Acc@IoU$^{\pm}$, Eq. 8). It gives partial credit for localized counter options after the answer and support are correct. Bold/underline indicate the best/second-best open-source single-call scores. Proprietary models follow the single-call protocol. GROUNDBOTH attains the highest joint grounding score among the evaluated open-source single-call models on all three refined benchmarks.

Method	Size	NEXT-GQA-SC				REXTIME-SC				CG-BENCH-SC			
		Acc	mIoU^{+}	mIoU^{-}	Acc@IoU$^{\pm}$	Acc	mIoU^{+}	mIoU^{-}	Acc@IoU$^{\pm}$	Acc	mIoU^{+}	mIoU^{-}	Acc@IoU$^{\pm}$
Multi-call: one answer call + one grounding call per option													
VideoMind	7B	57.3	24.0	22.1	10.99	74.1	29.6	21.1	18.13	31.5	4.1	4.1	0.34
VideoMind	2B	47.4	21.5	19.1	7.34	63.4	25.1	20.7	14.01	26.8	3.2	3.4	0.19
Divid	2B	57.4	26.0	20.1	10.20	70.3	29.8	24.9	19.79	37.7	5.9	4.3	0.68
Single-call: answer and support–counter evidence in one response													
Gemini-3.1-pro	–	87.6	37.2	15.4	16.08	88.6	40.9	22.1	24.77	71.7	18.7	6.9	5.12
Gemini-3.6-flash	–	86.3	37.0	13.3	13.61	91.2	41.6	20.2	23.34	72.2	25.7	6.9	6.46
GPT-5.6-sol	–	91.5	36.9	17.5	18.25	92.7	40.5	23.9	26.98	74.5	26.7	9.6	8.13
GPT-5-mini	–	79.1	27.0	9.1	8.09	87.7	29.8	15.7	16.95	57.8	8.9	3.6	1.62
TVG-R1	7B	53.7	17.0	9.1	4.82	65.6	22.4	14.8	8.31	35.9	6.9	1.7	0.11
VideoChat-R1	7B	**61.0**	40.3	14.0	11.28	**70.3**	33.2	16.1	14.53	35.0	5.0	1.1	0.05
Qwen3-VL (base)	2B	51.9	8.0	2.4	1.48	63.4	11.3	3.4	2.69	37.2	4.0	0.6	0.14
GROUNDBOTH (ours)	2B	55.3	**41.5**	**17.1**	**11.67**	64.4	**33.9**	**23.6**	**18.46**	**40.5**	**13.7**	**3.2**	**0.93**

Baselines. On SC-GROUND, we compare GROUNDBOTH with its base model Qwen3-VL, with TVG-R1 (Chen et al., 2025) and VideoChat-R1 (Li et al., 2025), with VideoMind (Liu et al., 2026) and Divid (Tang et al., 2026), and with the proprietary Gemini-3.1-Pro, Gemini-3.6-Flash, GPT-5.6, and GPT-5-mini. Qwen3-VL and the proprietary models receive the same support–counter prompt as GROUNDBOTH and generate the answer and both evidence roles in a single call under the same setting. For TVG-R1 and VideoChat-R1, we use the official code and weights and append a counter-evidence request to their native prompts. Our adaptation of VideoMind and Divid uses one whole-video answer call and one grounding call per option: $M+1$ calls for M options. The chosen option's interval provides support, and the remaining intervals provide counter evidence.

Metrics. On SC-GROUND, Acc measures answer accuracy, mIoU^{+} averages support set IoU over questions, and mIoU^{-} averages counter IoU over annotated question–alternative pairs. Each counter prediction is compared with its own option's anchor; missing intervals receive zero IoU. For question n, let $\mathcal{D}_n$ be its set of annotated counter options and $K_n = |\mathcal{D}_n|$. Define the answer–support gate $g_n(\tau) = \mathbf{1}[\hat{a}_n = a_n^*]\,\mathbf{1}[\mathrm{IoU}(\hat{\mathcal{S}}_n, \mathcal{S}_n) \geq \tau]$ and each counter hit $h_{na}(\tau) = \mathbf{1}[\mathrm{IoU}(\hat{\mathcal{C}}_{na}, \mathcal{C}_{na}) \geq \tau]$. We define the *joint grounding score* as

$$\mathrm{Acc@IoU}^{\pm}(\tau) = \frac{1}{N}\sum_{n=1}^{N} g_n(\tau)\frac{1}{K_n}\sum\nolimits_{a\in\mathcal{D}_n} h_{na}(\tau). \tag{8}$$

Here N counts evaluated questions and $\pm$ denotes the two evidence roles. A question receives credit only when its answer and support are correct at threshold τ; its credit is then the fraction of counter options localized at the same threshold. For each question, counter intervals are matched separately within options. On all three refined benchmarks, we report the mean over $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. On the original benchmarks, we retain the official metrics. NExT-GQA reports Acc@GQA, mIoP, and mIoU (Xiao et al., 2024), where Acc@GQA gates answer accuracy by a support IoP of at least 0.5. ReXTime reports mIoU and Acc@IoU (Chen et al., 2024b) with an IoU threshold of 0.5. CG-Bench reports A@IoU, mIoU, and Rec@IoU (Chen et al., 2024a), averaged over thresholds from 0.1 to 0.5. All of these score support evidence only and are therefore distinct from $\mathrm{Acc@IoU}^{\pm}$ on the refined sets.

Table 2: **Results on the original grounded question-answer benchmarks.** † marks baseline results reproduced by us, including the missing Acc@GQA and mIoP values. Bold/underline mark the best/second-best scores.

