MEASURING AND AUDITING RUBRIC-INDUCED CONFABULATION IN VLM-BASED ASSESSMENT OF HUMAN PROCEDURAL SKILLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Procedural skills constitute core practical capabilities with pivotal importance across K-12, vocational, and higher education systems. Nevertheless, their assessment has long relied on experts scoring students' live demonstrations against standardized rubrics, which is hardly scalable. Vision-language models (VLMs) enable scalable expert-grade observation via long-video understanding. However, we find that VLMs fail to target graded operations and frequently misjudge critical devices without rubric guidance; conversely, equipped with rubrics, the models falsely mark planned procedural steps as completed actions. We name this rubric-induced confabulation a machine analogue of rater expectation bias, and trace it to recitation: the model repeats the rubric instead of reading the clip. We therefore audit each claim: deciding whether each claimed step came from the video or the rubric, formalized as conditional pointwise mutual information (CPMI) under two counterfactual interventions (removing the rubric procedure; swapping the video), and estimated either white-box from token log-probabilities or black-box from how often each claim reappears under resampling. As an outcome-validity check, higher white-box visual grounding corresponds to lower sampled support for absent steps on all four open-weight models (Spearman $\rho = -0.47$, $N = 3492$). Computed from sampled claim recurrence alone, the trust signal needs no internal probabilities and extends to closed-source models. We release a benchmark of four domains spanning K-12 science practicals and vocational skills (874 sixty-second clips, 14k step-level occurrence labels) and evaluate nine VLMs, four open-weight and five closed-source. On the four open-weight models, the rubric raises step coverage from 22% to 68% while 60–86% of the claimed steps never happened, and the false-claim rate rises from no to full rubric for every open-weight model–domain pair (+20 points on average), with an increase in 15 of 20 closed-source model–domain cells. Rubric-recited false claims recur almost as reliably as true ones, limiting detection by cross-sample consistency; under severe imbalance where true-claim prevalence is 13–27% by domain, our trust signal achieves AP at 2.0–4.0 times the corresponding prevalence baseline on the four domain averages.

1 INTRODUCTION

Paper-and-pencil examinations alone are insufficient for granting certification to mechanics and chefs. Students' procedural competency should be evaluated via direct observation of their stepwise task execution, with performance graded against standardized rubrics (Anderson, 1982; Linn et al., 1991). This type of performance-based assessment is widely adopted across K-12 laboratory science courses and vocational training, exemplified by OECD's PISA-VET framework (OECD, 2024). Nevertheless, large-scale deployment is constrained by the necessity of deploying professionally trained human raters to score each candidate's hands-on operation. The latest long-video vision-language models (VLMs) offer a promising path toward automated scoring systems: these models first generate descriptive captions for video segments and subsequently conduct logical reasoning on the captioned records to reconstruct real operational behaviors (Rege et al., 2026).

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

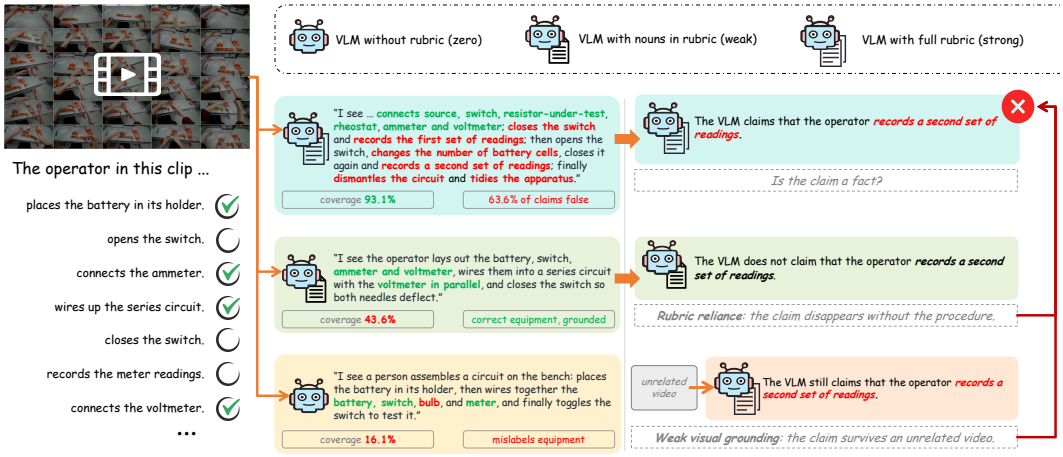

Figure 1: Rubric-induced confabulation and its claim-level audit. *Left*: on one resistance clip, Qwen3.6-27B without a rubric misses scorable steps and calls the resistor a bulb; the full rubric improves coverage but recites absent steps (green/red: occurred/absent). Percentages summarize Qwen3.6-27B across Resistance. *Right*: *records a second set of readings* disappears without the rubric procedure but survives a video swap.

For a VLM, this task is not generic video question answering but step-by-step verification against a rubric. Supplying the rubric during captioning focuses the model on scorable steps, yet it may also lead the model to report prescribed steps whether or not they occur. Existing hallucination benchmarks rarely treat such a required task prior as a controllable cause. This resembles human rater expectation bias (Rosenthal & Fode, 1963): a rater sees what the rubric leads them to expect. Does the AI rater inherit it? Unlike human expectations, a model's task prior can be removed and tested.

We caption the same videos with zero, weak, and strong task priors and check each claim against step-occurrence labels. For Qwen3.6-27B on Resistance (Figure 1, left), the zero prior recovers only 16.1% of occurred steps; the full rubric raises coverage to 93.1%, but 63.6% of its claims never happen. Across four open-weight models and four domains, coverage rises from 22% to 68%, while the full-rubric false-claim rate reaches 60–86%; closed-source models show the same overall tendency, with an increase in 15 of 20 model–domain cells. The errors exhibit *recitation*: the model reports rubric steps instead of reading the clip. Withholding the rubric from captioning is no cure: in the zero and weak conditions, scorable observations are lost even though the downstream judge still aligns captions to the rubric. We therefore ask which claims to trust, not how to remove the rubric.

We audit claims with two controlled tests: stripping the procedure from the rubric while keeping its task nouns, and swapping in an unrelated video while keeping the rubric fixed. A claim that disappears under the first shows *rubric reliance*; one that survives the second has weak *visual grounding*. Together they form the recitation signature, and a recited claim whose step did not occur is confabulation (Figure 1, right). We formalize these tests with CPMI, estimated from token log-probabilities on open models and approximated by recurrence frequencies on any model. The white-box rubric gap tracks its recurrence contrast, while the inverse association between white-box visual grounding and false-support frequency provides separate outcome-validity evidence. We evaluate the resulting sampling-only score on five closed-source models and use it to construct a per-claim trust signal.

Directly cleaning suspicious text trades fewer false claims for lost valid observations. Cross-sample consistency (Manakul et al., 2023) offers limited separation here: rubric-recited false claims recur almost as reliably as true ones. These findings motivate ranking claims by trust rather than rewriting the caption; the source attribution itself comes from the counterfactual interventions.

- **Phenomenon and benchmark.** We identify and measure rubric-induced confabulation, a machine analogue of rater expectation bias, and release a four-domain benchmark with 874 clips and ∼14k step-occurrence labels.

- **Counterfactual measurement.** We formalize a claim-level audit through two counterfactual interventions, with white-box CPMI and a sampling proxy that requires no model internals.
- **Claim-level auditing.** We show that rubric-recited false claims remain difficult to identify through consistency alone, and provide a source-aware trust signal that achieves AP at 2.0–4.0 times the true-claim prevalence baseline on the four domain averages.

2 Related Work

We review the three lines of work that motivate our source audit: procedural assessment and rater effects, caption-based memory for long videos, and visual hallucination detection.

2.1 Procedural Assessment and Rater Effects

Human performance is scored against rubrics, yet ratings vary with both raters and rubric design (Uto, 2021), including severity, leniency, and halo effects (Myford & Wolfe, 2004). Vision-based action quality assessment (AQA) seeks to automate this process by decomposing actions into steps (Xu et al., 2022; Bai et al., 2022; Xu et al., 2024), combining modalities (Zeng & Zheng, 2024), or attaching visual evidence to action reports (Okamoto & Parmar, 2024). These methods improve scoring and interpretability but generally assume that a recognized step came from the video. We instead audit whether the model observed the step or recited it from the rubric, a source question that conventional rater-effect designs cannot test counterfactually on the same observation.

2.2 Caption-Based Memory for Long Videos

Long-video systems compress observations through visual memory (Song et al., 2024; He et al., 2024), frame selection or retrieval (Wang et al., 2024a; Rege et al., 2026), or textual records built from clip captions (Zhang et al., 2024a; Kahatapitiya et al., 2025). Benchmarks such as LongVideoBench, MLVU, and LVBench primarily score final answers rather than these intermediate records (Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025). In procedural assessment, however, the caption is itself the evidence used to decide which rubric steps occurred. Conditioning it on the rubric highlights relevant tools and actions but also reveals the expected procedure, creating recitation that existing memory and retrieval methods do not distinguish from visual observation.

2.3 Visual Hallucination and Source Attribution

Image and video hallucination benchmarks test whether generated objects, facts, events, or temporal relations are visually supported (Rohrbach et al., 2018; Li et al., 2023; Jing et al., 2024; Wang et al., 2024b; Zhang et al., 2024b; Li et al., 2025). Existing methods then correct outputs (Yin et al., 2024), contrast clean and corrupted visual inputs during decoding (Leng et al., 2024), or detect uncertainty from repeated generations (Manakul et al., 2023; Farquhar et al., 2024). Intervention-based work is closest to ours: CFPO suppresses visual cues to train for grounding (Yu et al., 2026), while ContextCite ablates text passages to attribute statements to context (Cohen-Wang et al., 2024). These approaches modify a model or intervene on one source. Our inference-time audit instead separates two competing sources—the necessary rubric and the video—and uses sampling to rank claims even when model internals are closed.

