VID-EVI-QA: A CONTROLLED BENCHMARK FOR EVIDENCE-ADAPTIVE VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing video question answering (VideoQA) benchmarks assess whether models answer or refuse correctly, but not whether decisions adapt to changes in relevant visual evidence. We introduce VID-EVI-QA, a controlled evidence-intervention benchmark spanning Sufficient, Insufficient, Partial, and Conflicting evidence, with 24,357 condition-level instances from 5,422 questions and 1,739 videos. It intervenes on temporal evidence while holding questions fixed and separately constructs cases with multiple visually supported options. Same-question pairs and matched non-evidence controls distinguish answering ability, refusal tendency, and evidence sensitivity. Evaluations of ten open-weight video-language models across all four evidence conditions reveal a clear separation between these capabilities: the strongest answerer attains 70.53% Sufficient accuracy but only 1.87% strict Evidence Withdrawal Rate (EWR; correct-reason withdrawal conditional on initial correctness) and 1.32% joint Evidence-Contingent Success (ECS); the highest Evidence-Specific Success in matched-control experiments is just 12.68%. Unstable decisions under graded evidence and failures to recognize and appropriately reject multiple supported options reveal further limitations. We also introduce two training-free inference methods that explicitly assess option-level evidence support. Across all 9,209 non-Partial instances, two-pass visual verification improves the reason-correct paired success of the backbone from 1.45% to 18.01% and conflict detection from 5.44% to 50.95%. These results demonstrate the potential of explicit support assessment to improve evidence-aware perception and decision making, and show that video reasoning evaluation must go beyond static accuracy and refusal rates to directly test whether decisions adapt appropriately to changes in supporting evidence.

1 INTRODUCTION

Recent advances in video large language models (Video-LLMs) have brought growing attention to fine-grained temporal reasoning, where models must not only answer questions about actions and events but also identify when the supporting evidence occurs (Wu et al., 2025a; Jung et al., 2025; Xiao et al., 2025). Yet correct answers do not guarantee that models actually rely on the relevant visual evidence (Xiao et al., 2024). More critically, even when necessary evidence is absent, models often fail to withhold their answers and instead produce confident but unsupported responses (Yoon et al., 2025; Yu et al., 2026). Together, these findings indicate that the answering behavior of a model is not reliably controlled by the presence or absence of supporting evidence, and that accuracy alone cannot reveal this failure. This motivates a complementary evaluation question: do models respond appropriately across different states of supporting evidence?

As summarized in Table 1, existing benchmarks approach evidence use in VideoQA from two complementary directions: evidence localization and integration, and answerability and abstention. The former examines whether models can locate supporting temporal segments or integrate evidence distributed across time (Xiao et al., 2024; Liu et al., 2024; Ben-Ami et al., 2026). However, these evaluations typically use fixed video–question samples rather than systematically testing how answers change when the relevant evidence changes. Even a correct answer with accurate localization therefore does not establish whether the model will adjust its answer when that evidence is weakened or removed. The latter examines abstention using questions whose answers are unavailable in the video, questions rewritten to be ill-posed, or inputs obtained at different frame-sampling levels

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.

Table 1: **Benchmark design comparison.** ✓: supported; △: partial; –: not reported.

Benchmark	Temporal evidence	Abstention related	Evidence removal	Multi-level evidence	Multi-option support	Matched controls
Evidence-grounded VideoQA						
NExT-GQA (Xiao et al., 2024)	✓	–	–	–	–	–
E.T. Bench (Liu et al., 2024)	✓	–	–	–	–	–
HERBench (Ben-Ami et al., 2026)	✓	△	△	–	–	–
Answerability and abstention						
UVQA (Yoon et al., 2025)	–	✓	–	–	–	–
VirtueBench (Yu et al., 2026)	✓	✓	–	✓	–	–
AbstainEQA (Wu et al., 2025b)	✓	✓	–	–	–	–
Hallucination detection						
VideoHallucer (Wang et al., 2024b)	–	✓	△	–	–	–
VID-EVI-QA (ours)	✓	✓	✓	✓	✓	✓

Evidence removal denotes a same-question intervention that selectively removes question-relevant visual support while preserving the question, options, and remaining context; generic information unavailability alone does not satisfy this criterion.

(Yoon et al., 2025; Wu et al., 2025b; Yu et al., 2026). These settings expose failures to abstain, but do not necessarily isolate the effect of question-relevant evidence: question changes introduce linguistic differences, while frame sampling jointly changes relevant evidence, surrounding context, and input-frame distributions. A more controlled evaluation is therefore needed, holding the question, answer options, and remaining video content fixed while selectively modifying the supporting evidence. A detailed discussion of related work is provided in Appendix A.

More broadly, evidence conditions in VideoQA cannot be fully characterized by a binary distinction between sufficient and insufficient. As illustrated in Figure 1, the left example shows answer persistence after support disappears, whereas the right contains clear evidence for two different options, invalidating the unique-answer premise. Evidence-aware evaluation must therefore consider both the availability of supporting evidence and its distribution across candidate answers, as well as how behavior changes when the evidence becomes less clear. To evaluate these dimensions jointly, we introduce VID-EVI-QA, covering four temporal-evidence conditions: *Sufficient*, *Insufficient*, *Partial*, and *Conflicting*. The benchmark combines evidence interventions that preserve questions and answer options with independently constructed multiple-support scenarios, testing whether model decisions adapt to changes in question-relevant evidence.

We evaluate ten open-weight video-language models on the complete non-Partial benchmark and separately assess GPT-4o and Gemini-2.5 Pro on sampled subsets. Additional diagnostics examine responses to graded evidence degradation and matched evidence/non-evidence interventions. The results show that current models still struggle to make consistently correct decisions across evidence conditions. High answer accuracy or refusal rates do not guarantee reliable paired performance: although LLaVA-OneVision achieves 70.53% Sufficient accuracy, its strict Evidence Withdrawal Rate (EWR)—the proportion of initially correct answers correctly withdrawn for the missing-evidence reason—is only 1.87%, corresponding to 1.32% strict Evidence-Contingent Success (ECS). Meanwhile, Qwen3-VL-32B correctly refuses 73.91% of Insufficient inputs for the missing-evidence reason but achieves only 11.49% strict ECS. Even Qwen2.5-VL-32B, with relatively balanced answering and refusal performance of 50.49% and 51.39%, satisfies both requirements—answering correctly with sufficient evidence and refusing for the correct reason after its removal—on only 16.24% of question pairs, the highest strict ECS among the evaluated open-weight models. Our error analysis further identifies cases in which models retain question-driven judgments rather than revise their decisions as the supporting evidence changes (Appendix I). Multiple supported options pose another challenge: apart from Qwen3-VL-32B, which achieves 85.14% Conflict Detect, the other open-weight models perform poorly, with five of the ten models achieving at most 0.31% Conflict Detect. Most models therefore struggle to recognize support for multiple options and reject the

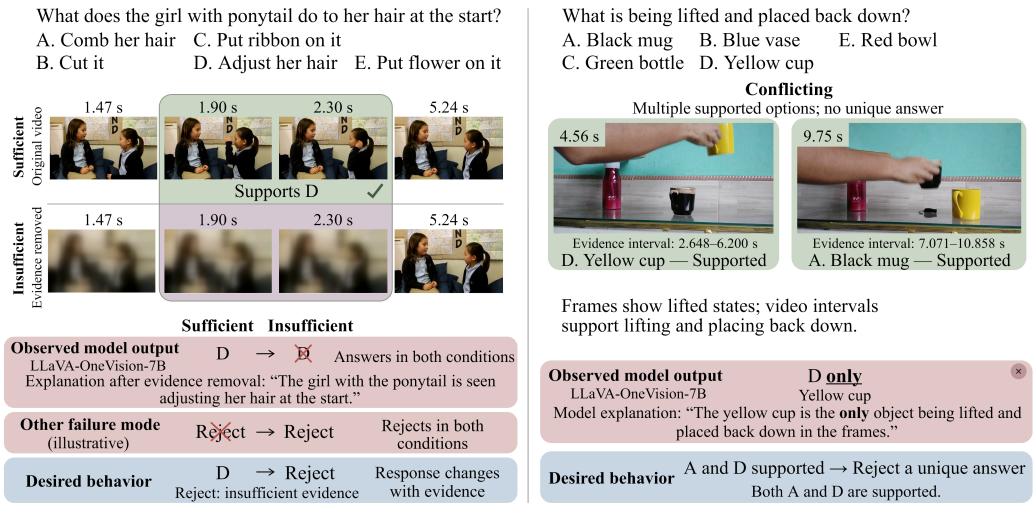


Figure 1: Two evidence-awareness failures.

unique-answer premise. Together, these findings underscore the need for evaluation across evidence states and reveal substantial limitations in the evidence-aware decision-making of current models.

Motivated by these failure modes, we investigate how to align the answers and refusal types of models with the conclusions supported by the available evidence. We design two training-free inference pipelines that decompose direct answering into visual evidence acquisition, question-conditioned verification, and final decision-making. The evidence-first pipeline (EF) first organizes visible facts and their temporal locations into structured textual evidence, then uses this evidence to verify the visual prerequisites and relations required by each candidate option, aiming to reduce reasoning-stage hallucinations arising from scene-based commonsense assumptions. The fixed-frame two-pass visual verification pipeline (fixed32 VV) performs an independent second verification pass over the same visual input to reduce evidence omissions and initial judgment biases. Both pipelines check question-relevant event identity and temporal conditions, and use the set of sufficiently supported options to distinguish answering from refusal due to insufficient or conflicting evidence. Through these two designs, we explore whether structured evidence and independent visual verification can improve decisions across evidence states, rather than merely increase the overall tendency to refuse.

Our main contributions are threefold:

- **An evidence-state benchmark.** We introduce VID-EVI-QA, covering Sufficient, Insufficient, Partial, and Conflicting evidence through same-question interventions and a separately constructed multi-option support track, with residual-evidence auditing, language-prior screening, and option-specific support verification.
- **Controlled evaluation and empirical diagnosis.** We combine paired answer–refusal evaluation, graded evidence interventions, matched non-evidence controls, and conflict-support diagnostics to distinguish question-answering competence, refusal tendency, and evidence sensitivity, revealing limitations obscured by standalone accuracy and refusal rates.
- **Training-free evidence-first inference.** We design two complementary pipelines based on structured textual evidence and independent visual verification. Both perform question-conditioned evidence verification before mapping supported options to an answer or an evidence-specific refusal.

2 THE VID-EVI-QA EVIDENCE-INTERVENTION BENCHMARK

2.1 BENCHMARK OVERVIEW

VID-EVI-QA comprises two complementary evaluation tracks. The **paired evidence-intervention track** starts from single-answer questions and constructs one Sufficient instance, one Insufficient in-

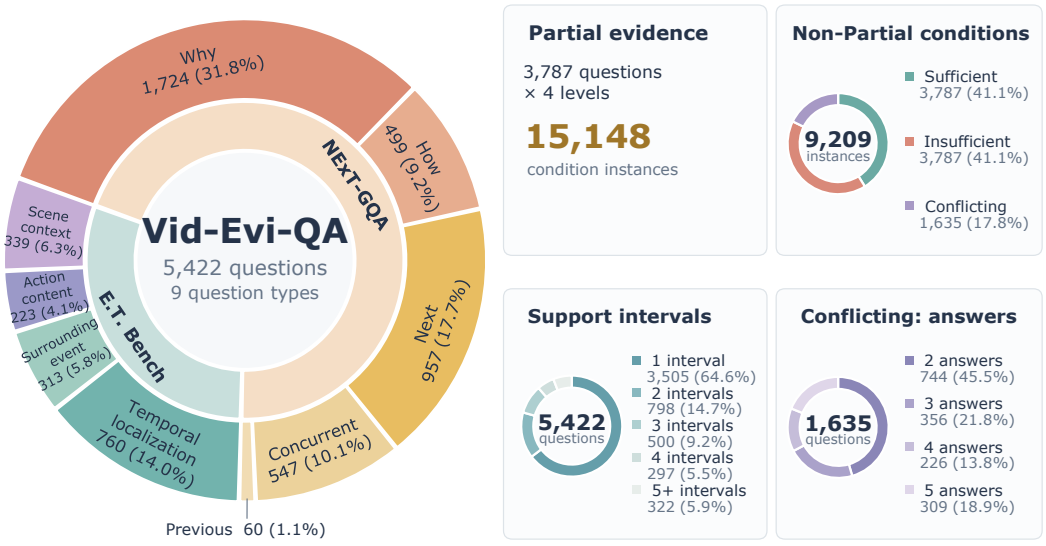


Figure 2: VID-EVI-QA composition and evidence structure.

stance, and four Partial variants for each question, all sharing the same question and answer options. Sufficient retains clear evidence supporting a unique answer and requires a correct answer with temporal grounding. Insufficient removes verified answer-supporting evidence and requires refusal due to insufficient evidence. Partial degrades the question-relevant visual content at four fidelity levels to trace responses as evidence clarity changes, without imposing a shared answer/refuse target at each level.

The **multiple-support track** separately constructs Conflicting questions in which distinct events support different answer options, with each valid option linked to its temporal evidence. This track requires models to recognize multiple supported options and refuse to select a unique answer: "conflict" denotes a violation of the unique-answer premise, not necessarily a logical contradiction between events. Conflicting questions are not a fourth variant of the paired-track questions. Figure 2 summarizes the benchmark composition and evidence-support structure. We report performance on the tracks separately; additional dataset statistics appear in Appendix D.

2.2 Dataset Construction and Quality Control

We build paired evidence interventions from temporally grounded NExT-GQA questions (Xiao et al., 2024) and the multiple-support track from E.T. Bench event annotations (Liu et al., 2024). Construction must ensure that the resulting evidence matches the intended condition: source annotations may omit alternative support, while a correct model answer or reported interval does not establish that a segment supports the answer. We therefore separate candidate discovery, visual observation, and semantic confirmation.

Residual-evidence discovery and verification. We mask annotated evidence and scan the remaining video for candidate windows. Given question q, option set $\mathcal{O}$, and correct option y, the answer-support score of model m is $M_m(X) = \log p_m(y \mid X, q, \mathcal{O}) - \log \sum_{a \in \mathcal{O} \setminus \{y\}} p_m(a \mid X, q, \mathcal{O})$. We measure visual contribution relative to a nonvisual baseline using $\Delta_m(X) = M_m(X) - M_m(B(X))$, where $B(X)$ replaces frames with gray images while preserving the question, options, frame count, and timestamps. The normalized score of candidate window W is $s_m(W) = \Delta_m(W)/\Delta_m(V)$, where V is the intact video; the correct option is used only for scoring and is not supplied to the model. Known positive controls calibrate candidate thresholds. Scoring requires recognition of known evidence and positive visual gains; other items receive a dense semantic audit, avoiding confusion between model failure and evidence absence. High-scoring windows are only candidates: visual observation withholds options and the correct answer, and subse-

quent semantic verification checks support for the complete queried event or relation, not merely an answer-related object.

Language-prior screening and paired variants. Multiple text models answer without video under different option permutations to filter questions reliably solvable from language priors. The same filter applies to all paired conditions. For retained questions, Sufficient uses the original video; Insufficient replaces evidence spans with blurred safe frames from outside those spans, preserving the timeline while removing the targeted visual content. Partial blends original frames with fixed noise only within evidence spans at four fidelity levels, leaving other frames unchanged. Levels were selected in a pilot before the full evaluation to diagnose response changes, not define a universal semantic-answerability threshold.

