ASSEMBLE: ATOMIC SKILLS FOR EVIDENCE-GROUNDED VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex video reasoning often depends on evidence scattered across distant moments, entities, and events, yet a correct answer alone does not reveal whether a model relied on the right parts of the video. We introduce ASSEMBLE, a framework that makes supporting evidence explicit throughout long-video reasoning. It organizes local observations and cross-clip narratives into timestamped evidence catalogs traceable to the source video. A grounding-aware reader then composes question-specific atomic skills whose structured outputs include explicit evidence references and support-assessment fields. We turn correctness-gated citation alignment into a direct evidence-grounding signal: after teacher-supervised fine-tuning, Group Relative Policy Optimization (GRPO) jointly optimizes answer correctness and citation alignment. The resulting intermediate traces are inspectable, and final predictions remain linked to explicit supporting evidence. With a 9B reader supervised by a 235B teacher and shared precomputed evidence catalogs, ASSEMBLE achieves 59.2% macro-averaged answer accuracy across three long-video reasoning benchmarks, compared with 58.3% for Gemini-2.5-Pro, and exceeds it by 6.7% in macro-averaged overlap-based Grounded accuracy, with gains on all three benchmarks. Ablations further show that, with the same post-trained reader and inference budget, structured skill inference improves overlap-based Grounded accuracy over free-form reasoning. Together, these results make evidence grounding an integral part of long-video reasoning.

1 INTRODUCTION

Answering complex questions about long videos requires integrating clues across distant moments, tracking entities and events, and connecting local observations to broader context (Cheng et al., 2025; Chen et al., 2025b; Yu et al., 2025). Failures can arise from missing evidence, incorrect entity or temporal grounding, or faulty reasoning over relevant observations (Yu et al., 2026; Meng et al., 2026; Tsuchiya et al., 2026). Answer-level supervision provides little indication of where these failures occur: an incorrect answer does not reveal where the reasoning failed, while a correct answer does not show whether the model relied on the relevant evidence (Xie et al., 2025). This motivates making evidence use explicit and evaluating grounding alongside answer correctness.

Prior work has explored reusable operations for multimodal reasoning by extracting visually grounded skills from trajectories (Jiang et al., 2026; Zhang et al., 2026b), jointly learning hierarchical skills and policies (Li et al., 2026d; Zhang & Qiu, 2026), and representing skills as executable structures with verification and local repair (Xia et al., 2026c). These approaches motivate making reasoning operations and the evidence supporting them explicit. For long-video reasoning, however, the challenge is not only whether an operation is valid, but whether it is applied to the right evidence across entities, events, and distant time intervals (Di & Xie, 2024; Xiao et al., 2024; Wei et al., 2026; Zhang et al., 2026c). Studies on grounded VideoQA further show that strong answer accuracy does not necessarily imply accurate localization of supporting evidence (Xiao et al., 2024; Wei et al., 2026). This raises a practical question: how can explicit reasoning operations be tied to supervision that encourages a reader to ground its answers in relevant video evidence?

To connect explicit reasoning operations with grounding supervision, we use correctness-gated citation alignment as a measurable evidence-grounding signal. Atomic skills provide structured interfaces for reasoning operations, evidence references, and support assessments, while timestamped

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.

references allow cited evidence to be compared with annotated video intervals. When an answer is correct, the reader is further rewarded for citing the moments that support it; conditioning this signal on correctness prevents aligned citations from compensating for an incorrect prediction. Post-training encourages correct, evidence-grounded answers, while atomic skills structure evidence selection, reasoning, and integration at inference time.

To implement this idea, we introduce ASSEMBLE, which combines a timestamped evidence catalog, a fixed atomic-skill vocabulary, and a grounding-aware reader. The catalog pairs textual descriptions with source intervals, making evidence reusable across questions and citations traceable to specific moments. Atomic skills define structured reasoning operations with explicit evidence references and support assessments, and their ordered traces expose intermediate outputs, cited evidence, and model-generated diagnostic signals. We first fine-tune the reader on teacher-generated responses with supervised fine-tuning (SFT), then apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with rewards for answer correctness, correctness-gated citation alignment, and output validity. At inference time, the same reader performs planning, ordered skill application, integration, and final answer generation through a multi-call pipeline, ending with an answer and supporting citations. We evaluate answer accuracy and citation grounding separately across three long-video reasoning benchmarks, allowing us to assess both whether the model answers correctly and whether it relies on the supporting video evidence. Our contributions include:

- We introduce ASSEMBLE, an atomic-skill framework for complex video reasoning that makes reasoning operations and their supporting evidence explicit through timestamped evidence catalogs and structured skill traces.
- We develop a grounding-aware reader with teacher-response supervision and GRPO using correctness-gated citation alignment. At inference time, a structured multi-call pipeline produces answers with explicit citations to supporting video evidence.
- We evaluate answer accuracy and overlap-based Grounded accuracy across three long-video reasoning benchmarks. ASSEMBLE achieves 59.2% macro-averaged answer accuracy versus 58.3% for Gemini-2.5-Pro and improves Grounded accuracy by 6.7%. Ablations further show that grounding-aware post-training improves evidence grounding, while structured skill inference outperforms free-form reasoning in Grounded accuracy.

2 Related Work

Skill Learning and Compositional Program Induction. Early visual-programming approaches showed that complex visual reasoning can be decomposed into executable compositions of reusable modules. Visual Programming (Gupta & Kembhavi, 2023) generates modular programs that invoke pretrained vision and language tools, while ViperGPT (Surís et al., 2023) uses generated Python programs to compose visual APIs for reasoning. More recent agents learn and maintain reusable skills from interaction: SkillRL (Xia et al., 2026a) jointly evolves skills and policies through RL, XSkills (Jiang et al., 2026) distills multimodal experience into reusable skills, and COS-PLAY (Wu et al., 2026) continually updates a skill library from trajectories. Other systems organize or refine skills for composition, including MemSkill (Zhang et al., 2026a), GraSP (Xia et al., 2026c), SkillGraph (Li et al., 2026c), Skill-Pro (Mi et al., 2026), and SkillOS (Ouyang et al., 2026). Our work brings this modular view to evidence-grounded video reasoning: a shared Reader applies structured atomic-skill interfaces over a reusable evidence catalog and returns answers with citations to the supporting video intervals.

Evidence-Grounded Video Reasoning. Evidence-grounded video reasoning requires models to identify the temporal support for an answer, not just predict the answer itself. NExT-GQA (Xiao et al., 2024) and EG-VQA (Huang et al., 2026) add answer-aligned temporal annotations, while VITED (Lu et al., 2025), Grounded-VideoLLM (Wang et al., 2024a), and TimeChat (Ren et al., 2024) improve joint reasoning and temporal grounding. GroundVQA (Di & Xie, 2024) and MultiHop-EgoQA (Chen et al., 2025c) extend this setting to long egocentric videos and evidence spread across multiple intervals, while EV^2-Bench (Wei et al., 2026) and Video-in-the-Loop (Wang et al., 2025a) evaluate or predict explicit supporting evidence. VideoMind (Liu et al., 2026) structures reasoning into planning, grounding, verification, and answering, while Video-Grounded Entailment Tree (Liu et al., 2025) decomposes hypotheses and verifies them against video evidence. Evidence-aware training has also advanced rapidly: Video-VER (Luo et al., 2026) rewards visually

grounded reasoning traces, Open-o3-Video (Meng et al., 2025) predicts explicit spatio-temporal evidence, SER (Xia et al., 2026b) uses semantic evidence verification, and EG-Reasoner (Huang et al., 2026) trains for grounded answering. TimeThink (Li et al., 2026a) and LOVER (Chen et al., 2026a) further introduce temporal process and grounding-aware rewards. Our method instead builds a reusable, question-independent timestamped catalog and supervises citation alignment to its entries. A shared post-trained Reader composes structured skill interfaces over this catalog while keeping its outputs linked to cited video intervals.

Long-Video Understanding and Evidence Retrieval. Long-video understanding requires finding sparse, question-relevant evidence across extended temporal contexts, as highlighted by CG-Bench (Chen et al., 2025b) and Video-Holmes (Cheng et al., 2025). LLoVi (Zhang et al., 2024) aggregates dense clip captions with an LLM, while VideoAgent (Wang et al., 2024b), VideoTree (Wang et al., 2025c), and T* (Ye et al., 2025) retrieve relevant visual evidence through iterative, hierarchical, or adaptive temporal search. ReWind (Diko et al., 2025) and SEAL (Wang et al., 2025b) use compact memory or semantic entities to guide selection, while LVAgent (Chen et al., 2025a) and DrVideo (Ma et al., 2025) reason over selected segments or searchable video representations. In contrast, we build a reusable, question-independent timestamped evidence catalog once per video and use a post-trained reader to compose atomic skills over it, linking answers back to the supporting video intervals.

3 Problem Formulation

Complex Video Reasoning. We study complex video reasoning across social-reasoning and long-video understanding tasks. Each instance consists of a video V, a question q including its answer options, a reference answer y, and, when available, annotated supporting evidence $\mathcal{E} = \{e_j\}_{j=1}^{M}$ with associated temporal spans. Using a set of perception tools $\mathcal{U}$, we construct an evidence catalog $\mathcal{C}(V)$ whose entries contain textual observations and their temporal provenance. Given q and $\mathcal{C}(V)$, the reader predicts an answer $\hat{y}$ and supporting citations $\hat{Z}$, where each citation indexes a catalog entry. Reference answers y and evidence annotations $\mathcal{E}$ are used for training and evaluation, but are excluded from the reader's inference inputs.

Atomic Skills and Structured Planning. We define a fixed vocabulary $\mathcal{S} = \{s_1, \ldots, s_K\}$ of atomic skills for evidence retrieval and structured reasoning. Each skill has an explicit interface:

$$s_k = \langle \mathcal{I}_k, \mathcal{O}_k, \ell_k, \mathcal{A}_k \rangle , \tag{1}$$

where $\mathcal{I}_k$ and $\mathcal{O}_k$ specify the input and output schemas, ℓ_k is the stage-specific instruction, and $\mathcal{A}_k$ specifies evidence-support assessment fields, including a support score, a status label, and an optional diagnostic label.

Given a question q and an evidence catalog $\mathcal{C}(V)$, the reader's planning call produces an ordered skill sequence:

$$\mathcal{P} = \mathrm{Plan}_{\pi_\theta}\left(q, \mathcal{C}(V); \mathcal{S}\right) = (a_1, \ldots, a_L). \tag{2}$$

Each planned skill application $a_i = (s_{k_i}, \boldsymbol{\alpha}_i)$ specifies a selected skill and its arguments. Arguments may reference the question, catalog entries, or outputs of preceding calls. Thus, $\mathcal{P}$ is a structured skill plan serialized in application order, rather than an independently executable program or dependency graph. Here, L counts planned skill applications within the structured multi-call pipeline. It does not determine the number of reader calls.

Skill Outputs and Support Assessment. For each planned skill, the reader resolves its arguments from the question, evidence catalog, and preceding outputs, then follows the corresponding instruction:

$$(o_i, \rho_i, u_i, d_i) \sim \pi_\theta(\cdot \mid q, \mathcal{C}(V), \mathbf{o}_{<i}, s_{k_i}, \boldsymbol{\alpha}_i, \ell_{k_i}) , \tag{3}$$

where o_i is the structured output, ρ_i contains its evidence references, u_i records the associated support assessment and status, and d_i is an optional diagnostic label. We retain the ordered records $\mathcal{R} = \{(o_i, \rho_i, u_i, d_i)\}_{i=1}^{L}$ for trace inspection. Support assessments are generated by the reader and are not independently verified.