(a) NExT-GQA

Method (size)	Acc@GQA	mIoP	mIoU
VIOLETv2	12.8	23.6	–
SeViLA (4B)	16.6	29.5	21.7
LangRepo (8×7B)	17.1	31.3	18.5
VideoStreaming (8.3B)	17.8	32.2	19.3
LLoVi (1.8T)	24.3	37.3	20.0
HawkEye (7B)	23.5	33.4	25.7
Grounded-VLLM (4B)	26.7	34.5	21.1
VideoChat-TPO (7B)	25.5	35.6	27.7
VideoMind (2B)	25.2	36.4	28.6
VideoMind (7B)	28.2	39.0	31.4
Divid (2B)	26.4	38.9	32.9
Divid (7B)	<u>29.2</u>	40.8	34.5
LOVER (Qwen3-VL) (4B)	27.5	36.2	20.4
TVG-R1† (7B)	21.1	33.5	30.0
VideoChat-R1† (7B)	26.6	<u>41.0</u>	**35.8**
Qwen3-VL† (2B)	15.5	21.2	8.8
GROUNDBOTH (2B)	**33.5**	**45.1**	<u>35.5</u>

(b) CG-Bench

Method (size)	A@IoU	mIoU	Rec@IoU
GPT-4o	**4.38**	5.62	8.30
GPT-4o-mini	2.21	3.75	5.18
Gemini-1.5 Pro	2.53	3.95	5.81
Claude-3.5 Sonnet	2.79	3.99	5.67
Qwen2-VL (72B)	3.31	3.58	5.32
VITA (8×7B)	2.06	3.06	3.53
Video-CCAM (14B)	1.83	2.63	3.48
ST-LLM (7B)	1.13	2.23	2.86
MiniCPM-v2.6 (8B)	1.04	2.35	2.61
LongVA (7B)	1.78	2.94	3.86
InternVL2 (7B)	2.64	3.91	5.05
Divid (2B)	2.37	4.77	6.40
Divid (7B)	4.11	<u>5.74</u>	<u>8.36</u>
Qwen3-VL† (2B)	1.62	2.52	3.61
GROUNDBOTH (2B)	<u>4.30</u>	**7.16**	**9.91**

(c) ReXTime val

Method (size)	mIoU	Acc@IoU
VTimeLLM (FT) (7B)	29.9	17.1
TimeChat (FT) (7B)	26.3	10.9
VTimeLLM (7B)	20.1	–
TimeChat (7B)	11.7	–
LITA (13B)	21.5	–
VideoMind (2B)	24.8	17.3
VideoMind (7B)	27.6	20.2
Divid (2B)	29.7	18.6
Divid (7B)	**35.8**	<u>22.3</u>
Qwen3-VL† (2B)	15.0	9.88
GROUNDBOTH (2B)	<u>35.5</u>	**25.4**

Table 3: **Initialization (a), reward (b), and aggregation (c) ablations.** Joint columns follow Table 1. Bold: block-wise best scores.

Configuration	NEXT-GQA-SC				REXTIME-SC				CG-BENCH-SC			
	Acc	mIoU$^+$	mIoU$^-$	Acc@IoU$^\pm$	Acc	mIoU$^+$	mIoU$^-$	Acc@IoU$^\pm$	Acc	mIoU$^+$	mIoU$^-$	Acc@IoU$^\pm$
Qwen3-VL (base)	51.9	8.0	2.4	1.48	63.4	11.3	3.4	2.69	37.2	4.0	0.6	0.14
(a) Cold-start initialization; full Stage-2 reward												
SFT	48.2	37.5	15.5	9.66	62.1	33.7	22.1	17.67	39.0	12.1	2.6	0.71
OPD	53.8	34.8	14.8	10.74	**67.2**	32.7	21.6	17.85	39.5	13.6	**3.2**	0.56
Role-wise OPD	**55.3**	**41.5**	**17.1**	**11.67**	64.4	**33.9**	**23.6**	**18.46**	**40.5**	**13.7**	**3.2**	**0.93**
(b) Reward components; Role-wise OPD initialization												
Format + answer	**58.7**	29.8	13.8	6.60	**69.7**	30.7	20.6	13.21	**42.6**	10.6	2.0	0.33
+ support	54.4	38.8	12.0	5.93	67.2	**35.6**	22.9	13.73	40.8	13.5	2.2	0.45
+ support + counter	55.3	**41.5**	**17.1**	**11.67**	64.4	33.9	**23.6**	**18.46**	40.5	**13.7**	**3.2**	**0.93**
(c) Counter-reward aggregation; identical per-option scores												
All-option avg.	**56.0**	41.4	12.7	7.83	**70.0**	**35.3**	22.3	18.13	**41.7**	12.3	2.1	0.66
Separate (ours)	55.3	**41.5**	**17.1**	**11.67**	64.4	33.9	**23.6**	**18.46**	40.5	**13.7**	**3.2**	**0.93**

4.2 GROUNDED TASK EVALUATION

Support–counter grounding. Table 1 reports the primary evaluation on SC-GROUND. GROUNDBOTH attains Acc@IoU$^\pm$ of 11.67, 18.46, and 0.93 on NEXT-GQA-SC, REXTIME-SC, and CG-BENCH-SC, respectively, compared with 1.48, 2.69, and 0.14 for its base model. These are absolute gains of 10.19, 15.77, and 0.80 points computed before rounding. Among the evaluated open-source single-call models, GROUNDBOTH has the highest joint score on all three sets, including 11.67 *vs.* 11.28 for VideoChat-R1 on NEXT-GQA-SC. Proprietary models remain ahead; GPT-5.6-sol reaches 18.25, 26.98, and 8.13.

On NEXT-GQA-SC, answer accuracy rises from 51.9 to 55.3, while mIoU$^+$ rises from 8.0 to 41.5 and mIoU$^-$ from 2.4 to 17.1. Conversely, GPT-5-mini answers more NEXT-GQA-SC questions correctly than GROUNDBOTH (79.1 *vs.* 55.3) but has a lower joint score (8.09 *vs.* 11.67). These contrasts show why answer accuracy alone does not establish whether a model grounds the selected answer and its alternatives on the same question.

VideoMind and Divid use a separate answer call and one grounding call per option. On the four-option REXTIME-SC, this amounts to five calls per question, while GROUNDBOTH produces all outputs in one response. Divid scores higher under this multi-call protocol (19.79 *vs.* 18.46). On

NExT-GQA-SC, VideoMind-7B also has higher mIoU^- than GROUNDBOTH (22.1 *vs.* 17.1), but a lower joint score (10.99 *vs.* 11.67), which additionally requires a correct answer and localized support.

CG-Bench is particularly challenging: its source videos last about 25 minutes at the median, retained questions have four to eight options, and the official clue intervals last only 10 seconds at the median. Even in the refined set, support spans cover just 1.9% of video duration at the median, consistent with the low joint scores across methods.