Multiple-support construction and quality control. For Conflicting questions, we identify differences between repeated events that support distinct answers, separate visual evidence recording from question writing, and bind each valid option to existing events and temporal intervals. Checks cover synonymous or duplicate options, unsupported answers, and option–evidence consistency, requiring independent support for multiple options rather than event repetition alone. Two reviewers, one involved in the experiments and one not involved, jointly checked evidence alignment, intervention quality, and Conflict support through iterative review, including all before/after candidates. Coverage is reported by question and review stage. Variants of a source video share a split, and we record questions, options, evidence intervals, and interventions. Complete construction procedures, calibration rules, and audit coverage appear in Appendices C and C.7.

3 Evaluating Evidence Sensitivity

We evaluate model responses to evidence removal, matched perturbations, graded evidence, and multiple supported options.

3.1 Responses to Evidence Removal

We first evaluate whether models answer or refuse correctly under each evidence condition: Sufficient inputs require the correct answer, whereas Insufficient inputs require refusal due to insufficient evidence. (`INSUFFICIENT_EVIDENCE` indicates that the model judges the input to lack answer-supporting evidence, whereas `CONFLICTING_EVIDENCE` indicates that it claims evidence supports multiple options.) However, separate accuracy and refusal rates do not establish whether a model adjusts its decision correctly for the same question as evidence changes. We therefore evaluate decisions on Sufficient–Insufficient question pairs.

Let C_i^S denote a correct answer on the Sufficient input for question i, and $R_{i,\text{ins}}^I$ denote refusal on its Insufficient counterpart with the `INSUFFICIENT_EVIDENCE` label. We define *Evidence-Contingent Success* (ECS) as $\text{ECS} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[C_i^S \wedge R_{i,\text{ins}}^I]$, where N is the number of question pairs. ECS requires both a correct answer with sufficient evidence and refusal for the correct reason after evidence removal.

To measure whether models withdraw answers they originally got right, we report the *Evidence Withdrawal Rate* (EWR). Let $N_C = \sum_{i=1}^{N} \mathbb{1}[C_i^S]$ be the number of correctly answered Sufficient questions. Then $\text{EWR} = \frac{1}{N_C}\sum_{i=1}^{N} \mathbb{1}[C_i^S \wedge R_{i,\text{ins}}^I]$ (we set EWR to zero when $N_C = 0$). ECS measures joint success over all questions, whereas EWR measures correct-reason withdrawal among questions originally answered correctly. Unusable outputs remain in the overall evaluation denominators and do not count as refusals.

On Sufficient inputs, we report question-answering (QA) accuracy and temporal grounding as mean intersection over ground truth (mIoG) and prediction (mIoP). Appendix E adds metrics jointly requiring correct answering, grounding, and withdrawal.

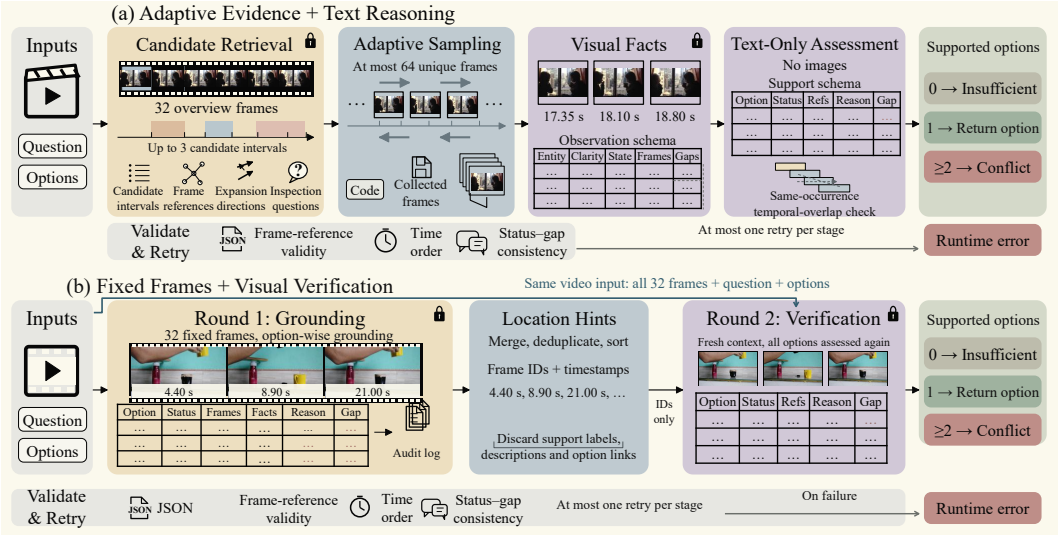


Figure 3: Adaptive EF and Two-pass VV pipelines.

3.2 Responses to Relevant and Irrelevant Perturbations

To separate evidence sensitivity from generic perturbation sensitivity, we identically corrupt evidence frames and an equal-size matched non-evidence set. *Evidence-Specific Success* (ESS), $P(C^{\text{orig}} \wedge R^{-E} \wedge C^{-N})$, requires an original correct answer, refusal after evidence corruption, and a correct answer after non-evidence corruption. Its conditional form (cESS) conditions on the original answer being correct. Unlike strict ECS, both are reason-agnostic behavioral diagnostics. Details appear in Appendix E.

3.3 Responses to Graded Evidence

Under Partial evidence, we track correct answers, wrong answers, and refusals for the same question as evidence fidelity increases. Alongside accuracy changes, we report *recovery*, the proportion of questions answered incorrectly or refused at low fidelity that become correct at high fidelity, and *regression*, the proportion initially correct that become incorrect or refused. Because fidelity does not correspond to a shared semantic-answerability threshold, we impose no uniform answer/refuse target at each level.

3.4 Responses to Multiple Supported Options

Under Conflicting evidence, models should recognize multiple supported options and reject the unique-answer premise. *Conflict Detect* measures refusal explicitly attributed to `CONFLICTING_EVIDENCE`, with a uniform loose parser accepting equivalent reasons. A three-model support-set audit additionally elicits all supported options and their temporal evidence. We report Set F1, exact-set accuracy, and grounded exact-set accuracy requiring intersection over prediction (IoP) ≥ 0.5 for every supported option (G-Set IoP). Full definitions appear in Appendix E.

4 Evidence-Aware Inference

4.1 Error-Informed Inference Framework

Before designing the pipelines, we audited sampled errors on Sufficient and Insufficient inputs and conducted a three-model diagnostic study on Conflicting inputs (Section 5). These analyses motivated separating evidence acquisition, option-level support assessment, and final decision making.

Both pipelines check whether visual evidence supports the complete claim formed by the question and each option, including the required subjects, actions, and temporal relations. This permits rea-

Table 2: Main evidence-awareness results (%; full 95% source-video-clustered confidence intervals (CIs) are reported in Appendix Tables 10–12). Pipeline localization (†) is conditional on nonempty frame-cell evidence; details are provided in Appendix E.6.

	Sufficient	Evidence localization		Insufficient	Paired S→I		Conflict
Model	QA acc.	mIoG	mIoP	Insuff. rec.	Strict ECS	Strict EWR	Conflict det.
Open-weight direct models							
Qwen2.5-VL-7B	60.65	9.55	36.09	32.56	14.44	23.81	0.06
Qwen3.6-35B-A3B	51.44	28.42	23.91	31.58	11.04	21.46	18.65
GLM-4.5V	60.68	32.59	22.60	31.05	13.10	21.58	27.16
VideoLLaMA3-7B	63.22	6.53	19.95	24.56	9.11	14.41	0.00
LLaVA-NeXT-Video-7B	33.96	2.67	15.75	10.85	0.66	1.94	0.06
LLaVA-OneVision-7B	**70.53**	7.65	28.32	1.87	1.32	1.87	0.00
Qwen3-VL-32B	21.65	34.25	28.32	**73.91**	11.49	**53.05**	**85.14**
Qwen2.5-VL-32B	50.49	40.24	29.46	51.39	**16.24**	32.17	0.31
InternVL3.5-8B	25.09	35.91	41.37	16.90	4.28	17.05	30.89
Qwen3-VL-8B	39.82	**42.95**	**50.86**	2.85	1.45	3.65	5.44
Training-free pipelines (Qwen3-VL-8B backbone)							
Qwen3 + adaptive EF†	19.57	34.78	45.80	**71.96**	11.49	**58.70**	27.83
Qwen3 + two-pass VV†	**43.65**	**60.48**	**53.62**	55.56	**18.01**	41.26	**50.95**
Closed models (fixed samples)							
GPT-4o	43.33	25.36	30.04	66.67	20.33	46.92	0.00
Gemini-2.5 Pro	69.67	44.37	43.91	28.33	16.67	23.92	44.00

sonable inference from observed events, but the presence of a related object or scene alone does not establish support. The predicted support set determines the output: zero supported options yield an insufficient-evidence refusal, one yields that answer, and multiple options yield a conflicting-evidence refusal. This rule ensures consistency with the predicted set, not its correctness. We instantiate this framework through adaptive evidence collection and fixed-frame two-pass visual verification (Figure 3), using a frozen model without gold answers or evidence intervals.

4.2 Adaptive Evidence Collection

Adaptive EF collects additional observations to address evidence missed by the initial sampling. The model first inspects 32 uniformly sampled overview frames and proposes candidate windows guided by the question and options. Additional frames from these windows are converted into structured observations indexed by subject, event, and time. A final text-only stage assesses each option using these records and applies the shared decision rule. This separates visual evidence collection from option reasoning, but the final assessment remains limited by the completeness and accuracy of the recorded observations.

4.3 Fixed-Frame Two-Pass Visual Verification

Two-pass VV checks the same 32 frames in two rounds. The first examines support for each option and marks frames for review. Only deduplicated, time-sorted frame IDs pass between rounds; answers, support labels, and explanations do not. A fresh context receives the question, all options, and all 32 frames again, verifies each option directly against the images, and applies the shared rule. Restricting cross-round information limits direct guidance from the initial conclusions. The number of unique frames is unchanged, while model calls and image presentations increase.

5 Experiments

5.1 Experimental Setup

All models follow a common evaluation protocol. We uniformly sample 32 timestamped frames per video and decode at temperature zero. Each model returns a structured answer or refusal, a

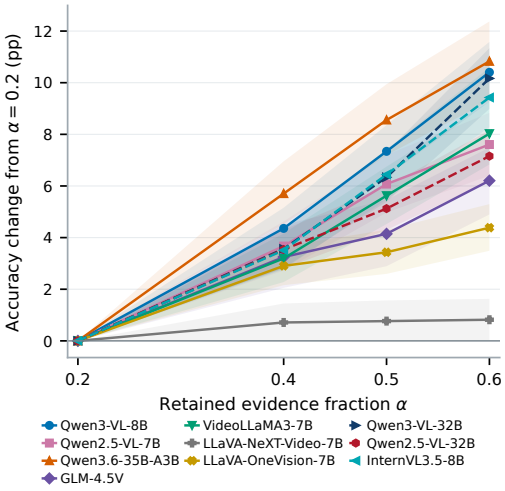


Figure 4: Accuracy response.

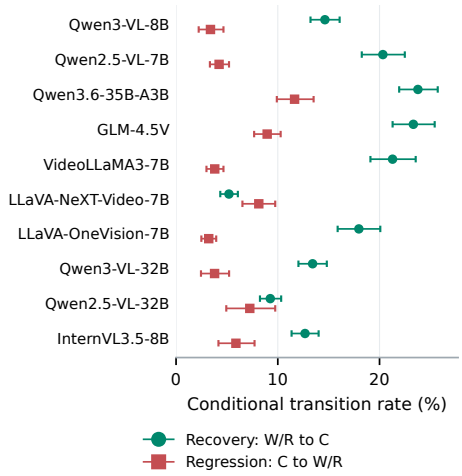


Figure 5: Recovery and regression.

refusal reason, and temporal evidence. A shared rule-based parser extracts only explicit outputs; missing or unparsable outputs count as errors. Open-weight inference runs on NVIDIA H800 GPUs. Checkpoint variants, software environments, and preprocessing details are given in Appendix E.

All ten open-weight models participate in all four evidence conditions under this protocol, targeting 24,357 condition-level records per model. The Partial analysis includes all ten models on the same 3,787 questions at each of four levels ($\alpha \in \{0.2, 0.4, 0.5, 0.6\}$), with 15,148 outputs per model. Full-denominator evaluation retains execution and parsing failures as errors; it does not imply that every output is valid. GPT-4o and Gemini-2.5 Pro are reported separately on fixed samples. Both methods from Section 4 use a frozen Qwen3-VL-8B backbone and are compared with its direct inference on all 9,209 non-Partial records. Confidence intervals use a source-video-clustered bootstrap.

5.2 Main Results: Accuracy Does Not Imply Evidence Awareness

Table 2 shows that answering ability, refusal tendency, and evidence awareness do not align. The strongest Sufficient answerer rarely withdraws its answer after evidence removal, whereas the model most responsive to insufficient and conflicting evidence performs weakly on answerable inputs. Even the best strict ECS remains low. Because ECS is a joint population rate, we interpret it together with EWR, which conditions correct-reason withdrawal on an initially correct Sufficient answer. The sampled closed models and the grounding results in Appendix E.6 exhibit the same separation: neither accuracy, refusal frequency, nor temporal localization alone establishes that a decision is controlled by visual evidence. The pipeline localization entries isolate the quality of emitted citations: they condition on a nonempty frame-cell reconstruction and therefore should be read together with the corresponding valid-evidence rates in Appendix E.6, not as unconditional end-to-end grounding success.

5.3 Responses Across Evidence Conditions

Graded evidence. Across all ten open-weight models, accuracy increases from $\alpha = 0.2$ to 0.6, but recovery and regression coexist (Figures 4 and 5). Qwen3-VL-32B, Qwen2.5-VL-32B, and InternVL3.5-8B gain 10.17, 7.16, and 9.43 percentage points, respectively; their conditional regression rates remain 3.79%, 7.25%, and 5.88%. Average improvement therefore does not imply stable question-level updating. Because visual fidelity is not a shared semantic-answerability scale, unrecovered cases alone do not establish evidence neglect.

Matched perturbations. Applying the same corruption to evidence and matched non-evidence regions reveals low evidence-specific success (Table 3). Thus, some withdrawal after evidence corruption reflects sensitivity to visual degradation rather than a response specific to the queried event.

Table 3: Matched-control and conflict-support diagnostics (%; CIs in Appendix Table 12).

	Matched controls ($N = 800$)			Conflict support audit ($N = 1{,}635$)			
Model	ESS	cESS	Gap	Exact set	Set F1	Multi-detect	G-Set IoP
Qwen3-VL-8B	**12.68**	22.62	20.93	**24.89**	69.27	47.03	**12.60**
Qwen3.6-35B-A3B	2.96	3.83	3.35	20.24	64.22	36.21	5.87
LLaVA-OneVision-7B	3.48	5.17	4.73	0.18	45.08	0.37	0.00

Table 4: Error diagnosis across evidence conditions.