Language-prior research in VQA likewise shows that a model can answer from textual regularities without using the image, motivating balanced datasets that make visual evidence necessary (Goyal et al., 2017). Our setting shares this text-dominance mechanism but differs in one operational constraint: the prior cannot simply be balanced away because the rubric defines what must be assessed. We therefore intervene on both the rubric and the video to attribute each assessment-record claim to its source.

3 Preliminaries

In this section, we formalize rubric reliance and visual grounding through CPMI-style controlled contrasts, combine them with occurrence ground truth to define rubric-induced confabulation, and

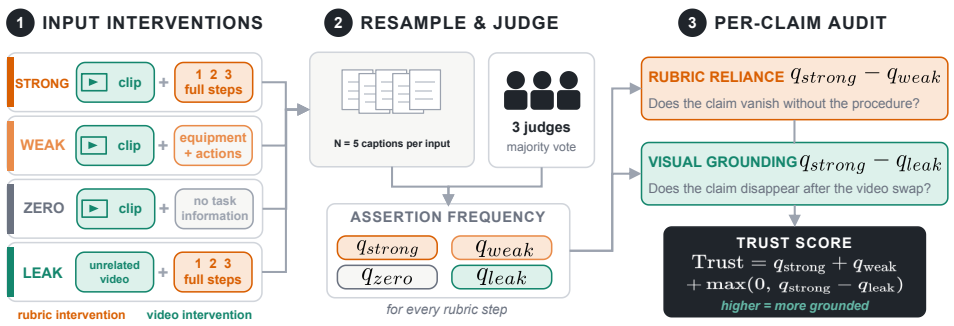

Figure 2: Claim-level audit pipeline. For each of the four inputs—strong, weak, zero, and video-swap (leak)—we generate five captions and map them to rubric claims using a three-judge majority. The resulting assertion frequencies form black-box proxies for rubric reliance and visual grounding; the trust score τ rewards claims that survive removal of the procedure and disappear after the video swap.

introduce a white-box measure and black-box proxy. We also establish the expectation bound used by the single-replacement white-box approximation.

3.1 Defining Confabulation with CPMI

Let v be a video input and r the full scoring rubric. The VLM generates a caption c with conditional distribution $P(c \mid v, r)$; a judge extracts the asserted step claims from c, and step-occurrence ground truth marks each claim as true or false. We write $\mathcal{W}(r)$ for the noun rubric: the result of removing all procedural content from r and keeping only the nouns for equipment and actions.

Both tests ask the same underlying question: if we change one input, how much does the probability of the video caption change? We use the log-ratio form of conditional pointwise mutual information (Church & Hanks, 1990; Cover & Thomas, 2006), but with controlled counterfactual references rather than full marginalization over every possible video or rubric. The resulting quantities are best read as counterfactual evidence gaps. For a piece of text y (the whole caption, or a single claim in it), we define its *visual grounding* as

$$\mathrm{CPMI}_v(y) = \log P(y \mid v, r) - \log P(y \mid r), \tag{1}$$

where $\log P(y \mid v, r)$ is the log-probability of writing y when the model sees both the matching video and the rubric, and $P(y \mid r) = \mathbb{E}_{v'}[P(y \mid v', r)]$ averages over unrelated replacement videos. Their difference measures how much additional evidence the matching video provides beyond the rubric. If it is positive, the video supports y beyond the rubric; if not, y survives the replacement distribution. A second controlled contrast measures *rubric reliance* by intervening on the procedural content:

$$\mathrm{CPMI}_r(y) = \log P(y \mid v, r) - \log P\big(y \mid v, \mathcal{W}(r)\big), \tag{2}$$

where $\log P(y \mid v, \mathcal{W}(r))$ is the log-probability of writing y once the procedure is stripped from the rubric, leaving only the nouns $\mathcal{W}(r)$. A positive gap means y becomes less likely when the procedural content is removed.

Definition 1 (rubric-induced confabulation)**.** An asserted claim s is rubric-induced confabulation if it did not occur in the video and exhibits a recitation signature: $\mathrm{CPMI}_r(s) > 0$ and $\mathrm{CPMI}_v(s) \leq 0$.

At the level of the whole caption, we call a clip *recitation-dominant* if $\mathrm{CPMI}_r(c) > \mathrm{CPMI}_v(c)$: the caption relies on the rubric more than on the video.

3.2 White-Box and Black-Box Estimators

Both quantities are built from the model's conditional probabilities, and there are two ways to read them out. On open models, the conditional probabilities can be read directly:

$$\log P(y \mid \cdot) = \sum_{i=1}^{T} \log P(y_i \mid y_{<i}, \cdot), \tag{3}$$

where $y = (y_1, \ldots, y_T)$ and each token conditions on its predecessors; for a claim embedded in a caption, the conditioning context also includes the fixed caption prefix preceding that claim. This makes Eq. (2) exact. In Eq. (1), $P(y \mid r)$ involves an average over replacement videos, which cannot be enumerated; in practice we substitute a single unrelated real video $\tilde{v}$, approximating $\log P(y \mid r)$ by $\log P(y \mid \tilde{v}, r)$. Each target uses one fixed replacement rather than repeatedly resampling $\tilde{v}$. Proposition 2 therefore concerns the expectation over the replacement-video distribution, not a deterministic guarantee for an individual target–replacement pair.

Proposition 2. *With a single $\tilde{v}$ in place of the average, the estimated visual-grounding gap is, in expectation over the replacement-video distribution, no lower than the corresponding distribution-averaged gap (proof in Appendix H).*

Closed models expose no token probabilities, but claim recurrence can still be estimated by repeated sampling. Let the model regenerate N times under the same input. The frequency of claim s,

$$\widehat{q}_N(s \mid \cdot) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\big[s \in A(c^{(n)})\big], \tag{4}$$

is the empirical recurrence frequency under our five-run protocol, where $c^{(n)}$ is the n-th regenerated caption and $A(c^{(n)})$ is the set of claims the judge extracts from it. Sample 0 is decoded at temperature 0 and samples 1–4 at temperature 0.5 when the endpoint exposes temperature control, so we do not interpret this average as an unbiased estimate for one homogeneous decoding distribution. The resulting frequency differences provide black-box proxies for rubric reliance and visual grounding using only sampled outputs.

4 METHODOLOGY

In this section, we treat rubric-conditioned assessment as a source-attribution problem: the rubric directs the model toward scorable actions but may substitute for visual evidence when it generates a claim. We separate these influences with interventions on the rubric procedure and video input, extract rubric claims from the resulting captions, and combine black-box frequency proxies into a claim-level trust score. White-box log-probabilities characterize recitation dominance on open models, while sampling extends the audit to closed models (Figure 2).

4.1 CONSTRUCTING THE COUNTERFACTUAL INPUTS

We realize the rubric intervention with three levels of task information and the video intervention with an unrelated real replacement video.

The *strong* level supplies the full rubric r. The *weak* level supplies $\mathcal{W}(r)$: equipment and action nouns without ordering, step boundaries, or conditions, directing the model to relevant objects without revealing the procedure. The *zero* level provides no task information. We use weak as the rubric-reliance contrast because zero may omit scorable steps regardless of whether they occurred; zero serves as the dose baseline. All three prompts share the same visibility, anti-guessing, and output-field constraints, so the controlled difference is the amount of task information supplied before that common instruction.

The video replacement keeps the full rubric r but replaces v with an unrelated real video $\tilde{v}$. The leak input reuses the strong prompt verbatim—including the scenario, rubric, caption instruction, and output schema—so only the video source changes. This gives $\log P(y \mid \tilde{v}, r)$, the single-video contrast in Eq. (1). If a claim persists, its frequency reflects rubric leakage rather than evidence specific to the matching clip. Unlike the rubric-dose comparison, this visual-axis contrast is not confounded by prompt-framing differences.

We use unrelated real videos rather than blank screens so that the intervention preserves a normal visual input while breaking its correspondence with the target rubric.

4.2 MEASURING RUBRIC RELIANCE AND VISUAL GROUNDING

Both quantities compare the same text under the inputs above. Open models expose token log-probabilities directly; repeated sampling provides a black-box proxy.

Reducing an assessment record to claims. A judge model maps each generated assessment record—the free-text caption together with its structured action field—to the numbered rubric claims under three roles: the lenient judge matches activities of the same kind, the strict judge requires specific textual support, and the reverse judge identifies unsupported steps, whose complement we take. We count a step as claimed when at least two judges agree. Applying the same protocol to all four inputs yields comparable claim frequencies. On a human audit of 200 records, majority extraction reaches $\kappa = 0.936$ and F1 $= 0.942$; after excluding every claim with action-field support, caption-only claims retain a 75.6% false-claim rate. Full validation and source-attribution results, together with the prompts, are reported in Appendix C.

White-box: teacher forcing. For open-weight models, we teacher-force the caption c generated under the strong input and score the same tokens under the strong, weak, zero, and leak inputs using Eq. (3). Holding c and every token prefix fixed isolates the input intervention: strong-minus-weak measures rubric reliance, strong-minus-leak measures visual grounding, and strong-minus-zero provides the bottom dose contrast. We average the token-level log-probability differences over c, yielding length-normalized caption-level scores, and verify token alignment before scoring. The evaluated closed APIs did not expose the required token log-probabilities.

