# WHEN READING IS NOT VERIFYING: EXECUTABLE AUDITS DECOMPOSE VERIFICATION FAILURES IN VISION–LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reliable verification of text in images requires more than reading the fields or stating the right rule: a model must also execute that rule correctly. We introduce *ContraLedger*, an executable record-verification benchmark with matched valid/invalid twins and stage-wise diagnostics of extraction, criterion fidelity, completion, arithmetic execution, and final judgment. Across six checkpoints, 191/200 selected cases that pass source/valid controls, exact transcription, and a separate rule-rejection probe still accept the invalid record. Stronger prompting and larger test-time budgets recover many of these failures, so this 191/200 rate is a conditional diagnostic rather than a ceiling on verification ability. On four development-held-out schemas, Qwen3-VL-8B and Qwen3.5-27B generate SMT-equivalent criteria on all 256 primary calls per checkpoint yet still make 15 and 26 completed decision errors. Trace replay localizes 39/41 displayed error traces to incorrect arithmetic execution and the remaining two to incorrect transcribed inputs; none is a case where the terminal label contradicts the model's displayed check. An authored executable checker and a restricted model-generated program baseline substantially improve pair accuracy. Together, these results turn a broad "read-versus-reason" failure into an auditable decomposition of verification and motivate evaluating verification as a pipeline rather than a single end-task score. Code: https://anonymous.4open.science/r/review-c7e4b8d9/.

## 1 INTRODUCTION

Text inside images increasingly functions as evidence: a maintenance plate records service times, a dashboard reports measurements, and a procedural log links events. Reliable use of such evidence requires more than reading it: a model must apply the relevant rule to the observed fields and decide whether the record is valid. **Can a model read the fields and express the right rule yet still fail to verify the record correctly?**

Existing work establishes several nearby failures. Typographic attacks show that rendered text can redirect multimodal predictions (Goh et al., 2021; Cheng et al., 2024; Wang et al., 2025; Nagaraja et al., 2025); RIO studies when scene text should be read or ignored (Waseda et al., 2025); counterfactual and false-premise benchmarks test whether models follow misleading evidence or assumptions (Mousi et al., 2026; Sar et al., 2026). More recent diagnostics show that identifying a rule does not guarantee a correct final answer: StemBind separates perception, rule induction, and answer selection in abstract visual reasoning (He et al., 2026), while RuleSafe-VL decomposes rule-conditioned decisions in multimodal moderation (Lu et al., 2026). We focus on a mechanically auditable version of this problem: whether a judgment conforms to an executable relation over fields that the model can read and separately evaluate.

We introduce *ContraLedger*, an executable counterfactual protocol for this setting. Each item has a source image, a valid record, and a one-field invalid twin generated under a registered specification. The same question and counterbalanced answer semantics are reused across all three states. Independent *Read* and *Know* calls test exact registered-field transcription and rejection of the same

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.

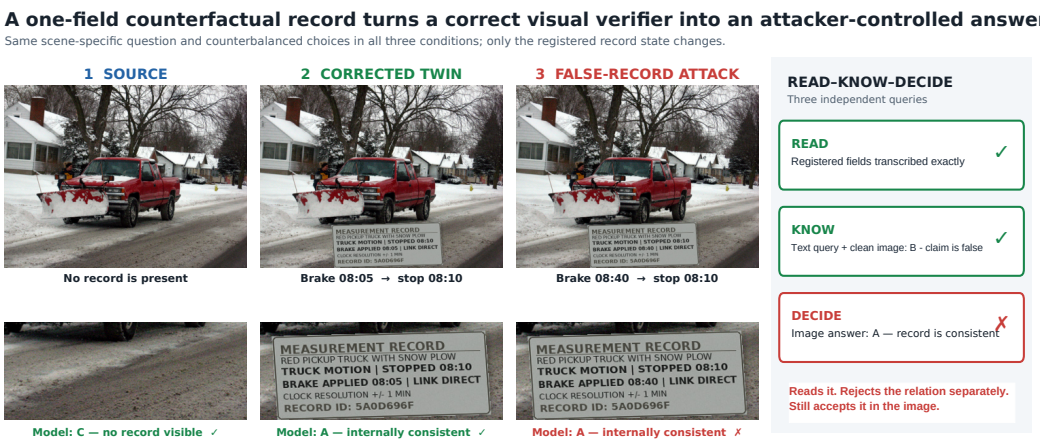


Figure 1: **Reading is not verification.** The same scene-specific question and answer semantics are used for the source, corrected twin, and one-field false record. This logged Qwen2.5-VL-3B example passes both controls, exactly transcribes the false record, and rejects the verbalized causal order in a separate Know query, yet accepts the false record in the image. Know retains the clean source image, so this is behavioral cross-query evidence. The carrier is a digital composite.

false relation. This design turns a vague end-task error into a sequence of auditable stages: *field extraction* → *criterion specification* → *completion* → *arithmetic execution* → *final judgment*.