Sufficient attribution	Share	Insufficient attribution	Share	Conflict diagnosis from ablations
Visual–temporal grounding failure	27%	Ungrounded option matching	49%	**Support representation:** option-level support improves state decisions over direct prediction.
Visual-conditioned simple reasoning failure	26%	Temporal/action confusion	29%	**Decision-rule application:** Direct prediction still fails with an explicitly supplied support set.
Over-strict refusal despite inferable evidence	12%	Context/option-prior completion	19%	**Evidence binding:** correct support sets do not ensure accurate temporal localization.
Multi-option conflict refusal	10%	Question-premise hallucination	2%	**Protocol reliability:** prompt wording and localization-output requirements affect results and cause format failures.
Hallucinated/unsupported wrong answer	24%	Decision–rationale inconsistency	1%	
Decision–rationale inconsistency	1%			

Multiple supported options. Conflict refusal varies sharply across models, and a high refusal rate can coexist with over-refusal on answerable inputs. The support-set audit further shows that detecting apparent conflict does not ensure correct option-level support attribution or grounding.

5.4 Evidence-Aware Inference Results

Both pipelines improve paired decisions: VV balances answering and conflict recognition; EF trades accuracy for insufficient-evidence rejection. Decision rules also contribute, so gains alone do not establish better visual understanding. Comparison with the matched-image one-pass control (2.30% strict ECS; 14.68% Conflict Detect) assesses the full second-pass bundle— extra call, fresh context, neutral frame hints, verification prompting—not repeated images alone.

5.5 Error Analysis Across Evidence Conditions

Table 4 summarizes two 100-case error audits (Appendix I) and Conflict diagnostics, motivating separate evidence acquisition, support assessment, and decision stages.

Conflicting evidence: ablation diagnosis. Across four conditions on 100 reused questions, support prediction improves state macro accuracy by 7.50–18.00 points over direct prediction for three core models; the supplementary gain is nonsignificant. Explicit-support-set failures, incomplete grounding, and prompt/output sensitivity implicate decision rules, evidence binding, and protocol reliability without isolating causal mechanisms (Appendix G).

6 Conclusion

We introduced Vid-Evi-QA to evaluate whether video-language models adapt answers and refusals to changing visual evidence. Paired interventions and the multiple-support track distinguish answering accuracy, refusal tendency, and evidence sensitivity. Results show that neither high accuracy nor frequent refusal ensures correct evidence-contingent decisions. Two training-free pipelines improve selected behaviors through explicit support assessment, but substantial errors remain. Future work should improve visually grounded reasoning and support–decision consistency; Appendix B discusses limitations and directions.

AI USE STATEMENT

During the research and preparation of this manuscript, we used generative AI tools to assist with identifying related work, writing code, generating test cases, and debugging. We also used these tools to provide feedback and revision suggestions on the Chinese-language draft of the authors and to refine the grammar and wording of the final English manuscript. The authors are responsible for the research design, methodological choices, interpretation of experimental results, and final conclusions. The authors are responsible for reviewing AI-assisted content and verifying the relevant references, code, and experimental results, and take full responsibility for the accuracy, originality, and integrity of the manuscript.

ETHICS STATEMENT

VID-EVI-QA is intended to support research on whether video-language models answer and abstain appropriately given the available visual evidence. It builds on NExT-GQA and E.T. Bench through derived annotations and controlled evidence interventions. We do not currently redistribute the source videos or their modified versions. We plan to release benchmark annotations and reconstruction code, subject to applicable upstream licenses, and direct users to the original providers for video access. We may additionally release source videos if and when the necessary redistribution permissions are secured from the relevant rights holders. Any such release will be limited to authorized materials and will comply with the applicable license terms. Access to and use of the videos, including modifications, remain subject to their respective licenses and access conditions; our work does not grant additional rights over the source media. Users must also respect the privacy of individuals depicted in the videos. Model-assisted construction and targeted human verification do not eliminate annotation errors or biases inherited from the source datasets. Model-generated explanations are treated as post-hoc outputs rather than faithful internal reasoning. Benchmark results should not be interpreted as a guarantee of safe or reliable real-world deployment.

REPRODUCIBILITY STATEMENT

The paper and appendix describe the benchmark construction process, evidence interventions, evaluation protocols, model configurations, prompt design, and statistical analysis methods. To support reproducibility, we recorded question identifiers, evidence intervals, video transformation parameters, sampled-frame timestamps, generation settings, and raw model responses. The code and data release materials are being organized and are not yet publicly available. Once these materials are ready, we plan to provide benchmark annotations, video reconstruction code, evaluation scripts, and the necessary configuration instructions through a public code repository, subject to applicable upstream licenses. The current release plan excludes source videos and their modified versions. Users will need to obtain videos from the original providers and run the reconstruction procedure only to the extent permitted by the applicable licenses. If the necessary redistribution permissions are secured, we may subsequently release authorized source videos.

REFERENCES

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A benchmark for multi-evidence integration in video question answering. arXiv:2512.14870v2, 2026.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11287–11297, June 2021.

Ayush Gupta, Anirban Roy, Rama Chellappa, Nathaniel D. Bastian, Alvaro Velasquez, and Susmit Jha. TOGA: Temporally grounded open-ended video qa with weak supervision, 2025. URL https://arxiv.org/abs/2506.09445.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2758–2766, July 2017.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, June 2025.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.T. Bench: Towards open-ended event-level video-language understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Fnu Pramono, John Cai, and Sourabh Kulkarni. TRAPSBench: Vision-language models encode but fail to express epistemic restraint, 2026. URL https://arxiv.org/abs/2608.13167.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.470. URL https://aclanthology.org/2024.emnlp-main.470/.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent, 2024a. URL https://arxiv.org/abs/2403.10517.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. arXiv:2406.16338v1, 2024b.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model, 2025a. URL https://arxiv.org/abs/2508.10922.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say "i don't know": Benchmarking abstention in embodied question answering. arXiv:2512.04597v1, 2025b.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.

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

Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, and Angela Yao. Videoqa in the era of llms: An empirical study. *Int. J. Comput. Vision*, 133(7):3970–3993, February 2025. ISSN 0920-5691. doi: 10.1007/s11263-025-02385-8. URL https://doi.org/10.1007/s11263-025-02385-8.

Eunseop Yoon, Hee Suk Yoon, Mark Hasegawa-Johnson, and Chang D. Yoo. Can video LLMs refuse to answer? alignment for answerability in video large language models. In *International Conference on Learning Representations*, 2025.

Xueqing Yu, Bohan Li, Yan Li, and Zhenheng Yang. VirtueBench: Evaluating trustworthiness under uncertainty in long video understanding. arXiv:2603.07071v2, 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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1209. URL https://aclanthology.org/2024.emnlp-main.1209/.

A Extended Related Work

A.1 Video Question Answering and Evidence Grounding

Video question answering and spatiotemporal reasoning. Video question answering (VideoQA) uses natural-language questions to assess the understanding of video content and its spatiotemporal relations by models. TGIF-QA incorporates tasks involving repeating actions, repetition counts, and state transitions, extending question answering from static content recognition to spatiotemporal video understanding (Jang et al., 2017). NExT-QA further emphasizes temporal and causal relations between actions, using descriptive, temporal, and causal questions to assess the ability of models to explain events (Xiao et al., 2021). AGQA constructs compositional questions about objects, actions, and their relations, and evaluates generalization to novel compositions, indirect references, and additional reasoning steps (Grunde-McLaughlin et al., 2021). Together, these studies broaden the reasoning demands of VideoQA and highlight the importance of event relations and their composition in evaluating video understanding.

Connecting answers to visual evidence. Grounded VideoQA links answer prediction to evidence localization to examine whether answers have relevant visual support. NExT-GQA augments question–answer pairs with temporal grounding annotations, requiring models to provide visual evidence alongside their answers, and shows that strong QA performance can coexist with weak evidence localization (Xiao et al., 2024). E.T. Bench extends evaluation to fine-grained, time-sensitive event-level understanding through tasks assessing event content and temporal location (Liu et al., 2024). At the method level, TOGA uses pseudo temporal labels and consistency constraints to learn joint generation of open-ended answers and temporal intervals in a weakly supervised setting without manual temporal annotations (Gupta et al., 2025). HERBench explicitly increases evidential requirements by requiring the integration of at least three non-overlapping cues from distinct video segments. Its analysis distinguishes retrieval deficits, where necessary evidence is not found, from fusion deficits, where models fail to integrate the information even when the necessary evidence is available (Ben-Ami et al., 2026). These efforts provide a foundation for evaluating answer correctness, evidence localization, and integration across events.

In real-world use, however, the available visual evidence may not suffice to answer every question. Reliable VideoQA therefore requires both correct answers when evidence is sufficient and recognition of insufficient evidence to avoid unsupported guesses. Moreover, even a correctly localized evidence interval does not, by itself, establish that the answer decision of the model depends on that evidence. This motivates examining whether answering and abstention decisions change appropriately when the supporting evidence is altered, complementing evaluations of answer and localization accuracy.

A.2 Answerability, Abstention, and Evidence Sensitivity

Answerability and appropriate abstention. Judgments about available visual evidence should also inform whether a model chooses to answer. UVQA and its associated answerability alignment framework address questions beyond the informational scope of a video, showing that strong VideoQA performance does not guarantee appropriate abstention and improving this ability through dedicated alignment training (Yoon et al., 2025). AbstainEQA extends this issue to embodied question answering, covering information unavailability, referential underspecification, and false presuppositions, among other cases requiring abstention. It pairs abstention cases with the original answerable questions to evaluate both answering and abstention behavior (Wu et al., 2025b). Relatedly, VideoHallucer uses binary QA evaluation with basic and hallucinated questions to assess whether models identify content that contradicts the video or cannot be verified from it (Wang et al., 2024b). These studies highlight the need to evaluate how models judge the limits of the information available in their inputs, distinguishing appropriate abstention, unsupported answers, and excessive abstention.

Behavioral consistency and input sensitivity. Another line of work varies inputs or question formulations to examine model behavior beyond standard QA performance. The empirical study by Xiao et al. (2025) finds that Video-LLMs respond weakly to some video perturbations while remaining sensitive to simple changes in questions and answer options, exposing a gap between

QA performance and robust video understanding. Jung et al. (2025) further study consistency in temporal comprehension, varying video content, language queries, and task settings to test whether models maintain appropriate grounding predictions and can verify events within their own predicted intervals. These findings show that a single answer or localization result does not fully characterize video understanding. Appropriate behavior depends on the nature of the change: transformations that preserve relevant information should induce corresponding consistency, whereas interventions that alter answer-supporting evidence may require a change in decision.

Decision evaluation under changing evidence. Several studies explicitly incorporate evidence availability into evaluation. VirtueBench introduces multiple frame-sampling levels and distinguishes answerable from unanswerable cases according to the information present in the actual input, avoiding the treatment of appropriate abstention under missing evidence as an ordinary QA error (Yu et al., 2026). In its evidence-integration analysis, HERBench removes oracle evidence frames one at a time and measures changes in the probability of the chosen answer to assess individual frame contributions (Ben-Ami et al., 2026). More recently, TRAPSBench uses procedurally generated physics videos to construct matched answerable and unanswerable conditions. Its PECS metric combines answer accuracy, abstention recall, and false abstention (Pramono et al., 2026). These studies extend static output evaluation to behavior under changing evidence conditions, but differ in their intervention targets, whether questions remain fixed, and their success criteria.

Relation to our work. Despite progress in evaluating abstention, temporal consistency, and evidence interventions, several gaps remain. First, observing response changes after evidence loss or perturbation alone does not distinguish sensitivity to question-relevant support from a general response to visual degradation; matched interventions on evidence and non-evidence regions are needed to separate these effects. Moreover, evaluation at a single condition or through aggregate averages can obscure recovery, regression, and persistent errors within the same question as evidence becomes clearer. Finally, abstaining because information is insufficient and declining to select a unique answer because multiple options have clear visual support require different support judgments; evaluating the former alone does not cover the latter. To address these gaps, VID-EVI-QA combines paired interventions with fixed questions and options, correct-reason requirements, matched non-evidence controls, trajectories across evidence-fidelity levels, and option-level support diagnostics to distinguish answering ability, abstention tendency, and evidence sensitivity.

A.3 Evidence Seeking and Verification at Inference Time

Several studies improve visual question answering by modifying the inference process. VideoAgent iteratively retrieves and collects video information conditioned on the question (Wang et al., 2024a), LLoVi uses a language model to aggregate and reason over clip descriptions (Zhang et al., 2024), and Pelican detects and corrects visual hallucinations through claim decomposition and tool-assisted verification (Sahu et al., 2024). These methods relate to our evidence acquisition, reasoning over observation records, and support verification, respectively.

Our two training-free pipelines are motivated by error-attribution analyses and designed to address missed evidence, substitution of local cues for complete support, and inconsistencies between support judgments and final decisions. Improvements in QA or claim correction reported by prior work do not by themselves establish whether a model appropriately abstains when evidence is insufficient or declines to select a unique answer when multiple options are supported. We explicitly assess the complete support requirements of each option and map the predicted support set to an answer or abstention. Adaptive EF collects additional observations and assesses support from structured records; two-pass VV directly reinspects the fixed images in a fresh context, passing only frame-location hints to limit the direct influence of earlier conclusions. We evaluate these pipelines through paired decisions across evidence conditions and conflict detection, examining whether they improve the correspondence between decisions and available evidence.

B Discussion and Limitations

Interpretation boundaries of Partial evidence. The main limitation of the Partial condition is that α controls the degree to which original visual content is retained within evidence regions, rather

than a shared semantic-visibility scale. The point at which an event changes from recognizable to unrecognizable varies across questions, depending on factors such as event duration, motion, contrast, spatial extent, and the visual detail required to answer the question. The current setting therefore does not establish a single α threshold that distinguishes answerable from unanswerable inputs across the benchmark, nor does it assign a uniform answer/refuse target to any particular Partial level.

We therefore use α as a diagnostic variable to examine model behavior as evidence becomes clearer, including when models begin to answer correctly, whether they retain incorrect judgments despite improved evidence, and whether transitions between correct and incorrect answers are non-monotonic. Compared with a single fixed Partial test set, this setting reveals a more complete response trajectory. Its results should nevertheless be interpreted as sensitivity to different levels of evidence fidelity, rather than as a universally applicable perceptual or answerability threshold.

Conflicting evidence concerns the unique-answer premise. Here, "conflict" means that at least two options receive clear visual support, violating the unique-answer premise, rather than that the propositions expressed by those options are logically contradictory. Accordingly, this condition tests whether models can identify multiple supported options and refuse to select a unique answer.

Training-free pipelines: remaining challenges and future directions. The two training-free pipelines improve some evidence-contingent answering and refusal behaviors through explicit evidence collection, option-level support assessment, and decision rules. However, these improvements do not eliminate existing model errors, leaving substantial room for progress. We identify three directions for further improvement.

First, **reasoning should be adequately constrained by visual evidence**. Our error analysis suggests that models sometimes use local cues, commonsense knowledge, or option priors to fill in information that is not supported by the video, leading them to answer despite insufficient evidence. In other cases, models fail to integrate available evidence or apply overly strict support criteria, resulting in incorrect refusals on answerable questions. The key challenge is to distinguish inferences reasonably supported by visible evidence from those that rely on unverified assumptions. Structured prompting and multi-stage verification can mitigate some errors, but whether these mechanisms alone can reliably address such problems remains an open question. Training models to reason under visual evidence constraints is a promising direction.