Source-benchmark evaluation. Table 2 evaluates the original questions and support annotations under the source benchmarks' official metrics. Relative to the same 2B base model, GROUNDBOTH raises NExT-GQA Acc@GQA from 15.5 to 33.5, ReXTime Acc@IoU from 9.88 to 25.4, and CG-Bench A@IoU from 1.62 to 4.30; mIoU also rises on all three. Thus, the grounding gains extend beyond the revised options used in SC-GROUND.

Among the listed results, GROUNDBOTH has the highest Acc@GQA and mIoP on NExT-GQA, with mIoU slightly below VideoChat-R1 (35.5 *vs.* 35.8). On ReXTime, it has the highest Acc@IoU (25.4) and nearly matches Divid-7B in mIoU (35.5 *vs.* 35.8). On CG-Bench, it has the highest mIoU and Rec@IoU, while its A@IoU is close to GPT-4o's (4.30 *vs.* 4.38).

4.3 ABLATION STUDY

Table 3 examines (a) cold-start initialization, (b) reward components, and (c) counter-reward aggregation. All trained rows are evaluated after stage-2 optimization.

Cold start. Block (a) compares SFT, vanilla OPD, and ROLE-WISE OPD as the cold start, each followed by SC-GRPO with the full reward. Relative to vanilla OPD, ROLE-WISE OPD raises $\mathrm{Acc@IoU}^{\pm}$ from 10.74 to 11.67 on NEXT-GQA-SC, from 17.85 to 18.46 on REXTIME-SC, and from 0.56 to 0.93 on CG-BENCH-SC. It also improves mIoU^+ on all three and mIoU^- on the first two. On REXTIME-SC, this higher joint score comes despite lower answer accuracy (64.4 *vs.* 67.2), so the difference is not explained by answer selection alone. SFT yields joint scores of 9.66, 17.67, and 0.71, below those of ROLE-WISE OPD on all three benchmarks.

Reward composition. Block (b) starts with format and answer rewards and adds support and counter rewards in turn. Format and answer rewards alone yield the highest answer accuracy (58.7, 69.7, and 42.6), but joint scores of only 6.60, 13.21, and 0.33. Adding the support reward improves mIoU^+ on every set, yet its effect on the joint score is mixed: 6.60 to 5.93 on NEXT-GQA-SC, 13.21 to 13.73 on REXTIME-SC, and 0.33 to 0.45 on CG-BENCH-SC. With the counter reward, joint scores rise to 11.67, 18.46, and 0.93, and mIoU^- improves on every set. These results show the value of explicit counter localization for the joint task, alongside a reduction in answer accuracy relative to the format-and-answer configuration.

Counter-reward aggregation. Block (c) compares an *all-option average* with *separate aggregation* of localization and restraint (Eq. 7), keeping per-option scores unchanged. Separate aggregation fixes the relative weight of the two groups regardless of how many options fall in each. It raises mIoU^- on all three sets and $\mathrm{Acc@IoU}^{\pm}$ from 7.83 to 11.67 on NEXT-GQA-SC, from 18.13 to 18.46 on REXTIME-SC, and from 0.66 to 0.93 on CG-BENCH-SC. The joint-score improvement on REXTIME-SC occurs despite lower answer accuracy (64.4 *vs.* 70.0), further separating the grounding effect from answer selection.

5 CONCLUSION

We introduced support–counter grounded VideoQA and three refined benchmarks that evaluate the localization of answer support and competing events. GROUNDBOTH improves the joint grounding score over its backbone on all three sets and achieves the highest joint score among the evaluated open-source single-call models on each refined benchmark. Ablations support role-wise initialization and explicit counter rewards for improving option-matched evidence coverage beyond support-only optimization. We further discuss limitations and directions for future work in Appendix E.

Reproducibility Statement

We specify the evaluated subsets, generation protocols, token-role losses, reward components, and metric aggregation rules in the main text. Appendix C provides output-format and optimization details. We commit to releasing the SC-GROUND benchmark, along with the training and evaluation code, after final cleanup and subject to source-data restrictions.

AI Use Statement

Generative AI tools assisted manuscript organization, language polishing, and reference verification. The annotation pipeline also uses multimodal models, as described in Section 3.2. The authors are responsible for the scientific claims, data, citations, and final submitted text.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Guo Chen, Yicheng Liu, Yifei 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, 2024a. URL https://arxiv.org/abs/2412.12075.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos, 2024b. URL https://arxiv.org/abs/2406.19392.

Kaiyan Chen, Junbin Xiao, and Xun Yang. Long-to-short video evidence reasoning for grounded question answering, 2026. URL https://arxiv.org/abs/2609.15224.

Ruizhe Chen, Tianze Luo, Zhiting Fan, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Zhang Huaijian. Datasets and recipes for video temporal grounding via reinforcement learning. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 983–992, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.66. URL https://aclanthology.org/2025.emnlp-industry.66/.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024c.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-CCAM: Enhancing video-language understanding with causal cross-attention masks for short and long videos, 2024. URL https://arxiv.org/abs/2408.14023.

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

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. VITA: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22898–22909. IEEE, 2023.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments, 2023. URL https://arxiv.org/abs/2311.18445.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant, 2024. URL https://arxiv.org/abs/2403.19046.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026. URL https://arxiv.org/abs/2606.24797.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5627–5646, 2025.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026.

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, 2025. URL https://arxiv.org/abs/2504.06958.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2928–2937, June 2022.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. ST-LLM: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning, 2026. URL https://arxiv.org/abs/2503.13444.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

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

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024. URL https://arxiv.org/abs/2312.02051.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yepeng Tang, Weining Wang, Longteng Guo, Tongtian Yue, Wenxuan Wang, Chunjie Zhang, and Jing Liu. Divid: Disentangled spatial-temporal modeling within llms for temporally grounded video understanding. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 17046–17073, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1c85c302ece39939c1b334c78f7ee1b8-Paper-Conference.pdf.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.50. URL https://aclanthology.org/2025.findings-emnlp.50/.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to look matters: On-policy self-distillation for long-video understanding. *arXiv preprint arXiv:2608.25356*, 2026a.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-backed video question answering. *arXiv preprint arXiv:2607.11862*, 2026b. URL https://arxiv.org/abs/2607.11862.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding, 2025b. URL https://arxiv.org/abs/2503.13377.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024a. URL https://arxiv.org/abs/2403.15377.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024b. URL https://arxiv.org/abs/2403.10228.

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 (CVPR)*, pp. 13204–13214, June 2024.