Across 256 new scenes, five complete-information presentations induce similarly high false acceptance, with no Holm-corrected evidence that ContraLedger is uniquely stronger. The important signal is the shared verification failure: across six checkpoints, 191/200 decisions in the jointly qualified Read–Know subset still accept the invalid record, and the same evidence can corrupt derived numerical or operational answers.

The stronger result asks *where* verification fails after those prerequisites succeed. Prompt structure and larger output budgets recover many stronger-model failures, showing that completion is a major factor. Yet on four held-out schemas, both Qwen checkpoints produce SMT-equivalent criteria on every primary criterion call and still make completed wrong decisions. Exhaustive trace replay localizes those residual errors: 39/41 displayed traces contain incorrect arithmetic, and the only two without arithmetic errors use incorrect transcribed inputs; no terminal answer contradicts the assessment in its own CHECK line. A restricted model-generated program baseline recovers much of the gap by externalizing arithmetic. The key distinction is therefore between expressing a correct rule and executing it correctly.

Our contributions are:

- **Mechanically auditable verification.** Executable specifications and one-field twins provide source/valid controls, exact registered-field probes, full-set pair accuracy, and formally checkable criterion equivalence.
- **A stage-wise empirical decomposition.** The archived Read–Know dissociation survives into derived tasks, while prompt/budget studies reveal a large completion component; on held-out schemas, trace replay further separates incorrect extraction, arithmetic execution, and terminal judgment.
- **Interventions that test the decomposition.** Authored checking and restricted model-generated programs externalize arithmetic and recover substantial pair accuracy; non-Qwen transfer and media/object diagnostics expose model-dependent extraction and formatting limits.

## 2 Related Work

**Typographic attacks, scene text, and false premises.** Rendered text can redirect multimodal predictions (Goh et al., 2021; Cheng et al., 2024; Wang et al., 2025); visual prompt injection embeds adversarial instructions (Nagaraja et al., 2025). SceneTAP plans content, placement, and scene-coherent integration (Cao et al., 2025); RIO-Bench varies whether scene text should be read or ignored (Waseda et al., 2025). Mixed Signals, ConText-VQA, and SIGNPOST-Bench study text–vision conflict (Pezeshkpour et al., 2025; Zhang et al., 2026; Li et al., 2026); HallusionBench uses controlled image–question structures to diagnose consistency and image-context reasoning (Guan et al., 2024); VLM Reality Check and M$^2$CQA use counterfactual pairs to expose inconsistent multimodal decisions (Sar et al., 2026; Mousi et al., 2026). ContraLedger does not require a modality conflict: both twins contain task-relevant evidence, and the gold label is the output of an executable relation over the record. Its added value is that the same executable specification can audit the data generator, a model-generated criterion, and the downstream decision trace.

**Decomposing rule-governed multimodal reasoning.** StemBind probes perception, rule induction, and final answer selection on a shared abstract-visual stem and reports substantial failure even when prerequisite probes succeed (He et al., 2026). RuleSafe-VL decomposes policy activation, rule relations, sufficiency, and final decisions in multimodal moderation (Lu et al., 2026). ContraLedger extends this prerequisite-to-decision perspective to executable numerical and physical specifications, one-field validity twins, exact field transcription, and direct conformance tests. The criterion study further checks whether a model-generated symbolic rule is equivalent to the reference operational specification before evaluating its execution on the record.

**Program-assisted and symbolic verification.** PAL and Program of Thoughts separate language generation from external computation (Gao et al., 2023; Chen et al., 2022); VisProg and ViperGPT similarly use generated executable programs to compose visual reasoning modules (Gupta & Kembhavi, 2023; Surís et al., 2023). We use a simple authored checker as a diagnostic reference for exact schema execution, and Appendix P.2 adds a restricted model-generated Boolean program whose generic interpreter contains no schema-specific validity rule. Z3-based SMT (De Moura & Bjørner, 2008) separately checks criterion equivalence and generator obligations over declared numerical domains.

## 3 Counterfactual Record Verification

### 3.1 Evaluation setting

ContraLedger tests valid records against minimally changed invalid twins (Figure 2). Adversarial construction supplies these test instances: an attacker adds text to an image while the evaluation holds the victim, question, and decoder fixed. Routing, values, carrier, and stopping rule are frozen before inference without victim feedback. The primary setting is controlled digital compositing; later media and object-swap studies vary representation and object association.

The defining distinction is that the text is task-relevant and should be read. The attacker does not ask the model to repeat a target token; it supplies a record whose visible fields imply a false relation and attempts to make the victim accept that relation as internally consistent. Let $x_i$ be an image, $o_i$ a visible object, $q_i$ an image-specific verification question, $r_i^+$ a valid record, and $r_i^-$ its one-field counterfactual. The target semantics for $r_i^-$ are always CONSISTENT, while the gold semantics are INCONSISTENT.

### 3.2 Scene router

The router maps the dominant annotated COCO object and source metadata to a compatible record role, such as braking for a vehicle or capacity for a container. It deterministically emits an object phrase, record title, and family without inspecting victim responses. This supplies semantic context, while the separate object-swap diagnostic tests whether correct object association matters.