Reader Inference. Given the question q, evidence catalog $\mathcal{C}(V)$, and skill records $\mathcal{R}$, the reader π_θ first produces an integration summary h and then generates the final response:

$$h \sim \pi_\theta(\cdot \mid q, \mathcal{C}(V), \mathcal{R}), \tag{4}$$

$$z \sim \pi_\theta(\cdot \mid q, \mathcal{C}(V), \mathcal{R}, h), \qquad (\hat{y}, \hat{Z}) = \mathrm{Parse}(z), \tag{5}$$

where z contains an answer rationale and label, and $\hat{Z}$ identifies the catalog entries cited in the rationale. Generating z is the final call of the inference pipeline. Reference answers and evidence annotations are excluded from its inputs.

Learning Objective. We warm-start the reader π_θ with supervised fine-tuning on teacher-generated rationales and answer labels, using cross-entropy loss on response tokens only. Starting from this initialization, we optimize the reader with GRPO toward

$$J(\theta) = \mathbb{E}_{\substack{(V,q,y,\mathcal{E})\sim\mathcal{D}_{\mathrm{RL}} \\ z\sim\pi_\theta(\cdot|q,\mathcal{C}(V))}} \left[R(z; y, \mathcal{E}, \mathcal{C}(V)) \right]. \tag{6}$$

The reward evaluates answer correctness and correctness-gated citation alignment, together with auxiliary output checks. Training optimizes responses generated from the question and catalog; skill traces provide context only at inference time. Both stages update the reader's LoRA parameters. Training and reward details are provided in Appendix D.

4 METHODOLOGY

Figure 1 provides an overview of ASSEMBLE, which combines a reusable video evidence catalog, an atomic-skill reader, and grounding-aware post-training. Videos are converted into timestamped textual catalogs that preserve temporal provenance, while a fixed atomic skill vocabulary structures evidence selection, candidate assessment, and integration of intermediate findings. We train the reader with a supervised warm start on teacher-generated responses with explicit citations, followed by GRPO on answer correctness and correctness-gated citation alignment. At inference time, the same post-trained reader dynamically composes and applies atomic skills to gather evidence, integrate intermediate findings, and produce a grounded final answer.

4.1 VIDEO EVIDENCE CATALOG CONSTRUCTION

The reader operates on a timestamped textual catalog $\mathcal{C}(V) = \{c_i\}_{i=1}^{N}$, where each entry $c_i = (x_i, I_i)$ pairs a textual observation x_i with its source interval I_i. Motivated by recent evidence that strong vision-language models can provide reliable grounded descriptions for downstream reasoning Li et al. (2026b), we use a large vision-language model to generate grounded descriptions from sampled video segments, including observable scenes, entities, events, and interactions. Each catalog entry is assigned a unique identifier that maps back to its corresponding video interval, providing an explicit link between textual evidence and the source video. Catalog construction is question-independent, so the same evidence representation can be reused across different queries.

To capture evidence at different temporal scales, we construct both local descriptions and longer-range narratives. Local entries preserve fine-grained observations within short temporal windows, while narrative entries connect events across longer spans and help maintain consistent entities and relationships over time. Together, they form a multi-scale evidence representation that supports localized retrieval and reasoning over temporally distributed observations. The resulting catalog serves as the shared evidence interface for subsequent planning, skill application, and answer grounding. Construction settings, serialization details, and representative examples are provided in Appendix A.

4.2 ATOMIC-SKILL READER

Skill interfaces. We define a fixed library $\mathcal{S}$ of reusable operations for evidence retrieval, hypothesis assessment, temporal and causal reasoning, and answer support. Each skill specifies its arguments, instruction, structured output, and evidence references. Skills are executed by the reader rather than separate models. The full vocabulary is given in Appendix B.

Question-dependent composition. Given a question q, answer options $\mathcal{O}$, and catalog $\mathcal{C}(V)$, the reader constructs an ordered plan $\mathcal{P} = (a_1, \ldots, a_L)$, where $a_i = (s_{k_i}, \boldsymbol{\alpha}_i)$ specifies a skill and its

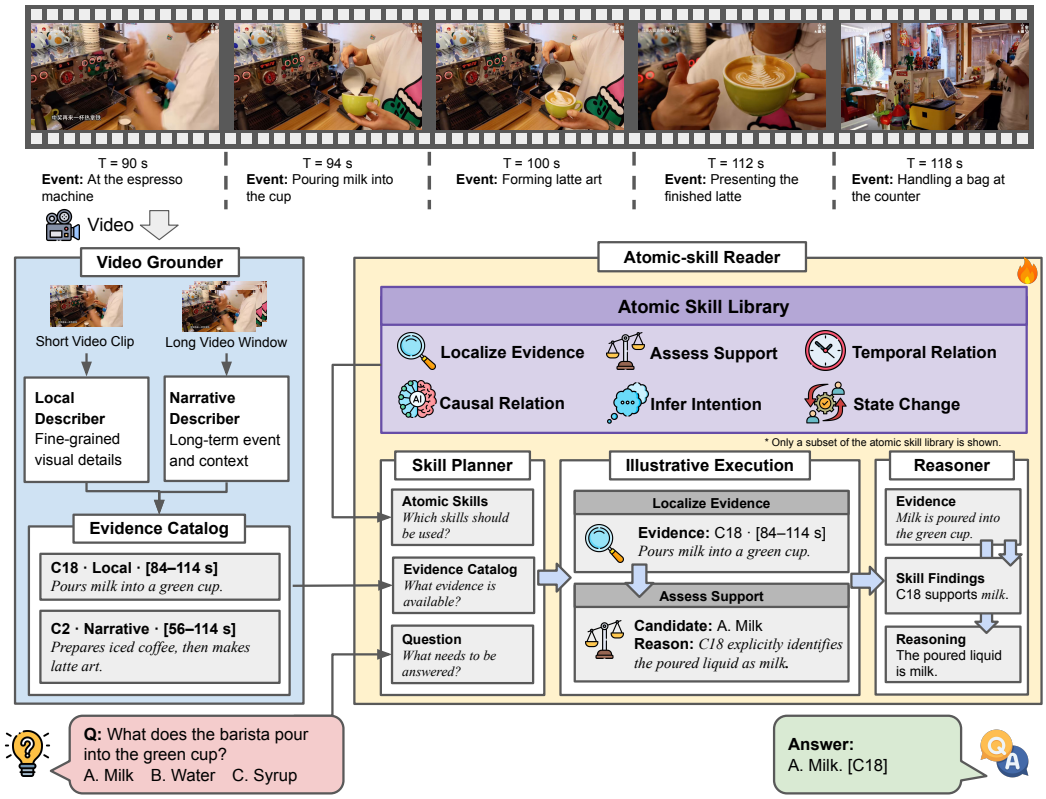


Figure 1: **Overview of ASSEMBLE.** ASSEMBLE combines a reusable timestamped evidence catalog with dynamically composed atomic skills for evidence-grounded long-video reasoning. The Video Grounder (**Blue Box**) converts the input video into a timestamped catalog of local observations and, when available, longer-range narrative context. For each question, the Atomic-skill Reader (**Yellow Box**) selects and composes structured operations from the Atomic Skill Library (**Purple Box**) over this catalog. In the illustrated example, the selected skills localize relevant evidence and assess its support for a candidate answer. The reader then integrates the resulting evidence and intermediate findings to produce a final answer with explicit catalog citations (**Green Box**). Only representative skills are shown.

arguments. The planning call also selects a catalog subset $\mathcal{C}_{\text{sel}} \subseteq \mathcal{C}(V)$. This subset is distinct from the annotated evidence $\mathcal{E}$, which is not an inference input. The selected skills and their order depend on the question and available evidence.

Skill execution and integration. Each skill application returns a structured finding, evidence references, and a support assessment, forming an ordered trace $\mathcal{R}$. Support assessments are internal reader outputs rather than independent verification, and final citations map back to source video intervals. Representative prompts are provided in Appendix H. Algorithm 1 summarizes the inference flow. The reader plans question-specific skills, applies them sequentially over selected catalog evidence, and builds an evidence-linked trace. The same post-trained reader is reused across stages to integrate intermediate findings and produce the final cited answer.

4.3 Grounding-Aware Reader Training

We train π_θ to produce a rationale, catalog citations, and a final answer from the question and evidence catalog. Training uses SFT followed by GRPO. Intermediate skill outputs are used only at inference time.

Supervised warm start. A teacher model generates cited rationales and answers from the question, answer choices, and catalog. We retain correct responses with valid rationales and, when evidence annotations are available, filter them by citation quality. The resulting responses are used for completion-only SFT.

Algorithm 1 Dynamic atomic-skill inference.

Require: Question q, options $\mathcal{O}$, precomputed catalog $\mathcal{C}(V)$, skill library $\mathcal{S}$, reader π_θ
Ensure: Answer $\hat{y}$, cited catalog identifiers $\hat{Z}$, intermediate records $\mathcal{R}$
1: $p \leftarrow \textsc{Reader}(\pi_\theta, \textsc{Plan}, q, \mathcal{O}, \mathcal{C}(V), \mathcal{S})$
2: $(\mathcal{P}, \mathcal{C}_{\text{sel}}) \leftarrow \textsc{ParsePlan}(p)$
3: $\mathcal{R} \leftarrow \textsc{ApplyPlannedSkills}(\pi_\theta, \mathcal{P}, q, \mathcal{O}, \mathcal{C}(V), \mathcal{C}_{\text{sel}})$
4: $h \leftarrow \textsc{Reader}(\pi_\theta, \textsc{Integrate}, q, \mathcal{O}, \mathcal{C}(V), \mathcal{R})$
5: $z \leftarrow \textsc{Reader}(\pi_\theta, \textsc{Answer}, q, \mathcal{O}, \mathcal{C}(V), \mathcal{R}, h)$
6: $(\hat{y}, \hat{Z}) \leftarrow \textsc{ParseAnswerAndCitations}(z)$
7: **return** $\hat{y}, \hat{Z}, \mathcal{R}$

Correctness-gated grounding reward. For a generated response z, let A denote answer correctness and P_{cite} denote citation precision, measured by overlap between cited catalog intervals and annotated evidence. Our full reward is

$$R_{\text{full}} = A + 0.5AP_{\text{cite}} - 0.2T + 0.1F - 0.05L_{\text{len}}, \tag{7}$$

where T penalizes temporal-order errors, F rewards valid response structure, and L_{len} penalizes excessive length. The product AP_{cite} makes citation reward conditional on a correct answer, encouraging grounding without rewarding well-cited incorrect predictions. Citation precision measures temporal alignment with annotated evidence rather than semantic entailment.

GRPO optimization. Starting from the SFT checkpoint, we apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each question, multiple sampled responses are scored with the reward above and optimized using group-relative advantages. Reference answers and evidence annotations are used only for reward computation and are not provided to the reader. Training details, data statistics, and reward ablations are given in Appendix D.

5 Experiments

Benchmarks. We evaluate on three complementary video reasoning benchmarks: CG-Bench (Chen et al., 2025b) for clue-grounded long-video understanding, VRBench (Yu et al., 2025) for multi-step reasoning over long narrative videos, and Video-Holmes (Cheng et al., 2025) for complex social reasoning that requires locating and integrating visual clues. Dataset splits and evaluation statistics are provided in Appendix C.