Second, **identification and localization of question-relevant visual evidence should be improved**. Models must identify events relevant to the question and candidate options and accurately localize their temporal extent. Locating a relevant segment, however, does not establish that it sufficiently supports an answer: models must also verify whether the actions, entities, and temporal relations within that segment satisfy the full meaning of the question and option. Future work could explore question-guided adaptive frame sampling, re-examination of local video segments, and targeted retrieval of missed evidence.

Finally, **support assessment and final decisions should be more consistent**. Even after collecting relevant evidence, models may incorrectly judge whether an option is sufficiently supported or fail to identify multiple simultaneously supported options. Deterministic decision rules specify which answering or refusal behavior corresponds to each support state, but their reliability still depends on the accuracy of upstream support assessment. Further improvements could focus on option-level evidence verification and mechanisms that trigger additional checks when evidence is incomplete or support judgments are uncertain.

C Dataset Construction and Quality Assurance

This section supplements the construction procedures in Section 2.2 with implementation details, calibration rules, and the scope of quality-control checks.

C.1 Construction Records and Sample Consistency

Paired instances are linked by source-video and question identifiers. Filtering operates at the question level: an exclusion applies jointly to the Sufficient, Insufficient, and all Partial variants of that

Table 5: Residual-evidence window settings.

Window length (seconds)	Stride (seconds)
4	2
8	4
16	8

question. Variants retain identical questions and options, and variants of the same source video share a data split.

Construction records associate source annotations, evidence intervals, interventions, and the corresponding filtering and audit outcomes. Conflicting instances additionally link each valid option to its supporting events and temporal intervals, making its visual basis traceable. Section C.3 specifies the language-prior filter, and Section C.7 reports manual review coverage.

C.2 Implementation and Calibration of Residual-Evidence Detection

We use the visual gain Δ_m and normalized window score s_m defined in Section 2.2. The following settings specify candidate discovery and semantic verification.

Windows and detector eligibility. Official evidence intervals are padded by one second on each side before masking. Sliding windows use the length–stride pairs in Table 5.

A model–question pair is eligible for probability-based discovery only if the model answers correctly on both the intact video and the official-evidence-only input, with visual gains greater than $\epsilon = 10^{-6}$ on both.

Positive controls and thresholds. For each model and window scale, positive controls are windows from the original, unmasked video that contain all official evidence intervals; partial overlap is insufficient. Let $\mathcal{S}^{+}_{m,\ell}$ denote their normalized scores. The candidate threshold is

$$\tau_{m,\ell} = \max\left\{0, \mathrm{Quantile}_{0.05}(\mathcal{S}^{+}_{m,\ell})\right\}. \tag{1}$$

Each scale requires at least 20 positive controls and at least 95% of control visual gains to be nonnegative. A masked window becomes a candidate when $s_m(W) \geq \tau_{m,\ell} - 0.05$, where the tolerance retains near-threshold windows. Smaller positive windows are preferred over enclosing positive windows. Calibration targets approximately 95% recall on known positive controls; it does not estimate recall over unknown residual evidence.

Dense inspection. Ineligible questions are inspected using gap-free ten-second cores covering the residual video, with two seconds of context added on each side. The cores determine coverage, while the extensions provide event context. Missing option probabilities and failed calls are recorded separately and are not treated as evidence of absence.

Semantic decisions and supplemental inspection. Visual records contain timestamps, subjects, actions, objects, and event relations, separating direct observations from interpretations. A semantic judge first produces an open answer from these records and then compares it with the reference answer. Verification checks the required conditions, temporal anchors, and relations; windows containing only answer-correlated content are classified separately.

For questions still answered correctly after evidence removal, reported intervals serve as additional candidates for fresh visual inspection of the actual Insufficient video. The answer and explanation of the original answering model are withheld during this inspection to limit their influence. The ensemble-checking mechanism and actual audit coverage are reported in Section C.7.

C.3 Language-Prior Screening

Three text models answer each question without video under five option permutations. Predictions are mapped back to the original option identities to reduce fixed-position effects. A question is

excluded when at least two models achieve accuracy ≥ 0.8 and the pooled accuracy across all model–permutation trials is also ≥ 0.8.

The 5,279-question candidate pool yields 3,787 retained questions after excluding 1,492 prior-positive questions. Two questions with incomplete trial records were retained after manual review. The same exclusion list applies to every paired condition. This screen reduces strong language priors but does not guarantee that retained questions cannot be answered by guessing or other nonvisual cues.

C.4 Construction of Paired Evidence Conditions

Sufficient uses the original video, question, and options. Insufficient replaces evidence intervals, padded by one second on each side, with blurred safe frames from outside those intervals and removes audio. Replacement preserves the timeline rather than shifting subsequent events through clip deletion. Residual-evidence checks and manual inspection assess whether the targeted support has been removed.

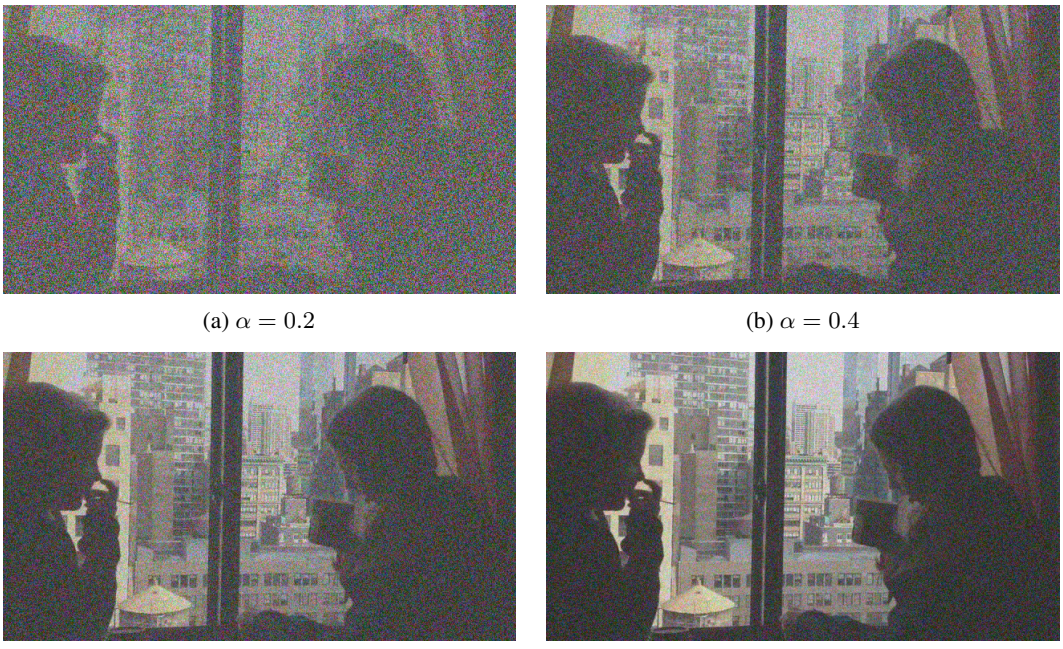
Partial blends original frames with fixed noise within the annotated evidence intervals:

$$V_t^{(\alpha)} = \begin{cases} \alpha V_t + (1-\alpha)N_t, & t \in G, \\ V_t, & t \notin G, \end{cases} \tag{2}$$

where G is the evidence-interval set, N_t is fixed noise, and $\alpha \in \{0.2, 0.4, 0.5, 0.6\}$. Questions, options, and non-evidence frames remain unchanged across levels.

The coefficient α controls retained pixel fidelity rather than the fraction of semantic information preserved. Since visibility thresholds differ across questions, Partial has no uniform must-answer or must-refuse label. Noise blending and safe-frame replacement are different operators, so Partial and Insufficient are not strict endpoints of one continuous intervention.

Figure 6 illustrates the visual progression across the four Partial levels retained for the benchmark.



(a) $\alpha = 0.2$ (b) $\alpha = 0.4$

(c) $\alpha = 0.5$ (d) $\alpha = 0.6$

Figure 6: A real NExT-GQA Partial-condition example at the four selected evidence-fidelity levels (video 6660008425, question 1, timestamp 18.10 s, within the official 17.3–19.0 s evidence interval). All panels use the same source frame and the repository's random-noise blending operator; only the blend coefficient changes. Larger α retains more of the original pixels. The example visualizes intervention strength and does not imply a universal semantic-visibility or answerability threshold.

Table 6: Partial-grid reconstruction errors. Lower is better.

Levels	Candidate grid	Reconstruction error
3	$\{0.2, 0.3, 0.6\}$	0.024616
3	$\{0.2, 0.4, 0.6\}$	0.022300
3	$\{0.2, 0.5, 0.6\}$	0.026496
4	$\{0.2, 0.3, 0.4, 0.6\}$	0.014010
4	$\{0.2, 0.3, 0.5, 0.6\}$	0.016361
4	$\{0.2, 0.4, 0.5, 0.6\}$	**0.008289**

C.5 Selection of Partial-Evidence Levels

Before the full evaluation, a pilot used 500 matched questions and three model families at nine levels, $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. Selection considered answer accuracy, answer coverage, and grounded-QA accuracy, aiming to retain the main behavioral changes under a limited compute budget. After range-normalizing each model–metric curve, we computed the mean adjacent response magnitude:

$$D_j = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \left| \widetilde{f}_{m,k}(\alpha_{j+1}) - \widetilde{f}_{m,k}(\alpha_j) \right|. \tag{3}$$

This identified $[0.2, 0.6]$ as the principal response window. With both boundaries fixed, we enumerated internal knots and compared the normalized reconstruction errors of piecewise-linear interpolation of the dense curves (Table 6).

The selected four-level grid has the lowest reconstruction error. It characterizes the principal response window of the pilot, not universal semantic answerability thresholds. Overlap between the pilot and the current 3,787-question paired set has not been fully checked; the pilot is therefore treated as development calibration rather than independent held-out validation.

C.6 Verification and Re-grounding of Conflicting Evidence

Original event annotations locate candidate segments, but their grounding may omit actions, participants, or relations needed by the question. An annotation may, for example, record an object or scene without fully describing the relevant action. It therefore does not by itself establish sufficient support for an option.

For candidates with missing actions, incomplete descriptions, or unclear support relations, we inspect the corresponding video segments again and perform re-grounding to supplement or correct the visible actions, subjects, objects, and temporal relations. Re-grounding uses the actual video to check whether a segment supports the complete claim formed by the question and candidate answer. A candidate is not used as valid support if the required visual evidence remains unconfirmed.

Questions and options are written from the verified event content. Valid options are then linked to their supporting events and temporal intervals, using both source annotations and records verified through re-grounding. Support is not freely generated during question writing.

Final checks require clear, separate support for at least two semantically distinct options and consistency between the verified records and the option–event–interval mapping. Event repetition, related objects, or multiple synonymous options alone do not establish a valid Conflicting instance.

C.7 Human Review and Quality-Control Coverage

Review units and reviewer roles. The 24,357 evaluation instances are not 24,357 independently annotated questions: they comprise 3,787 underlying video–question pairs, each with Sufficient, Insufficient, and four Partial variants, plus 1,635 Conflict questions ($6 \times 3{,}787 + 1{,}635 = 24{,}357$). Quality control therefore distinguishes checks of underlying evidence annotations, intervened visual content, and multiple-option support relations.

Table 7: Human construction-quality checks by stage. Counts describe reviewed items within each stage, not a pooled sample of independent release instances.

Audit group	N	Scope
Development: round 1	40	Evidence-removal checks motivated multi-scale window screening
Development: round 2	40	Further checks informed refinement of the current pipeline
Insufficient: targeted	147	Model-guided diagnosis motivated freezing combined with blurring
Sufficient: final	100	Annotated evidence supports the answer; no issues detected
Insufficient: final	100	Random sample checked for residual evidence and intervention quality; no issues detected
Rejected candidates	10	Spot-check of pipeline exclusions for possible over-filtering
Partial: final	20	Graded visual clarity across noise levels matched the intended manipulation
Conflict: before/after	315	Entire candidate subset reviewed; 2 excluded and 313 retained
Conflict: final sample	100	Multiple-option support and option validity; all passed

Two reviewers jointly conducted the human quality checks: one participated in the experiments and the other did not. Review followed an iterative process of inspection, problem diagnosis, pipeline revision, and final spot-checking. Table 7 summarizes the review stages.

Paired conditions and intervention checks. During initial construction, we inspected a sample of 40 questions to assess whether the source evidence intervals were adequate for evidence removal. The issues identified motivated multi-scale window screening. A second 40-question inspection informed further refinement of residual-evidence detection into the current pipeline. These rounds served method development, rather than estimation of the error rate of the final release.

We subsequently selected 147 Insufficient questions that models continued to answer frequently for targeted inspection. Even when evidence frames were correctly located, freezing alone could introduce distracting visual cues or ambiguity. We therefore revised the intervention to combine frame freezing with blurring. Continued model answering was a diagnostic cue for inspection, not itself a criterion for declaring evidence leakage.

For the final checks, we separately reviewed 100 Sufficient questions and a random sample of 100 Insufficient questions. The former checked whether the annotated evidence supported the answer; the latter checked for residual answer-supporting evidence and whether freezing and blurring were applied as intended. Neither final review identified issues under these criteria. We also inspected ten pipeline-rejected questions for possible over-filtering.

Partial intervention checks. Partial was developed after the Sufficient and Insufficient conditions. Its human inspection focused on whether the noise-blending operator produced the intended graded change in visual clarity as α varied. After implementation, we reviewed 20 questions. In the inspected examples, the range $\alpha = 0.2$–0.6 exhibited a clear progression from heavily degraded evidence toward visually discernible evidence as α increased. These checks supported the intended manipulation over the chosen range; they did not assign a universal answerability threshold to any α value.

Conflict review and revision. Conflict construction also underwent repeated human review and revision. Checks addressed missing `official_action` annotations in E.T. Bench, overly revealing answer-option priors, and options combining multiple claims. We inspected the videos to supplement or correct evidence records, revised option wording, and separated composite claims into individually verifiable alternatives, checking that each valid option had clear visual support.

All 315 before/after candidates were reviewed for temporal direction, event visibility, and option-specific support; two were excluded, leaving 313. Other Conflict questions also underwent multiple construction-time reviews and revisions. In the final 100-question Conflict sample, all items passed the multiple-support and option-validity checks, with no further exclusions.

Interpretation of the audit. This process combines development-time error discovery, targeted diagnosis, and final-version checks. We report stages separately rather than pooling reviews from different pipeline versions or sampling schemes into a single release-wide error rate. Because the reviewers worked jointly rather than producing independent labels, we do not report an inter-annotator agreement statistic. No issues detected in the final samples does not imply that the entire release is error-free.

Verification scope and inherited annotations. VID-EVI-QA is built on existing benchmarks rather than annotated from scratch. The paired track derives from temporally grounded NExT-GQA questions, and the multiple-support track derives from E.T. Bench event annotations whose grounding has undergone human verification. Our verification burden is therefore not proportional to the total number of condition-level records, but to the construction steps we newly introduce.

For the paired track, the underlying video–question pairs (3,787) are filtered rather than re-annotated: a pair is retained only when the upstream annotation is credible and our residual-evidence detection confirms that the targeted support can be removed without leaving independently answerable content. Because filtering operates at the question level, the retained set applies jointly to all paired variants. We do not claim to re-verify the upstream grounding of every retained pair; the targeted inspections described above are designed to detect failures introduced by our own intervention pipeline.