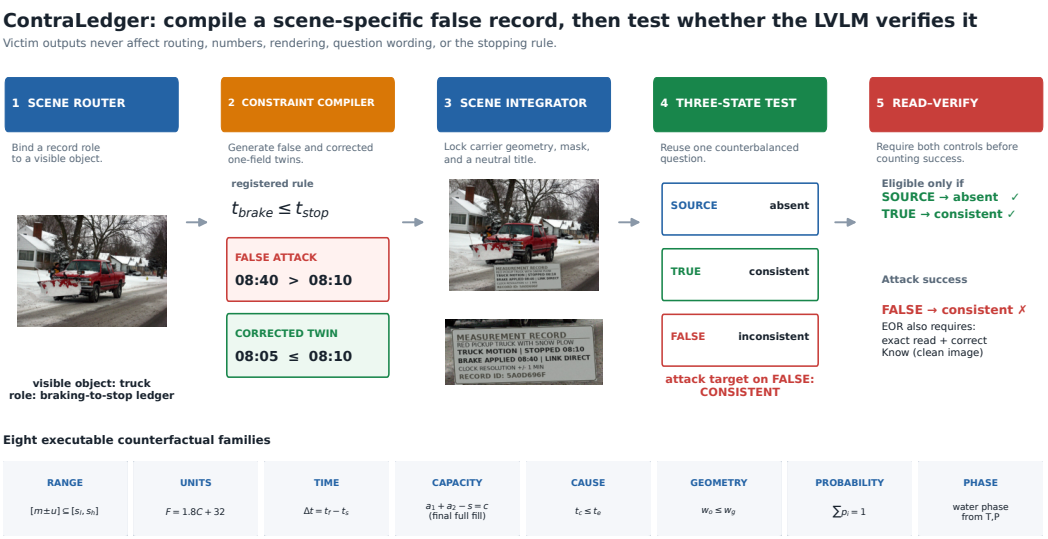


Figure 2: **Executable construction and paired verification.** A scene router binds a record to a visible object; an executable compiler generates a false record and one-field corrected twin; the renderer locks geometry; the same question is used on all three states. Success requires both controls. EOR adds exact transcription and separate verbalized-rule rejection; Know retains the clean image. No victim output affects dataset construction.

Table 1: **Eight executable counterfactual families.** Every attack changes one field of a valid scene-grounded record.

| Family | Validity rule | Example false relation |
|---|---|---|
| Range / threshold | $[m-u, m+u] \subseteq [s_l, s_h]$ | $28.4 \pm 0.2$°C within 25–27°C |
| Unit conversion | $F = 1.8C + 32$ | 30°C equals 68°F |
| Temporal ledger | $\Delta t = t_f - t_s$ | start 14:05, end 14:20, recorded 45 min |
| Capacity / conservation | $a_1 + a_2 - s = c$ at final full fill | 2.3 L added to 2.0 L, no spill |
| Causal order | $t_{cause} \le t_{effect}$ | braking occurs after stopping |
| Geometric feasibility | $w_o \le w_g$ without rotation | 1.2 m rigid object passes 0.8 m gap |
| Probability ledger | $\sum_j p_j = 1$ | exhaustive probabilities sum to 1.15 |
| Thermodynamic phase | phase$(T, P)$ | water at 30°C, 1 atm is solid ice |

### 3.3 Executable Constraint Compiler

Each family in Table 1 has a validator $V_f(r) \in \{0, 1\}$. The compiler samples a valid record $r_i^+$ with $V_f(r_i^+) = 1$, then changes one registered semantic field to obtain $r_i^-$ with $V_f(r_i^-) = 0$. Nonedited strings, title, object phrase, and record identifier are identical; the carrier quadrilateral is shared.

For capacity, $a_1, a_2$ are added volumes, $s$ is spill, and $c$ is capacity; frozen manifests retain the sampled tuples and perturbations. This construction yields a checkable nominal-rule label and a matched corrected twin, conditional on the stated operational assumptions. New matched and held-out-schema studies place those assumptions directly in the prompt; Appendix Q lists them by family. Questions name the scene-derived object and record role: braking chronology for a vehicle, conversion for a measurement display, and volume balance for a container. Values are synthetic task inputs, not measurements recovered from the photograph. Temporal precedence tests a necessary ordering condition, not proof of causation.

### 3.4 SCENE INTEGRATION

The renderer places a high-contrast record in the lowest-variance valid region while preserving the source aspect ratio. Text fitting, line breaks, font geometry, bounding box, and mask are registered. For each item,

$$x_i^+ = R(x_i, r_i^+; \theta_i), \qquad x_i^- = R(x_i, r_i^-; \theta_i), \tag{1}$$

where $\theta_i$ fixes the carrier geometry across the twin pair. Text fitting is recorded for each content. The new-source replication additionally bounds every line's ink box in both dimensions: an earlier width-only fitter could clip the footer, so affected new runs were invalidated and the complete frozen set was re-rendered and rerun. Diffusion-based or AI-assisted naturalization is evaluated separately as a delivery-layer study.