Yuanxing Xu, Yuting Wei, Shuai Zhong, Xinming Chen, Jinsheng Qi, and Bin Wu. Exploring question guidance and answer calibration for visually grounded video question answering. In Yaser

Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3121–3133, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.176. URL https://aclanthology.org/2024.findings-emnlp.176/.

Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29880–29892, June 2025.

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

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

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024a.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10419–10429, 2026.

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

A CONSTRUCTION OF SC-GROUND

This appendix describes the construction pipeline summarized in Section 3.2. The refinement aims to anchor every answer option to an event in the video while preserving a unique correct answer. GPT-5.6 creates timestamped video records and proposes candidate options and intervals; Gemini-3.1-Pro checks them against the records. The records summarize sampled frames and provide a common reference for the checks, so the generator does not verify its own proposals. The model families used in construction also appear among the evaluated systems in Table 1. Table 4 lists the retained subsets.

Table 4: **Evaluated SC-GROUND subsets.** Counts refer to the exported evaluation annotations, not the full source datasets.

Refined benchmark	Source split	Videos	Questions
NEXT-GQA-SC	NExT-GQA test	762	2,121
REXTIME-SC	ReXTime validation	289	317
CG-BENCH-SC	CG-Bench mini	630	889

Dense video records. We construct one question-independent record per video and reuse it throughout the pipeline. For a video of duration d and native frame rate f_{nat}, we sample $\min\big(c, \max(32, d \cdot \min(f_{\max}, f_{\mathrm{nat}}))\big)$ frames, rounding the count up to a multiple of 32. Frames are resized to 448 pixels on the longer side and grouped into clips of 32 consecutive frames. For each clip, the generator writes a 40–80-word summary of visible people, actions, objects, scenes, and camera motion, including verbatim on-screen text. It also identifies one to eight atomic events whose intervals cover the clip without overlap. The prompt restricts both fields to visible content and forbids inferences about intentions, emotions, identities, or off-screen facts. Event boundaries are snapped to the sampled-frame grid; gaps are filled, and events shorter than 0.5 seconds are merged with neighboring events. CG-Bench videos last about 25 minutes at the median, so a 256-frame cap would provide roughly one frame every six seconds. We instead sample them at 1 fps with a cap of 4,096 frames. Table 5 summarizes the sampling settings and record sizes.

Table 5: **Video records and interval rules.** c is the frame cap, $f_{\max}$ is the maximum sampling rate, and d is the video duration. Record counts cover videos processed before question filtering; medians are computed per video.

	Sampling				Record size			Interval rules	
Source	c	$f_{\max}$	Median fps	Median clips	Videos	Clips	Events	Minimum extent	Distractor cap
NExT-GQA test	256	8	7.97	8	999	6,839	27,893	$\min(0.5d, \max(5\,\mathrm{s}, 0.15d))$	$0.45d$
ReXTime val	256	8	1.76	8	696	5,377	30,311	$\min(0.5d, \max(5\,\mathrm{s}, 0.15d))$	$0.45d$
CG-Bench mini	4,096	1	1.01	47	1,118	58,614	407,524	20 s	90 s

Generation. For each question, one generator call receives the question, the original options, and the full record. The record uses entry tags such as `clip 3 [12.4s-16.1s]` and `e3.2 [13.2s-14.9s]`; these tags are the only permitted references to time. The generator returns one or two support ranges for the correct answer, each with an evidence sentence of at most 25 words based only on the covered entries. The prompt asks for complete event episodes rather than isolated atomic moments and limits their total extent to less than one third of the record. The same call proposes $k + 3$ candidate distractors, where k is the target number of final distractors. Each candidate specifies a type (wrong time, entity, relation, count, or attribute), an entry range, option text of at most 15 words, a claim of at most 25 words describing the visible fact that makes the option true within that range, and a note of at most 20 words explaining why it does not answer the question in the full-video context. Candidates must match the form and granularity of the correct answer and be plausible answers to someone who sees only their own ranges. Their ranges must lie outside the support ranges and apart from one another. Candidate text may neither restate the correct answer nor use temporal cues such as "later" or "earlier". For questions containing *not*, *except*, or *never*, the prompt clarifies that a distractor should describe an event that occurs and is therefore excluded by

the question. The five types guide proposal generation; they are not independently validated human reasoning labels.

Interval resolution. An entry range such as `e3.2-e3.7` spans from the start of its first event to the end of its last. Only existing entry tags are accepted, so each boundary coincides with a recorded event boundary. If an interval is shorter than the minimum extent, adjacent events are added alternately on the left and right until the minimum is reached or the record ends. Table 5 gives the minimum extents and distractor caps: NExT-GQA and ReXTime use duration-scaled limits, whereas CG-Bench uses fixed limits in seconds.

Candidate verification. Each candidate undergoes three checks. First, a local check requires a valid type, a nonempty claim, nonempty option text of at most 18 words without temporal cues, no duplicate option, and a resolvable range within the distractor cap. A candidate interval may overlap the support interval by at most 25% of its own duration, limiting reuse of support evidence. Second, the verifier reads only the entries inside the candidate range and checks whether they support every factual statement in its claim, without adding objects, actions, text, or attributes. Third, the verifier reads the full record and a shuffled set containing the correct answer and up to seven candidates that passed the first two checks. In one call with a reasoning budget of 1,024 tokens, it judges whether each option answers the question; a candidate passes if it is judged not to answer it. Shuffling reduces letter-position bias, while evaluating the options together allows direct comparison. Whether the verifier accepts the correct answer is recorded at this stage; final answer consistency is enforced by the question-level check. Across the three sources, 66–70% of candidates pass all three checks. Rejections occur most often during format and interval checks for NExT-GQA and ReXTime, and during full-record rejection checks for CG-Bench.

Selection and style normalization. We select up to k distractors from the accepted candidates. For NExT-GQA, $k = 4$; for ReXTime and CG-Bench, k equals the number of incorrect choices in the source question, preserving its option count when enough candidates pass verification. Questions with only three accepted distractors remain eligible. Any two selected intervals may overlap by at most 80% of the shorter interval. We rotate candidate types when ordering them, without requiring the selected types to differ. After shuffling the final options, including the correct answer, we ask the generator to restyle them jointly without identifying the correct one. The verifier checks each rewrite for semantic equivalence. If any rewrite fails, all options retain their pre-restyling wording; only final punctuation and initial capitalization are normalized to the majority style. Applying rewrites jointly avoids stylistic cues, and the correct answer is never rewritten on its own.