Black-box: counting. For the black-box path, we count empirical claim recurrence across $N = 5$ captions for each of the strong, weak, zero, and leak inputs, as in Eq. (4); write $\widehat{q}_{\mathrm{str}}$, $\widehat{q}_{\mathrm{wk}}$, $\widehat{q}_{\mathrm{zr}}$, and $\widehat{q}_{\mathrm{lk}}$ for the resulting frequencies. The differences $\widehat{q}_{\mathrm{str}} - \widehat{q}_{\mathrm{wk}}$ and $\widehat{q}_{\mathrm{str}} - \widehat{q}_{\mathrm{lk}}$ serve as proxies for rubric reliance and visual grounding. High reliance together with low grounding indicates a recitation signature; when occurrence ground truth also shows that the step did not occur, it is rubric-induced confabulation under Definition 1.

The two paths are complementary: white-box scores expose the recitation mechanism where model weights are open, while black-box frequency proxies require only sampled outputs. We treat their observed relations as validity evidence rather than equality between estimators. Across the four open models, white-box rubric reliance has a weak positive association with the black-box strong-minus-weak frequency lift ($\rho = 0.195$ over $N = 3496$ caption cases). Separately, higher visual grounding predicts lower false-support frequency ($\rho = -0.47$, source-clustered 95% CI $[-0.516, -0.419]$), providing outcome-validity evidence. The audit itself needs no gradients, fine-tuning, or occurrence ground truth, so it can rank claims whose correctness is unknown.

4.3 The Trust Score

The black-box audit yields three observable forms of support: recurrence under the full assessment input, recurrence after procedural instructions are removed, and target-specific recurrence that disappears when the video is swapped. We combine them into a fixed claim-ranking score

$$\tau(s) = \widehat{q}_{\mathrm{str}}(s) + \widehat{q}_{\mathrm{wk}}(s) + \max\big(0,\ \widehat{q}_{\mathrm{str}}(s) - \widehat{q}_{\mathrm{lk}}(s)\big), \tag{5}$$

where $\widehat{q}_{\mathrm{str}}$, $\widehat{q}_{\mathrm{wk}}$ and $\widehat{q}_{\mathrm{lk}}$ are the strong, weak and leak claim frequencies from the black-box path.

Here $\widehat{q}_{\mathrm{wk}}$ acts as a witness after the procedure is removed, while $\max(0, \widehat{q}_{\mathrm{str}} - \widehat{q}_{\mathrm{lk}})$ rewards support specific to the matching video. All terms lie in $[0, 1]$ and use fixed unit coefficients without fitting or normalization. Computing τ requires the strong, weak, and leak arms and therefore uses three times the sampling budget of a single-input score. We evaluate τ as a ranking score rather than a calibrated probability; threshold-based use requires selecting an operating point.

5 Experiments

Our experiments ask three questions across nine models and four domains: whether stronger rubrics induce more false claims, whether those claims come from the rubric rather than the video, and whether they can be detected without model internals. The results converge on recitation: rubric dose raises false claims, counterfactual interventions expose their source, and the resulting trust score ranks grounded above recited claims even for closed-source models.

5.1 EXPERIMENTAL SETUP

Datasets. Our benchmark spans four domains. Two are K-12 physics practicals recorded in schools: measuring an unknown resistance with an ammeter and a voltmeter, and investigating how current varies with voltage. One is cooking, taken from the public CaptainCook4D dataset (Peddi et al., 2024). One is a vocational auto-repair task, removing and refitting an engine cylinder head. Each video is cut into fixed 60-second clips, 874 in total. For our three collected domains, multiple domain experts annotate and adjudicate the time segment in which each rubric step occurs, yielding a present/absent label for every (clip, step) pair (collection and annotation details in Appendix A). This lets us check each claim against what happened in that clip: a step recited into a clip where it does not appear is counted as false, even if it occurs elsewhere in the video. Table 1 reports the counts. For Kitchen, we use the official video-level annotations from CaptainCook4D and broadcast them to the constituent clips. This can understate the false-claim rate and overstate coverage in this domain.

Table 1: The four-domain benchmark. Each clip is a fixed 60-second window; *Occurred*/*Absent* count labeled (clip, step) pairs. Kitchen uses video-level labels from CaptainCook4D, broadcast to its clips.

Domain	Setting	Clips	Occurred	Absent
Resistance	K-12 physics	203	548	1482
Voltage	K-12 physics	238	476	1904
Kitchen	CaptainCook4D	168	294	2033
Auto-repair	Vocational	265	1013	6407
Total		874	2331	11826

Table 2: Claim-level detection AP across nine models and four domains. *zero*, *weak*, and *strong* use the corresponding input frequency; *ours* uses the trust score τ. Bold marks scores that exceed all three single-input baselines.

	Resistance				Voltage				Kitchen				Auto-repair			
Model	zero	weak	strong	**ours**	zero	weak	strong	**ours**	zero	weak	strong	**ours**	zero	weak	strong	**ours**
Open-weight																
Qwen3.6-27B	0.40	0.55	0.56	**0.63**	0.29	0.37	0.28	**0.41**	0.35	0.47	0.51	**0.55**	0.22	0.23	0.27	**0.30**
Qwen3.6-35B	0.45	0.50	0.51	**0.57**	0.29	0.35	0.30	**0.37**	0.38	0.44	0.40	**0.45**	0.24	0.24	0.25	**0.28**
InternVL3.5-30B	0.35	0.33	0.51	0.49	0.22	0.24	0.23	**0.25**	0.38	0.41	0.36	**0.43**	0.20	0.19	0.17	0.19
InternVL3.5-38B	0.35	0.35	0.49	0.48	0.26	0.26	0.29	**0.32**	0.39	0.46	0.37	**0.49**	0.19	0.18	0.24	**0.25**
Mean	0.39	0.43	0.52	**0.54**	0.27	0.30	0.28	**0.34**	0.37	0.44	0.41	**0.48**	0.21	0.21	0.23	**0.25**
Closed-source																
Qwen3.6-Plus	0.53	0.53	0.64	**0.65**	0.42	0.44	0.48	**0.55**	0.42	0.43	0.55	**0.57**	0.24	0.24	0.26	**0.32**
GPT-4.1	0.30	0.32	0.41	**0.47**	0.18	0.27	0.22	**0.35**	0.28	0.33	0.43	**0.44**	0.19	0.17	0.26	0.26
GPT-5.4	0.27	0.32	0.53	0.53	0.26	0.29	0.44	**0.53**	0.21	0.28	0.47	0.45	0.20	0.22	0.23	**0.31**
Gemini-2.5-Pro	0.45	0.47	0.57	**0.58**	0.31	0.31	0.34	**0.38**	0.43	0.51	0.53	**0.61**	0.24	0.24	0.26	**0.28**
Gemini-3.1-Pro	0.45	0.45	0.51	**0.52**	0.31	0.32	0.33	**0.38**	0.43	0.51	0.54	**0.61**	0.24	0.23	0.26	**0.28**
Mean	0.40	0.42	0.53	**0.55**	0.30	0.33	0.36	**0.44**	0.36	0.41	0.50	**0.53**	0.22	0.22	0.25	**0.29**

Models. We evaluate nine VLMs: four open-weight models (Qwen3.6-27B, Qwen3.6-35B-A3B, InternVL3.5-30B-A3B, and InternVL3.5-38B) and five closed-source models (Qwen3.6-Plus, GPT-4.1, GPT-5.4, Gemini-2.5-Pro, and Gemini-3.1-Pro). White-box measurements are restricted to the open-weight models, whereas the black-box audit covers all nine.

Captioning and decoding. We generate five captions for each of the strong, weak, zero, and leak inputs. Each model uses its supported video interface, held fixed across the four inputs; the counterfactual results are therefore interpreted within model rather than as controlled comparisons between model families. Runtime caption and judge prompts are Chinese; Appendix C provides English versions and the Code and Data Supplement provides the executable templates. We map captions to claims using the three judge roles defined above, run locally with a shared gpt-oss-120b backbone.

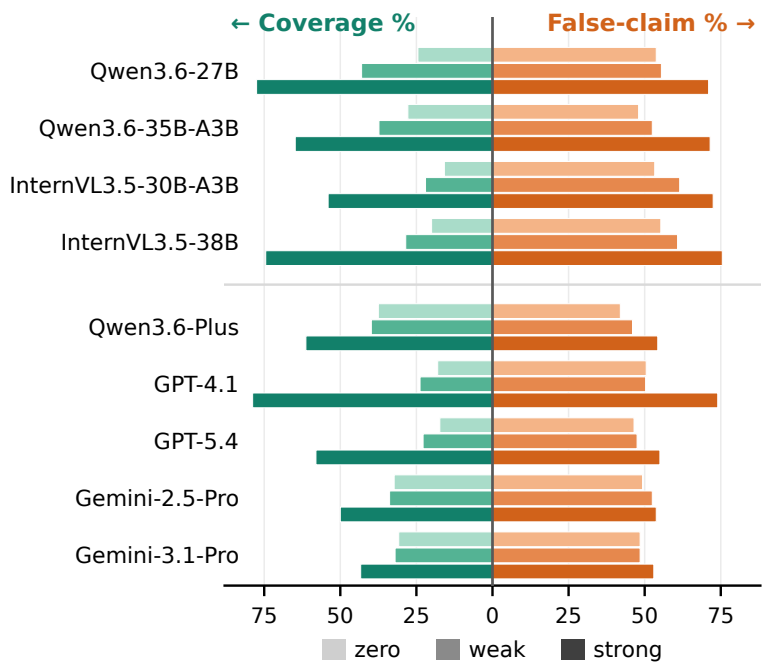

Figure 3: Rubric dose increases both coverage (left) and the false-claim rate (right). Each row is one model, with zero, weak, and strong inputs shown from light to dark; values are averaged across four domains.