Baselines and comparison protocol. For Table 1, external video-reasoning models retain their original training setups but are evaluated as readers of the same precomputed textual evidence catalogs used by our reader. We compare ASSEMBLE with three reinforcement-learning-based methods: GRPO-CARE (Chen et al., 2026b), which adds a reasoning–answer consistency reward to outcome-based training; VideoChat-R1 (Li et al., 2025), which applies GRPO to spatio-temporal video reasoning; and Video-R1-7B (Feng et al., 2026), which combines a chain-of-thought cold start with temporal-aware reinforcement learning. We also include Qwen2.5-VL-7B (Bai et al., 2025) as a reference baseline. The Qwen3.5-9B (Qwen Team, 2026) ablations in the same table provide controlled comparisons of training objectives.

Evaluation metrics. We report answer accuracy, Grounded accuracy, and generation length. A response is Grounded if its answer is correct and at least half of its valid, distinct citations have positive-duration overlap with at least one annotated evidence span. Duplicate citations are removed and invalid identifiers are ignored. Responses with no valid citations or no evidence annotations are counted as ungrounded, and all evaluated questions remain in the Grounded-accuracy denominator. Grounded accuracy therefore measures temporal alignment with annotated evidence rather than semantic entailment and does not require citations to cover every annotated span (Gao et al., 2023). Generation length is the average number of generated tokens per question, including both the rationale and final answer. Reported gains are absolute differences between percentage-valued scores. For each checkpoint, we generate one response per question and report 95% bootstrap confidence intervals for answer and Grounded accuracy. Appendix F reports complementary benchmark-native temporal and semantic grounding metrics, while Appendix G tests whether predictions depend on cited catalog content through evidence interventions.

Table 1: **Answer accuracy and evidence grounding across three video-reasoning benchmarks.** All methods are evaluated as catalog readers using the same default catalog within each benchmark's main comparison. Catalog ablations vary only the catalog while holding the reader checkpoint fixed. The default VRBench catalog uses approximately 30-second local entries; narrative augmentation is evaluated separately. **Grounded** requires a correct answer and citation precision of at least 0.5 against annotated evidence spans. Values are percentages with 95% bootstrap confidence intervals; bold indicates the best result within each large- or small-model group. **Takeaway:** ASSEMBLE achieves the highest overlap-based Grounded accuracy among the evaluated small models while maintaining competitive answer accuracy relative to large-model baselines.

	Video-Holmes		VRBench		CGBench	
Method	**Accuracy ↑**	**Grounded ↑**	**Accuracy ↑**	**Grounded ↑**	**Accuracy ↑**	**Grounded ↑**
Large-model baselines						
GEMINI-2.5-PRO	$\mathbf{51.39 \pm 2.29}$	$\mathbf{6.75 \pm 1.12}$	$\mathbf{77.78 \pm 3.64}$	$\mathbf{29.70 \pm 4.04}$	$\mathbf{45.81 \pm 2.29}$	$\mathbf{19.70 \pm 1.82}$
QWEN3-235B-A22B	45.62 ± 2.29	4.35 ± 0.93	69.09 ± 4.04	23.64 ± 3.74	41.89 ± 2.24	13.74 ± 1.57
DEEPSEEK-V4-FLASH	41.97 ± 2.26	4.84 ± 0.98	61.01 ± 4.24	23.84 ± 3.74	26.93 ± 2.04	11.64 ± 1.46
Video-reasoning models evaluated as catalog readers						
Qwen2.5-VL-7B	31.30 ± 2.12	0.22 ± 0.19	37.78 ± 4.24	2.83 ± 1.52	19.43 ± 1.79	2.59 ± 0.75
GRPO-CARE (Chen et al., 2026b)	32.44 ± 2.12	0.93 ± 0.44	45.25 ± 4.34	4.24 ± 1.72	25.61 ± 1.99	3.70 ± 0.86
VIDEOCHAT-R1 (Li et al., 2025)	32.61 ± 2.12	0.27 ± 0.24	45.86 ± 4.34	2.22 ± 1.31	24.67 ± 1.96	3.70 ± 0.86
VIDEO-R1-7B (Feng et al., 2026)	28.91 ± 2.07	0.05 ± 0.08	37.17 ± 4.24	0.00 ± 0.00	21.41 ± 1.85	0.50 ± 0.30
ASSEMBLE and ablations (Qwen3.5-9B)						
QWEN3.5-9B	43.44 ± 2.26	7.62 ± 1.22	69.09 ± 4.04	28.08 ± 3.94	38.85 ± 2.24	18.98 ± 1.79
+ SFT	51.99 ± 2.29	7.19 ± 1.17	74.95 ± 3.74	23.43 ± 3.74	39.96 ± 2.24	20.75 ± 1.85
+ SFT w/o narrative	46.82 ± 2.29	2.18 ± 0.68	74.55 ± 3.84	25.25 ± 3.74	39.46 ± 2.25	17.49 ± 1.75
+ GRPO (full reward; w/o narrative)	51.06 ± 2.29	1.47 ± 0.57	75.96 ± 3.74	30.10 ± 4.04	39.01 ± 2.25	17.49 ± 1.75
+ GRPO (OUTCOME-ONLY REWARD)	54.44 ± 2.29	6.64 ± 1.12	76.36 ± 3.74	24.04 ± 3.74	39.80 ± 2.25	22.46 ± 1.90
+ GRPO (OUTCOME + CITATION)	$\mathbf{55.69 \pm 2.29}$	11.59 ± 1.44	77.98 ± 3.64	26.46 ± 3.84	42.83 ± 2.26	21.63 ± 1.85
+ GRPO (OUTCOME + TEMPORAL)	55.31 ± 2.29	5.50 ± 1.03	76.77 ± 3.74	23.43 ± 3.74	43.98 ± 2.26	17.94 ± 1.77
ASSEMBLE (FULL)	54.06 ± 2.26	$\mathbf{16.49 \pm 1.69}$	$\mathbf{78.99 \pm 3.64}$	$\mathbf{35.15 \pm 4.24}$	$\mathbf{44.59 \pm 2.29}$	$\mathbf{24.50 \pm 1.99}$

Training setup and controlled ablations. We train the Qwen3.5-9B reader in two stages: supervised fine-tuning (SFT) followed by Group Relative Policy Optimization (GRPO). During SFT, the model learns from filtered teacher responses containing rationales, evidence citations, and final answers. All GRPO variants start from the same SFT checkpoint and share the answer-correctness reward, format bonus, and excess-length penalty. The citation variant adds a correctness-gated citation-alignment reward, the temporal variant adds a temporal-order penalty, and the full model includes both. Both stages update only the LoRA parameters. Detailed reward definitions and optimization settings are provided in Appendix D.

5.1 MAIN RESULTS

ASSEMBLE maintains competitive answer accuracy while improving grounding. As shown in Table 1, ASSEMBLE achieves macro-averaged answer accuracy of 59.2% and Grounded accuracy of 25.4% across the three benchmarks. Compared with Gemini-2.5-Pro, ASSEMBLE improves answer accuracy by 0.9% and overlap-based Grounded accuracy by 6.7%, with higher Grounded accuracy on all three benchmarks. Relative to the Qwen3.5-9B base model, answer accuracy and Grounded accuracy improve by 8.8% and 7.2%, respectively. Because these comparisons use the same evidence catalogs and output budget, the grounding gap suggests that performance depends not only on access to relevant evidence, but also on how the reader uses and cites that evidence during reasoning. Controlled comparisons of training and inference are provided by the Qwen3.5-9B ablations and inference-policy controls below.

Grounding gains hold across different reasoning settings. Across all three benchmarks in Table 1, ASSEMBLE maintains answer accuracy close to Gemini-2.5-Pro while achieving higher overlap-based Grounded accuracy. The gain is largest on Video-Holmes at 9.7%, followed by 5.5% on VRBench and 4.8% on CG-Bench. These benchmarks cover different evidence demands, from connecting clues across distant video segments to multi-step reasoning over long narratives and broader mixtures of perception and reasoning.

Table 2 further evaluates temporal alignment using Union mIoU, citation IoU-F1, and Acc@GQA (Xiao et al., 2024). Union mIoU measures overlap between the temporal unions of pre-

dicted and annotated evidence, while citation IoU-F1 evaluates interval-level matching. Acc@GQA jointly evaluates answer correctness and evidence localization. ASSEMBLE achieves higher Acc@GQA point estimates on all three benchmarks, with the largest gain on VRBench. Union mIoU and citation IoU-F1 are mixed across benchmarks, showing that gains in grounded QA do not always coincide with finer temporal localization. Overall, these results suggest that the main benefit is more consistent alignment between correct predictions and relevant evidence, rather than uniformly tighter temporal boundaries. Full metric definitions are provided in Appendix E. Appendices F and G report transfer and evidence-intervention results.

5.2 Ablation

The ablation block of Table 1 compares supervised fine-tuning, narrative evidence, and GRPO reward design. Qwen3.5-9B is the unadapted reader, while + SFT is trained on filtered teacher responses containing rationales, citations, and answers. For Video-Holmes and CG-Bench, the default catalogs include narrative entries, and w/o narrative removes these entries at inference time while keeping the reader checkpoint fixed. For VRBench, the default catalog contains approximately 30-second local entries, with narrative augmentation evaluated separately under the same fixed reader. This catalog choice is held constant across the main VRBench comparisons. All GRPO variants start from the same SFT checkpoint and include the answer-correctness reward. outcome + citation adds citation alignment, outcome + temporal adds temporal-order supervision, and the full model includes both. Table 3 separately compares inference policies using the post-trained reader.

Table 2: **Grounding diagnostics (%).** Comparison between ASSEMBLE and Gemini-2.5-Pro. Bold indicates the better result. ASSEMBLE has higher Acc@GQA on all three benchmarks, while the temporal-alignment metrics are mixed.

Metric ↑	Model	Video-Holmes	VRBench	CG-Bench
Union mIoU	Gemini	**1.41**	**17.73**	7.92
	Ours	1.40	11.76	**9.15**
IoU-F1	Gemini	**1.86**	6.34	7.87
	Ours	1.42	**7.04**	**8.68**
Acc@GQA	Gemini	0.11	9.70	0.99
	Ours	**0.16**	**16.77**	**1.71**

SFT and GRPO play complementary roles. Relative to Qwen3.5-9B, SFT improves answer accuracy by 8.6% on Video-Holmes and 5.9% on VRBench, but does not consistently improve Grounded accuracy, motivating explicit citation-alignment supervision. Starting from the SFT model, outcome-only GRPO further improves accuracy to 54.4% and 76.4%, respectively, again with limited gains in Grounded accuracy. Together, these results show that SFT transfers the teacher's answering behavior and outcome-only GRPO sharpens answer selection, while citation alignment requires an explicit training signal.

Narrative evidence is most useful for dispersed clues. We keep the reader checkpoint fixed and vary only the evidence catalog at inference time. On Video-Holmes and CG-Bench, the *w/o narrative* rows remove narrative entries while retaining the local structured descriptions and atomic skills. The largest effect appears on Video-Holmes: removing narratives from the SFT model reduces Accuracy and Grounded accuracy by 5.2 and 5.0 percentage points, respectively, while the corresponding GRPO comparison reduces Grounded accuracy by 15.0 points. This pattern is consistent with the need to integrate clues across distant parts of the video. On VRBench, adding narrative entries to the default local-only catalog changes SFT Accuracy from 74.55% to 74.95% and Grounded accuracy from 25.25% to 23.43%, while the changes on CG-Bench are also smaller than on Video-Holmes. Overall, narrative evidence appears most useful when relevant clues are temporally dispersed rather than as a uniform improvement across benchmarks.