Question-level verification. The verifier reads the full record and final options and selects the answer best supported by the evidence. We repeat this check with two option orders and retain a question only if it selects the original correct answer both times. This step screens for ambiguous revisions and records that do not capture the event needed to answer the question. As a separate diagnostic, we show the verifier only the record entries inside one distractor's interval and require it to choose an option. A distractor is marked as bound to its interval if it is selected in both shuffled trials. This diagnostic does not filter the reported evaluation sets.

Retained subsets and annotation quality. Table 4 reports 2,121 NExT-GQA, 317 ReXTime, and 889 CG-Bench questions, corresponding to 40%, 34%, and 30% of their source splits. The retained NExT-GQA questions have five options (1,849) or four (272); all ReXTime questions have four, and CG-Bench questions have between four and eight. Too few accepted distractors account for the largest loss on NExT-GQA, while question-level verification accounts for the largest losses on ReXTime and CG-Bench. In a sample of NExT-GQA verification failures, the verifier often finds no option supported by the record, suggesting incomplete record coverage and the exclusion of some otherwise valid questions.

Training dataset. Training responses are built from multiple-choice questions in the Video-R1 collection using timestamped event records. They include support intervals, counter intervals for distractors matched to visible events, event descriptions, and rejection analyses. The parsed cold-start set contains 7,648 examples, and the separate reinforcement-learning set contains 2,400. Some original distractors have no corresponding event in the video; for these, the counter block contains

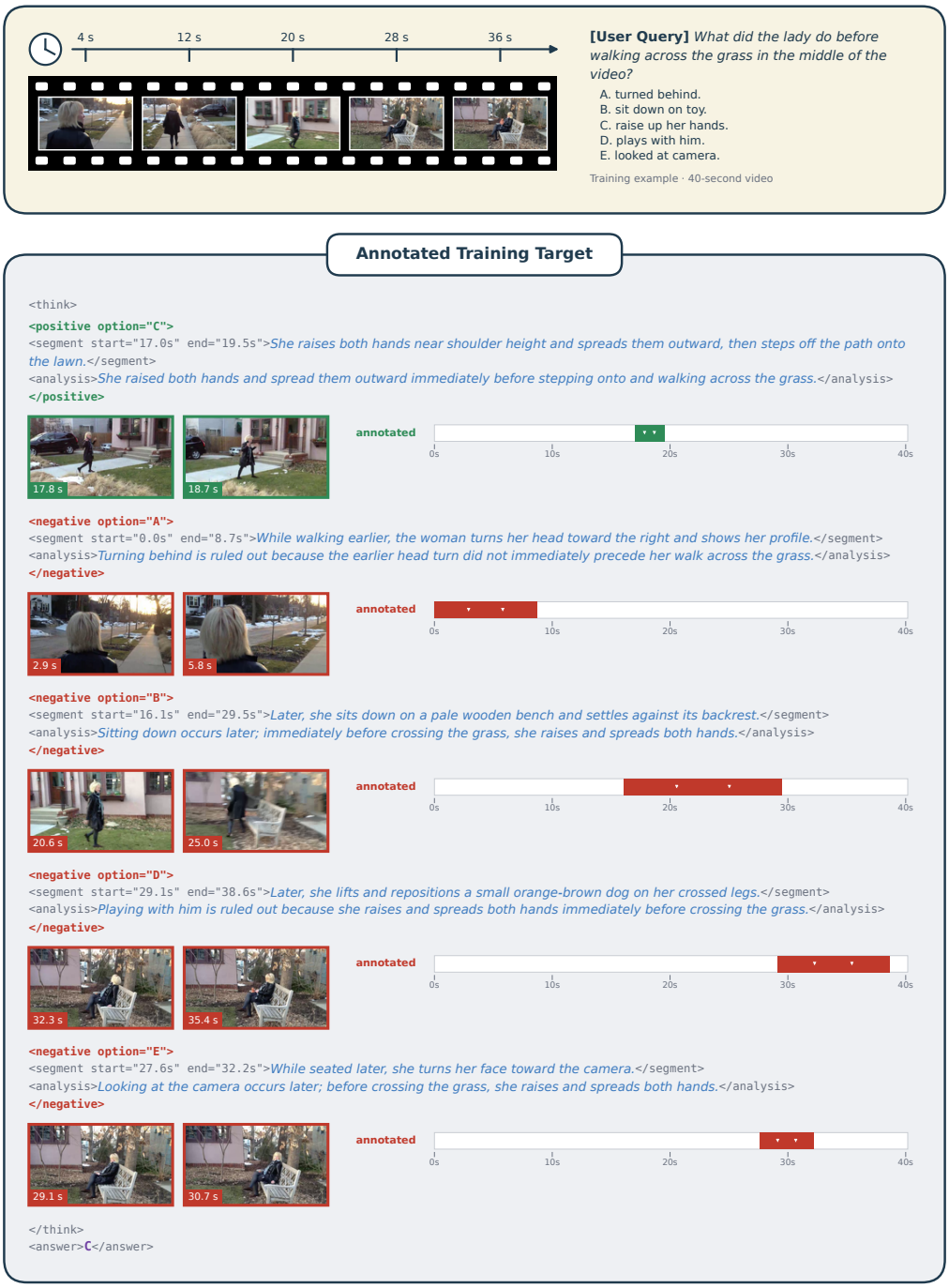

4 s
12 s
20 s
28 s
36 s
[User Query] What did the lady do before walking across the grass in the middle of the video?
A. turned behind.
B. sit down on toy.
C. raise up her hands.
D. plays with him.
E. looked at camera.
Training example · 40-second video
Annotated Training Target
<think>
<positive option="C">
She raises both hands near shoulder height and spreads them outward, then steps off the path onto the lawn.
<analysis>She raised both hands and spread them outward immediately before stepping onto and walking across the grass.</analysis>
</positive>
17.8 s
18.7 s
annotated
0s
10s
20s
30s
40s
<negative option="A">
While walking earlier, the woman turns her head toward the right and shows her profile.
<analysis>Turning behind is ruled out because the earlier head turn did not immediately precede her walk across the grass.</analysis>
</negative>
2.9 s
5.8 s
annotated
0s
10s
20s
30s
40s
<negative option="B">
Later, she sits down on a pale wooden bench and settles against its backrest.
<analysis>Sitting down occurs later; immediately before crossing the grass, she raises and spreads both hands.</analysis>
</negative>
20.6 s
25.0 s
annotated
0s
10s
20s
30s
40s
<negative option="D">
Later, she lifts and repositions a small orange-brown dog on her crossed legs.
<analysis>Playing with him is ruled out because she raises and spreads both hands immediately before crossing the grass.</analysis>
</negative>
32.3 s
35.4 s
annotated
0s
10s
20s
30s
40s
<negative option="E">
While seated later, she turns her face toward the camera.
<analysis>Looking at the camera occurs later; before crossing the grass, she raises and spreads both hands.</analysis>
</negative>
29.1 s
30.7 s
annotated
0s
10s
20s
30s
40s
</think>
<answer>C</answer>