Metrics. Dose metrics count steps extracted in any of the five samples per input. We report false-claim rate, $\mathrm{false}/(\mathrm{false} + \mathrm{true})$, and coverage, $\mathrm{true}/\mathrm{occurred}$. Detection uses average precision (AP) over all labeled clip–step pairs, with occurred steps as positives and their prevalence as the reference baseline (Manakul et al., 2023).

5.2 Main Results

More rubric, more claims. We caption the same clips under zero, weak, and strong rubric inputs and check each claimed step against the clip. On the four open-weight models, Figure 3 shows the false-claim rate rising from about 52% under zero to 73% under the full rubric, while coverage rises from 22% to 33% to 68%. Every prompt retains the same visible-only, no-guessing instruction, so the monotonic change tracks added task information rather than removal of an anti-confabulation caveat. Strong has the highest domain-average error for every model; cell-level and closed-source results are in Appendix F. Yet the coverage gain confirms that the rubric is informative: discarding it is not a solution, so the audit must distinguish what the model saw from what it recited.

The claims are recited, not seen. Counterfactual CPMI provides the primary source test: removing the rubric procedure measures rubric reliance (CPMI_r), while swapping the video measures visual grounding (CPMI_v). A clip is recitation-dominant when $\mathrm{CPMI}_r > \mathrm{CPMI}_v$. The rubric dominates resistance, voltage, and auto-repair on 54%–99% of clips (Table 3); Kitchen reverses this pattern in three of four models, with CPMI_v reaching 0.72. Because CPMI uses no occurrence labels, the reversal is independent of Kitchen's video-level label broadcast. The signature therefore separates rubric- and video-dominant cases rather than penalizing all rubric-conditioned captions.

On the two Qwen models, 70% and 79% of caption tokens place more attention mass on rubric than video tokens. This complementary analysis is correlational and limited to the two physics domains; Appendix Figure 4 gives the full plot.

Sampling-based auditing extends to closed models. White-box CPMI requires token log-probabilities, so sample recurrence serves as a black-box proxy. On open models, rubric reliance weakly correlates with the strong-minus-weak lift ($\rho = 0.195$, $N = 3496$ caption cases), while

Table 3: Caption-level white-box recitation signature. Here r/v denote $\mathrm{CPMI}_r/\mathrm{CPMI}_v$, and *Rec.* is the percentage of clips where $r > v$. Bold marks the larger CPMI and, within each domain, the highest *Rec.*; ↑ / ↓ mark rubric-/video-dominant pairs.

	Resistance			Voltage			Kitchen			Auto-repair		
Model	r	v	Rec.	r	v	Rec.	r	v	Rec.	r	v	Rec.
Qwen3.6-27B	**0.39**	0.18	68.5↑	**0.74**	0.11	80.3↑	0.39	**0.54**	31.0↓	**0.47**	0.36	58.9↑
Qwen3.6-35B	**0.36**	0.26	60.6↑	**0.76**	0.14	88.7↑	**0.78**	0.55	**55.4**↑	**0.44**	0.35	53.6↑
InternVL3.5-30B	**1.41**	0.26	**98.5**↑	**1.37**	0.25	**93.7**↑	0.66	**0.72**	43.5↓	**1.00**	0.20	**97.4**↑
InternVL3.5-38B	**0.33**	0.21	69.0↑	**0.45**	0.19	82.8↑	0.40	**0.59**	36.9↓	**0.75**	0.30	87.5↑

stronger visual grounding predicts lower sampled support for absent steps ($\rho = -0.47$, $N = 3492$ valid model–clip pairs; source-clustered 95% CI $[-0.516, -0.419]$). Across the four domain averages, τ achieves AP at 2.0–4.0 times the true-claim prevalence baseline; Table 2 gives all model–domain AP results. The sampling-only score also applies to five closed models; zero is used only for dose analysis. Appendix Figure 5 illustrates retrospective cross-clip scoring.

5.3 Analysis and Ablations

The auxiliary terms have domain-dependent effects. Component ablations show that the witness term contributes most on resistance, voltage, and Kitchen, and the grounding reward on auto-repair. Using the weak rather than zero witness improves Kitchen AP for all nine models (Appendix Table 13).

Recitation limits consistency checking. Rubric-recited false claims are nearly as stable as true ones: their consistency gap is only 0.08–0.18, and SelfCheckGPT's false-claim AP exceeds the base rate by just 0.01–0.11 across nine models (Appendix Table 14). The three-of-five rule is an operating point of the strong-input recurrence score, not a separate ranking method. Recurrence alone therefore offers limited discrimination.

Correction and filtering expose different operating points. On Resistance, Voltage, and Kitchen, Woodpecker (Yin et al., 2024) reaches a 46%–54% false-claim rate at 22%–73% coverage and can retain more true claims than fixed claim filters. VCD (Leng et al., 2024) lowers the rate by under ten points but retains only 5%–9% coverage; weak-witness filtering reaches 46%–58% at 19%–46% coverage (Appendix Table 15). Correction rewrites captions, whereas τ ranks claims for a selectable review budget.

6 Conclusion

Performance-based assessment needs the rubric, yet that rubric can make a VLM report steps it never saw. Across four domains and 874 clips, we observe rubric-induced confabulation alongside a clear recitation pattern: resistance, voltage, and auto-repair are recitation-dominant on 54%–99% of clips, whereas Kitchen reverses the pattern in three of four models. The contrast identifies a claim's source rather than penalizing rubric conditioning itself, providing a machine analogue of rater expectation bias. More broadly, the task prior that improves coverage can also weaken the link between generated claims and visual evidence.

Recitation also explains why standard remedies are incomplete: false claims recur almost as consistently as true ones, while mitigation reduces them with model-dependent coverage losses. Our sampling-only counterfactual score instead ranks asserted steps for human review and extends to closed models without internal probabilities.

Provider-specific interfaces and annotation granularity limit comparisons, so our contrasts remain within-model and mitigation results method-dependent. The uncalibrated ranking supports, rather than replaces, human scoring.

Ethics Statement

Collection of the three non-public datasets was approved through the authors' institutional ethics process. The institution and protocol identifier are withheld during double-blind review and will be restored in the camera-ready version. Participating institutions authorized collection for research use. The recordings use first-person cameras and contain neither participant faces nor direct personal identifiers. Data release is restricted to the authorized research scope, and CaptainCook4D remains governed by its original terms of use. The proposed audit is intended to support human review, not to automate consequential certification decisions.

AI Use Statement

Generative AI tools aided writing, LaTeX formatting, and minor revisions to supplementary code. They were not used to generate research data or experimental results. The authors reviewed all AI-assisted edits and take responsibility for the paper's final content.

Reproducibility Statement

The experimental setup, prompt design, estimators, evaluation metrics, and statistical procedures are documented in the main paper and Appendix. The accompanying Code and Data Supplement provides representative clips, runtime prompts, model configurations, implementation materials, and selected aggregate results. The complete research repository will be released upon acceptance.

References

John R Anderson. Acquisition of cognitive skill. *Psychological review*, 89(4):369, 1982.

Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European Conference on Computer Vision*, pp. 422–438, 2022.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. ContextCite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514. IEEE, 2024.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FAITHScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Robert L Linn, Eva L Baker, and Stephen B Dunbar. Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher*, 20(8):15–21, 1991.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Carol M Myford and Edward W Wolfe. Detecting and measuring rater effects using many-facet rasch measurement: Part ii. *Journal of applied measurement*, 5(2):189–227, 2004.

OECD. *PISA Vocational Education and Training (VET): Assessment and Analytical Framework*. OECD Publishing, Paris, 2024. doi: 10.1787/b0d5aaf9-en. URL https://doi.org/10.1787/b0d5aaf9-en.

Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 3204–3213, 2024.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, et al. Captain-Cook4D: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems*, 37:135626–135679, 2024.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Robert Rosenthal and Kermit L Fode. The effect of experimenter bias on the performance of the albino rat. *Behavioral science*, 8(3):183–189, 1963.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Masaki Uto. A multidimensional generalized many-facet rasch model for rubric-based performance assessment. *Behaviormetrika*, 48(2):425–457, 2021.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025.

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

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

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

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. FineDiving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2949–2958, 2022.

Jinglin Xu, Yongming Rao, Jie Zhou, and Jiwen Lu. Procedure-aware action quality assessment: Datasets and performance evaluation. *International journal of computer vision*, 132(12):6069–6090, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual policy optimization for multimodal reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 33:1600–1613, 2024.

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

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*, 2024b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Appendix

This appendix documents the dataset and experimental protocol, provides reproducibility and statistical details, reports complete results and additional analyses, and concludes with a worked audit example and proof.

A Dataset Documentation

A.1 Dataset Selection and Scientific Role

The benchmark combines four domains chosen to vary both application setting and visual observability. Resistance and voltage are K–12 physics practicals in which scorable evidence often appears in small instrument readings and fine-grained wiring operations. Auto-repair is a longer vocational procedure with a 28-step rubric and substantial visual occlusion. Kitchen, drawn from the public CaptainCook4D dataset (Peddi et al., 2024), contains larger and more visually legible actions. Together, the domains span educational and vocational assessment, short and long rubrics, and actions ranging from subtle instrument manipulations to conspicuous object interactions.

Existing public procedural-video datasets do not provide all of the supervision required by our source audit. In particular, the audit needs a rubric and a clip-level occurrence label for every rubric step, including steps that are absent from a clip. CaptainCook4D supplies public procedural videos and official action annotations; for Kitchen, we use its video-level labels. The three collected domains provide the interval-level labels needed to test whether a rubric claim occurred in a particular clip.