Citation and temporal rewards capture different aspects of supervision. Citation supervision gives the clearest single-component Grounded accuracy gain on Video-Holmes and VRBench, raising the metric over outcome-only GRPO by 5.0% and 2.5%, respectively. Temporal supervision alone does not consistently improve Grounded accuracy, but provides an ordering signal absent from citation alignment. The combined reward yields the highest measured Grounded accuracy among our models on all three benchmarks: 16.5% on Video-Holmes, 35.2% on VRBench, and 24.5% on CGBench. These results indicate that citation alignment and temporal ordering target different aspects of the measured answer-and-overlap criterion, with dataset-dependent effects.

Table 3: **Inference-time control study for atomic-skill application.** The reader checkpoint and evidence catalog are fixed across all settings. FREE TEXT uses unconstrained reasoning, FIXED SKILLS applies the same predefined skill sequence to every question, and DYNAMIC SKILLS selects a question-specific sequence. We report Accuracy, Grounded accuracy, and output tokens. **Takeaway:** With the reader and inference budget fixed, DYNAMIC SKILLS achieves the highest overlap-based Grounded point estimate on all three benchmarks, while accuracy remains similar to FIXED SKILLS.

	Video-Holmes			VRBench			CGBench		
Inference policy	**Accuracy ↑**	**Grounded ↑**	**Tokens**	**Accuracy ↑**	**Grounded ↑**	**Tokens**	**Accuracy ↑**	**Grounded ↑**	**Tokens**
FREE TEXT	53.40 ± 2.29	11.76 ± 1.47	662.7 ± 7.0	72.73 ± 3.94	26.67 ± 3.94	765.2 ± 28.3	38.91 ± 2.21	21.14 ± 1.88	722.0 ± 17.2
FIXED SKILLS	53.78 ± 2.29	12.63 ± 1.53	715.0 ± 12.4	76.57 ± 3.74	31.31 ± 4.04	731.9 ± 19.8	44.54 ± 2.29	23.90 ± 1.99	760.8 ± 19.1
DYNAMIC SKILLS (ASSEMBLE)	54.06 ± 2.26	$\mathbf{16.49 \pm 1.69}$	670.5 ± 10.1	78.99 ± 3.64	$\mathbf{35.15 \pm 4.24}$	669.0 ± 15.0	44.59 ± 2.29	$\mathbf{24.50 \pm 1.99}$	723.5 ± 16.1

5.3 STRUCTURED SKILL INFERENCE

Inference controls. Table 3 compares three policies using the same post-trained reader, inputs, decoding settings, and matched reader-call and output-token budgets. FREE TEXT uses unconstrained planning and evidence analysis without predefined skills. FIXED SKILLS applies the same sequence of evidence localization and support assessment to every question. DYNAMIC SKILLS selects a question-specific skill sequence during planning and executes it in order, following Section 4.2 and Algorithm 1. Reported token counts include completions from all reader calls; parsing, schema, and API failures are scored as incorrect. The atomic skill bank and qualitative execution traces are provided in Appendix B and I, respectively.

Dynamic Skills improves overlap-based grounding under matched inference budgets. Dynamic Skills achieves the highest Grounded point estimate on all three benchmarks, averaging 5.5% above FREE TEXT. Accuracy is similar to FIXED SKILLS, suggesting that the main gain is in evidence grounding rather than answer accuracy. Because FREE TEXT and DYNAMIC SKILLS also differ in how intermediate reasoning is structured, this comparison captures the effect of the overall inference design rather than skill selection alone. Temporal-localization metrics are mixed, so the grounding gains do not consistently translate into tighter temporal boundaries. Overall, structured inference helps the reader connect its predictions more consistently to relevant evidence without increasing model capacity or inference budget.

6 CONCLUSION

We introduced ASSEMBLE, an evidence-grounded video reasoning approach that combines grounding-aware reader training with structured atomic-skill inference over timestamped evidence catalogs. The method organizes long-video reasoning around reusable evidence, question-specific skill composition, and citation-aware post-training that ties predictions back to supporting video content. Using a 9B reader over precomputed evidence catalogs, ASSEMBLE achieves answer accuracy competitive with much larger models while improving overlap-based Grounded accuracy across three benchmarks. Controlled ablations show that SFT and outcome optimization mainly improve answer accuracy, whereas narrative evidence and grounding-aware rewards strengthen evidence grounding. Structured skill inference further improves grounding under matched inference budgets, highlighting the value of explicitly organizing how evidence is selected, combined, and used during reasoning.

Limitations. The video grounder remains an important bottleneck. Because the reader operates on precomputed evidence catalogs, evidence that is missed or misrepresented during catalog construction may be unavailable to downstream reasoning. Our experiments therefore evaluate reasoning over structured evidence rather than end-to-end retrieval from raw video.

Future work. We plan to extend atomic skills to video grounding for active evidence retrieval, refinement, and verification, and to make the skill library adaptive across new tasks and failure modes. More broadly, this points toward long-video reasoning systems that can refine both their evidence and reasoning processes as they operate.

AI USE STATEMENT

We used generative AI tools to assist with experimental design, method implementation, dataset processing, qualitative analysis, interpretation of results, figure generation, related-work search, English editing, LaTeX formatting, and code development. We did not use generative AI to generate synthetic datasets, develop theoretical frameworks, formulate mathematical claims or proofs, refine hypotheses, or perform translation. All AI-assisted outputs were reviewed by the authors, and generated code was tested for correctness.

REFERENCES

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

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20237–20246. IEEE, 2025a.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025b.

Kaiyan Chen, Junbin Xiao, and Xun Yang. Long-to-short video evidence reasoning for grounded question answering. *arXiv preprint arXiv:2609.15224*, 2026a.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025c.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4305–4320, 2026b.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12943. IEEE, 2024.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. Rewind: Understanding long videos with instructed learnable memory. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13734–13743. IEEE, 2025.

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. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2023.

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.

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R Fung. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026.

Handong Li, Longteng Guo, Zikang Liu, Dongze Hao, Yepeng Tang, Zijia Zhao, Jie Jiang, Zhiwei Jin, Chen Chen, Haonan Lu, et al. Timethink: Reasoning with time for video llms. *arXiv preprint arXiv:2607.05089*, 2026a.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026b.

Xiaoyuan Li, Moxin Li, Keqin Bao, Yubo Ma, Wenjie Wang, Dayiheng Liu, and Fuli Feng. Skillgraph: Skill-augmented reinforcement learning for agents via evolving skill graphs. *arXiv preprint arXiv:2605.12039*, 2026c.

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.

Zongxia Li, Dawei Liu, Fuxiao Liu, Yuhang Zhou, Xiyang Wu, Jingxi Chen, Jing Xie, Xiaomin Wu, and Lichao Sun. Comfyclaw: Self-evolving skill harnesses for image generation workflows. *arXiv preprint arXiv:2607.01709*, 2026d.

Huabin Liu, Filip Ilievski, and Cees GM Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3262–3271. IEEE, 2025.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 57481–57506, 2026.

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. Vited: Video temporal evidence distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8501–8511. IEEE, 2025.

Romy Luo, Zihui Sherry Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *Advances in Neural Information Processing Systems*, 38:83696–83727, 2026.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

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.

Qirui Mi, Zhijian Ma, Mengyue Yang, Haoxuan Li, Yisen Wang, Haifeng Zhang, and Jun Wang. Skill-pro: Learning reusable skills from experience via non-parametric ppo for llm agents. *arXiv preprint arXiv:2602.01869*, 2026.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, et al. Skillos: Learning skill curation for self-evolving agents. *arXiv preprint arXiv:2605.06614*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

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.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11854–11864. IEEE, 2023.

Fumihiko Tsuchiya, Taiki Miyanishi, Mahiro Ukai, Nakamasa Inoue, Shuhei Kurita, Yusuke Iwasawa, and Yutaka Matsuo. Ec-bench: Enumeration and counting benchmark for ultra-long videos. *arXiv preprint arXiv:2603.29943*, 2026.

Chendong Wang, Donglin Bai, Yifan Yang, Xiao Jin, Anlan Zhang, Rui Wang, Shiqi Jiang, Yuqing Yang, Hao Wu, Qi Dai, et al. Video-in-the-loop: Span-grounded long video qa with interleaved reasoning. *arXiv preprint arXiv:2510.04022*, 2025a.

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. *arXiv preprint arXiv:2410.03290*, 2024a.

Lan Wang, Yujia Chen, Du Tran, Vishnu Naresh Boddeti, and Wen-Sheng Chu. Seal: Semantic attention learning for long video representation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26192–26201. IEEE, 2025b.

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

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025c.

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

Xiyang Wu, Zongxia Li, Guangyao Shi, Alexander Duffy, Tyler Marques, Matthew Lyle Olson, Tianyi Zhou, and Dinesh Manocha. Co-evolving llm decision and skill bank agents for long-horizon tasks. *arXiv preprint arXiv:2604.20987*, 2026.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026a.

Sheng Xia, Zhengqin Lai, Tianxiang Jiang, Kanghui Tian, Shoujun Zhou, Bin Li, and Yi Wang. Ser: Learning to ground video reasoning with semantic evidence rewards. *arXiv preprint arXiv:2606.24726*, 2026b.

Tianle Xia, Lingxiang Hu, Yiding Sun, Ming Xu, Lan Xu, Siying Wang, Wei Xu, and Jie Jiang. Grasp: Graph-structured skill compositions for llm agents. *arXiv preprint arXiv:2604.17870*, 2026c.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

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

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21655–21666. IEEE, 2025.

Rongyi Yu, Chenyuan Duan, and Wentao Zhang. Longvidsearch: An agentic benchmark for multi-hop evidence retrieval planning in long videos. *arXiv preprint arXiv:2603.14468*, 2026.

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, 2024.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. Memskill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*, 2026a.

Kangning Zhang, Shuai Shao, Qingyao Li, Jianghao Lin, Lingyue Fu, Shijian Wang, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, et al. Mmskills: Towards multimodal skills for general visual agents. *arXiv preprint arXiv:2605.13527*, 2026b.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, Rajat Saini, Mustafa Erdogan, Hsuan-Kung Yang, Caixin Kang, Yifei Huang, Yoichi Sato, and Quan Kong. Cast-bench: Benchmarking causal chain-grounded spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2605.23216*, 2026c.

Ru Zhang and Weijie Qiu. Spyce: Skill-policy co-evolution for multimodal agents. *arXiv preprint arXiv:2607.13854*, 2026.

A Evidence Catalog Construction Details

Local descriptions. We use Qwen3.5-9B to describe sampled video windows and serialize the outputs as textual evidence entries. Video-Holmes uses approximately 4-second windows, while CG-Bench and VRBench use approximately 30-second windows.

Narrative construction. For Video-Holmes and CG-Bench, we additionally generate narrative entries with Qwen3-VL-235B-A22B-Instruct over 30- and 60-second windows, respectively, using 16 uniformly sampled frames per window. Frames are resized to a maximum width of 448 pixels. Each request includes the current frames, overlapping timestamped dialogue, the preceding narrative, and a running character list. The preceding narrative is truncated to 1,500 characters and the character list to 20 entries. Generation uses temperature 0 and a maximum of 1,200 output tokens. Dialogue is obtained from Whisper transcription for Video-Holmes and the available subtitle tracks for CG-Bench.

Catalog composition is fixed for each benchmark across all main comparisons. The default VR-Bench catalog contains only its approximately 30-second local entries; teacher-generated narratives are evaluated separately as a catalog augmentation with the reader checkpoint fixed. Video-Holmes and CG-Bench include narrative entries by default, and their W/O NARRATIVE variants remove only those entries at inference time.