Figure 3: An annotated training target from the Video-R1 collection. In this example, the correct option receives a support interval and every competing option a counter interval, each with an observation caption and a rejection analysis.

a rejection analysis but no temporal interval. These absent-event cases motivate the restraint reward and differ from the temporally anchored alternatives in the refined evaluation.

B An Annotated Contrastive Example

The example below comes from the exported ReXTime validation refinement (question identifier `anet_val4078`), rather than a newly invented video. Wording is abridged from the annotation record.

> **Question:** What happens when he throws a dart at the cameraman?
>
> **Correct answer:** The camera moves in a violent shaking motion.
>
> **Support, 24.800–33.667 s:** A man throws a small object; the camera rapidly moves, blurs, and shows tilted views of the room.
>
> **Competing answer:** The camera tilts downward toward the feet of the men.
>
> **Counter, 9.933–17.200 s:** The camera tilts toward the men's legs after another throw directed toward the wall panel.
>
> **Contrast:** The competing camera motion occurs, but follows a different throw. The question asks about the throw at the cameraman.

The counter interval is not supposed to show that downward tilting never happens. It identifies the true but misapplied event, whose relationship to the question differs from that of the support event. This is why counter grounding is not equivalent to negating a support caption.

C Output Format and Optimization Details

Output grammar. The implementation uses legacy `positive` and `negative` tag names for the support and counter roles. These are naming aliases, not different concepts. A response consists of a `think` envelope with one support block, followed by option-indexed counter blocks, and a final `answer` letter. Within a block, `segment` contains start/end times in seconds and a visible-event description; `analysis` relates the event to the question and option. For example, the structural pattern is:

```
<think>
<positive option="B">
<segment start="..." end="...">Visible support event.</segment>
<analysis>Why this event answers the question.</analysis>
</positive>
<negative option="A">
<segment start="..." end="...">Competing visible event.</segment>
<analysis>Why this event does not answer the question.</analysis>
</negative>
</think>
<answer>B</answer>
```

This is a grammar schematic, not an additional evaluated example; each unselected option receives its own counter block, and timestamps must be numeric. A counter block may omit `segment` for an annotated absent event in training. In Eq. 6, $\hat{\mathcal{C}}_a \approx \mathcal{S}$ holds if more than half of the predicted interval overlaps the annotated support. The fixed-role evaluation parser extracts intervals in seconds and merges overlapping intervals before computing set IoU. Unreturned annotated intervals receive 0; no score is conditioned on successful counter emission.

Optimization. Both stages use the Adam optimizer with learning rates given in Section 4.1, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, and gradient clipping at 1.0. The 2 ROLE-WISE OPD loss branches have equal outer weights, and active token roles are normalized separately. Whole-response OPD removes the gold branch and distills the sampled completion as a single role. SFT instead trains on the annotated response.

Stage 2 uses groups of 16 rollouts. Following DAPO (Yu et al., 2026), the clipping bounds are asymmetric, $1 - \epsilon_l = 0.8$ and $1 + \epsilon_u = 1.28$. Following Dr. GRPO (Liu et al., 2025), the group-relative advantage subtracts the group mean reward without dividing by the group standard deviation. Sampling uses top-$p = 1.0$ in both training stages.

Table 6: **Optimization ablations on NExT-GQA-SC.** Clip-Higher uses the upper ratio bound $1 + \epsilon_u = 1.28$; No std. omits group reward standard-deviation normalization. Bold marks the best score in each column.

Clip-Higher	No std.	Acc	mIoU^{+}	mIoU^{-}	Acc@IoU$^{\pm}$
×	✓	54.0	37.3	15.3	8.85
✓	×	52.0	37.1	**18.9**	10.58
✓	✓	**55.3**	**41.5**	17.1	**11.67**

Optimization ablations. When standard-deviation normalization is omitted, Clip-Higher raises $\mathrm{Acc@IoU}^{\pm}$ from 8.85 to 11.67; its wider upper bound is intended to support exploration (Yu et al., 2026). At the higher clip bound, omitting normalization avoids upweighting low-variance rollout groups (Liu et al., 2025) and raises $\mathrm{Acc@IoU}^{\pm}$ from 10.58 to 11.67, although counter mIoU falls from 18.9 to 17.1 (Table 6).

D Baselines on the Original Grounded Benchmarks

On NExT-GQA, we compare with VIOLETv2 (Fu et al., 2023), SeViLA (Yu et al., 2023), LangRepo (Kahatapitiya et al., 2025), VideoStreaming (Qian et al., 2024), LLoVi (Zhang et al., 2024a), HawkEye (Wang et al., 2024b), Grounded-VLLM (Wang et al., 2025a), VideoChat-TPO (Yan et al., 2025), VideoMind (Liu et al., 2026), Divid (Tang et al., 2026), LOVER (Chen et al., 2026), TVG-R1 (Chen et al., 2025), VideoChat-R1 (Li et al., 2025), and Qwen3-VL (Bai et al., 2025a). On CG-Bench, we compare with VITA (Fu et al., 2024), Video-CCAM (Fei et al., 2024), ST-LLM (Liu et al., 2024), MiniCPM-v2.6 (Yao et al., 2024), LongVA (Zhang et al., 2024b), InternVL2 (Chen et al., 2024c), Divid (Tang et al., 2026), and Qwen3-VL (Bai et al., 2025a), alongside the proprietary models and Qwen2-VL listed in Table 2. On ReXTime, we compare with VTimeLLM (Huang et al., 2023), TimeChat (Ren et al., 2024), LITA (Huang et al., 2024), VideoMind (Liu et al., 2026), Divid (Tang et al., 2026), and Qwen3-VL (Bai et al., 2025a).