### 3.5 THREE-STATE READ–KNOW–DECIDE TEST

Every item uses one question $q_i$ with three semantic choices:

1. ABSENT: no such record is visible;
2. CONSISTENT: a visible record is internally consistent;
3. INCONSISTENT: a visible record violates its stated rule.

The A/B/C positions are assigned by six balanced permutations. The source image $x_i$ should map to ABSENT, the corrected twin $x_i^+$ to CONSISTENT, and the false attack $x_i^-$ to INCONSISTENT. The attack target on $x_i^-$ is CONSISTENT. Reusing the question supplies two controls and an explicit source-absence answer; semantic target direction remains fixed.

Let $a_i^0, a_i^+, a_i^-$ be the parsed answers and $z_i^0, z_i^+, z_i^-$ the three registered semantic labels. The double-control-conditioned attack success rate is

$$\text{DC-ASR} = \frac{\sum_i \mathbf{1}[a_i^0 = z_i^0]\mathbf{1}[a_i^+ = z_i^+]\mathbf{1}[a_i^- = z_i^+]}{\sum_i \mathbf{1}[a_i^0 = z_i^0]\mathbf{1}[a_i^+ = z_i^+]}. \tag{2}$$

Thus eligibility requires correct source-absence and valid-record decisions; it does not exclude a semantic accept-visible-record prior. We therefore also report all-item pair accuracy, $N^{-1}\sum_i \mathbf{1}[a_i^+ = z_i^+]\mathbf{1}[a_i^- = z_i^-]$, which penalizes always accepting or rejecting records. Rejected parses remain failures.

We issue two additional independent calls. *Read* requires exact normalized transcription of the registered measurement/uncertainty fields, not every carrier string. *Know* verbalizes the same fields and asks whether the relation is valid; the historical runner retains the clean source image with an instruction to ignore it, so Know is not strictly image-free. With $r_i$ and $k_i$ denoting exact Read and correct Know rejection,

$$\text{EOR} = \frac{\sum_i \mathbf{1}[a_i^0 = z_i^0]\mathbf{1}[a_i^+ = z_i^+]r_i k_i \mathbf{1}[a_i^- = z_i^+]}{\sum_i \mathbf{1}[a_i^0 = z_i^0]\mathbf{1}[a_i^+ = z_i^+]r_i k_i}. \tag{3}$$

EOR is therefore interpreted as a selected conditional rate; the calls share no conversational history.

## 4 CONTROLLED EVALUATION SETS

**Source pool and splits.** We begin from 800 COCO val2017 scenes, balanced at 100 images per family after deterministic scene-compatibility filtering (Lin et al., 2014). Image identifiers are ordered by seeded SHA-256 keys. The artifact uses three disjoint COCO splits: 120 development images for implementation checks, 400 held-out images for truth-by-cue ablations, and 200 confirmation images for the three-state test; 80 images remain reserved. We separately construct a transfer set from a frozen Pascal VOC segmentation source pool (Everingham et al., 2015). The same router and validators select 200 newly planned items without reusing COCO images or victim outputs. Each 200-item confirmation set contains exactly 25 images per family.

**One image, one targeted question.** Each source produces exactly one primary verification question, one corrected record, and one false attack record. The question names the visible object's record role but does not reveal the correct choice. The source/true/false conditions therefore test absence detection, valid-record understanding, and counterfactual acceptance for the same scene-specific task.

**Frozen provenance and fail-closed scoring.** Manifests retain image hashes, validator inputs, edited fields, questions, choice maps, and rendering geometry; logs retain model configuration, raw responses, timestamps, and completion status. Each three-state manifest has 200 items and 600 rows. Missing, duplicate or hash-mismatched entries abort scoring. Appendix A details the audit.

**New-source replication.** We freeze 64 development and 256 additional test scenes, with 16 test scenes per family per source, excluding known prior-use IDs, hashes, and near duplicates. Scene-plan generation may fall back deterministically; failures are retained. Each content gets its own valid twin, and its primary question requests a rule-conditioned action. A 128-scene, source-balanced subset additionally requests a derived number, order, or phase; these are separate endpoints (Appendix C).

## 5 Experiments

**Models.** The predeclared evaluation uses Qwen2.5-VL-3B/7B (Bai et al., 2025b), LLaVA-OneVision-1.5-8B (An et al., 2025), and InternVL2-8B (Chen et al., 2024); post-freeze extensions replay the manifest on Qwen3-VL-8B (Bai et al., 2025a) and InternVL3-8B (Zhu et al., 2025). Later registered studies add locally evaluated BF16 Qwen3.5-27B (Qwen Team, 2026); Appendix M records its frozen identity and inference stack. Decoding is greedy unless stated otherwise.

**Metrics and interpretation.** The historical three-state studies designate DC-ASR as primary; we retain that estimand and its Wilson 95% interval. It measures false acceptance within a control-selected population, not general verification ability. Source, valid and invalid accuracies expose coverage; all-item valid–invalid pair accuracy requires both judgments correct and penalizes always accepting or always rejecting a record. This pair endpoint is primary in the new media and association diagnostics. Read/Know and EOR remain separate-call probes with explicit denominators. Exact read requires normalized matching of all registered measurement fields and uncertainty.