A.2 Composition and Clip Construction

Table 4 expands the benchmark inventory in the main paper. The 874 clips are non-overlapping 60-second windows drawn from 104 source videos. Across all clips, the annotation matrix contains 2,331 occurred and 11,826 absent (clip, step) pairs. A claim is labeled true only when its rubric step occurs in the corresponding clip; occurrence elsewhere in the same source video does not make the claim true for that clip. For Kitchen, the official video-level CaptainCook4D labels are broadcast to the constituent clips. This can understate the measured false-claim rate and overstate coverage, so all Kitchen results should be read under this conservative broadcast protocol.

The three collected datasets are first-person recordings acquired at 4K/60 fps. Experiment clips are 640×360 at 30 fps for the two physics domains and 24 fps for auto-repair; Kitchen is 640×360 at $30000/1001$ fps. The two physics rubrics contain 10 steps each, auto-repair contains 28, and the four Kitchen recipes contain 9, 12, 17, and 19 steps.

A.3 Annotation Protocol

For the three collected domains, multiple domain experts label the interval in which each rubric step occurs, and disagreements are adjudicated before mapping intervals to 60-second clips. Raw inter-annotator agreement ranges from 87.4% to 94.6% across these domains. The final reconciled annotations are used as ground truth. Kitchen instead uses the official video-level annotations from CaptainCook4D, broadcast to its constituent clips.

A.4 Release Scope and Documentation

For anonymous review, the accompanying Code and Data Supplement includes representative clips, code, and other materials documenting the experimental workflow, including prompt templates, model configurations, selected aggregate results, dataset summaries, machine-readable schemas, and design manifests. The complete research repository will be released upon acceptance.

Table 4: Dataset inventory. *Sources* counts source videos.

Domain	Sources	Clips	Occurred	Absent
Resistance	35	203	548	1482
Voltage	38	238	476	1904
Kitchen	18	168	294	2033
Auto-repair	13	265	1013	6407
Total	104	874	2331	11826

B Assumptions and Scope

Information-theoretic background. Pointwise mutual information is the log ratio between the probability of an outcome under joint conditioning and its probability under the corresponding marginal; conditional mutual information averages such pointwise ratios while holding a third variable fixed (Church & Hanks, 1990; Cover & Thomas, 2006). Our quantities use this log-ratio form with controlled references. Visual grounding compares the matching video with a distribution of unrelated replacement videos, while rubric reliance compares the full rubric with a controlled noun-only version. They should therefore be read as CPMI-style counterfactual evidence gaps, not as estimates obtained by marginalizing over every possible video or rubric.

Distributional intervention. Visual grounding is defined against the distribution of unrelated replacement videos, whereas the white-box experiment substitutes one fixed replacement per target. The bound in the Proofs section is therefore an expectation statement over the replacement-video distribution, not a guarantee about the thresholded detection set or any individual target–replacement pair. A valid replacement must be unrelated to the target observation while leaving the rubric and caption-scoring procedure unchanged.

Sampling proxy. The black-box path records empirical claim recurrence under the five-run protocol. These frequencies summarize the stated sampling procedure, and their relationship to the white-box evidence gaps is evaluated empirically rather than assumed to be an exact equality.

Scope of claims. The single-replacement result is a limited expectation bound for the white-box approximation and is proved in the Proofs section. The trust score is a fixed-form ranking rule whose behavior is evaluated empirically rather than asserted as an optimal or calibrated estimator. The recitation results characterize the evaluated models, prompts, rubrics, and four domains.

C Detailed Experimental Protocol

C.1 Counterfactual Inputs

The four inputs implement two contrasts. Along the rubric axis, the strong input contains the complete ordered scoring rubric, the weak input retains only equipment and action nouns while removing ordering, step boundaries, and conditions, and the zero input supplies no task information. All three arms end with the same instruction to report only directly visible content, avoid guessing, and exclude prompt-only items from every JSON field. The rubric-axis contrast therefore varies the task information preceding this common instruction. Along the video axis, the leak input retains the strong input's scenario, complete target rubric, caption instruction, and output schema verbatim, but replaces the matching target clip v with an unrelated real video $\tilde{v}$. Strong versus leak therefore changes only the video source.

Across all four domains, each replacement is a visually substantive real recording drawn from a different procedure and source recording and is unrelated to the target procedure. The assignment is fixed before generation: each target–replacement pair is reused for all five black-box samples and for the corresponding white-box contrast. Consequently, sampling variation reflects the model's conditional generation rather than repeated changes to the replacement itself.

C.2 Why the Video Intervention Uses Real Videos

An empty or blank visual input changes both the evidential content and the nature of the multimodal task. An unrelated real video instead preserves a normal, describable visual stream while selectively breaking its correspondence with the target rubric. If a target-rubric claim persists after this swap, the persistence cannot be attributed to visual evidence specific to the original clip; if it disappears, the matching video contributed evidence for the claim.

C.3 Caption Generation and Repeated Sampling

For each model and clip, we generate five captions under each of the strong, weak, zero, and leak inputs. Sample 0 uses temperature 0 and samples 1–4 use temperature 0.5 where the endpoint exposes this control. Open models use a maximum of 1,200 output tokens; GPT and Qwen3.6-Plus use 1,024, and Gemini uses 2,048. Unspecified top-p, top-k, and related decoding fields retain their framework defaults. Qwen disables thinking; GPT-5.4 uses minimal reasoning effort. The same claim-extraction protocol is applied to every generated caption, yielding assertion frequencies in increments of $1/5$ for the black-box audit.

The GPT models receive 20 uniformly indexed frames, including both endpoints, with frame index $\mathrm{round}(i(F-1)/19)$ for $i = 0, \ldots, 19$ and F source frames; if $F < 20$, every frame is used. Each frame is resized so its longest side is 512 pixels and encoded as JPEG at quality 85. All other evaluated models pass the video to their multimodal processor. InternVL permits at most 48 image inputs and uses repetition penalty 1.0. Table 7 gives the local serving settings, and Table 8 records API-specific controls.

C.4 Prompt Templates and Claim Extraction

Across the four inputs, the visibility constraint and output requirements are held fixed while rubric content or video source changes as described above. A judge model reads the generated assessment record together with the numbered rubric steps and returns the steps asserted by that record. Here, an assertion means content that enters the model-produced record—whether as free text or in its action field—rather than a claim about the model's inaccessible internal belief. The lenient role counts an activity of the same kind; the strict role requires specific textual support; and the reverse role identifies unsupported steps, whose complement we take. A step is claimed when at least two of the three roles agree. All roles use the same locally run gpt-oss-120b backbone, and the same protocol is used for all models, domains, and counterfactual inputs.

The blocks below are English versions of the Chinese runtime prompts. Each condition-specific prefix is followed immediately by the shared suffix; no system message is set. Angle-bracketed fields are filled from the experiment configuration.

Shared Caption-Prompt Suffix

Describe only what is visible; do not judge correctness or assign a score. Do not invent anything unclear or absent. **Do not use technical terms or action names that cannot be verified from the image** (when an identity is unclear, describe only shape, color, and position).
In `caption`, `tools`, `parts`, and `actions`, include only content that can be directly verified from this clip. Do not copy any equipment, action, or rubric step mentioned in the prompt unless it appears in the video.
Output JSON only: `{"caption":"...","tools":[],"parts":[],"actions":[]}`

Zero- and Weak-Input Prefixes

Zero input:
Objectively describe the actions and visual content **actually seen** in this video clip, in the order in which the actions occur.

Weak input:
This is a clip from a video of `<SCENARIO>`.
The following nouns are **naming references only** (use them only when directly visible; they are not necessarily present in this clip):
Equipment: `<TOOLS>`
Actions: `<ACTIONS>`
Objectively describe the actions and visual content **actually seen** in this clip, in temporal order.

Strong-Input Prefix and Leak Intervention

Strong-input prefix:
This is a clip from a video of `<SCENARIO>`. The **standard correct procedure** for this task includes:
`- <RUBRIC STEP 1>`
`- <RUBRIC STEP 2>`
`- ...`
Objectively describe the actions and visual content **actually seen** in this clip, in temporal order.

Leak input: The caption prompt is the strong-input prefix and shared suffix verbatim, instantiated with the target scenario and target rubric. The matching target clip is replaced by an unrelated real video; prompt wording, JSON schema, decoding configuration, number of samples, and downstream claim-judging pipeline remain fixed. Thus only the identity of the visual input changes.

The resulting zero template inserts neither scenario nor rubric text. The weak template inserts the scenario and configuration-derived equipment and action nouns as naming aids. The strong template inserts the scenario and complete ordered rubric. The leak input uses that same strong template verbatim.

Judge outputs are parsed deterministically, restricted to valid step IDs, and reduced by the same two-of-three rule. The reverse set is complemented over the active rubric-step universe before majority reduction.

All three judge prompts insert the same `CAPTION`, ordered `RUBRIC STEPS`, and `SCENARIO` fields. The three cards below give English versions of their runtime prompts. The Chinese templates are also supplied as plain-text files.