The paired variants are deterministic transformations of retained pairs. Sufficient, Insufficient, and Partial share the same question, options, and timeline, and differ only in the visual content within the annotated evidence interval. Given a verified evidence interval, frame freezing combined with blurring does not introduce new answer-supporting content; the relevant verification target is therefore the operator, not each output. Early construction used freezing alone, which could leak evidence when non-conservative frames were selected; this was revised to combined freezing and blurring and confirmed on the targeted Insufficient sample. Partial applies the same logic to noise blending: although the four α levels produce 15,148 records, they are repeated transformations of the same 3,787 questions, and the verification target is whether the blending operator produces the intended progression from heavily degraded to visually discernible evidence across $\alpha = 0.2$–0.6. The 20-question inspection confirmed this progression; we do not assign a universal answerability threshold to any α value.

The Conflicting track requires denser verification because "at least two semantically distinct options independently supported" is not directly inherited from upstream event annotations. We therefore reviewed all 315 before/after candidates, excluded 2, and retained 313, and further checked a final 100-question sample for multiple-option support and option validity. This density reflects the additional construction step rather than a larger instance count.

The total number of evaluation records is large, but the number of independently constructed units is smaller: 3,787 paired questions, four deterministic Partial levels per question, and 1,635 Conflicting questions. The sampling described above is intended to match the verification requirement of each construction step, not to estimate a release-wide error rate.

The implementation also supports ensemble checks based on independent visual observations. A leak is confirmed only when at least two distinct grounding models independently support it; a single positive, uncertainty, or an incomplete grounder set triggers review. Available records do not establish that every released instance passed this mechanism, so it is not claimed as release-wide audit coverage.

D Temporal Characteristics of Videos and Evidence

Counting conventions. Video duration is measured once per source video within each dataset. Evidence duration is measured once per unique question as the union of its annotated support intervals, clipped to the video boundaries and merged to avoid counting overlap twice. For Conflicting questions, this union includes support for all valid options. Evidence coverage is the union duration divided by the corresponding video duration. These quantities describe annotated support before intervention, excluding masking padding; paired conditions and Partial levels do not create additional observations.

Table 8: Temporal statistics by source. Entries are medians with interquartile ranges.

Measure	NExT-GQA	E.T. Bench
Video duration (s)	33.00 [21.00, 52.00]	27.50 [23.50, 33.67]
Evidence union duration (s)	4.90 [3.20, 8.50]	11.73 [7.07, 16.74]
Evidence coverage (%)	15.67 [8.87, 28.28]	43.27 [25.57, 62.07]

Duration and coverage distributions. Table 8 summarizes the temporal characteristics by source, and Figure 7 shows their binned distributions. NExT-GQA source videos have a wider duration range (10–180 seconds) than E.T. Bench source videos (20–35.83 seconds). Question-level evidence duration has a median of 4.90 seconds for the paired track and 11.73 seconds for the multiple-support track; their median evidence coverage is 15.67% and 43.27%, respectively. The latter summarizes the combined support for multiple valid options rather than the duration of a single supporting event. These are descriptive differences between the two source-specific tracks, not estimates of relative task difficulty.

E Evaluation Protocol and Implementation Details

This appendix supplements Sections 3 and 5.1 with the exact evaluation matrix, direct-inference protocol, output handling, grounding definitions, and statistical procedures. Appendix H separately specifies the additional stages used by the two training-free inference pipelines.

E.1 Evaluation Populations

The ten open-weight models listed in the main paper are evaluated on the full benchmark across all four evidence conditions, using all questions and all retained Partial levels rather than sampled subsets. Figures 4 and 5 include all ten models, including Qwen3-VL-32B, Qwen2.5-VL-32B, and InternVL3.5-8B, on the same 3,787 questions at each of the four Partial levels. The final Partial outputs contain all 15,148 expected identifiers per model, with no duplicates. The accuracy plot reports changes relative to $\alpha = 0.2$; recovery and regression compare $\alpha = 0.2$ with 0.6. C, W, and R denote correct answers, wrong answers, and refusals. Bands and whiskers show 95% source-video-clustered bootstrap intervals under the frozen v2.2 protocol. The closed-model evaluations instead use fixed samples: each contains 300 complete Sufficient–Insufficient pairs and 200 Conflicting-evidence questions. The two closed models share the Conflict questions but have different paired question subsets. Their equal sample sizes therefore do not imply fully paired cross-model comparisons. A separate closed-model Partial sample provides cross-sectional observations, not complete same-question trajectories across all fidelity levels.

The auxiliary experiments have distinct populations. Matched perturbation controls use 800 eligible questions with three corruption seeds per question; the support-set audit uses all 1,635 Conflicting-evidence questions for three representative models. Both training-free methods use the same frozen backbone on all 9,209 non-Partial instances. These auxiliary prompts and inference procedures are evaluated separately from the direct-prompt benchmark runs.

The ten open-weight models in the full evaluation matrix are Qwen3-VL-8B, Qwen2.5-VL-7B, Qwen3.6-35B-A3B, GLM-4.5V, VideoLLaMA3-7B, LLaVA-NeXT-Video-7B, LLaVA-OneVision-7B, Qwen3-VL-32B, Qwen2.5-VL-32B, and InternVL3.5-8B. GPT-4o and Gemini-2.5 Pro are the two sampled closed models. Model size is not treated as an experimental factor: the study compares frozen released checkpoints under the same behavioral task, with checkpoint-specific visual processors applied after shared frame extraction.

Table 9 summarizes the units entering each reported evaluation. Counts are condition-level records per model or pipeline arm. The four Partial levels contribute $4 \times 3{,}787 = 15{,}148$ records to each open-weight direct evaluation. The closed-model Partial diagnostic is a separately sampled cross-section and is therefore not added to the 800-record non-Partial sample in the table.

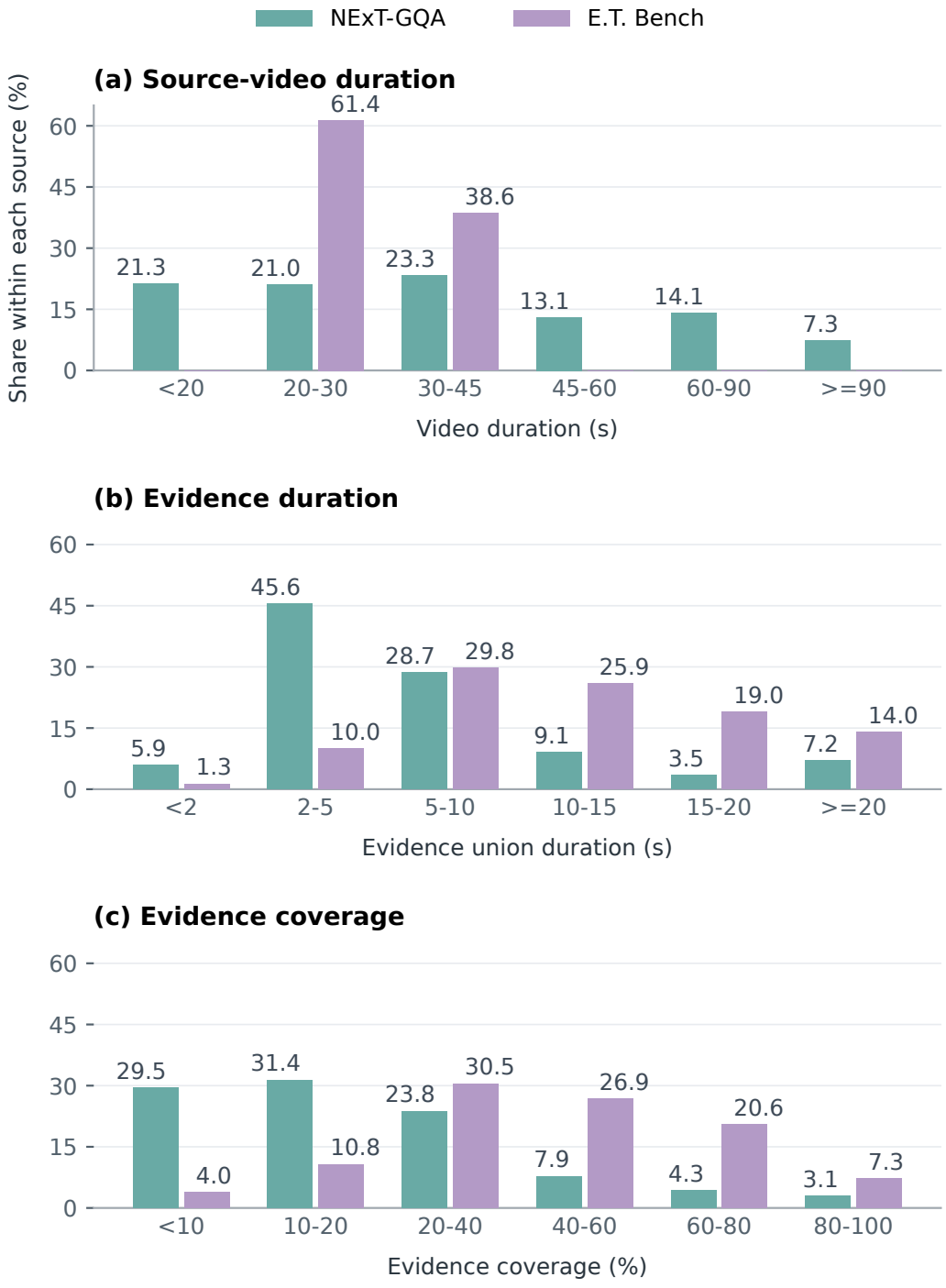


Figure 7: Video duration, evidence duration, and evidence coverage distributions.

E.2 Confidence Intervals for the Main Results

Tables 10 and 11 report the complete uncertainty intervals corresponding to the point estimates in Table 2. We split the columns across two tables so that all entries remain at the 9-point table size of the template. Intervals are 95% source-video-clustered percentile bootstrap intervals. The

Table 9: Evaluation matrix. S, I, C, and P denote Sufficient, Insufficient, Conflicting, and Partial evidence.

Protocol	S	I	C	P	Total
Open-weight direct (each of 10 models)	3,787	3,787	1,635	15,148	24,357
Closed-model direct (each model)	300	300	200	separate	800*
Training-free pipeline (each arm)	3,787	3,787	1,635	–	9,209

*Excludes the separate closed-model Partial sample. All missing or unusable records remain in the relevant denominator.

Table 10: Main results I: answering, localization, and insufficient-evidence recall (%; 95% CIs).

Model	Sufficient QA acc.	mIoG	mIoP	Insuff. recall
Open-weight direct models				
Qwen3-VL-8B	39.82 [37.95,41.70]	**42.95** [41.52,44.39]	**50.86** [49.25,52.44]	2.85 [2.13,3.65]
Qwen2.5-VL-7B	60.65 [58.82,62.44]	9.55 [8.94,10.22]	36.09 [34.41,37.76]	32.56 [30.68,34.43]
Qwen3.6-35B-A3B	51.44 [49.51,53.37]	28.42 [26.87,30.03]	23.91 [22.53,25.30]	31.58 [29.93,33.27]
GLM-4.5V	60.68 [58.82,62.54]	32.59 [30.90,34.37]	22.60 [21.25,24.02]	31.05 [29.13,32.92]
VideoLLaMA3-7B	63.22 [61.47,64.97]	6.53 [5.93,7.18]	19.95 [18.52,21.47]	24.56 [22.92,26.21]
LLaVA-NeXT-Video-7B	33.96 [32.39,35.57]	2.67 [2.40,2.95]	15.75 [14.48,17.08]	10.85 [9.73,12.01]
LLaVA-OneVision-7B	**70.53** [69.01,72.07]	7.65 [7.11,8.21]	28.32 [26.70,30.03]	1.87 [1.41,2.38]
Qwen3-VL-32B	21.65 [20.13,23.25]	34.25 [32.76,35.76]	28.32 [26.91,29.78]	**73.91** [72.15,75.62]
Qwen2.5-VL-32B	50.49 [48.72,52.29]	40.24 [38.71,41.84]	29.46 [28.06,30.93]	51.39 [49.45,53.31]
InternVL3.5-8B	25.09 [23.39,26.84]	35.91 [34.48,37.37]	41.37 [39.83,42.94]	16.90 [15.36,18.48]
Closed models (fixed samples)				
GPT-4o	43.33 [37.63,49.04]	25.36 [21.20,29.59]	30.04 [25.26,35.02]	66.67 [61.17,72.04]
Gemini-2.5 Pro	69.67 [64.56,74.68]	44.37 [39.97,48.91]	43.91 [39.26,48.65]	28.33 [23.13,33.66]
Training-free pipelines (Qwen3-VL-8B backbone)				
Qwen3 + adaptive EF†	19.57 [18.21,20.94]	34.78 [32.96,36.58]	45.80 [43.45,48.05]	**71.96** [70.40,73.52]
Qwen3 + two-pass VV†	**43.65** [41.87,45.44]	**60.48** [58.79,62.16]	**53.62** [51.80,55.44]	55.56 [53.79,57.32]

population and conditioning rules are identical to the main table, including conditional frame-cell localization for the two pipelines.

Table 11: Main results II: paired evidence removal and conflict detection (%; 95% CIs).

Model	Strict ECS	Strict EWR	Conflict detect
Open-weight direct models			
Qwen3-VL-8B	1.45 [0.96,1.99]	3.65 [2.43,4.98]	5.44 [4.30,6.66]
Qwen2.5-VL-7B	14.44 [13.17,15.77]	23.81 [21.77,25.93]	0.06 [0.00,0.19]
Qwen3.6-35B-A3B	11.04 [9.93,12.21]	21.46 [19.45,23.56]	18.65 [16.67,20.72]
GLM-4.5V	13.10 [11.86,14.38]	21.58 [19.58,23.66]	27.16 [24.97,29.34]
VideoLLaMA3-7B	9.11 [8.11,10.15]	14.41 [12.88,16.02]	0.00 [0.00,0.00]
LLaVA-NeXT-Video-7B	0.66 [0.42,0.93]	1.94 [1.22,2.72]	0.06 [0.00,0.19]
LLaVA-OneVision-7B	1.32 [0.95,1.73]	1.87 [1.34,2.45]	0.00 [0.00,0.00]
Qwen3-VL-32B	11.49 [10.33,12.70]	**53.05** [49.34,56.99]	**85.14** [83.33,86.86]
Qwen2.5-VL-32B	**16.24** [14.93,17.64]	32.17 [29.78,34.66]	0.31 [0.06,0.61]
InternVL3.5-8B	4.28 [3.50,5.12]	17.05 [14.16,20.04]	30.89 [28.54,33.23]
Closed models (fixed samples)			
GPT-4o	20.33 [15.81,25.08]	46.92 [38.10,55.65]	0.00 [0.00,0.00]
Gemini-2.5 Pro	16.67 [12.29,21.19]	23.92 [17.96,30.14]	44.00 [37.06,51.02]
Training-free pipelines (Qwen3-VL-8B backbone)			
Qwen3 + adaptive EF†	11.49 [10.42,12.56]	**58.70** [54.94,62.41]	27.83 [25.76,29.88]
Qwen3 + two-pass VV†	**18.01** [16.64,19.42]	41.26 [38.62,43.93]	**50.95** [48.36,53.43]

Table 12: Matched-control and conflict-support diagnostics (%; complete point estimates and 95% CIs where available).