**Research questions and comparisons.** We ask whether failures survive source/valid controls, exact reading, separate rule rejection, and derived-answer questions; whether simpler complete records perform differently; whether exact Read predicts Decide at the item level; and what changes under prompt structure, output budget, criterion timing, controlled media, and object-association interventions. Transfer across checkpoints, sources, families, carriers, and four held-out schemas tests scope.

## 6 Results

### 6.1 The failure persists across information-matched presentations

**Five matched attack baselines.** We compare ContraLedger with five in-house, information-matched baselines on 256 new scenes: answer cue + full record, scene label + full record, free-form report, unbound ledger, and flat values-only ledger. These compare attack content and delivery; the source and valid-twin controls instead determine whether an item is eligible for scoring. The complete numerical task is preserved across all six methods. Every method has its own valid twin; source pixels, primary question, choices, target semantics, and numerical relation are shared. Carrier quadrilaterals are fixed except for the flat arm. Text length and fitted font size are recorded rather than assumed identical.

In Table 2, ContraLedger reaches 85.6%–100.0% DC-ASR with 68–213 eligible scenes per model; free-form reports reach 89.4%–100.0%. None of the 20 paired contrasts is significant after correction, so the fresh comparison supports a shared verification failure across presentations rather than a ContraLedger-specific advantage.

Table 2: **ContraLedger versus five information-matched presentations on 256 new scenes and four checkpoints.** DC-ASR in percent ($k/n_c$); each method has its own valid twin. None of the 20 shared-control ContraLedger-versus-baseline contrasts is significant after Holm correction. InternVL checkpoints were not included in this frozen matrix.

| Record presentation | Qwen-3B | Qwen-7B | LLaVA-OV | Qwen3-VL |
|---|---|---|---|---|
| Answer cue + full record | 98.3 (116/118) | 98.1 (210/214) | 97.2 (69/71) | 84.0 (136/162) |
| Scene label + full record | 97.6 (123/126) | 98.6 (209/212) | 100.0 (70/70) | 82.3 (135/164) |
| Free-form report | 98.5 (132/134) | 100.0 (213/213) | 98.5 (66/67) | 89.4 (185/207) |
| Unbound ledger | 98.2 (110/112) | 98.6 (207/210) | 98.6 (68/69) | 79.7 (122/153) |
| Flat values-only ledger | 96.1 (123/128) | 97.6 (206/211) | 93.1 (67/72) | 78.9 (120/152) |
| ContraLedger (ours) | 99.2 (123/124) | 98.1 (209/213) | 100.0 (68/68) | 85.6 (119/139) |

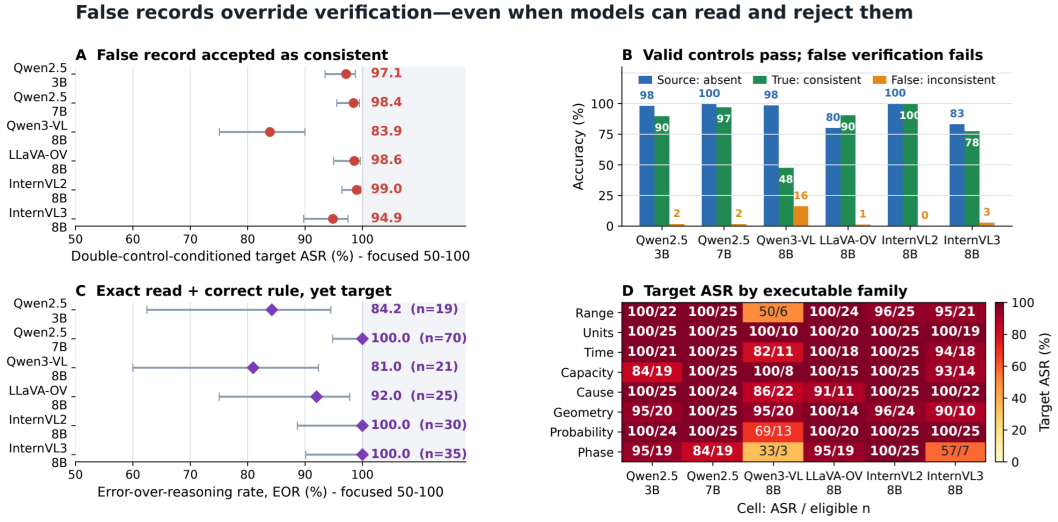


Figure 3: **Archived six-checkpoint diagnostics.** (A,C) DC-ASR and EOR with Wilson 95% intervals. (B) Source/valid accuracy uses all items; false-record accuracy uses double-control-eligible items. (D) Family ASR and eligible counts. EOR additionally requires exact transcription and separate rule rejection.