Lenient Judge Prompt (English Version)

Below is an **objective description** of a video clip showing `<SCENARIO>` (it may be coarse, mix Chinese and English, or use an incorrect tool name):
"`<CAPTION>`"
Below is the list of **standard steps** for this task:
`<NUMBERED RUBRIC STEPS>`
For each step, determine whether the objective description says that **an activity of the same kind actually occurred in this clip**.
Be lenient at the activity level: count the step as **described** if the caption does not separate its substeps, names a tool incorrectly, or uses different wording for the same kind of activity. Count it as **not described** only when this kind of activity is wholly absent, without even a coarse mention.
Analyze every step first: write one line per step, quoting the corresponding sentence or stating that it was not mentioned. Give the decision only after this analysis.
Finally, on a new line, give the conclusion in a JSON code block containing only the IDs of described steps:
`{"described":[<STEP IDS>]}`

Strict Judge Prompt (English Version)

Below is an **objective description** of a video clip showing `<SCENARIO>`:
"`<CAPTION>`"
Below is the list of **standard steps** for this task:
`<NUMBERED RUBRIC STEPS>`
Count a step as **described** only when the caption contains **a sentence that clearly corresponds to that step**, and copy that sentence verbatim.
Strict rules:

- Generic phrases (e.g., "record data," "organize equipment," or "perform the operation") cannot be mapped to a specific step (e.g., "record the second set of data") and do not count.
- Merely writing a table header or preparation action does not mean that the step occurred.
- The supporting sentence must identify what distinguishes this step from others (which trial, which object, or which order).

Analyze every step first: state whether a clearly corresponding sentence can be found. If so, copy it and explain why it specifically supports this step rather than a generic activity.
Finally, on a new line, give the conclusion in a JSON code block:
`{"items":[{"step":<ID>,"quote":"<SUPPORT>"}]}`

Reverse Judge Prompt (English Version)

You are a strict verifier. Below is an **objective description** of a video clip showing `<SCENARIO>`:
"`<CAPTION>`"
List of **standard steps**:
`<NUMBERED RUBRIC STEPS>`
Check every step: for which steps does the description **provide no evidence that the step actually occurred in this clip**?
Note: if the description mentions a corresponding activity of the same kind, count that as evidence. A generic reference, or related nouns without the step's action, does not count as evidence.
Analyze every step first: state whether evidence exists and identify the supporting sentence.
Finally, on a new line, give the conclusion in a JSON code block containing only the IDs of steps for which no evidence was found:
`{"no_evidence":[<STEP IDS>]}`

C.5 Human Validation and Output-Field Audit

We validate the caption-to-claim reduction on a fixed, human-audited sample of 200 generated records. For each record, the auditor reads exactly the caption and structured action field supplied to the three judge roles, compares their majority step set with the numbered rubric, and supplies the complete corrected set when they disagree. This yields 3,113 binary record–step decisions. Table 5 reports agreement against the human sets. The machine sets are conservative: all 36 disagreements at the decision level are omissions, with no unsupported additions.

Table 5: Human validation of the three-role claim extractor ($N = 200$ records; 3,113 binary record–step decisions). Exact-set, precision, and recall intervals are Wilson intervals; F1 and κ intervals use 10,000 bootstrap resamples of complete records (seed 0627).

Metric	Estimate	95% CI
Exact-set agreement	0.875	[0.822, 0.914]
Precision	1.000	[0.987, 1.000]
Recall	0.891	[0.852, 0.920]
F1	0.942	[0.916, 0.965]
Cohen's κ	0.936	[0.907, 0.961]

Because the judge receives both output fields, we additionally test whether the observed false-claim rate is merely an artifact of rubric content being reproduced in the structured `actions` field. We partition each of the 329 human-confirmed claims in the validation sample by whether it is textually

supported by the free-text caption only, the action field only, or both, and then join these claims to the existing occurrence labels. Table 6 gives the resulting mutually exclusive accounting. Caption-only and action-only claims have nearly identical false-claim rates. More importantly, after conservatively excluding every claim with action-field support, the 45 remaining caption-only claims still have a 75.6% false-claim rate. The phenomenon is therefore not driven solely by copying into the structured field.

Table 6: Post-hoc output-field audit on the 329 human-confirmed claims. Categories are mutually exclusive; FCR uses the existing clip–step occurrence labels.

Exclusive textual source	Claims	False	FCR
Caption only	45	34	75.6%
Actions only	126	96	76.2%
Both fields	158	93	58.9%

C.6 White-Box Teacher Forcing

On an open-weight model, we freeze the strong-input caption c and teacher-force its tokens under the strong, weak, zero, and leak inputs. We compute the mean token log-probability

$$\ell(c \mid \cdot) = \frac{1}{|c|} \sum_i \log P(c_i \mid c_{<i}, \cdot). \tag{6}$$

The text and every preceding caption token remain fixed while only the input changes. Differences in ℓ give the reported length-normalized caption-level contrasts: strong-minus-weak measures rubric reliance, strong-minus-leak measures visual grounding, and strong-minus-zero supplies the bottom dose contrast. Before aggregation, we verify that scored tokens and offsets exactly match the generated caption.

Freezing the caption prevents input and output variation from being conflated: independently regenerating under each intervention would change both at once. Teacher forcing instead conditions every score on the same caption prefix, isolating the input intervention without requiring gradients or fine-tuning.

Implementation uses natural log-probabilities. A fixed sample-0 strong caption is retokenized without special tokens and teacher-forced after the multimodal prompt. Only caption tokens enter the mean: prompt, rubric, video, chat-template, BOS, and EOS positions are excluded. InternVL token-span checks confirm that the scored offsets match the generated caption.

C.7 Black-Box Frequencies and Trust Score

Let $\widehat{q}_{\mathrm{str}}$, $\widehat{q}_{\mathrm{wk}}$, $\widehat{q}_{\mathrm{zr}}$, and $\widehat{q}_{\mathrm{lk}}$ denote the frequencies with which a rubric step is extracted from the five captions under the corresponding inputs. Strong-minus-weak and strong-minus-leak serve as black-box proxies for rubric reliance and visual grounding. The claim-level ranking score is

$$\tau(s) = \widehat{q}_{\mathrm{str}}(s) + \widehat{q}_{\mathrm{wk}}(s) + \max\big(0, \widehat{q}_{\mathrm{str}}(s) - \widehat{q}_{\mathrm{lk}}(s)\big). \tag{7}$$

All terms lie in $[0, 1]$ and use fixed unit coefficients without fitted weights or normalization. The score is evaluated as a ranking signal, not as a calibrated probability. The score itself uses three sampled inputs—strong, weak, and leak. The zero input is retained for the rubric-dose characterization but is not an additional term in τ.

D Reproducibility Details

D.1 Models and Computing Infrastructure

We evaluate four open-weight VLMs and five closed-source VLMs. Table 7 lists the open-model serving configurations; exact checkpoint revisions are given below. Closed APIs are identified by their provider model names because versioned checkpoint revisions are not exposed.

Table 7: Open-model serving settings. All use bfloat16 without additional quantization. TP is tensor parallel size; Qwen3.6-27B also uses data parallel size 2.

Model	Runtime	TP	Context	Output
Qwen3.6-27B	SGLang 0.5.11	4	262144	1200
Qwen3.6-35B	SGLang 0.5.11	4	262144	1200
InternVL3.5-30B	vLLM 0.24.0	2	16384	1200
InternVL3.5-38B	vLLM 0.24.0	4	16384	1200

The repositories and revision prefixes are `Qwen/Qwen3.6-27B` (`6a9e13bd6f`), `Qwen/Qwen3.6-35B-A3B` (`995ad96eac`), `OpenGVLab/InternVL3_5-30B-A3B` (`ba1762d830`), and `OpenGVLab/InternVL3_5-38B` (`de99855be3`). The complete 40-character revisions are included in the accompanying configuration files.

Table 8: API settings. Output is the token cap; $T = 0/0.5$ means sample 0 uses 0 and samples 1–4 use 0.5. No API seed is requested.

API model	Input	Output	Control
GPT-4.1	20 frames	1024	$T = 0/0.5$
GPT-5.4	20 frames	1024	minimal reasoning; $T = 0/0.5$
Gemini-2.5-Pro	video	2048	defaults
Gemini-3.1-Pro	video	2048	defaults
Qwen3.6-Plus	video	1024	$T = 0/0.5$

Claim extraction uses `openai/gpt-oss-120b` revision `b5c939de8f`, served with SGLang using the model's native MXFP4 weights and bfloat16 compute. Its material settings are TP=4, context length 32,768, temperature 0, maximum output 4,096 tokens, and server seed 204759902.

Experiments run on two machines, each with eight NVIDIA RTX 4090 GPUs (24 GiB each), an Intel Xeon Gold 6530 CPU, and 125 GiB RAM, under Ubuntu 24.04 LTS. Video processing uses FFmpeg 6.1.1. The Qwen serving environment uses Python 3.12.13, PyTorch 2.11.0+cu130, CUDA 13.0, cuDNN 9.19.0, Transformers 5.6.0, torchvision 0.26.0, SGLang 0.5.11, OpenCV 4.10.0.84, Pillow 11.3.0, NumPy 2.3.5, pandas 3.0.2, and SciPy 1.17.1. InternVL is served with vLLM 0.24.0.

D.2 Parameter Selection and Randomness

The black-box sample count is fixed at five per input, and the trust-score coefficients are not fitted to the reported test results. The Qwen servers use seeds 820670465 and 32873699 for the 27B and 35B models, respectively; InternVL uses seed 0. Closed APIs are called without a requested seed, so provider-side nondeterminism prevents bitwise reproduction. Statistical resampling uses seed 0627.

D.3 Run Counts and Baseline Configurations

The complete black-box design contains $9 \times 874 \times 4 \times 5 = 157{,}320$ model–clip–input–sample cells. Each caption is processed independently by three judge roles, giving 471,960 role-level evaluations before majority reduction. The five samples remain separate until claim extraction; a claim frequency is then the number of samples in which the majority extracts that step, divided by five. Derived frequencies, the trust score, and its component ablations require no additional VLM generation.

White-box scoring is restricted to the four open models. For each clip, it evaluates one frozen sample-0 strong caption under all four inputs without regenerating the target text, giving $4 \times 874 = 3{,}496$ scored captions. The pooled outcome-validity analysis contains 3,492 valid paired observations.