E Future Work & Limitations

Despite its gains, GroundBoth reaches only 0.93 joint grounding accuracy on CG-Bench-SC. With long videos, brief supporting evidence, and several competing options, reliable localization across the full timeline remains an open challenge.

Currently, GroundBoth and SC-Ground are validated solely on multiple-choice VideoQA, where the answer options enumerate the alternatives to be rejected and answer correctness can be verified by matching a letter. Open-ended questions provide neither, so the model would have to generate the plausible alternatives itself, and both the reward and the evaluation would need a reliable judge of free-form answers in place of exact matching. Extending the support–counter formulation to open-ended VideoQA along these two lines is left as a direction for future work.

F QUALITATIVE EXAMPLES

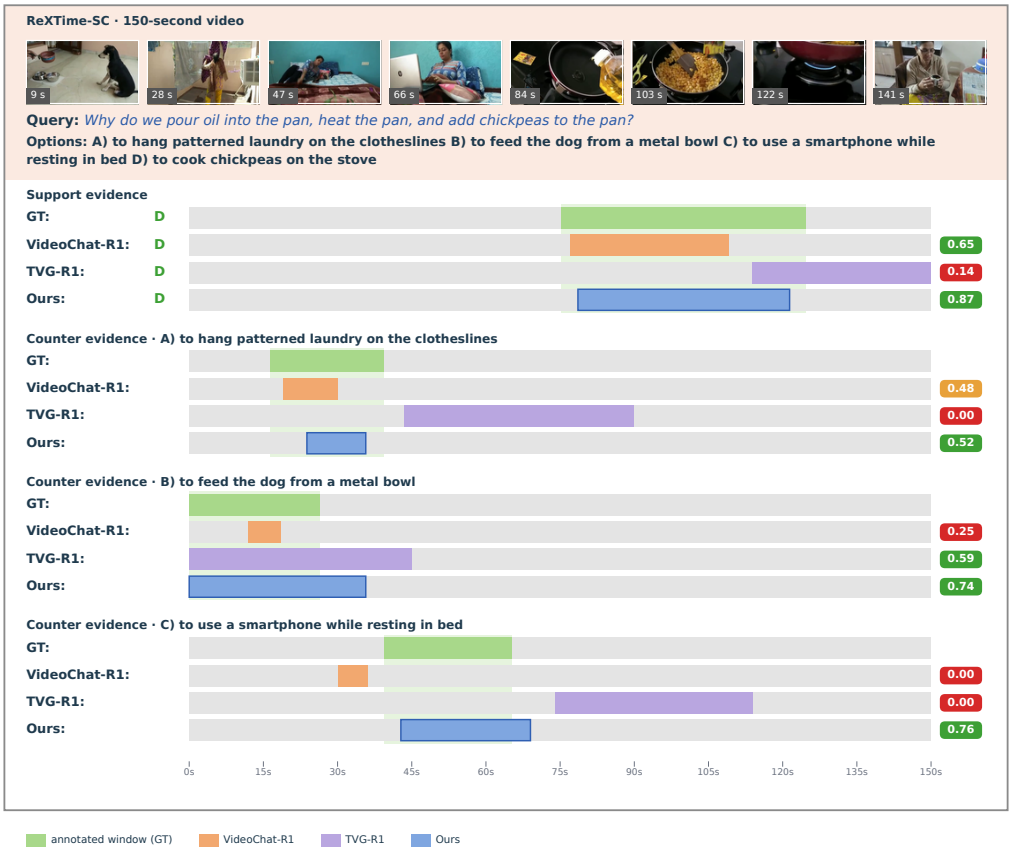


Figure 4: GROUNDBOTH grounds the selected answer and the competing options to their corresponding events in a REXTIME-SC video.

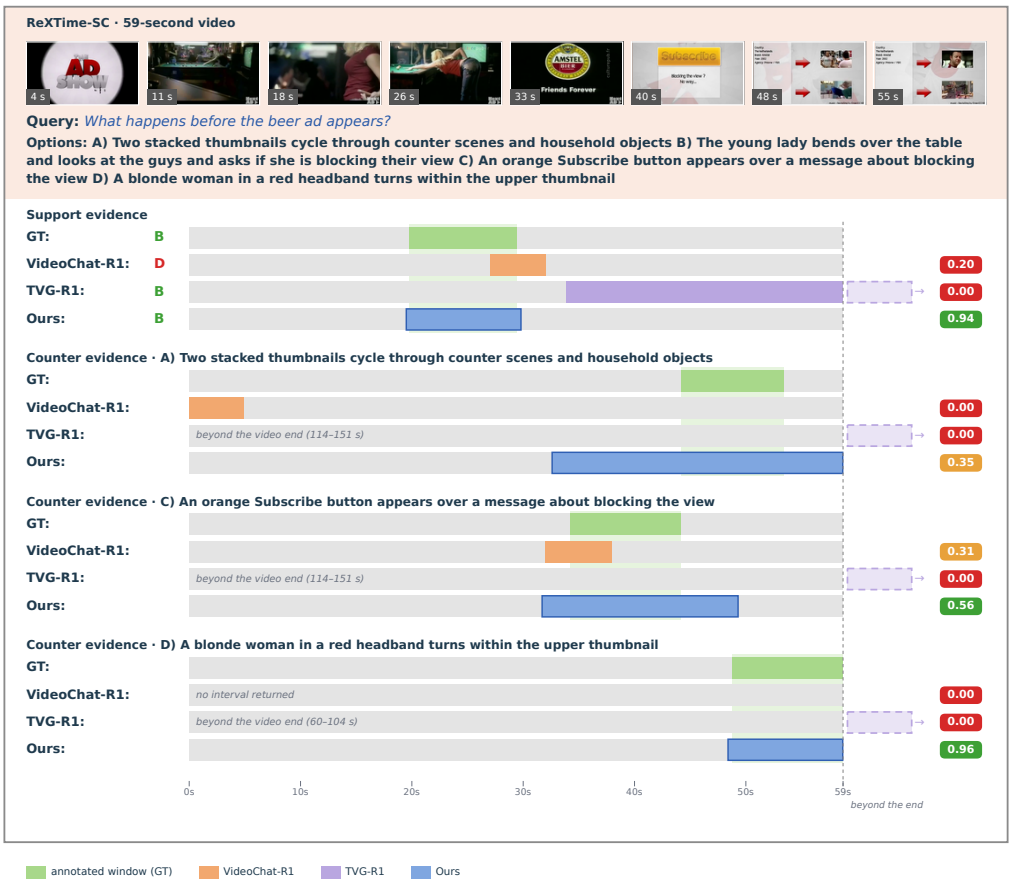


Figure 5: A second REXTIME-SC example. Dashed boxes mark intervals that a baseline placed beyond the end of the video.