Matched controls ($N = 800$)			
Model	ESS	cESS	Gap
Qwen3-VL-8B	**12.68** [10.29,15.14]	22.62 [18.66,26.75]	20.93 [16.53,25.37]
Qwen3.6-35B-A3B	2.96 [1.79,4.31]	3.83 [2.31,5.58]	3.35 [1.76,5.12]
LLaVA-OneVision-7B	3.48 [2.20,4.94]	5.17 [3.29,7.34]	4.73 [2.62,7.03]

Conflict support audit ($N = 1{,}635$)				
Model	Exact set	Set F1	Multi-detect	G-Set IoP
Qwen3-VL-8B	**24.89** [22.92,26.94]	69.27	47.03	**12.60** [11.08,14.18]
Qwen3.6-35B-A3B	20.24 [18.37,22.13]	64.22	36.21	5.87 [4.66,7.17]
LLaVA-OneVision-7B	0.18 [0.00,0.43]	45.08	0.37	0.00 [0.00,0.00]

Table 13: Open-weight checkpoint inventory and inference backends.

Display name	Released checkpoint repository	Backend
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	vLLM 0.11.0
Qwen3-VL-32B	`Qwen/Qwen3-VL-32B-Instruct`	vLLM 0.11.0
Qwen2.5-VL-7B	`Qwen/Qwen2.5-VL-7B-Instruct`	vLLM 0.11.0
Qwen2.5-VL-32B	`Qwen/Qwen2.5-VL-32B-Instruct`	vLLM 0.11.0
Qwen3.6-35B-A3B	`Qwen/Qwen3.6-35B-A3B-FP8`	vLLM 0.19.1
GLM-4.5V	`zai-org/GLM-4.5V-FP8`	vLLM 0.19.1
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B`	vLLM 0.11.0
VideoLLaMA3-7B	`DAMO-NLP-SG/VideoLLaMA3-7B`	native Transformers
LLaVA-NeXT-Video-7B	`llava-hf/LLaVA-NeXT-Video-7B-hf`	vLLM 0.11.0
LLaVA-OneVision-7B	`llava-hf/llava-onevision-qwen2-7b-ov-hf`	vLLM 0.11.0

E.3 Checkpoints and Computing Environment

Open-weight inference is performed on NVIDIA H800 GPUs with 80 GB of memory per GPU. We evaluate released checkpoints without task-specific fine-tuning. Table 13 gives the exact repository identifiers used by the main study; the shorter names used elsewhere in the paper are display labels only.

The vLLM 0.11.0 environment records PyTorch 2.8.0 and Transformers 4.57.0; the vLLM 0.19.1 environment records PyTorch 2.10.0 and Transformers 4.57.6. Qwen3.6 and GLM use the released FP8 weights. The two 32B Qwen models and GLM use tensor parallelism over two H800 GPUs; the other open-weight models use one H800. VideoLLaMA3 runs through its checkpoint-native Transformers adapter with scaled dot-product attention (SDPA) rather than an OpenAI-compatible serving layer. All local snapshots are frozen before evaluation. Release manifests retain repository revisions when recorded and SHA-256 hashes of model configuration files; we do not replace missing historical revision metadata with a later Hub HEAD.

The separate support-decomposition study uses the environments specified in Appendix G; its preprocessing and generation budgets should not be conflated with the main evaluation. Closed-model requests use the API identifiers `gpt-4o` and `gemini-2.5-pro`. These are mutable service aliases rather than pinned weight revisions, so their results are timestamped service evaluations and not exactly reproducible local checkpoint evaluations.

E.4 Frozen Direct-Inference Protocol

For every direct-prompt record, the evaluator receives only the intervened video, the unchanged question, and the unchanged answer options. It does not receive the gold option, reference evidence

interval, paired-condition output, or the prediction of another model. Reference intervals are used to construct the evidence intervention and later to score localization; they are not included in the inference prompt. Sufficient and Insufficient records sharing a `pair_id` retain the same question and options. Conflicting records are scored as a separate population because their annotated target is a set of visually supported options rather than a unique answer.

For a video of duration T, the shared uniform sampler uses

$$t_j = (j + \tfrac{1}{2})T/32, \qquad j = 0, \ldots, 31. \tag{4}$$

Frames are decoded independently at these source timestamps with FFmpeg, resized to width 560 while preserving aspect ratio, and encoded as JPEG. The ordered timestamps are retained with the request and raw result. Depending on the model interface, the same sampled images are supplied either as individual timestamp-labelled images or as an ordered JPEG-frame video; this transport difference does not change the timestamp plan. Equal frame counts therefore do not imply equal visual-token counts, because model processors can differ in resizing, tiling, video pooling, and tokenization.

The frozen direct prompt asks for one of options A–E or an explicit refusal, an explicit refusal reason when applicable, and temporal evidence for a non-refusal answer. Direct runs use the same v2.2 task definition and temperature-zero decoding. Open-weight direct runs allow up to 8,192 output tokens; the pipeline and diagnostic budgets are specified separately in their respective appendices. The structured schema is requested when supported by the serving backend; the shared parser described below is applied to the text returned by each backend. Prompt version, checkpoint tag, sampled timestamps, raw response, parsed response, and request metadata are retained per record.

The planned code release will include the complete system and user prompt templates, output schemas, and model-specific input wrappers, with an explicit mapping from each experiment to its prompt and inference configuration. It will also include the available checkpoint revision identifiers, configuration hashes, and environment requirements. We describe the task and stage-level prompting logic here and in the pipeline appendix, rather than reproducing the full templates in the manuscript.

E.5 Output Handling and Run Integrity

A shared deterministic loose parser recovers explicit answer or refusal decisions from outputs that do not strictly follow the requested structure. It does not infer missing reasons, probabilities, or evidence spans. Decision parsing and evidence localization are scored separately: an explicitly recoverable answer may receive decision credit even if its temporal evidence is missing, in which case its grounding score is zero. Unrecoverable outputs and execution errors receive no success credit and remain in the applicable evaluation denominators; they are not treated as refusals. In particular, a correct Sufficient answer followed by an unusable Insufficient output is a failed withdrawal. Reason-aware scores additionally require an accepted explicit reason; generic uncertainty alone is insufficient.

Runs are partitioned into frozen manifests with unique record identifiers. Completion checks compare the merged identifier set against the manifest, reject duplicates, and retain execution and parsing failures in the audit records and evaluation denominators. Records are not rerun to obtain a more favorable prediction. Shard outputs retain the frozen input and code hashes, raw model responses, parsed outputs, and execution logs. Aggregate reports are generated from the merged manifest-aligned records rather than from only successful API calls. Consequently, infrastructure failures, exhausted parsing failures, and missing outputs lower the reported full-denominator metrics instead of silently changing the evaluation population.

Full evaluation coverage refers to the intended population, not error-free execution. The non-Partial outputs used in the main table number 9,115 of 9,209 for Qwen3-VL-32B and 8,951 of 9,209 for InternVL3.5-8B; their 94 and 258 missing records remain errors. Qwen2.5-VL-32B and the other seven open-weight models have all 9,209 non-Partial outputs. All ten models have 15,148 Partial outputs. Of these, 33, 64, and 231 responses from Qwen3-VL-32B, Qwen2.5-VL-32B, and InternVL3.5-8B, respectively, fail strict schema checks. These responses are retained and scored with the same loose decision parser; schema failure alone does not erase an explicitly recoverable answer or refusal. The main four-condition evaluation includes all ten models, while the auxiliary experiments retain the model subsets stated for each analysis.

E.6 Temporal Grounding and Grounded Paired Metrics

Let G_i and $\hat{G}_i$ be the unions of reference and predicted temporal intervals, and let $|\cdot|$ denote duration. We compute

$$\mathrm{IoP}_i = \frac{|\hat{G}_i \cap G_i|}{|\hat{G}_i|}, \qquad \mathrm{IoG}_i = \frac{|\hat{G}_i \cap G_i|}{|G_i|}, \qquad \mathrm{IoU}_i = \frac{|\hat{G}_i \cap G_i|}{|\hat{G}_i \cup G_i|}. \tag{5}$$

Intervals are clipped to the video duration and overlapping intervals are merged. Invalid or zero-duration intervals are discarded; a missing or empty prediction receives zero overlap credit. For direct-prompt models, localization means use the entire Sufficient population, without conditioning on a correct answer or a refusal. IoP measures temporal precision, IoG measures coverage of the reference, and IoU penalizes both missed support and excessive predicted duration. The pipeline-specific conditional localization protocol is defined below.

For $m \in \{\mathrm{IoP}, \mathrm{IoU}\}$, grounded answer accuracy is

$$\mathrm{Acc@GQA}^m_\tau = \frac{1}{N} \sum_i \mathbb{1}[C^S_i \wedge m_i \geq \tau]. \tag{6}$$

Using the correct-answer and correct-reason refusal events defined in the main paper, grounded evidence-contingent success and withdrawal are

$$\text{G-ECS}^m_\tau = \frac{1}{N} \sum_i \mathbb{1}[C^S_i \wedge m_i \geq \tau \wedge R^I_{i,\mathrm{ins}}], \tag{7}$$

$$\text{G-EWR}^m_\tau = P(R^I_{\mathrm{ins}} \mid C^S,\ m \geq \tau). \tag{8}$$

We use $\tau = 0.5$ for the reported thresholded metrics. All models receive interventions on the reference evidence. These metrics test consistency between reported evidence and decisions; they do not establish that the reported intervals faithfully trace internal reasoning.

Direct-prompt models provide predicted temporal intervals. The pipelines are designed to return the most specific localization available in their trace: the exact sparse frames cited for the final supported options, without artificially widening them into temporal windows. Treating an isolated citation at time t as $[t, t]$ nevertheless gives it zero duration and therefore zero overlap even when the cited frame lies inside the reference event. Among Sufficient outputs with at least one recoverable final citation, single-frame citations account for 1/1,119 (0.09%) for adaptive EF and 323/2,385 (13.54%) for two-pass VV, or 324/3,504 (9.25%) when the two pipeline arms are pooled. Relative to all 3,787 Sufficient examples per arm, the corresponding rates are 0.03% and 8.53%. These are nonempty evidence predictions whose zero raw span is a point-representation artifact, not a failure to cite evidence. To avoid this artifact, we reconstruct a temporal cell for every observed frame. Let $0 < t_{i1} < \cdots < t_{iK} < T_i$ be the distinct timestamps shown to the model, with boundaries

$$b_{i0} = 0, \qquad b_{ij} = \frac{t_{ij} + t_{i,j+1}}{2}\ (1 \leq j < K), \qquad b_{iK} = T_i. \tag{9}$$

The cell represented by frame j is $B_{ij} = [b_{i,j-1}, b_{ij}]$. For each pipeline, $\widetilde{G}_i$ is the union of cells whose frames are cited by the final supported-option evidence of the model. We do not add retrieval windows, bridge uncited cells, or use the gold answer or reference interval to select citations.

The main table reports pipeline citation quality conditional on $\mathcal{V} = \{i : |\widetilde{G}_i| > 0\}$:

$$\mathrm{mIoG}_{\mathrm{pipe}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathrm{IoG}_i, \qquad \mathrm{mIoP}_{\mathrm{pipe}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathrm{IoP}_i. \tag{10}$$

The valid-evidence rate $|\mathcal{V}|/N_S$ is 29.55% for adaptive EF and 62.98% for two-pass VV. Refusals, errors, and outputs without final citations are not silently reclassified as successful localization; they reduce this coverage rate. The conditional means separate the temporal quality of emitted citations from the decision of the pipeline to emit evidence, but they are not unconditional end-to-end grounding scores and are not directly comparable to the full-population means of the direct models.

The cell construction already expands a cited timestamp both backward and forward to the neighboring midpoints. As a sensitivity check, we additionally pad every reconstructed interval symmetrically by $T_i/64$ (half of one base 32-frame stride) or $T_i/32$ (one base stride). Conditional nonzero-overlap rates for Qwen3 direct, adaptive EF, and two-pass VV change from 78.77/85.25/88.55% without extra padding to 81.64/86.68/90.65% with $T_i/64$ padding and 83.46/88.38/91.82% with $T_i/32$ padding. This gain is partly mechanical temporal dilation: over the same settings, mIoP changes from 51.80/45.80/53.62 to 50.56/45.84/51.83 and then 48.74/44.48/49.72. We therefore use midpoint cells without extra padding in the main table and treat padded variants only as a sensitivity analysis.

E.7 Matched Controls and Conflict Support Diagnostics

Matched non-evidence controls. Each evidence-frame set is matched to a disjoint non-evidence set containing the same number of sampled frames. Both receive the same noise operator. The eligible population contains 3,410 questions, or 90.04% of the paired track. Questions without enough non-evidence frames for a disjoint matched control are outside this population. We proportionally stratify 800 questions by question type, evidence-frame count, and video duration, and evaluate three deterministic corruption seeds with sampling weights. In addition to ESS, we report its conditional version $\mathrm{cESS} = P(R^{-E} \wedge C^{-N} \mid C^{\mathrm{orig}})$ and the specificity gap

$$\mathrm{Gap} = P(R^{-E} \mid C^{\mathrm{orig}}) - P(R^{-N} \mid C^{\mathrm{orig}}). \tag{11}$$

These are reason-agnostic behavioral diagnostics: R denotes an explicit refusal regardless of its reason.

Conflict detection. Consistent with the main paper and the current scorer, Conflict Detect is

$$\text{Conflict Detect} = \frac{1}{N_{\mathrm{conf}}} \sum_i \mathbb{1}[R_i \wedge r_i \in \mathcal{R}_{\mathrm{conf}}], \tag{12}$$

where r_i is the normalized explicit refusal reason. The scorer accepts `conflict`, `conflicting`, `conflicting_evidence`, and `multiple_valid_answers` after normalization of case and separators. A refusal without an accepted conflict reason does not qualify. This metric does not require two predicted evidence groups or correct option-specific grounding. Conflict Detect in the main table denotes this same criterion.

Separate support-set audit. An auxiliary prompt elicits the supported-option set $\hat{S}_i$ and temporal evidence for each selected option. Against the annotated set S_i, we report set precision, recall, F1, and exact-set accuracy. These assess support recognition separately from the refusal decision of the primary prompt. Grounded set accuracy additionally requires every true option to meet the temporal overlap threshold:

$$\mathrm{GroundedSetAcc}^m_\tau = \frac{1}{N_{\mathrm{conf}}} \sum_i \mathbb{1}\left[\hat{S}_i = S_i \ \wedge \bigwedge_{o \in S_i} m_{i,o} \geq \tau\right], \tag{13}$$

where $m_{i,o}$ compares the predicted and annotated interval unions for option o. We use IoP or IoU with $\tau = 0.5$. Neither exact support recovery nor grounded set accuracy is part of the primary Conflict Detect criterion.

E.8 Uncertainty Estimates and Comparison Scope

We use 95% percentile confidence intervals from 10,000 source-video-clustered bootstrap replicates with seed 20260916. Questions and condition variants from the same source video remain in one cluster, and conditional denominators are recomputed in every replicate. Matched-control resampling preserves the complete corruption-seed bundle and sampling weight of each question.

Paired comparisons require the same question population and aligned cluster draws. Equal sample sizes alone do not justify paired comparisons between separately sampled closed-model runs. Partial recovery and regression require matched observations of the same question at both levels; cross-sectional endpoints cannot substitute for such trajectories. Auxiliary experiments with different input or prompting protocols retain their own evaluation populations.