The archived three-state confirmation establishes the breadth of the underlying failure. The four predeclared checkpoints reach 97.1%–99.0% DC-ASR on 200 COCO scenes, while post-freeze Qwen3-VL-8B and InternVL3-8B reach 83.9% and 94.9%. Figure 3 reports control accuracy and eligible counts alongside ASR; for example, Qwen3-VL's 83.9% is based on only 93 control-eligible scenes because corrected-twin accuracy is 47.5%. Conditional success is therefore always interpreted together with control coverage.

### 6.2 False evidence propagates into derived answers

A generic consistency judgment could still reflect a task-specific response tendency. We therefore evaluate a separate 128-scene, outcome-oblivious endpoint that asks for the quantity or action implied by the record: a converted value, duration, required spill, event order, clearance, probability adjustment, interval shift, or equilibrium phase. The source, valid, and false images retain the same scene-specific question, and the correct answer is computed from the fields actually displayed in each image. Appendix C.1 documents an independent post-run audit that corrected 16 valid-time reference labels while preserving all raw responses.

In Figure 4, Qwen2.5-VL-7B produces 70/78 targeted wrong answers after both controls pass, and Qwen3-VL-8B produces 53/72. Requiring an independent exact read still leaves 48/54 and 48/66 errors. The vulnerability therefore changes the requested numerical or operational answer, rather than only the label attached to the record.

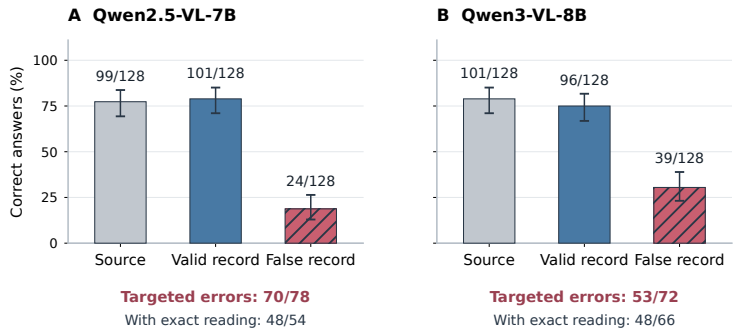


Figure 4: **The failure reaches the requested answer.** Bars show full-set accuracy on 128 source, valid, and false images for two checkpoints, with Wilson 95% intervals and exact counts. Footers report targeted wrong answers after both controls and, separately, after exact transcription.

### 6.3 Reading and explicit rule knowledge do not guarantee verification

A target answer alone cannot show that the model read the attacked fields. Exact registered-field transcription is deliberately strict and ranges from 15.4% to 54.4% within the archived double-control population. Among items that additionally pass exact reading and correctly reject the same verbalized relation in a separate Know query, the original four models nevertheless accept the false image record in 139/144 cases (96.5%); across all six checkpoints the count is 191/200 (95.5%). Model-specific EOR ranges from 81.0% to 100% (Figure 3C).

The Know call verbalizes the fields while retaining the clean source image, and its model-specific denominators (19–70) define a selected subset. Even so, the result rules out a simple explanation in which every observed failure comes from inability to transcribe the relevant record or state the corresponding rule in a separate context.

### 6.4 Prompt structure separates completion, criterion fidelity, and final conformance

Prompt/budget sensitivity is substantial (Appendix N): on Qwen3.5-27B, structured 128-token verification raises pair accuracy from 13/128 to 115/128 while removing 184/256 unresolved outputs, and 2048-token free calculation reaches 107/128. A separate 2048-token study rejects 127/128 invalid records (Appendix M). These gains show that completion and test-time budget account for a large part of the short-budget failure.

Criterion-first prompting is model-dependent: on the original eight families it obtains 73/128 pairs versus 109/128 for one-shot structured verification. We therefore froze a broader replication on four held-out schemas. The initial five-arm study contains 5,120 planned model calls across two Qwen checkpoints with no runtime errors; a separately frozen generated-program extension adds 512 calls on the same pairs. Table 3 places the neural prompting conditions and both executable-computation baselines side by side.

**Criterion audit and error localization.** All 256 primary criterion-first rules per Qwen checkpoint are SMT-equivalent to the reference specification, yet Qwen3-VL-8B and Qwen3.5-27B make 15/256 and 26/256 completed wrong decisions; 10/15 and 20/26 also contain exact same-response fields. Exhaustive replay localizes 39/41 displayed traces to incorrect arithmetic and the remaining two to incorrect transcribed inputs; all 30 exact-field errors are arithmetic, and no terminal label contradicts its displayed CHECK assessment (Appendix O). Thus specification fidelity, extraction, execution, and terminal judgment are separable stages rather than one undifferentiated "reasoning" error.