D.4 Baseline Implementations

All baselines are reduced to the same rubric-step universe and evaluated by the same occurrence labels and metrics as our method. Our rubric-step adaptation of SelfCheckGPT (Manakul et al.,

2023) reuses the five strong-input captions; for each step, cross-sample consistency is the fraction of those captions supporting the claim, and inconsistency is used as the detection score. Thus it requires no additional caption generation. The reported threshold diagnostic retains a claim when at least three of five strong samples support it.

Our video adaptation of Woodpecker (Yin et al., 2024) corrects each strong-input caption through claim extraction, visual verification, and rewriting. We generate five corrected outputs per evaluated clip and pass every nonempty correction through the same three-role claim extractor. Empty corrections contribute no claims rather than being silently dropped. This preserves a like-for-like false-claim-rate/coverage comparison between correction and the original captioning path.

VCD (Leng et al., 2024) is run with contrast weight $\alpha = 1.0$ and five decodes per clip at temperature 0.5 on Qwen3.6-27B and Qwen3.6-35B-A3B. It requires token-by-token control of the decoding loop, unavailable through the standard InternVL vLLM serving path used here. Its captions are judged identically to the other methods. Trust-score component variants reuse the four claim frequencies.

For the retrospective filtering comparison, *Ours-gate* quarantines a strong claim when zero-input support is zero and stored replacement persistence is at least 0.6; where persistence is unavailable, it uses the zero-support condition alone. *Ours-weak* retains strong claims with $\widehat{q}_{\mathrm{wk}} > 0$. These fixed filters are evaluated separately from the ranking score τ.

D.5 Failure Handling and Audit Accounting

An empty caption is retained as a zero-claim sample, contributing zero to the coverage numerator while the occurrence denominator stays fixed. Parser outputs are restricted to integer IDs in the active rubric; invalid IDs are discarded, and the reverse role is complemented only after this filtering. Because each valid role contributes a binary decision for every step, the two-of-three reducer always has a majority. Unsupported model–baseline combinations remain missing and are never converted to zero-valued performance cells.

E Statistical Analysis and Reliability

We report false-claim rate as $\mathrm{false}/(\mathrm{false} + \mathrm{true})$, coverage as $\mathrm{true}/\mathrm{occurred}$, and trust-score average precision (AP) with true claims as the positive class. We use the positive-class prevalence as the reference baseline. The SelfCheckGPT analysis instead evaluates false-claim detection and states its positive class explicitly. Direct proxy-validity relations compare the white-box gaps with their corresponding recurrence contrasts; outcome validity compares white-box visual grounding with black-box false-support frequency.

Because clips can share a source video, the pooled correlation uses source-clustered resampling. The analysis contains 35, 38, 18, and 13 source clusters in resistance, voltage, Kitchen, and auto-repair, respectively. It resamples these 104 source videos with replacement and retains all model–clip observations from each selected source. With 10,000 percentile resamples and seed 0627, the pooled Spearman estimate is $\rho = -0.4673$ for $N = 3492$ valid model–clip pairs, with a source-clustered 95% confidence interval of $[-0.5163, -0.4191]$.

Detection AP evaluates whether the trust score provides a useful ordering over the complete labeled clip–step universe under severe class imbalance. Comparisons with the individual input frequencies assess ranking quality.

The four model-specific pooled correlations are all negative: -0.6771 (Qwen3.6-27B), -0.3966 (Qwen3.6-35B-A3B), -0.3428 (InternVL3.5-30B-A3B), and -0.4409 (InternVL3.5-38B).

On the rubric axis, white-box rubric reliance has a weak positive association with the black-box strong-minus-weak claim-frequency lift ($\rho = 0.195$, $N = 3496$), providing direct proxy-validity evidence for the text contrast.

F Complete Results and Additional Analyses

F.1 Per-Domain False-Claim Rates and Coverage

Tables 9–12 report the complete model–domain results without compressing the model names or domain labels. Each entry is false-claim rate / coverage (%), and Z, W, and S denote zero, weak, and strong input, respectively.

Table 9: False-claim rate / coverage (%) on Resistance.

Model	Zero (Z)	Weak (W)	Strong (S)
Qwen3.6-27B	41/16	39/44	64/93
Qwen3.6-35B	30/22	37/36	60/73
InternVL3.5-30B	49/9	61/17	60/77
InternVL3.5-38B	50/10	53/17	67/89
GPT-4.1	27/11	45/16	72/97
GPT-5.4	18/6	35/12	43/46
Gemini-2.5-Pro	38/26	43/28	37/50
Gemini-3.1-Pro	37/23	38/24	40/39
Qwen3.6-Plus	30/41	32/40	41/66

Table 10: False-claim rate / coverage (%) on Voltage.

Model	Zero (Z)	Weak (W)	Strong (S)
Qwen3.6-27B	60/23	65/45	77/97
Qwen3.6-35B	55/21	56/34	77/79
InternVL3.5-30B	51/6	66/14	76/51
InternVL3.5-38B	55/13	65/24	77/74
GPT-4.1	50/1	33/13	79/94
GPT-5.4	41/10	31/14	59/72
Gemini-2.5-Pro	60/30	61/32	69/62
Gemini-3.1-Pro	57/30	55/30	67/50
Qwen3.6-Plus	38/34	38/36	62/76

Table 11: False-claim rate / coverage (%) on Kitchen.

Model	Zero (Z)	Weak (W)	Strong (S)
Qwen3.6-27B	54/36	54/56	70/73
Qwen3.6-35B	49/41	54/53	75/65
InternVL3.5-30B	46/37	53/47	68/57
InternVL3.5-38B	48/44	58/55	76/84
GPT-4.1	59/33	55/40	66/74
GPT-5.4	60/18	56/30	52/67
Gemini-2.5-Pro	41/47	46/58	52/68
Gemini-3.1-Pro	44/44	43/57	49/65
Qwen3.6-Plus	43/47	52/53	48/66

Table 12: False-claim rate / coverage (%) on Auto-repair.

Model	Zero (Z)	Weak (W)	Strong (S)
Qwen3.6-27B	60/23	64/27	73/46
Qwen3.6-35B	58/27	63/26	74/42
InternVL3.5-30B	67/11	66/10	86/31
InternVL3.5-38B	68/13	67/18	82/51
GPT-4.1	66/27	68/26	79/50
GPT-5.4	67/35	68/35	66/47
Gemini-2.5-Pro	58/26	60/17	57/20
Gemini-3.1-Pro	56/26	58/17	56/19
Qwen3.6-Plus	57/28	62/30	66/37

F.2 Attention-Based Evidence

On Qwen3.6, whose hybrid attention exposes a token-to-token matrix at its full-attention layers, we measure for each generated caption token the share of its attention over the video and rubric tokens that lands on the video, which we call the visual attention ratio (VAR). The analysis covers the two Qwen models on the resistance and voltage domains. Most tokens place more attention mass on the rubric than on the video: 70% and 79% fall below the VAR balance point of 0.5. This architecture-specific analysis is correlational and serves as complementary evidence to the counterfactual CPMI test in the main paper.

Attention is averaged across heads. Video positions are identified by the model's video-token ID, rubric positions by the located token span of the verbatim rubric block, and the VAR denominator contains only those two token sets. VAR compares total attention mass, without dividing by each modality's token count. These architecture-specific measurements are descriptive rather than an independent causal test.

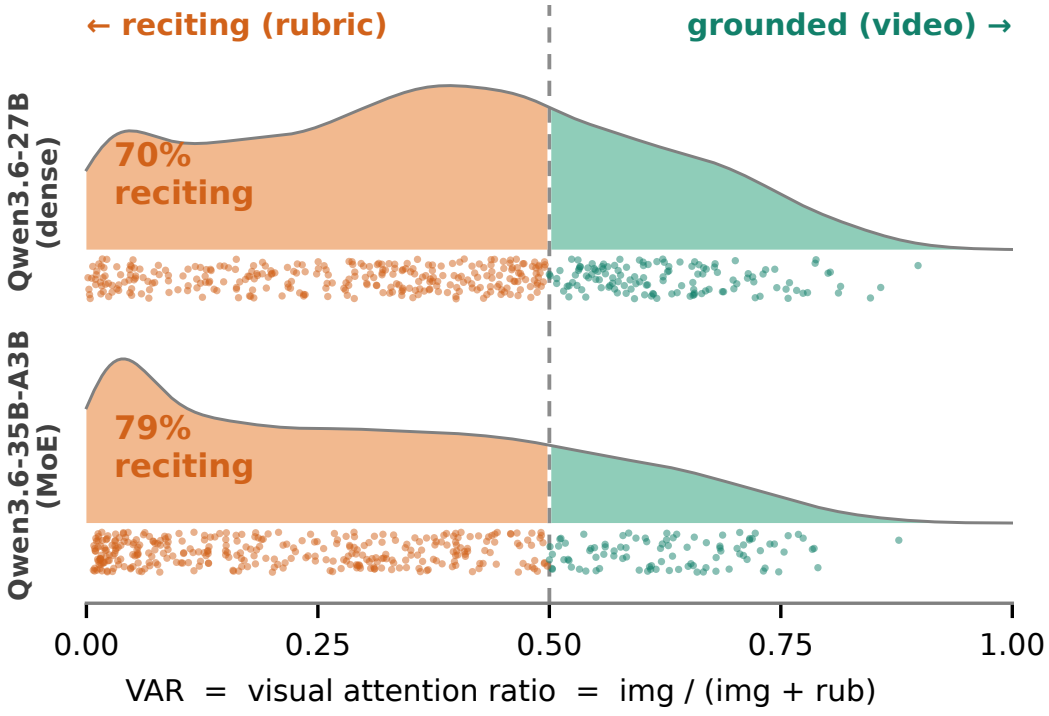

Figure 4: Each dot is one generated caption token. VAR is the share of attention over video and rubric tokens that lands on the video, so tokens left of 0.5 place more attention mass on the rubric. On the dense and MoE models, respectively, 70% and 79% of tokens fall below 0.5.