Table 14: Fixed-answer score diagnostics for nine models. N is the number of questions with one unchanged option and valid scores at all four levels. Δs is the mean score change from $\alpha = 0.2$ to 0.6 in percentage points; C and W restrict it to always-correct and always-wrong questions. Drop is the percentage of the $3N$ adjacent-level comparisons with a score decrease greater than one percentage point. Brackets show 95% source-video-clustered bootstrap intervals (10,000 replicates).

Model	N	Δs (pp)	Drop (%)	Δs_{C} (pp)	Δs_{W} (pp)
Qwen3-VL-8B	1,061	0.345 [0.032,0.669]	4.932 [4.055,5.869]	0.509	-0.267
Qwen2.5-VL-7B	2,573	0.999 [0.700,1.301]	22.347 [21.345,23.373]	1.814	-0.959
Qwen3.6-35B-A3B	1,261	1.212 [0.820,1.617]	11.340 [10.118,12.582]	1.542	-0.918
GLM-4.5V	2,045	0.000 [-0.000,0.001]	0.000 [0.000,0.000]	0.000	0.000
LLaVA-NeXT-Video-7B	2,748	0.371 [0.258,0.490]	21.846 [21.039,22.671]	0.825	0.070
LLaVA-OneVision-7B	3,286	1.614 [1.298,1.940]	24.701 [23.720,25.688]	2.567	-0.653
Qwen3-VL-32B	750	0.300 [0.012,0.612]	2.044 [1.364,2.812]	0.370	-0.246
Qwen2.5-VL-32B	532	0.709 [0.213,1.231]	14.912 [12.905,16.979]	0.957	-0.323
InternVL3.5-8B	734	0.178 [-0.215,0.587]	11.490 [9.871,13.148]	0.621	-1.789

Reason-aware ECS, EWR, and their grounded variants are recomputed from the original predictions with the corresponding bootstrap analysis. Statistical claims based on earlier, reason-agnostic versions are not transferred to the revised metrics without recomputation.

F Partial-Evidence Behavior and Answer-Score Diagnostics

The main paper reports how accuracy changes as evidence becomes clearer for ten open-weight models, together with recovery from wrong answers or refusals to correct answers and regression in the opposite direction. This section asks a further question: even when the final answer remains unchanged, does the score of the model for that answer change with the evidence?

F.1 Score Changes with a Fixed Answer

We select questions for which a model returns the same option at all four evidence levels, and separately examine the always-correct and always-wrong subsets. The answer score is computed from the logged probabilities of the tokens that generate the answer value, using their geometric mean when the value spans multiple tokens. It reflects the tendency to generate that answer, not the probability that the answer is correct or a calibrated confidence estimate.

The token-score audit covers nine open-weight models (Table 14). For Qwen2.5-VL-7B, the mean score increases by 1.81 percentage points from the lowest to the highest evidence level on always-correct questions, but decreases by 0.96 points on always-wrong questions. This distinction is not universal: LLaVA-NeXT-Video-7B still shows a small increase on always-wrong questions. Thus, an unchanged final answer need not imply an unchanged score, and score changes alone do not establish correct use of evidence.

F.2 Local Score Drops and Saturation

An increase in the average score can conceal drops between intermediate levels. We count a drop when the score decreases by more than one percentage point between adjacent evidence levels. Among questions with an unchanged answer and available scores at all four levels, the adjacent-comparison drop rates are 22.35% for Qwen2.5-VL-7B and 24.70% for LLaVA-OneVision-7B.

Score saturation also limits what this diagnostic reveals (Table 15). For example, GLM-4.5V has many answer scores close to one and very small score changes across evidence levels. Small changes therefore cannot by themselves be interpreted as stable decisions or reliable use of evidence.

Table 15: Token-score coverage and saturation across all four Partial levels. Each model has 15,148 outputs with token logs. Valid denotes outputs with an unambiguous, finite answer-value score, including refusals. Option N counts valid A–E option scores, excluding refusals; the last three columns use this population, pooled across levels. Scores are shown on a 0–100 scale; the two saturation columns are percentages of Option N.

Model	Valid	Option N	Mean score	$s \geq 0.99$	$s = 1$
Qwen3-VL-8B	15,138	5,641	97.325	82.70	37.51
Qwen2.5-VL-7B	15,146	12,463	80.562	14.83	0.00
Qwen3.6-35B-A3B	15,128	7,917	91.644	49.87	0.00
GLM-4.5V	15,117	11,606	99.998	99.97	60.49
LLaVA-NeXT-Video-7B	15,122	13,031	36.227	0.00	0.00
LLaVA-OneVision-7B	15,148	15,051	71.903	5.27	0.00
Qwen3-VL-32B	15,133	4,142	98.595	90.58	53.69
Qwen2.5-VL-32B	15,148	3,245	86.596	34.42	0.00
InternVL3.5-8B	15,148	4,186	92.218	51.15	0.00

G Support-Decomposition Ablations and Diagnostic Controls

G.1 Study Overview and Research Questions

The pipelines in Section 4 explicitly assess option-level support before choosing an answer or refusal. This study examines that design principle through three questions: (RQ1) does option-level support prediction improve evidence-state decisions over direct state or support-count prediction; (RQ2) do better state decisions also recover the correct support identities and temporal evidence; and (RQ3) how sensitive are the results to localization requirements, prompt wording, and output budget?

The experiment is a separate support-state diagnostic, not a component-wise evaluation of the complete Adaptive EF or Two-pass VV pipelines. Its reused, selected population neither replaces nor adds to the 24,357-instance benchmark denominator. The main comparison shows gains for three core models, while the grounding and control analyses identify limits to those gains.

G.2 Experimental Design

Diagnostic population. The fixed set contains 100 questions from 100 distinct videos: 78 action-content and 22 surrounding-event questions. An author participated in selection and review. These questions had prior prompt/data revisions and model-result exposure; they are not an independent held-out test or a representative population sample. No second independent human review or inter-annotator agreement is claimed for this diagnostic.

Four support conditions. For each question with two candidate options, A and B, we retain both supports $\{A, B\}$, only the support of A $\{A\}$, only the support of B $\{B\}$, or neither $\varnothing$. Questions, options, timestamps, and the frame grid stay fixed. Unlike the main uniform-width-560 protocol, the diagnostic uses 32 frames at width 448: eight inside the annotated evidence interval of each option and 16 in the complement. Frames outside the retained support intervals, including all context frames, are replaced by RGB$(127, 127, 127)$. The four conditions therefore contain 16, eight, eight, and zero real frames, respectively. The neither condition is all gray; the number of visible frames is coupled to the target support state. Reference intervals guide input construction, so this is not an evaluation of end-to-end evidence discovery in natural videos.

Three output protocols. Table 16 summarizes the representations being compared. Within each condition, all protocols receive the same question and visual input. The original prompt explicitly defines visual support and excludes object presence, plausibility, and inability to rule out an option as sufficient support.

Table 16: Support-decomposition protocols.

Protocol	Predicted representation	Scored evidence state
Direct	State, answer identity, and a short reason	Directly predicted state
Count	Support cardinality: ZERO, ONE, MULTIPLE	Predicted cardinality
Set	Two support booleans and option-specific frame-index intervals	Cardinality of the predicted support set

Table 17: Support-decomposition results on 100 questions. Brackets show 95% paired video-bootstrap intervals.

Model	Direct	Count	Set	Set–Direct [95% CI]	p_{Holm}
Molmo2	33.17	35.17	40.67	+7.50 [2.50, 12.50]	0.0051
InternVL3.5	60.50	57.50	77.50	+17.00 [12.67, 21.17]	$< .001$
Qwen3-VL	66.67	67.50	84.67	+18.00 [14.33, 21.67]	$< .001$
MiniCPM-V-4.5	59.00	55.17	62.33	+3.33 [-2.67, 9.50]	0.4278

Table 18: State accuracy by support condition (%).

Model	Protocol	$\varnothing$	$\{A\}$	$\{B\}$	$\{A, B\}$	Schema valid
Molmo2	Direct	0.00	100.00	99.00	0.00	99.50
Molmo2	Count	18.00	83.00	92.00	0.00	100.00
Molmo2	Set	19.00	38.00	26.00	71.00	91.25
InternVL3.5	Direct	100.00	98.00	65.00	0.00	100.00
InternVL3.5	Count	100.00	63.00	32.00	25.00	97.25
InternVL3.5	Set	82.00	71.00	78.00	76.00	95.50
Qwen3-VL	Direct	100.00	97.00	85.00	9.00	100.00
Qwen3-VL	Count	100.00	91.00	96.00	9.00	100.00
Qwen3-VL	Set	99.00	80.00	92.00	69.00	100.00
MiniCPM-V-4.5	Direct	89.00	96.00	44.00	18.00	100.00
MiniCPM-V-4.5	Count	35.00	82.00	87.00	46.00	99.25
MiniCPM-V-4.5	Set	44.00	61.00	89.00	68.00	100.00

Let a_c be state accuracy in condition c. Our common comparison uses

$$\mathrm{Macro} = \frac{1}{3}\left[a_{\varnothing} + \frac{a_{\{A\}} + a_{\{B\}}}{2} + a_{\{A,B\}}\right]. \tag{14}$$

The two single-support conditions are averaged before combining the three states, so single support does not receive double weight. A constant state predictor scores one third. State correctness alone does not require the correct answer identity, a consistent answer field, or valid localization; these are separate diagnostics. Schema-invalid or missing outputs count as incorrect in the intended denominator, rather than being removed.

G.3 Main Results: Does Explicit Support Prediction Help?

Set improves state macro accuracy over Direct by 7.50, 17.00, and 18.00 points for Molmo2, InternVL3.5, and Qwen3-VL, respectively; all three contrasts survive Holm correction (Table 17). Count does not consistently improve over Direct. The Set–Direct interval of MiniCPM spans zero, limiting a model-general claim. These are effects of the tested output protocols: Set changes support representation, response structure, localization requirements, and the deterministic state mapping together.

Table 18 separates the four conditions. For the three core models, Set increases both-support state accuracy relative to Direct, but does not improve every condition. For example, the both-support accuracy of Molmo2 increases from 0% to 71%, while its single-support accuracies decline. Aggregate improvement therefore does not imply uniformly better behavior across states.

Table 19: Support-set diagnostics (%).

Model	Schema valid	Exact set	Grounded positive	Trajectory
Molmo2	91.25	37.00	1.67	2.00
InternVL3.5	95.50	76.00	5.00	32.00
Qwen3-VL	100.00	83.75	50.33	50.00
MiniCPM-V-4.5	100.00	63.50	41.67	21.00

Table 20: Set ablation and prompt sensitivity on 60 questions (%).

Model	Original	Swap	Paraphrase 1	Paraphrase 2	Booleans only
Molmo2	43.61	42.50	16.67	21.94	56.67
InternVL3.5	76.39	79.44	63.89	59.44	83.61
Qwen3-VL	84.44	85.56	83.33	83.61	78.33

G.4 Beyond State Accuracy: Support Identity and Grounding

Exact-set accuracy requires the predicted support identities to match the gold set, not merely its cardinality. Grounded-positive accuracy additionally requires consistent support/interval fields, temporal IoU≥ 0.5 for each supported option, a visible frame from the evidence of that option, and no frame assigned to the event of the other option within the cited interval. Complete-trajectory accuracy requires exact support sets for every condition of a question.

Improved state decisions leave a substantial grounding gap. InternVL attains 77.50% state macro accuracy but only 5.00% grounded-positive accuracy; Qwen attains 84.67% and 50.33%, respectively. The measures have different weights and denominators, so their difference is not an error-category rate. Even Qwen recovers the exact support set in all four conditions for only half the questions. These results distinguish state prediction, support identity, and localization rather than treating them as interchangeable evidence of success.

G.5 Component Ablations and Diagnostic Controls

Matched 60-question population. We fix 60 questions by stratified sampling (47 action-content and 13 surrounding-event; seed 20260917), without selecting them by the new control results. Each variant is compared with its baseline on these same questions, not with the 100-question aggregate. Each core model contributes 3,540 control calls: explicit-set rules (720), option swaps (720), a neutral multiple-support label (240), two prompt paraphrases (1,440), question/options without visual input (180), and Set without localization fields (240).

Removing the localization component. The booleans-only Set variant removes the four start/end fields while preserving per-option support prediction. Table 20 shows its results. Molmo2 improves by 13.06 points relative to its matched baseline ($p_{\mathrm{Holm}} = .0018$). This change reduces both formatting and localization demands, so it does not isolate a visual-reasoning mechanism.

Diagnostic controls. The explicit-set control supplies the support set and asks each protocol to apply its decision rule, without visual inference. Count and Set achieve 100% state macro for all three core models, while Direct achieves 66.67%/33.33%/66.67% for Molmo2/InternVL/Qwen. This exposes failures in the rule-prompt interface without identifying an internal model mechanism. Repeated identical set inputs are not independent semantic problems, and this is not a human-confirmed factual-text control.

The question/options-only control removes all images and tests whether the model reports no support. Set succeeds in 11.67%/61.67%/96.67% of the 60 questions for Molmo2/InternVL/Qwen. These are zero-state accuracies, not three-state macros. Option swaps exchange option identities while preserving visual input and remapping the target identities. The neutral-label control replaces the `CONFLICTING` label of Direct with `MULTIPLE_SUPPORTED` in both prompt and schema. Table 21 reports the full protocol-wise state results; similar aggregate scores do not establish per-question invariance.

Table 21: Control and paraphrase results on 60 questions (%).

Model	Protocol	Explicit set	No video	Neutral	P1 valid	P2 valid
Molmo2	Direct	66.67	100.00	33.06	99.58	99.17
Molmo2	Count	100.00	63.33	–	100.00	100.00
Molmo2	Set	100.00	11.67	–	45.83	56.25
InternVL3.5	Direct	33.33	96.67	64.44	100.00	100.00
InternVL3.5	Count	100.00	85.00	–	99.58	96.25
InternVL3.5	Set	100.00	61.67	–	86.25	82.50
Qwen3-VL	Direct	66.67	100.00	73.61	99.58	99.58
Qwen3-VL	Count	100.00	98.33	–	100.00	100.00
Qwen3-VL	Set	100.00	96.67	–	100.00	100.00

Table 22: Output-budget sensitivity on 60 questions (%).

		State macro		Schema valid	
Model	Protocol	256 tokens	512 tokens	256 tokens	512 tokens
Molmo2	Direct	33.06	33.06	99.17	99.17
Molmo2	Count	37.22	37.22	100.00	100.00
Molmo2	Set	43.61	43.61	93.33	93.33
InternVL3.5	Direct	60.28	60.28	100.00	100.00
InternVL3.5	Count	56.94	56.94	97.08	97.08
InternVL3.5	Set	76.39	76.39	95.83	95.83
Qwen3-VL	Direct	66.67	66.67	100.00	100.00
Qwen3-VL	Count	68.33	68.33	100.00	100.00
Qwen3-VL	Set	84.44	84.44	100.00	100.00

G.6 Prompt and Output-Budget Sensitivity

The two paraphrases replace only the shared support-definition prefix, retaining the question, options, and protocol-specific instruction. The Set macro of Molmo2 falls by 26.94/21.67 points and that of InternVL by 12.50/16.94 points; all four declines survive model-wise Holm correction. The Set schema-valid rates of Molmo2 fall to 45.83%/56.25%, so its performance decline includes output-format failures. The changes of Qwen are smaller (1.11/0.83-point declines) and not significant, which does not demonstrate prompt equivalence. All 3,540 control calls per model stay in their applicable denominators, including 315/113/2 format failures for Molmo2/InternVL/Qwen.