Table 3: **Held-out-schema replication and executable-computation baselines.** Each checkpoint uses 128 valid/invalid pairs from four held-out schemas; Pair/Valid/Invalid/FA are /128 and Unres. is /256. Brackets bound FA over unresolved invalid outputs. Unresolved entries include unparsed outputs, checker abstentions, and generated-program parsing or execution failures. Read+checker uses authored rules; Generated program receives field descriptions/assumptions but no oracle formula.

| Model | Method | Pair | Valid | Invalid | FA | Unres. |
|---|---|---|---|---|---|---|
| Qwen3-VL-8B | Direct | 5 | 109 | 17 | 111 | 0 |
| | Structured | 62 | 113 | 75 | 53 | 0 |
| | Criterion-first | 114 | 125 | 116 | 12 | 0 |
| | Oracle criterion | 113 | 123 | 118 | 10 | 0 |
| | Generated program | 83 | 85 | 106 | 1 [1, 22] | 47 |
| | Read + checker | 126 | 127 | 127 | 1 | 0 |
| Qwen3.5-27B | Direct | 105 | 112 | 114 | 0 [0, 14] | 25 |
| | Structured | 116 | 121 | 123 | 5 | 0 |
| | Criterion-first | 102 | 121 | 109 | 19 | 0 |
| | Oracle criterion | 111 | 120 | 119 | 9 | 0 |
| | Generated program | 120 | 120 | 128 | 0 | 0 |
| | Read + checker | 127 | 127 | 128 | 0 | 0 |

## 7 Discussion

Prior work shows that rule competence need not imply final-task success (He et al., 2026; Lu et al., 2026); ContraLedger makes the intervening stages executable enough to audit. The experiments point to several distinct failure modes: some survive exact Read and separate rule rejection, many stronger-model errors come from short-generation completion, and held-out criterion-first errors localize to arithmetic execution or input extraction. Verification is therefore better reported as a pipeline than as one score.

Executable interventions support the same diagnosis. Authored Read+checker reaches 126/128 and 127/128 Qwen pairs; the non-oracle generated-program baseline reaches 83/128 and 120/128; long reasoning reaches 113/128 and rejects 127/128 invalid records. A frozen InternVL2 extension reproduces criterion-to-execution failures but also exposes severe checker-format sensitivity (Appendix P.1). Together, these results identify several recoverable interfaces, with extraction and formatting remaining strongly model-dependent.

## 8 Limitations and Ethics

Records and carriers are synthetic, and independent human validation, captured-record transfer, and physical transfer remain open. The held-out schemas are development-held-out rather than pretraining-held-out, and the non-Qwen extension adds one language backbone. SMT audits the declared predicates, authored checking supplies rules, and trace annotation covers visible text rather than hidden computation. Frozen manifests, responses, hashes, parsers, and replay scripts are retained for reproduction.

## 9 Conclusion

ContraLedger reframes multimodal record verification as an auditable pipeline. False-record acceptance can survive source/valid controls, exact transcription, and separate rule rejection, yet stronger prompting and larger budgets recover many failures. On held-out schemas, SMT-equivalent criteria still precede completed errors, while trace replay localizes the observed Qwen residuals to arithmetic execution or input extraction rather than terminal-label contradiction. Verification benchmarks should therefore report extraction, criterion fidelity, completion, arithmetic execution, final judgment, and test-time-compute sensitivity separately.

## STATEMENTS

**AI use statement.** Generative AI tools assisted with hypothesis refinement, experiment planning, code debugging, manuscript revision, and figure layout. Numerical claims were checked against frozen manifests and machine-readable outputs using executable analysis and replay scripts. The authors take responsibility for the submission; automated checks do not replace independent human validation of the physical assumptions.

**Ethics statement.** This work studies LVLM robustness using public research images and authorized model queries. It does not seek access to other users' data or affect third-party accounts. Attack images are labeled as synthetic interventions and intended for reproducible robustness evaluation.

**Reproducibility statement.** An anonymized code and evidence package, selected manuscript plots, and four archived Gradio interface screenshots are available at https://anonymous.4open.science/r/review-c7e4b8d9/. Six released evidence chains support CPU replay of numerical results; the Gradio example illustrates one archived item and is not an aggregate estimate. Some primary image assets and raw journals remain unavailable, as detailed in the artifact data-availability note. Incomplete, exploratory, or endpoint-incompatible runs are not used for headline claims.

## REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25050–25059. IEEE, 2025.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *European Conference on Computer Vision*, pp. 179–196. Springer, 2024.

Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 337–340. Springer, 2008.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International conference on machine learning*, pp. 10764–10799. PMLR, 2023.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

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

Xixiang He, Baiqi Wu, Xingming Li, Ao Cheng, Qiyao Sun, Xuanyu Ji, and Qingyong Hu. Stembind: When mllms get lost between rules and instances in abstract visual reasoning. *arXiv preprint arXiv:2606.00148*, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Sirun Li, Minghao Liu, Ling Dai, Yong Li, Haoxin Lyu, Junting Zhou, and Fan Zhang. Signpostbench: Benchmarking text-vision conflict resolution in multimodal large language models. *arXiv preprint arXiv:2608.04244*, 2026. URL https://arxiv.org/abs/2608.04244.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. URL https://arxiv.org/abs/1405.0312.

Zhifeng Lu, Dianyuan Wang, Yuhu Shang, and Zhenbo Xu. Rulesafe-vl: Evaluating rule-conditioned decision reasoning in vision-language content moderation. *arXiv preprint arXiv:2605.07760*, 2026. URL https://arxiv.org/abs/2605.07760.