F.3 TRUST-SCORE COMPONENT ABLATION

Table 13(a) separates the weak-input witness from the video-swap grounding reward. Panel (b) replaces only $\widehat{q}_{\mathrm{wk}}$ with $\widehat{q}_{\mathrm{zr}}$, holding the other score terms fixed. Per-model witness comparisons across all four domains are provided in the Code and Data Supplement (`trust_witness_ablation.csv`).

Table 13: Trust-score ablations. (a) Component AP averaged over nine models within each domain. (b) Zero- versus weak-input witness on Kitchen, with the other score terms unchanged.

(a) Component ablations

Variant	Resist.	Volt.	Kitchen	Auto
Trust score τ	0.546	0.393	0.509	0.273
without witness	0.522	0.354	0.481	0.267
without grounding	0.547	0.382	0.506	0.253
Strong input only	0.526	0.324	0.463	0.244

(b) Witness choice on Kitchen

Model	Zero witness	Weak witness
Qwen3.6-27B	0.507	0.551
Qwen3.6-35B	0.428	0.447
InternVL3.5-30B	0.423	0.432
InternVL3.5-38B	0.476	0.487
GPT-4.1	0.436	0.438
GPT-5.4	0.429	0.448
Gemini-2.5-Pro	0.581	0.609
Gemini-3.1-Pro	0.571	0.605
Qwen3.6-Plus	0.563	0.568

F.4 CONSISTENCY-BASED DETECTION

SelfCheckGPT treats cross-sample consistency as evidence against hallucination. Here, however, rubric-recited false claims recur nearly as consistently as true claims. Table 14 pools results over domains for each model and compares false-claim detection AP with the corresponding false-claim prevalence.

The three-of-five rule retains claims with $\widehat{q}_{\mathrm{str}} \geq 3/5$; it does not define a separate ranking score. Occurrence-detection AP uses $\widehat{q}_{\mathrm{str}}$ (Table 2), whereas the false-claim AP below uses $1-\widehat{q}_{\mathrm{str}}$. Thresholds affect retention, not the underlying score used for AP.

Table 14: Cross-sample consistency of true and false claims, their gap, and false-claim detection AP against the false-claim prior.

Model	true	false	gap	AP	false-claim prior
Open-weight					
Qwen3.6-27B	0.69	0.51	+0.18	0.82	0.72
Qwen3.6-35B	0.54	0.41	+0.13	0.79	0.73
InternVL3.5-30B	0.53	0.42	+0.11	0.77	0.76
InternVL3.5-38B	0.65	0.47	+0.18	0.84	0.77
Closed-source					
GPT-4.1	0.79	0.64	+0.15	0.84	0.76
GPT-5.4	0.71	0.63	+0.08	0.63	0.59
Gemini-2.5-Pro	0.61	0.46	+0.15	0.65	0.58
Gemini-3.1-Pro	0.63	0.48	+0.15	0.64	0.56
Qwen3.6-Plus	0.70	0.53	+0.17	0.68	0.57

F.5 Mitigation Results

Values are false-claim rate / coverage (%), pooled from raw counts over Resistance, Voltage, and Kitchen for all methods. Woodpecker corrects captions, while *Ours-gate* and *Ours-weak* apply the fixed claim filters described above, so the table compares their resulting operating points.

Table 15: Correction and filtering trade-off, pooled over Resistance, Voltage, and Kitchen. *Ours-gate* uses the two-axis gate; *Ours-weak* retains strong claims also supported under weak input. VCD is unavailable on the InternVL serving path (—).

	Qwen3.6		InternVL3.5	
Method	27B	35B	30B	38B
Naive	71/90	72/74	68/63	73/82
Woodpecker	54/73	46/52	50/22	53/31
VCD	66/9	63/5	—	—
Ours-gate	65/52	65/52	66/46	70/43
Ours-weak	52/46	46/36	54/19	58/27

F.6 Reconstructing the Reported Results

The anonymous Code and Data Supplement includes code and accompanying materials documenting the experimental workflow, including prompt templates, model configurations, selected aggregate results, dataset summaries, and representative clips. The code defines the run-level and claim-level schemas used by the audit, including the four sampled frequencies, trust score, white-box contrasts, and baseline outputs.

The supplied code implements the post-generation reductions. First, the three judge roles are reduced by a two-of-three vote for every caption and rubric step. Second, the five binary decisions within an input arm are averaged to obtain $\widehat{q}_{\mathrm{zr}}$, $\widehat{q}_{\mathrm{wk}}$, $\widehat{q}_{\mathrm{str}}$, or $\widehat{q}_{\mathrm{lk}}$. For the rubric-dose tables, a step is treated as claimed when its arm frequency is greater than zero. False-claim rate and coverage are then computed from the resulting claim set; no caption or clip is weighted by its number of words. For detection, all clip–step rows are ranked by the specified frequency or by τ. Model–domain cells are computed before the displayed macro averages; Table 15 pools raw counts over the same three domains for every method.

These materials show how the reported results are constructed from the model outputs. The complete research repository will be released upon acceptance.

G Worked Audit Example

Figure 5 provides a retrospective illustration of Eq. (5). The strong and weak frequencies each use $N = 5$ generations. For this case, $\widehat{q}_{\mathrm{lk}}$ is the mean strong-input support over other clips from the same source recording where the step is annotated as absent. This labeled cross-clip reference differs from the unrelated-video intervention in the main protocol, and its average need not lie on the $1/5$ frequency grid. A separate failure case follows in Figure 6.

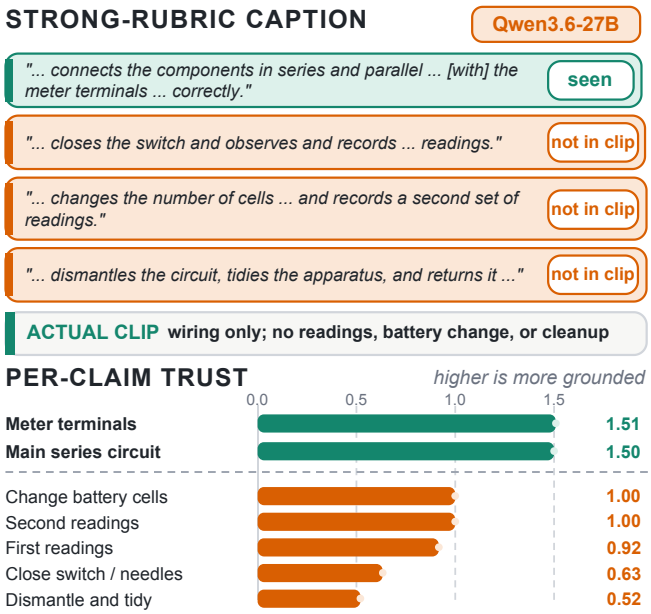

Figure 5: Retrospective claim-level scoring on a Qwen3.6-27B resistance clip using the cross-clip average reference described above. The score ranks two present claims (teal) above five absent claims (orange).

Worked Case: Procedure-shaped False Claim in Resistance

Anonymous clip: `RE-S001-C000`. **Evaluated step:** 5. **Ground truth:** absent.
Content-preserving summary of the selected strong caption: The caption describes circuit assembly and observations, followed by changing the battery, recording a second measurement, and cleaning up. The latter procedure-shaped actions extend beyond what occurs in the clip.
Matching-input frequencies for step 5: $\widehat{q}_{\mathrm{zr}} = 0$, $\widehat{q}_{\mathrm{wk}} = 0.4$, and $\widehat{q}_{\mathrm{str}} = 0.6$.
Parsed judge returns:

- Lenient `described`: [1, 2, 4, 5, 7, 8, 9, 10]
- Strict `items` step IDs: [4, 8, 9, 10]
- Reverse, after complementing `no_evidence`: [4, 5, 7, 8, 9, 10]
- Two-of-three majority: step 5 is extracted as claimed, although its occurrence label is absent.

White-box check: 122 caption tokens, $\mathrm{CPMI}_r = 0.5826$, $\mathrm{CPMI}_v = 0.1125$. Rubric reliance exceeds visual grounding, exhibiting a recitation-dominant signature at the caption level.

Figure 6: Resistance failure case. The displayed caption text is a content-preserving summary, while the judge ID sets are parsed returns.

H PROOFS

Proof of the Single-Replacement Expectation Bound. Let $\tilde{v}$ be drawn from the replacement-video distribution used to define $P(y \mid r) = \mathbb{E}_{\tilde{v}}[P(y \mid \tilde{v}, r)]$. Since $\log$ is concave, Jensen's inequality gives

$$\mathbb{E}_{\tilde{v}}[\log P(y \mid \tilde{v}, r)] \leq \log \mathbb{E}_{\tilde{v}}[P(y \mid \tilde{v}, r)] = \log P(y \mid r). \tag{8}$$

Subtracting either side from the fixed term $\log P(y \mid v, r)$ reverses the inequality:

$$\begin{aligned} &\mathbb{E}_{\tilde{v}}[\log P(y \mid v, r) - \log P(y \mid \tilde{v}, r)] \\ &\quad \geq \log P(y \mid v, r) - \log P(y \mid r). \end{aligned} \tag{9}$$

The left-hand side is the expected single-replacement visual-grounding gap and the right-hand side is its distribution-averaged counterpart. This proves only the stated expectation inequality; it does not imply a one-sided guarantee for an individual replacement or for the thresholded detection set. □