A separately frozen 512-token run contains all 720 calls per core model on the same 60 questions ($60 \times 4 \times 3$). Input and configuration checks pass, and every per-question state-correctness indicator matches its 256-token counterpart. Nevertheless, 18 Molmo2 and 17 InternVL outputs remain invalid, while Qwen has none. Table 22 reports state accuracy alongside schema validity. This checks a specific budget increase, not universal budget adequacy.

G.7 Reproducibility Details

Models and execution. The core models are Molmo2-O-7B, InternVL3.5-8B, and Qwen3-VL-8B-Instruct; MiniCPM-V-4.5 is supplementary. Each contributes 1,200 main-study calls ($100 \times 4 \times 3$). Molmo2 uses its native RGB processor and constrained decoding; InternVL and Qwen use vLLM 0.11 with JPEG-95 image transport. MiniCPM uses vLLM 0.19.1, `max_slice_nums=1`, and disabled thinking. The InternVL and Qwen environment uses PyTorch 2.8.0 and Transformers 4.57.0; Molmo2 uses PyTorch 2.8.0 and Transformers 4.57.1. The MiniCPM environment uses PyTorch 2.10.0 and Transformers 4.57.6. All local inference uses NVIDIA H800 GPUs. Decoding is deterministic with a 256-token cap, except in the budget sensitivity run. Source-frame hashes, timestamps, prompts, and schemas are checked, but native preprocessing is not assumed identical. Main-study invalid outputs number 37/29/0/3 for Molmo2/InternVL/Qwen/MiniCPM across all three protocols. There is no successful-output-only filtering or score-based retry. Prior failed engineering attempts are recorded separately and are not interpreted as semantic accuracy zero.

Scoring and statistical analysis. The diagnostic uses schema validation rather than the loose decision parser of the main benchmark. Direct state correctness depends on its decision field, Count on its cardinality, and Set on the number of predicted true booleans. Answer identity and internal consistency do not enter state macro. Exact-set accuracy averages the 400 condition records equally. For grounding, frame indices are mapped back to source timestamps; the predicted start/end span is compared with the annotated interval of the corresponding option. The binding checks described above reject cross-option event coverage and require visible own-option support. Schema failures receive zero success credit in every applicable metric.

We use 10,000 whole-video bootstrap replicates and 10,000 paired sign-flip randomizations (seed 20260917), preserving the four conditions and three protocols of each question. The three core Set–Direct contrasts form one primary Holm family; the six remaining core protocol contrasts form a separate secondary family. The three contrasts of MiniCPM form a supplementary family. For the controls, 11 frozen variant/protocol comparisons are corrected within each model using their own matched baseline. These are analyses of reused data, not preregistered tests on unseen examples.

Prompt organization. All three protocols share an evidence-only instruction, followed by the question, options A and B, and a protocol-specific output instruction. Direct predicts the evidence state and answer identity, Count predicts support cardinality, and Set independently assesses each option and localizes its support. Two paraphrase controls replace only the shared evidence instruction, preserving the protocol-specific instructions and schemas. Complete templates and paraphrases will be included in the planned evaluation-script release.

Schema contracts. All schemas are JSON objects, require every listed field, and forbid additional properties. Direct requires `decision` in {`ANSWERABLE`, `INSUFFICIENT`, `CONFLICTING`}, `answer` in {`A`, `B`, `R`}, and a string `reason` of 1–200 characters. Count requires only `support_count` in {`ZERO`, `ONE`, `MULTIPLE`}. Set requires booleans `A_supported`, `B_supported` and integers `A_start`, `A_end`, `B_start`, `B_end`, each between 0 and 32 inclusive. The schema enforces types and ranges; consistency checks separately require $1 \leq \text{start} \leq \text{end} \leq 32$ for a supported option and both indices zero otherwise. Booleans-only Set retains only the two required booleans and forbids other properties.

G.8 Interpretation and Limitations

Option-level prediction improves state decisions for the three core models under the original protocol. Grounding and complete trajectories remain harder, while prompt and localization requirements limit the robustness of the gains.

The questions were reused in development; gray-frame fraction tracks support state; sampling uses reference intervals; and the protocols jointly change representation, formatting, localization, and state mapping. Natural-condition, matched-editing, denser-sampling, and factual-text extensions remain unreported because their required review is unfinished. The findings therefore establish protocol-dependent benefits on this diagnostic, without isolating internal mechanisms, natural-input robustness, or component contributions to the full inference pipelines.

H Training-Free Inference Pipelines

Section 4 presents the design rationale and the high-level data flow of the two inference methods. This appendix records the operational contracts needed to reproduce the reported runs: what is allowed to cross stages, how support predictions become final decisions, how invalid outputs are handled, and the physical inference cost. Neither method receives a gold answer, a reference evidence interval, or additional training.

H.1 Adaptive Evidence Collection

Adaptive EF begins with 32 uniformly sampled overview frames. The frozen Qwen3-VL-8B model proposes at most three option-guided temporal windows. Each window is expanded by three seconds of context and sampled at six frames, subject to a cap of 64 distinct frames across the record. One

Table 23: Pipeline reliability and inference cost on 9,209 non-Partial records.

Pipeline	Strict Conflict	Errors	Calls	Image presentations	Prompt tokens	Output tokens
Adaptive EF	21.96%	1,108	39,850	1,327,733	356,327,965	24,684,971
Two-pass VV	41.90%	314	18,735	599,520	154,178,239	9,451,074

additional visual lookup is permitted before reasoning when the observation record marks an explicit coverage gap.

The visual stage writes a structured ledger whose entries identify the observed subject, event, temporal phase, and frame timestamps. The final stage receives this ledger rather than the video and independently labels every option as supported or unsupported. For a *during* relation, temporal overlap is accepted only when the observation refers to the same subject, action, and occurrence. No visual verdict is requested after this text-only assessment. Thus, missed or distorted observations cannot be repaired by the last stage and remain a characteristic failure mode of Adaptive EF.

H.2 Fixed-Frame Two-Pass Visual Verification

Two-pass VV keeps the visual sample fixed. Two fresh model contexts each receive the question, all answer options, and the same 32 timestamped frames. The first pass grounds candidate option support and marks frames that merit reinspection. The only information transferred to the second pass is the deduplicated, time-sorted list of cited or marked frame identifiers. First-pass answers, support labels, explanations, descriptions, missing-link judgments, and option–frame associations are discarded. The second pass reassesses every option directly from all 32 images; the transferred identifiers serve only as neutral location hints. A record without retries therefore uses two model calls and 64 image presentations.

This interface is intended to reduce direct anchoring on the first verdict. It does not isolate the causal effect of repeated viewing, because the second pass also introduces a fresh context and a verification-specific prompt.

H.3 Validation and Decision Contract

Both pipelines decode at temperature zero with a 3,500-token output limit and allow at most one validation retry per stage. A support claim is accepted only when it is attached to the evidence fields required by that stage. Exhausted schema validation, frame-extraction failures, and audited semantic degradation are recorded as unusable outputs. They remain in every task-metric denominator and are never converted into evidence-based refusals.

Let $\widehat{S}_i$ denote the set of options marked supported after validation. The final decision is deterministic:

$$d_i = \begin{cases} \texttt{INSUFFICIENT_EVIDENCE}, & |\widehat{S}_i| = 0, \\ o, & \widehat{S}_i = \{o\}, \\ \texttt{CONFLICTING_EVIDENCE}, & |\widehat{S}_i| \geq 2. \end{cases} \tag{15}$$

The mapping guarantees agreement between the emitted state and the predicted support-set cardinality; it does not guarantee that $\widehat{S}_i$ is correct. For the stricter Conflict audit reported below, a prediction succeeds only if it emits the conflict state, contains at least two options, and introduces no option outside the annotated support set. This criterion deliberately allows an incomplete subset of a larger gold support set and is therefore distinct from exact-set accuracy.

H.4 Reliability and Compute Accounting

Both methods were run on all 9,209 non-Partial records: 3,787 Sufficient, 3,787 Insufficient, and 1,635 Conflicting-evidence instances. Table 23 reports physical usage including retries. Token totals are those returned by the serving interface; image presentations count every image sent on every call, including repeated presentation of the same frame.

Adaptive EF averages 4.33 calls and 144.18 image presentations per record; two-pass VV averages 2.03 calls and 65.10 presentations. The latter exceeds its nominal two-call, 64-presentation cost only because of validation retries. Among Adaptive EF failures, 974 are audited semantic-degradation recoveries that were retained as unusable and 134 are frame-extraction failures. Two-pass VV failures are dominated by internally incomplete support records—for example, an option marked supported without the required evidence or with an unresolved missing link—with a smaller number of malformed or truncated JSON outputs. These are execution outcomes rather than evidence-state predictions.

Matched-image one-pass control. The predeclared fixed-frame study also includes a one-pass support-set method using the identical 32 images. It obtains 14.79% Sufficient accuracy, 20.78% correct-reason Insufficient rejection, 14.68% strict Conflict, and 2.30% strict ECS, with 36 unusable outputs. Its any-reason ECS is 9.88%. Comparison with two-pass VV therefore evaluates the complete second-pass protocol—the additional call, fresh context, neutral frame hints, and verification prompt—rather than repeated viewing alone.

I Qualitative Error Analysis

The main paper reports the aggregate category shares from two 100-case audits alongside controlled Conflict diagnostics in Table 4. The Conflict column draws on the ablation study in Appendix G; it does not report a third manual attribution sample or causal proportions. This appendix documents how the two manual samples were formed, what the labels mean, and which additional output patterns were observed. The categories describe visible failures in model outputs; they are not claims about latent computation.

I.1 Sampling and Coding Protocol

Each audit initially sampled 100 records stratified across Qwen3-VL-8B (35), Qwen2.5-VL-7B (35), and LLaVA-OneVision-7B (30), with coverage across multiple question types. The Sufficient audit contains 57 refusals and 43 wrong-option answers. The Insufficient audit contains high-confidence outputs labeled ANSWERABLE even though the required evidence had been removed. Reviewers inspected the same 32 timestamped frames supplied to the evaluated model rather than the unsampled source video. One primary label was assigned per record. Consequently, the reported shares describe the composition of these stratified error samples and are not population-weighted model error rates.

I.2 Sufficient-Input Coding Scheme

The Sufficient audit separates failures that occur before and after relevant content is observed. *Visual grounding/temporal-localization failure* means that the output misses the relevant event, attaches it to the wrong time, or grounds an unrelated event. *Visual-conditioned simple reasoning failure* means that relevant content is available but the output fails to bind the subject, action, temporal relation, or option at the required granularity. *Over-strict refusal despite inferable evidence* is used when the necessary visual premises are present but the model demands evidence beyond the question. *Multi-option conflict refusal* denotes an unsupported claim that multiple options are visually established. *Hallucinated/unsupported wrong answer* is assigned when an answered output adds content not established by the sampled frames. *Decision–rationale inconsistency* is reserved for a final label that conflicts with the observations or conclusion stated in the output itself.

These labels distinguish missing perception from failures of relational or decision-level use. They should not be interpreted as mutually exclusive mechanisms: for example, a localization mistake may also induce an unsupported option choice, although only the primary observable failure is counted.

I.3 Answered-Insufficient Coding Scheme

For Insufficient inputs, *ungrounded option matching* denotes a locally compatible object, action, or outcome used without verifying the prerequisite event, queried subject, temporal anchor, or relation

Table 24: Answered-Insufficient errors by model ($N = 100$). Parentheses show within-model percentages.

Model	n	Match	Temp.	Prior	Premise	D–R
Qwen3-VL-8B	35	21 (60.0)	9 (25.7)	5 (14.3)	0	0
Qwen2.5-VL-7B	35	16 (45.7)	12 (34.3)	7 (20.0)	0	0
LLaVA-OneVision-7B	30	12 (40.0)	8 (26.7)	7 (23.3)	2 (6.7)	1 (3.3)
Total	100	49	29	19	2	1

needed by the complete question–option proposition. *Temporal/action confusion* covers substitution among related actions, reactions, orderings, simultaneity, or temporal boundaries. *Context/option-prior completion* fills absent evidence from scene context, commonsense intent, or option wording. *Question-premise hallucination* explicitly treats an entity or event introduced by the question as visible. *Decision–rationale inconsistency* denotes a selected option that does not match the behavior described by the explanation of the model itself.

Table 24 gives the model-wise composition omitted from the main paper. Percentages are within each audited sample and must not be read as model-level error frequencies.

All 100 audited answered-Insufficient outputs assign at least 0.9 probability to ANSWERABLE. In 59 cases the model selects the benchmark-valid option but should still refuse because the complete proposition is not verifiable from the intervention; in the remaining 41 it both fails to refuse and selects another option. The combined share of ungrounded matching and temporal/action confusion is 77.5%, 78.3%, and 78.0% after 40, 60, and 100 annotations, respectively. This stability supports the sample-level conclusion that proposition verification is the dominant issue, while leaving the exact split between the two categories uncertain.

I.4 CROSS-CUTTING OUTPUT PATTERNS

Recurring patterns help interpret the primary labels. In *condition neglect*, a visible outcome is used without checking its removed prerequisite; in *surrogate-evidence substitution*, a nearby or similar event replaces the removed event. Other outputs complete a relation from scene context even when its decisive antecedent is absent. Conversely, some Sufficient outputs describe the relevant event but refuse because they demand unnecessary detail.

Generated explanations help locate output-level mismatches but are not treated as faithful traces: some acknowledge missing evidence and still answer, while others describe option-compatible evidence and then refuse. Reported results therefore rely on behavioral metrics and frame-based audits.

The stratified, single-label audits do not estimate full-benchmark category prevalence, causal dependence among categories, or inter-annotator agreement. Population-level claims require a larger independently annotated study.

J ILLUSTRATIVE FAILURE CASES

INSUFFICIENT

Question. what does the girl do after keeping the card in her pocket at the start

A. shake balloon
B. look around
C. turn away
D. play the arcade
E. drinks from cup

Video frames.

Original evidence: 0.10–4.90 s (obscured input).



0.39 s 2.75 s 4.33 s

Model-cited interval: 14.50–16.90 s (not ground truth).



14.55 s 15.34 s 16.12 s

Model response (Qwen3-VL-8B).

Answer: D – play the arcade.

Reason: The girl is shown actively playing the arcade game after the initial moment of holding the card.

Correct response: UNANSWERABLE – INSUFFICIENT_EVIDENCE.

Error analysis: The model overlooks and fails to verify the prerequisite of "holding the card," directly selecting an option that matches the visible action.

CONFLICTING

Question. What is being put on top of the table?

A. plate
B. book
C. glass
D. fork
E. notebook

Video frames.

Evidence for D (fork): 11.602–12.939 s.


11.90 s


12.31 s


12.72 s

Evidence for B (book): 20.696–23.204 s.


20.93 s


21.75 s


22.57 s

Model response (LLaVA-OneVision-7B, rerun).

Answer: D – fork.

Reason: The fork is the only object being placed on the table.

Supported options: B – book; D – fork.

Correct response: UNANSWERABLE – CONFLICTING_EVIDENCE.

Error analysis: Both a fork and a book are placed on the table at different times. The model selects only the fork and incorrectly claims it is the only object placed there, overlooking evidence that rules out a unique answer.