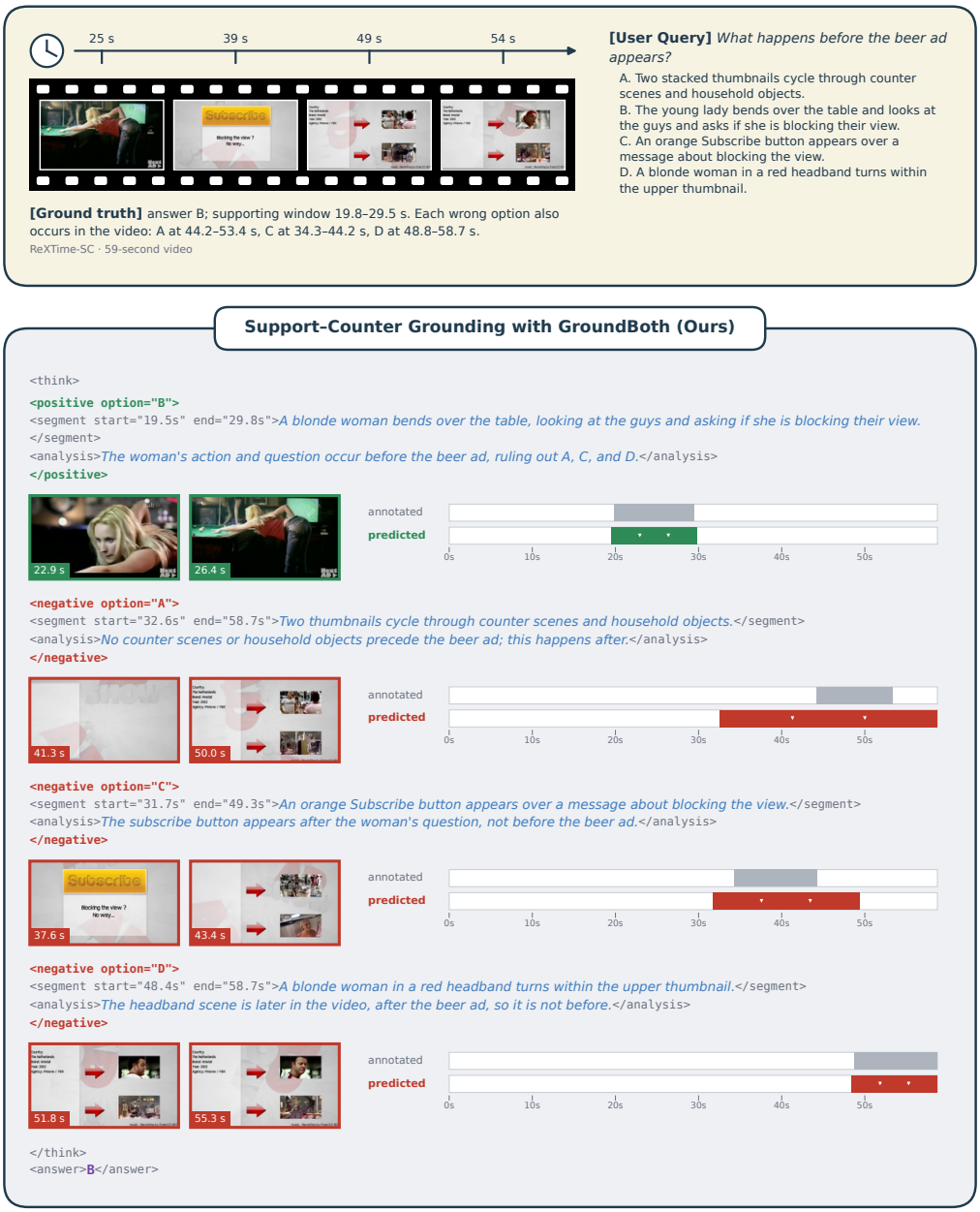

25 s
39 s
49 s
54 s
[User Query] What happens before the beer ad appears?
A. Two stacked thumbnails cycle through counter scenes and household objects.
B. The young lady bends over the table and looks at the guys and asks if she is blocking their view.
C. An orange Subscribe button appears over a message about blocking the view.
D. A blonde woman in a red headband turns within the upper thumbnail.
[Ground truth] answer B; supporting window 19.8–29.5 s. Each wrong option also occurs in the video: A at 44.2–53.4 s, C at 34.3–44.2 s, D at 48.8–58.7 s.
ReXTime-SC · 59-second video
Support–Counter Grounding with GroundBoth (Ours)
<think>
<positive option="B">
A blonde woman bends over the table, looking at the guys and asking if she is blocking their view.
<analysis>The woman's action and question occur before the beer ad, ruling out A, C, and D.</analysis>
</positive>
22.9 s
26.4 s
annotated
predicted
0s 10s 20s 30s 40s 50s
<negative option="A">
Two thumbnails cycle through counter scenes and household objects.
<analysis>No counter scenes or household objects precede the beer ad; this happens after.</analysis>
</negative>
41.3 s
50.0 s
annotated
predicted
0s 10s 20s 30s 40s 50s
<negative option="C">
An orange Subscribe button appears over a message about blocking the view.
<analysis>The subscribe button appears after the woman's question, not before the beer ad.</analysis>
</negative>
37.6 s
43.4 s
annotated
predicted
0s 10s 20s 30s 40s 50s
<negative option="D">
A blonde woman in a red headband turns within the upper thumbnail.
<analysis>The headband scene is later in the video, after the beer ad, so it is not before.</analysis>
</negative>
51.8 s
55.3 s
annotated
predicted
0s 10s 20s 30s 40s 50s
</think>
<answer>B</answer>

Figure 6: A complete GROUNDBOTH response pairs support and counter intervals with event descriptions and analyses.