Catalog serialization. Narrative entries precede local entries in the combined catalog. Each entry retains its source interval and is assigned a one-based `rank` i. The reader-facing citation `(clip i)` and case-study notation C_i refer to the same ranked entry and source interval; for example, `rank: 3`, `(clip 3)`, and C_3 identify the same entry. Descriptions are limited to 1,200 characters, and narratives are cached per video and reused across questions. Questions and reference answers are not used during catalog construction.

Input and output format. Teacher supervision, reader training, and evaluation use the same catalog serialization, with citation indices mapping directly to the corresponding source intervals.

Table 4: **Reader input and output format.** The reader consumes a question, answer options, and indexed catalog entries, then returns a citation-grounded rationale and answer label. The example is excerpted from Figure 5.

Field	Serialized example
Question	`What psychological state does the girl's action of gripping her phone tightly in the car reflect?`

Options	`A: Looking forward to news from friends.`
	`B: Afraid of being followed.`
	`C: Worry about the phone running out of power.`
	`D: Plan dinner menu.`
	`E: Recall childhood memories.`
	`F: Check time.`

Catalog entry	`rank: 3`
	`time_span: [63.0, 97.0] s`
	`description: The woman ... pulls out her phone and makes a call, her expression anxious as she asks, ''Hello? Will you come get me, please?''`

Rationale	`(clip 3) explicitly shows the woman making an anxious phone call. This supports a psychological state of fear about being followed. (clip 4) provides additional support.`
Answer	`label: B`

B Atomic Skills

Skill library. Long-video questions often require recurring operations such as locating relevant evidence, comparing hypotheses, tracking temporal or causal relations, identifying state changes,

and checking whether the retrieved evidence supports an answer. We capture these operations in a fixed library of 23 atomic skills, summarized in Table 5. Each skill defines an instruction together with structured inputs and outputs, including evidence references, support scores, status labels, and optional diagnostic fields. This gives the reader a consistent way to carry information across intermediate reasoning steps while keeping those steps inspectable.

Connection to the reader. The skills are prompt-defined interfaces rather than separate learned modules. At inference time, the same post-trained Qwen3.5-9B reader selects the skills relevant to a question, executes their instructions over the evidence catalog, and integrates the resulting intermediate records into its final prediction. SFT and GRPO train the reader for answer generation and citation alignment; no separate planner, executor, or verifier is introduced. Accordingly, *atomic* refers to the granularity of the interface rather than to an independently trained component or a guarantee of correctness. Names such as `verify_claim_support` simply identify particular operations in the library. In our experiments, inference uses five reader calls for planning, skill application, integration, and final answer generation. Multiple selected skills are handled within the skill-application stage, so the number of selected skills does not change the total reader-call count.

Table 5: **Atomic skill bank.** The 23 prompt-defined skill interfaces available to the planner, grouped by their primary function. The table documents the available operations and does not report usage frequency or individual causal contribution.

Skill interface	Description
Question analysis and answer construction	
`parse_question_target`	Extract target entities, events, constraints, and the required answer format.
`propose_evidence_roles`	Propose reusable evidence roles needed to answer the question.
`generate_answer_hypotheses`	Convert answer options or free-form targets into explicit candidate hypotheses.
`retrieve_evidence_for_hypothesis`	Retrieve supporting evidence for a single candidate hypothesis.
`score_hypothesis_support`	Score a hypothesis using supporting evidence and counterevidence.
`compare_hypotheses`	Compare scored hypotheses and select the best-supported answer candidate.
`verify_claim_support`	Assess whether an evidence chain supports a claim.
`commit_answer`	Map an assessed claim to the final answer and record its supporting evidence.
Retrieval and localization	
`retrieve_by_event`	Retrieve event or evidence nodes matching an event description.
`localize_clue`	Select the most relevant clue span or node for a requested role.
`retrieve_by_time`	Retrieve evidence around an anchor event or within a time window.
`retrieve_by_entity`	Retrieve an entity's timeline, history, or related evidence.
`retrieve_by_relation`	Query graph paths or relation edges.
Inference and specialized checks	
`verify_temporal_social_consistency`	Check temporal ordering and social plausibility in a hypothesis's evidence chain.
`bridge_evidence_hops`	Construct a short multi-hop evidence bridge from source references to a hypothesis.
`infer_intention_or_motive`	Infer an agent's intention, goal, or motive from actions and context.
`assign_evidence_role`	Bind evidence to a semantic role.
`infer_causal_relation`	Infer causal relations between events or states.
`extract_claim`	Extract a claim from dialogue, annotations, or evidence text.
`compose_evidence_chain`	Assemble role-labeled evidence into a chain supporting an answer.
`infer_temporal_relation`	Infer temporal relations among events, including precedence and overlap.
`infer_state_change`	Infer changes between an entity's or object's earlier and later states.
`infer_social_contradiction`	Infer conflicts between statements, alibis, or promises and later actions or evidence.

C Data Splits and Training Statistics

Dataset splits. Video-Holmes uses its official train and test splits, with 1,551 training questions from 233 videos and 1,837 test questions from 270 disjoint videos. VRBench has no official split, so we use a video-level partition: 480 questions from 60 videos form the SFT candidate pool, and 495 questions from a disjoint set of 60 videos are used for evaluation. For CG-Bench, we construct a training pool of 1,080 questions from 161 videos and evaluate on 1,812 questions from 823 videos in CG-Bench-mini, a commonly used subset of CG-Bench. Training and evaluation sets are disjoint at both the question and source-video levels across all three benchmarks. No evaluation question or video is used for SFT or GRPO. Reader inputs contain only the question, answer options, evidence

catalog, and output-format instruction. Gold answers and evidence annotations are used only for filtering, reward computation, and evaluation.

Training statistics. The combined SFT candidate pool contains 3,111 questions from 454 videos. Teacher-response filtering yields 4,520 response-level examples covering 1,879 unique questions. After length filtering, 4,416 responses covering 1,830 questions are used for SFT. All GRPO variants train only on the official Video-Holmes training split of 1,551 questions from 233 videos. VRBench and CG-Bench are not used for GRPO.

Evaluation protocol. All methods are evaluated on the same fixed question set for each benchmark using the same answer-parsing and scoring procedure. Evaluation sets are fixed before scoring and are not filtered based on model outputs or correctness.

D Reader Training and Reward Definitions

SFT warm start. We fine-tune Qwen3.5-9B on responses generated by Qwen3-VL-235B-A22B-Instruct, each containing a rationale, evidence citations, and a final answer. For each of the 3,111 candidate questions, we generate responses using the original answer-option order and two shuffled orders. We retain responses with a correct final answer and a nonempty rationale; when gold evidence spans are available, we additionally require citation precision of at least 0.5. This yields 4,520 response-level examples covering 1,879 unique questions. After length filtering, 4,416 examples covering 1,830 unique questions are used for SFT.

We train for one epoch with completion-only loss using LoRA with rank 16, scaling factor 32, and dropout 0.05. The learning rate is 10^{-4}, with a per-device batch size of one, gradient accumulation of eight, and a maximum sequence length of 32,768. Thinking mode is disabled.

Reward definitions. For a generated response z, let A denote answer-label correctness and P_{cite} denote citation precision, defined as the fraction of valid cited catalog entries whose time spans overlap at least one annotated evidence span. Citations are deduplicated, and invalid indices are ignored. The citation contribution is set to zero when no valid citations or evidence annotations are available.

Let F indicate that the output contains both a parsed answer and a nonempty rationale, and define the excess-length penalty as

$$L = \max\left(0, \frac{\ell_{\text{char}}(z) - 2500}{2500}\right).$$

Temporal-order questions are identified by dataset question-type codes beginning with `TA`. For these questions, $T = 1$ if the start times of valid cited entries, in their cited order after deduplication, are not nondecreasing; otherwise, $T = 0$. We also set $T = 0$ for other question types and for responses with fewer than two valid citations. We consider four GRPO reward configurations:

$$R_{\text{out}} = A + 0.1F - 0.05L, \tag{8}$$

$$R_{\text{out+cite}} = A + 0.5AP_{\text{cite}} + 0.1F - 0.05L, \tag{9}$$

$$R_{\text{out+temp}} = A - 0.2T + 0.1F - 0.05L, \tag{10}$$

$$R_{\text{full}} = A + 0.5AP_{\text{cite}} - 0.2T + 0.1F - 0.05L. \tag{11}$$

The citation reward is gated by answer correctness, so citation alignment cannot compensate for an incorrect answer. The four variants differ only in whether citation alignment and temporal-order consistency are included. Citation precision measures temporal overlap with annotated evidence rather than semantic entailment.

GRPO post-training. We run each GRPO configuration with three random seeds and report the mean across runs. All variants start from the same SFT checkpoint and train only on the 1,551 Video-Holmes training questions. Each batch contains four questions, with eight sampled responses per question. Rollouts use temperature 1.0 and a maximum of 1,024 generated tokens. We train for 384 optimization steps and evaluate the checkpoint saved at step 384.

Rewards are standardized within each question group, and groups with near-zero reward variance are skipped. We optimize a token-level clipped policy-gradient objective with clipping parameter

Table 6: **Temporal grounding and citation duration.** Grounding metrics are percentages; ASSEMBLE denotes Dynamic Skills. Sum and Union are mean citation durations before and after merging overlapping intervals. Video is the mean fraction of video duration covered by the citation union. Questions without valid citations remain in the denominator.

Benchmark	Inference policy	IoU-P	IoU-R	IoU-F1	Union mIoU	Sum (s)	Union (s)	Video (%)
Video-Holmes	Free Text	1.12	2.46	1.37	1.33	95.48	92.21	53.18
	Fixed Skills	1.22	2.86	1.56	1.40	108.69	105.00	60.22
	ASSEMBLE	1.18	2.32	1.42	1.40	93.20	90.78	52.45
VRBench	Free Text	5.72	9.28	6.03	12.15	233.19	222.83	12.26
	Fixed Skills	6.34	10.81	7.28	14.13	220.91	212.11	11.78
	ASSEMBLE	7.21	9.04	7.12	11.74	171.16	164.60	9.06
CG-Bench	Free Text	7.36	11.68	8.34	8.78	126.85	119.31	8.77
	Fixed Skills	6.56	15.38	8.70	8.92	167.61	151.39	10.68
	ASSEMBLE	6.57	14.73	8.62	9.10	165.13	149.83	10.65

0.2. LoRA parameters are updated with AdamW using a learning rate of 10^{-5}, betas $(0.9, 0.99)$, ten warm-up steps, and gradient-norm clipping at 1.0. We use neither a KL penalty nor an auxiliary SFT loss. Gold answers and evidence annotations are used only to compute rewards and are never included in the reader input. SFT and GRPO optimize complete generated responses for answer correctness and citation alignment; skill selections and intermediate outputs are not separately labeled or optimized. At inference time, the skill interfaces structure the operations performed by the same reader.

E Temporal Grounding and Citation Analysis

We further evaluate temporal grounding and citation behavior beyond the overlap-based Grounded metric used in the main experiments. We rescore the existing predictions on Video-Holmes, VRBench, and CG-Bench for Free Text, Fixed Skills, and ASSEMBLE (Dynamic Skills), using the same reader checkpoint and inference budget.

Temporal grounding metrics. Let P and G denote the valid predicted and annotated intervals after removing exact duplicates, and let n_{invalid} be the number of invalid citation IDs. We compute a maximum-weight one-to-one matching M with interval IoU as the edge weight:

$$W = \sum_{(p,g)\in M} \text{IoU}(p, g).$$

We then define

$$\text{IoU-P} = \frac{W}{|P| + n_{\text{invalid}}}, \qquad \text{IoU-R} = \frac{W}{|G|}, \qquad \text{IoU-F1} = \frac{2\,\text{IoU-P}\,\text{IoU-R}}{\text{IoU-P} + \text{IoU-R}}.$$

Invalid citations receive zero matching weight but remain in the precision denominator. Metrics are computed per question and then averaged. Union mIoU is the IoU between the temporal unions of the predicted and annotated intervals.

Citation duration and evidence type. The inference policies produce different citation profiles. Fixed Skills generally cites broader temporal regions, whereas Dynamic Skills uses shorter citation unions on Video-Holmes and VRBench. The corresponding precision and recall differences show that citation breadth changes the balance between focused evidence and broader temporal coverage.

We also separate local and narrative entries by catalog type. Type-specific scores use the full benchmark denominator, assigning zero when a question cites no entry of that type. Finally, we group questions by total citation-union duration to examine how grounding quality varies with citation breadth.

Table 7: **Local and narrative citation analysis for ASSEMBLE.** Grounding scores are percentages and use the full benchmark denominator. A question may cite both entry types. Type-specific union durations need not sum to the overall union because local and narrative intervals may overlap. The default VRBench catalog contains only local entries; adding narrative entries is evaluated separately in the catalog ablation.

Benchmark	Citation type	Questions citing type	Union mIoU	IoU-F1	Union (s)
Video-Holmes	Local	208	0.36	0.35	0.92
	Narrative	1778	1.38	1.26	90.54
VRBench	Local	476	11.74	7.12	164.60
CG-Bench	Local	705	5.27	4.85	19.28
	Narrative	1641	8.93	8.57	143.10

Table 8: **Citation-duration stratification for ASSEMBLE.** IoU-F1 is reported as a percentage. Each question is assigned to one bin according to its total citation-union duration.

Citation union (s)	Video-Holmes		VRBench		CG-Bench	
	n	IoU-F1	n	IoU-F1	n	IoU-F1
0	38	0.00	9	0.00	106	0.00
$(0, 30]$	38	2.75	5	10.03	17	9.48
$(30, 60]$	388	1.73	178	8.32	406	12.78
> 60	1363	1.33	293	6.56	1273	7.99

Results and interpretation. These analyses complement the main evaluation by separating citation alignment from answer correctness. Dynamic Skills improves citation IoU-F1 over Free Text on all three benchmarks, while Fixed Skills obtains slightly higher IoU-F1 with generally broader citation coverage. Union mIoU shows no consistent ordering across inference policies, indicating that no single policy dominates fine-grained temporal localization.

The duration analysis further shows that broader citations are not uniformly better: intermediate citation spans often achieve higher IoU-F1 than citations extending beyond 60 seconds. The type analysis also highlights the importance of narrative evidence on Video-Holmes and CG-Bench, where most cited temporal coverage comes from narrative entries. Overall, these results suggest that inference structure shapes not only whether the model cites relevant evidence, but also how broadly that evidence is localized; improvements in answer–evidence grounding therefore need not correspond to tighter temporal boundaries.

F Additional Evidence-Grounding Benchmarks

We further evaluate whether the reader's grounding behavior transfers beyond the three main benchmarks. These experiments cover different answer formats, grounding annotations, and video domains. We compare the Base, SFT, and ASSEMBLE readers to examine the effect of post-training, with Gemini-2.5-Pro included as a shared-catalog large-model reference.

Benchmark coverage and evaluation setup. We evaluate on three additional evidence-grounding benchmarks. NExT-GQA (Xiao et al., 2024) combines multiple-choice video QA with temporal localization in untrimmed videos; EG-VQA (Huang et al., 2026) evaluates open-ended answers with fine-grained temporal and semantic evidence; and GroundVQA (Di & Xie, 2024) evaluates grounding in long egocentric videos.

We use 5,553 questions from 990 videos on NExT-GQA, 2,889 questions from 336 videos on EG-VQA, and 500 questions from 148 videos under the GroundVQA CloseQA protocol. Each GroundVQA question is evaluated under five answer-option permutations. Base, SFT, ASSEMBLE, and Gemini-2.5-Pro use the same frozen catalogs and output budget. Base, SFT, and ASSEMBLE addi-

Table 9: **Answer accuracy and evidence grounding on three additional benchmarks.** We compare Base, SFT, ASSEMBLE, and Gemini-2.5-Pro on NExT-GQA, EG-VQA, and GroundVQA under the same evidence catalogs and the same output budget. Base, SFT, and ASSEMBLE additionally share the inference procedure and differ only in the reader checkpoint. **Takeaway:** Relative to SFT, ASSEMBLE shows stronger grounding on NExT-GQA and improves relaxed accuracy and EG-F1 on EG-VQA, while differences on GroundVQA are small.

Method	NExT-GQA			EG-VQA			GroundVQA		
	Accuracy ↑	mIoU ↑	Acc@GQA ↑	Relaxed Acc. ↑	Strict Acc. ↑	EG-F1 ↑	Accuracy ↑	mIoU ↑	Citation IoU-F1 ↑
Large-model baselines									
GEMINI-2.5-PRO	$\mathbf{67.24 \pm 1.42}$	$\mathbf{25.78 \pm 0.98}$	$\mathbf{11.94 \pm 1.19}$	$\mathbf{47.53 \pm 2.49}$	$\mathbf{35.31 \pm 2.36}$	$\mathbf{42.20 \pm 1.81}$	$\mathbf{51.32 \pm 4.18}$	$\mathbf{4.95 \pm 0.95}$	$\mathbf{6.01 \pm 0.81}$
ASSEMBLE and ablations (Qwen3.5-9B)									
QWEN3.5-9B	64.81 ± 1.35	24.50 ± 0.89	$\mathbf{12.08 \pm 1.08}$	45.50 ± 2.24	31.29 ± 2.25	38.53 ± 1.70	50.20 ± 3.63	5.49 ± 0.99	6.03 ± 0.86
+ SFT	$\mathbf{66.50 \pm 1.35}$	21.16 ± 0.80	9.51 ± 0.99	46.47 ± 2.26	32.88 ± 2.24	38.78 ± 1.71	52.56 ± 3.64	$\mathbf{5.58 \pm 1.08}$	6.11 ± 0.99
ASSEMBLE (FULL)	65.48 ± 1.37	$\mathbf{25.95 \pm 0.93}$	11.63 ± 1.16	$\mathbf{48.20 \pm 2.27}$	$\mathbf{34.20 \pm 2.28}$	$\mathbf{41.54 \pm 1.81}$	$\mathbf{53.24 \pm 3.32}$	5.22 ± 1.00	$\mathbf{6.43 \pm 0.97}$

tionally share the same Dynamic Skills inference procedure and decoding settings, differing only in the reader checkpoint. Prediction failures remain in the evaluation denominator.

Evaluation metrics. We use each benchmark's native metrics rather than the Grounded criterion from the main experiments.

NExT-GQA. We report answer accuracy, temporal mIoU, and Acc@GQA. Acc@GQA requires a correct answer and temporal intersection-over-prediction of at least 0.5, where intersection-over-prediction is the fraction of the predicted interval covered by the annotated evidence.

EG-VQA. We report relaxed and strict answer accuracy using the benchmark judge, together with EG-F1 for evidence grounding. Following the benchmark scorer, EG-F1 matches evidence using temporal IoU ≥ 0.3 and semantic similarity ≥ 0.5.

GroundVQA. Answer accuracy is averaged over five answer-option permutations per question. For temporal mIoU, cited intervals are merged into their enclosing interval and scored with the benchmark evaluator. We additionally report Citation IoU-F1 over individual cited intervals. Let M denote the total IoU under maximum-weight one-to-one matching between predicted and annotated intervals, with n_p predicted and n_g annotated intervals. We define citation precision and recall as $P = M/n_p$ and $R = M/n_g$, and report their harmonic mean, averaged over questions and option permutations. Unlike the thresholded EG-VQA evidence F1, this metric retains the continuous IoU of matched intervals.

Results. Post-training changes grounding behavior beyond the three main benchmarks, although the effects vary across datasets and metrics. On NExT-GQA, ASSEMBLE improves temporal grounding and joint grounded-QA performance over SFT, with a small decrease in answer accuracy. On EG-VQA, the clearest gains appear in relaxed answer accuracy and evidence grounding, while results are less consistent under the stricter criterion. On GroundVQA, SFT and ASSEMBLE perform similarly, indicating weaker transfer to the egocentric setting.

The results also show that SFT and GRPO do not improve every metric monotonically. GRPO is trained only on Video-Holmes, while these benchmarks use different answer formats and grounding criteria. The observed transfer is therefore strongest for some forms of answer–evidence alignment rather than for temporal localization or answer accuracy uniformly.

Comparison with Gemini-2.5-Pro. The comparison with Gemini-2.5-Pro also depends on the evaluation metric. Our reader performs better on some grounding measures, while Gemini retains advantages on others. Because both use the same evidence catalogs and output budget, these differences reflect how the readers use the available evidence under the shared evaluation protocol.

Table 10: **Evidence dependence under catalog interventions.** Accuracy (%) after retaining cited entries, removing cited entries, or removing a matched amount of catalog content. Each intervention reruns the complete inference pipeline on the modified catalog.

Catalog condition	Video-Holmes	VRBench	CG-Bench
Full catalog	54.06	78.99	44.59
Cited-only	52.49	79.79	45.89
Remove-cited	50.36	75.26	37.46
Matched random removal	52.49	77.94	43.06

G Evidence Dependence under Catalog Interventions

We test whether the reader's predictions depend on the catalog entries it cites. All conditions use the same five-call Dynamic Skills pipeline, reader checkpoint, decoding settings, and output budget (DeYoung et al., 2020).

Experimental setup. Starting from each full-catalog prediction, we construct three interventions based on its cited entries. **Cited-only** retains only the cited entries, **Remove-cited** removes them, and **Matched random removal** removes the same number of randomly selected entries. For the random control, we choose among 256 seeded candidate subsets to approximately match the cited entries in union duration, total duration, and text length; the selected subset may overlap with the cited entries.

After each intervention, catalog entries are renumbered and the full pipeline is rerun from scratch without reusing the original answer or intermediate states. The same questions are evaluated in every condition, and failures are scored as incorrect.

Results and interpretation. The interventions show a clear asymmetry between retaining and removing cited evidence. Keeping only the cited entries preserves most of the full-catalog accuracy, whereas removing them is consistently more harmful than removing a matched amount of other catalog content. This indicates that the cited evidence is behaviorally relevant to the reader's predictions rather than merely attached after the answer is formed.

Performance does not collapse when cited entries are removed, and cited-only inputs do not exactly reproduce the full-catalog results. This is expected when catalog entries contain overlapping information and the rerun pipeline can recover alternative evidence. The intervention therefore supports a dependence between citations and predictions without requiring each cited entry to be uniquely necessary for the final answer.

H Prompt Templates

Tables 11 and 12 show representative prompts for catalog construction, evidence assessment, and final answer generation. The local-describer prompt is shown in excerpted form, while the others are reproduced in full. Braced fields are populated at runtime.

I Qualitative Atomic-Skill Traces

We present six successful cases and two failure cases from CG-Bench, Video-Holmes, and VR-Bench. The successful cases (Figures 2–7) illustrate different combinations of evidence localization, causal and intention reasoning, and support assessment. The failure cases (Figures 8 and 9) show two common limitations: insufficient visual evidence and cases where an unsupported hypothesis still influences the final answer because support assessment is not a hard constraint. Figure 10 provides an additional case where the answer is correct and grounded, but the cited evidence does not cover all annotated intervals.

Table 11: **Representative prompts for evidence-catalog construction.** The local describer records clip-level observations, while the narrative describer summarizes events using frames, dialogue, and preceding context. The local prompt is excerpted and retains the original rule numbering.

Stage	Prompt

Local describer	`You convert one video clip span into a structured perception record.`
	`1. Use only information supported by the clip frames or provided subtitle/context text.`
	`2. Do not invent characters, objects, or events.`
	`3. Prefer clue-oriented noun phrases over generic captions: objects, colors, place, repeated-looking props, screen text, spoken clues, temporal changes.`
	`5. For social questions, record only visible social cues: facial expression, gaze direction, hesitation-like motion, distance, posture, gesture, or group interaction. Do not infer private motives unless the visual evidence is clear.`
	`7. If nothing is visible, return empty lists and a cautious scene_description.`

Narrative describer	`You are annotating one stretch of a short film for a detective-style question set. You are given evenly spaced frames from that stretch in time order, the dialogue heard in it (if any), the previous stretch's annotation and the running cast. Write ONE paragraph, the way a careful human annotator would, of what happens: who is present (stable descriptors reused from the cast, e.g. 'the man with the backpack'), what each person does and says, how they react, what changes, and what the sequence shows or implies about intentions, relationships and cause-and-effect when the frames and dialogue make it clear. Quote or paraphrase dialogue in English. Describe events, not camera work. Do not invent what is not shown or heard. Reply with JSON only: {"narrative": "<paragraph>", "cast": ["<stable descriptor>", ...]}.`

Table 12: **Representative prompts for evidence assessment and answer generation.** The first scores evidence for a candidate hypothesis; the second produces the final answer and citation references from the catalog and accumulated notes.

Stage	Prompt
Evidence support	`How well does the evidence support this hypothesis? Hypothesis: '{hypothesis}' Supporting evidence: {support} Counterevidence: {counter} Answer with JSON: {{"support_score": 0.0-1.0, "contradiction_score": 0.0-1.0, "reasoning": "..."}}`
Final answer	`Answer the original multiple-choice question using the catalog and the accumulated notes. Notes are fallible; use the evidence to decide. Return only JSON {"reasoning":"reason step by step, citing evidence as clip 1, clip 2, etc.","label":"option letter"}.`

CG-Bench: Is the coffee shop open?

Question: **In the video, what happened when the protagonist came to the "BEAR POND" store?**

Selected video frames



C32 / 1808.0 s
Approaching the coffee shop.



C32 / 1814.0 s
The door displays a CLOSED sign.



C105 / 1816.0 s
The sign and the shop's notice.



C105 / 1818.0 s
The subtitle says the shop is closed.

Evidence from the catalog

C32 / [1792, 1850] s
The narrator arrives at Bear Pond Espresso, finds it closed, and goes to another coffee shop.

C105 / [1792, 1822] s
The entry describes the shop name, a "CLOSED" sign on the door, and a subtitle saying that the shop seems closed.

Planner-selected skills

Plan: `localize_clue` → `score_hypothesis_support`

Why these skills: **Localization** finds the entry for this particular visit. **Support assessment** then evaluates whether it establishes that the shop is closed. This matches the question's request for an observed state.

Candidate: F. The shop is not open. **Starting references:** C32, C105.

1. Localize the clue. Given C32 and C105, the localizer selects only C32 as its `clue_refs`, with confidence **0.95** and status **assessed as supported**.

2. Score the hypothesis. Using C32 and the localizer's result, the scorer assigns "The shop is not open" a support score of **1.00** and a contradiction score of **0.00**. Its status is **assessed as supported**.

Support assessment and final answer

3. Support-assessment output: `verify_claim_support` checks C32 and the scoring result. It returns **assessed as supported**, score **1.00**, `passed=true`, and `supported_by_refs` = [C32].

4. Final answer: F. The shop is not open. [C32, C105]

Gold answer: F. **Logged result:** `ok=true`, `grounded=true`.

Discussion: **Localization changes what the next calls receive.** The reference set shrinks from C32, C105 to C32, and both scoring and support assessment use that selected entry. The resulting support score is tied to a specific description of the closure. The final-answer call still sees the full catalog and cites both entries. Recording these stages separately lets us trace which evidence was used to check the answer and which evidence appeared in the final response.

Figure 2: **CG-Bench: Is the coffee shop open?** Localization narrows the evidence passed to both downstream checks. The intermediate trace distinguishes those inputs from the citations in the final answer.

CG-Bench: Why is the player stuck?

Question: **Why couldn't the player move the next day in the video?**

Selected video frames



C9 / 516.0 s
Entering the narrow passage.



C9 / 524.0 s
The subtitle says the web can trap people.



C48 / 526.0 s
The player says he cannot move.



C48 / 527.0 s
A web spans the passage.

Evidence from the catalog

C9 [476, 534] s describes Max entering a narrow passage with a spiderweb and saying that he cannot move. **C48 [504, 534] s** also describes the web across the passage and a subtitle mentioning it.

Planner-selected skills

Plan: `localize_clue` → `infer_causal_relation`

Why these skills: The question asks for a cause. **Localization** selects the encounter in which the player becomes stuck; **causal inference** connects the web to his inability to move. Finding the scene and explaining the event are separate tasks.

Skill outputs: The planner proposes **B** using C9, C48. The localizer keeps C9 (confidence **0.95**). The causal skill then uses C9 and the localizer's result to infer that the web prevents the player from moving (confidence **0.98**). Both skills return **assessed as supported**.

Support assessment and final answer

Support-assessment output: **assessed as supported**; score **1.00**; supported references: C9.

Final answer: **B. Caught in a spider's web. [C9, C48]**

Gold answer: B. **Logged grounding:** `true`.

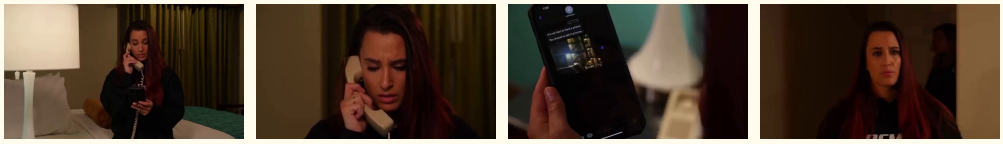
Discussion: **The same localization skill supports a different next step.** For the shop question, the planner checks a state claim; here it selects causal inference to explain a physical consequence. C9 contains both the web encounter and the player's complaint. Passing this entry and the localizer's result to the next call ties the explanation to the selected event. The causal output has empty `evidence_refs`, but the input and support-assessment output still identify C9, so the evidence used for the causal judgment remains traceable.

Figure 3: **CG-Bench: Why is the player stuck?** The planner reuses localization, then selects causal inference to explain why the player is stuck. The selected event supplies both the obstacle and its effect.

Video-Holmes: How was Kate's location exposed?

Question: **What is the direct reason for Kate's position being ultimately exposed?**

Selected video frames



C3 / 84.0 s
Kate holds a receiver and a mobile phone.

C37 / 93.0 s
Close-up during the phone call.

Supplement / 95.0 s
A message says phones are easy to hack.

Supplement / 145.0 s
A visitor appears at the doorway.

Supplementary context: The 95 s and 145 s frames are outside C3 and C37, which both end at 94 s. Neither frame belongs to the intervals cited in the original run.

Evidence from the catalog

C3 [63, 94] s and **C37 [90, 94] s** describe a threatening message: a phone is easy to hack and should have been left at home. The entries disagree about the attached image, so we retain only the message they both describe.

Planner-selected skills

Plan: `infer_causal_relation` → `score_hypothesis_support`

Why these skills: **Causal inference** turns the message about hacking into an explanation of the exposed location. **Support assessment** evaluates whether C3, C37 substantiate that explanation. The sequence separates proposing a cause from assessing its evidence.

Skill outputs: The planner proposes **C** using C3, C37. The causal skill infers that a phone hack revealed Kate's location (confidence **0.95**). Using that result and the same references, the scorer assigns support **0.95** and contradiction **0.00**. Both skills return **assessed as supported**.

Support assessment and final answer

Support-assessment output: **assessed as supported**; score **0.95**; supported references: C3, C37.

Final answer (condensed): **C. Her location was exposed through a phone hack. [C3, C37]**

Gold answer: C. **Logged grounding:** `true`.

Discussion: **The proposed cause becomes an explicit claim to check.** The first skill attributes the exposure to a phone hack; the second checks that claim against the same catalog entries, reporting support **0.95** and contradiction **0.00**. This exposes the inference being evaluated and the evidence used to evaluate it, beyond the final choice of **C**. The causal output itself has empty `evidence_refs`; the scoring and support-assessment outputs record both references. Because the intervals overlap, their agreement should not be read as two independent observations.

Figure 4: **Video-Holmes: How was Kate's location exposed?** Causal inference proposes the phone-hacking explanation, and the support assessment evaluates it against the cited message. The supplementary frames do not alter the recorded evidence.

Video-Holmes: Why does the girl seem afraid?

Question: What psychological state does the girl's action of gripping her phone tightly in the car reflect?

Selected video frames



C3 / 76.0 s
The girl talks on a phone outdoors.



C3 / 90.0 s
She turns toward the driver.



C4 / 99.0 s
Both people are seated in the car.



C4 / 115.0 s
She looks back after entering.

Evidence from the catalog

In **C3 [63, 97] s**, the girl asks to be picked up and says that she feels she is being followed. **C4 [96, 127.98] s** describes her conversation with the driver about seeking help if she feels unsafe. She remains nervous and is startled during the ride.

Planner-selected skills

Plan: `infer_intention_or_motive` → `score_hypothesis_support`

Why these skills: The grip described in the question could have several meanings. **Motive inference** uses the call and conversation to infer the girl's state of mind; **support assessment** evaluates the proposed fear of being followed against those entries.

Skill outputs: The planner proposes **B** using C3, C4. The first skill infers that she is afraid of being followed (confidence **0.95**). Using this result and the same references, the scorer assigns support **0.95** and contradiction **0.00**. Both skills return **assessed as supported**.

Support assessment and final answer

Support-assessment output: assessed as supported; score **0.95**; supported references: C3, C4.

Final answer: B. Afraid of being followed. [C3, C4]

Gold answer: B. **Logged grounding:** `true`.

Discussion: The answer draws on context across two entries. The request to be picked up and the feeling of being followed in C3 suggest a threat; the conversation in C4 adds her continued nervousness. The first skill also records alternatives: expecting news from friends and worrying about battery life. The next skill assigns support to the proposed fear explanation using the same evidence. This makes the interpretation and its support separately inspectable. The tight grip itself is not described in the catalog or clearly shown in these frames; the answer is supported by the dialogue.

Figure 5: **Video-Holmes: Why does the girl seem afraid?** Motive inference interprets the surrounding dialogue, while the support assessment evaluates the resulting fear explanation. The recorded evidence supports the emotional context, not the hand gesture itself.

VRBench: Why are some children told not to sing?

Question: What is the reason that certain kids in the choir have to be silent during the competition?

Selected video frames



C38 / 1046.0 s
Teacher: we cannot be the best if everyone sings.



C38 / 1052.0 s
Teacher: some children do not sing well.



C39 / 1066.0 s
The teacher asks who was told not to sing.



C39 / 1074.0 s
Children raise their hands.

Evidence from the catalog

C38 / [1036, 1066] s: The teacher gives her reason for keeping some children silent: "but if we let everyone sing, we will never be the best."

C39 / [1064, 1094] s: She asks the group: "Well kids, who did I tell not to sing out loud?"

Planner-selected skills

Plan: `infer_causal_relation` → `infer_intention_or_motive`

Why these skills: The question links a rule to the teacher's objective. **Causal inference** connects silencing weaker singers to her expectation of winning; **motive inference** then uses that relation to explain what she hopes to achieve.

Candidate: D. The teacher believes silencing weaker singers is necessary to win.

References: C38, C39 for both skills.

1. Infer the cause. The causal skill links the instruction to the teacher's stated goal of winning. It returns `causal=true`, confidence **0.95**, and status **assessed as supported**.

2. Infer the motive. Using the first result and C38, C39, the motive skill reads the rule as an effort to win by excluding weaker singers. It returns confidence **0.95** and status **assessed as supported**.

Support assessment and final answer

3. Support-assessment output: `verify_claim_support` checks C38, C39 and the second skill's result. It accepts the candidate with score **0.95**, `passed=true`, and `supported_by_refs` = [C38, C39].

4. Final answer (condensed): D. The teacher believes silencing weaker singers is necessary to win. [C38, C39]

Gold answer: D. **Logged result:** `ok=true`, `grounded=true`.

Discussion: The sequence connects an instruction, its expected effect, and the teacher's goal. C39 identifies the instruction, while C38 supplies her stated reason. The motive skill receives the causal result together with those references, forming an explicit dependency that can be inspected. The resulting answer explains why she imposes the rule, beyond identifying which children stay silent. Both intermediate outputs have empty `evidence_refs`; the input bindings and support-assessment output preserve C38, C39. The reasoning concerns one conversation and the teacher's expectations about winning.

Figure 6: **VRBench: Why are some children told not to sing?** The two skills connect the teacher's instruction to its expected benefit and her competitive goal. The second call uses the first result as well as the original evidence.

VRBench: Why end with a skidding car?

Question: **Why was a car skidding episode inserted at the end of the short movie?**

Selected video frames



C48 / 1335.0 s
Calendars show winter months.

C48 / 1343.0 s
Tire types under the "Important" heading.

C49 / 1348.0 s
Close-up of the M+S marking.

C49 / 1360.0 s
The heading reads "All-season tires."

Evidence from the catalog

C48 [1316, 1346] s describes police on a snowy street, followed by winter-tire graphics, a December–February calendar, and an M+S marking. **C49 [1344, 1374] s** describes more tire graphics, a June–August calendar, and text about all-season tires.

Planner-selected skills

Plan: `infer_intention_or_motive` → `score_hypothesis_support`

Why these skills: The question concerns the filmmaker's purpose. **Motive inference** explains why the skidding scene was included, using the surrounding tire advice. **Support assessment** evaluates whether that interpretation follows from the cited calendars and tire markings.

Skill outputs: The planner proposes **A** using C48, C49. The motive skill reads the ending as advice about tire safety (confidence **0.95**). Using this result and the same references, the scorer assigns support **0.95** and contradiction **0.00**. Both skills return **assessed as supported**.

Support assessment and final answer

Support-assessment output: **assessed as supported**; score **0.95**; supported references: C48, C49.

Final answer (condensed): **A. Explain road safety and remind viewers to change to winter tires in time. [C48, C49]**

Gold answer: A. **Logged grounding:** `true`.

Discussion: **Skill selection follows the kind of explanation requested.** A physical account of skidding would leave the reason for inserting the scene unanswered. The planner instead asks what the ending communicates to viewers. Winter months, the M+S marking, and all-season tire graphics connect the scene to practical advice across C48, C49. The scorer then checks the specific educational claim against those details. The creator's purpose remains an inference, and the intermediate records make clear which observations are used to support it.

Figure 7: **VRBench: Why end with a skidding car?** The planner treats the question as one about communicative intent. Motive inference links the ending to tire advice, and the support assessment evaluates that interpretation.

Failure / CG-Bench: An unsupported sticker count

Question: How many green puppy stickers are there on the coffee machine when the girl is making coffee in the video?

Selected video frames



C2 / 57.0 s
The machine is partly obscured.



C2 / 65.0 s
A glass with a dog graphic.



C2 / 94.0 s
Milk is poured beside the machine.



C18 / 100.0 s
Latte art in the green cup.

Evidence from the catalog

C2 [56, 114] s mentions coffee preparation and a cartoon dog on a glass; **C18 [84, 114] s** describes milk being poured into a green cup. Neither entry states how many green puppy stickers are on the coffee machine.

Planner-selected skills

Plan: `localize_clue` → `score_hypothesis_support`

Why these skills: Counting requires evidence about both the right object and its quantity. **Localization** searches for entries about the coffee machine; **support assessment** evaluates whether those entries justify the proposed count of seven. A relevant scene alone cannot settle that claim.

Skill outputs: The planner proposes **C. Seven** using C1, C2. The localizer returns C1, C2, C17, C18, C26, C31, C42, with status **assessed as supported** and confidence **0.85**. The scorer uses all seven references and the localizer's result. It assigns support **0.00** and contradiction **1.00**, returning **assessed as unsupported** with diagnostic label `no_hypothesis_support`.

Support assessment and final answer

Support-assessment output: **assessed as unsupported**; score **0.00**; `insufficient_evidence`; no supported references. The catalog does not give the requested count.

Final outcome: **No valid answer**. The response reaches its length limit (`finish_reason=length`).

Gold answer: **E. Five**.

Discussion: **The trace distinguishes successful retrieval from unsupported count support.** The planner uses a dog graphic on a glass to support a claim about stickers on the machine. The localizer returns seven related entries with status **assessed as supported**, yet the scorer assigns zero support and the support-assessment output marks the claim as unsupported. Adding references has not supplied the missing quantity. These records locate the problem in the evidence for the count, before the final call reaches its length limit. The assessment is advisory and triggers no abstention, replanning, or repair.

Figure 8: **Failure / CG-Bench: An unsupported sticker count.** Finding related scenes does not establish a count. Separate retrieval and support-assessment outputs reveal where the proposed answer loses evidential support.

Failure / VRBench: The final answer retains an unsupported claim

Question: **Why were the two boys in the video repeatedly extorted for money?**

Selected video frames



C4 / 88.0 s
Students face seated classmates.



C4 / 100.0 s
A close-up during the exchange.



C8 / 210.0 s
Students confront one another.



C8 / 216.0 s
One student steps between the others.

Evidence from the catalog

C4 [84, 114] s describes classroom extortion and resistance. **C8 [196, 226] s** describes a confrontation and physical restraint. Neither the catalog nor the selected frames supports the claim that the bullies expected the victims not to report them.

Planner-selected skills

Plan: `infer_causal_relation` → `infer_intention_or_motive`

Why these skills: Repeated extortion calls for an explanation of behavior over time. **Causal inference** relates the confrontation to the repeated targeting; **motive inference** examines the proposed belief that the victims will not report it. Evidence of intimidation alone does not establish that belief.

Skill outputs: The planner proposes **D**: the bullies believed the victims would not report them. Both skills use C4, C8; the second also uses the first result. Both return **assessed as supported** with confidence **0.85**, although their output `evidence_refs` lists are empty.

Support assessment and final answer

Support-assessment output: **assessed as unsupported**; score **0.00**; `insufficient_evidence`; no supported references. The support-assessment stage assesses the proposed motive as unsupported.

Final answer (condensed): **D. The bullies believed the victims would not report them. [C4, C8]**

Gold answer (condensed): **B. The victim felt unable to fight back after an earlier failed attempt.**

Logged result: `ok=false`, `grounded=false`.

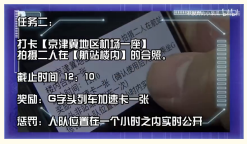
Discussion: **The trace exposes both an unsupported inference and a failure to use the support assessment.** The selected skills retain the same proposed motive, but the support-assessment output indicates insufficient evidence. The final call nevertheless returns **D**, with `grounded=false`. The wrong answer alone would not reveal that the lack of support had already been detected. Here the separate assessment provides a diagnostic signal, but that advisory signal does not gate the final decision or trigger abstention, replanning, or repair.

Figure 9: **Failure / VRBench: The final answer retains an unsupported claim.** The support-assessment output indicates insufficient evidence, but because this model-generated signal is advisory, the final generation may still retain the hypothesis. The intermediate trace exposes this system boundary.

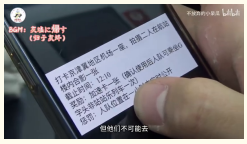
CG-Bench: Correct answer, incomplete citation coverage

Question: **In the video, when the subtitle "Task 2" appears, what is the status of the phone screen?**

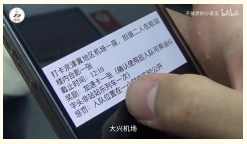
Selected video frames



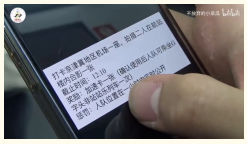
Cited / 580 s
"Task 2" over a lit phone screen.



Covered / 589 s
Task text in the first annotated span.



Missed / 592 s
The screen stays lit after the citations end.



Missed / 593 s
The same task text remains visible.

Frames are post-run illustrations; "missed" means outside the final cited intervals.

Evidence and planner-selected skills

C10 [532, 590] s describes the team reading Task 2 on a smartphone. **C53 [560, 590] s** describes a close-up of the screen displaying task text.

Plan: `localize_clue` → `score_hypothesis_support`

Why these skills: **Localization** finds the Task 2 episode; **support assessment** evaluates whether the screen description supports "Bright screen status." This separates finding the event from checking its state.

Skill outputs: The localizer keeps C10 from C10, C53 (confidence **0.95**). Using C10 and that result, the scorer assigns support **0.90**, contradiction **0.10**, and overall score **0.80**. Both return **assessed as supported**.

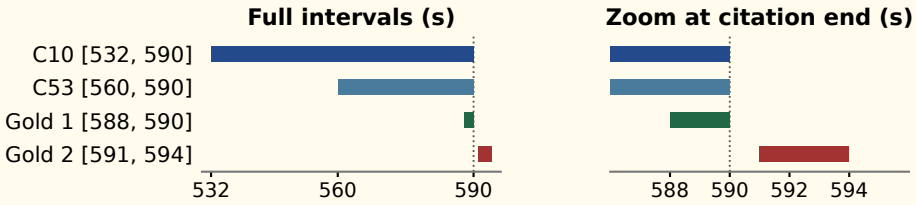
Support assessment and final answer

Support-assessment output: `verify_claim_support` accepts C10 and the scoring result: **assessed as supported**, score **0.90**, `passed=true`.

Final answer / Gold: D. Bright screen status. **Final citations:** C10, C53.

Logged metrics: `ok=true`; `citation_precision=1.00`; `grounded=true`.

Citation coverage against the saved annotations



Post-hoc diagnostics: **1/2 annotated spans (50%)** and **2/5 annotated seconds (40%)** are covered. The cited union is **58 s**; its intersection with the annotations is **2 s**. Temporal precision is **2/58 = 3.45%**; temporal IoU is **2/61 = 3.28%**.

Discussion: **A citation hit does not guarantee complete or precise temporal coverage.** Both references overlap [588, 590] s, so the logged citation precision is 1.00, yet neither reaches [591, 594] s. The nested C53 adds no temporal coverage beyond C10. This motivates checking coverage alongside citation precision. It does not make the answer unsupported: the screen is already lit at 580 s, and the annotations need not exhaust all valid evidence. The uncited C54 [588, 618] s covers both spans, but the advisory assessment does not trigger citation revision.

Figure 10: **Correct answer, incomplete citation coverage.** The final answer and logged grounding check pass, while both citations cover only the first annotated span. Coverage values are supplementary diagnostics.