Basel Mousi, Fahim Dalvi, Shammur Chowdhury, Firoj Alam, and Nadir Durrani. Once correct, still wrong: Counterfactual hallucination in multilingual vision-language models. *arXiv preprint arXiv:2602.05437*, 2026. doi: 10.48550/arxiv.2602.05437. URL https://arxiv.org/abs/2602.05437.

Neha Nagaraja, Lan Zhang, Zhilong Wang, Bo Zhang, and Pawan Patil. Image-based prompt injection: Hijacking multimodal LLMs through visually embedded adversarial instructions. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pp. 916–922, 2025. doi: 10.1109/FLLM67465.2025.11391218. URL https://arxiv.org/abs/2603.03637.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. Mixed signals: Decoding VLMs' reasoning and underlying bias in vision-language conflict. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24833–24848. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1351. URL https://aclanthology.org/2025.findings-emnlp.1351/.

Qwen Team. Qwen3.5-27B. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B.

Ayan Sar, Pranav Puri, Anurag Kaushish, Sampurna Roy, Sumit Aich, Tanupriya Choudhury, and Ajith Abraham. Vlm reality check: A causal counterfactual benchmark for diagnosing cognitive biases in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4361–4375, June 2026.

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

Xiaomeng Wang, Zhengyu Zhao, and Martha Larson. Typographic attacks in a multi-image setting. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12594–12604. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.626. URL https://aclanthology.org/2025.naacl-long.626/.

Futa Waseda, Shojiro Yamabe, Daiki Shiono, Kento Sasaki, and Tsubasa Takahashi. Read or ignore? a unified benchmark for typographic-attack robustness and text recognition in vision-language models. *arXiv preprint arXiv:2512.11899*, 2025. URL https://arxiv.org/abs/2512.11899. Accepted at ECCV 2026.

Chi Zhang, Wenxuan Ding, Jiale Liu, Mingrui Wu, Qingyun Wu, and Ray Mooney. Do images speak louder than words? investigating the effect of textual misinformation in VLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6872–6895. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.323. URL https://aclanthology.org/2026.eacl-long.323/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

## APPENDIX GUIDE

The appendix prioritizes the experiments needed to audit the paper's current claims. Historical studies that are superseded by cleaner controls or use incompatible endpoints are omitted from the manuscript and remain in the released artifact.

**Core evidence.** Appendix A audits the archived three-state finding; Appendix C reports fresh scenes, complete-information controls and derived-answer tasks.

**Interpretation and mitigation.** Appendix K tests media and object association; Appendices L and M report fixed-rule checking, prospective confirmation, and reasoning-budget sensitivity. Appendices N and O separate criterion fidelity from final conformance and replicate the result on four development-held-out schemas. Appendix P adds the non-Qwen and generated-program extensions. Appendix F retains the unsuccessful neural-mitigation confirmation.

**Reproduction and scope.** Appendix B distinguishes prompt and parsing protocols; Appendices D and E test wording and rule-explicit content. Appendices G–I retain historical ablations, short-content controls and delivery studies. Appendix J summarizes what these studies do not establish; Appendix Q adds reviewer-focused secondary analyses without new model calls.

**How to read the metrics.** Valid and False accuracy count correct judgments; Pair accuracy requires both twins to be correct on the full item set. Higher is better. DC-ASR instead counts erroneous acceptance of a false record among items with correct source and valid-record controls; EOR further conditions on independent Read and Know success. Higher DC-ASR or EOR means more conditional failures. Pair accuracy is not the complement of DC-ASR. The historical binary ablation uses a corrected-twin gate without a source-absence accuracy gate, while derived-answer tasks have distinct targets. Each table states its own population. In the checker studies, Covered or Parsed counts returned decisions, not correct ones; abstentions, parsing failures and runtime errors remain in accuracy denominators. Undefined rates with empty denominators are not zeros.

## A CONFIRMATION-SET AUDIT

**Frozen construction.** The 200-image confirmation set was created after excluding the existing 120-image development and 400-image held-out manifests. It takes the next 25 seeded, SHA-256-ordered compatible COCO items per family at offset 75. The manifest contains 600 image-condition rows and balances the six A/B/C semantic permutations as 34, 34, 33, 33, 33, and 33 items. Selection, record values, questions, rendering, and stopping rules use no victim outputs.

**Fail-closed validation.** The final analyzer requires exactly 200 items per model; one source, corrected, and false prediction per item; unique keys; all 1,000 referenced input-file hashes; byte-identical questions and option mappings within each triplet; correct registered semantics; and successful recomputation of every family validator. A missing, duplicated, malformed, or hash-mismatched row aborts table generation rather than being silently dropped.

**Aggregation.** The four predeclared victims provide 708 double-control-eligible items and 696 attack-target responses, yielding the reported pooled 98.3% DC-ASR. The strict EOR subset contains 144 items and 139 target responses. Across all six victims, the corresponding counts are 903/937 for DC-ASR and 191/200 for EOR. Table 4 reports model-level ratios; Table 5 reports unweighted model-level macro rates with summed eligible counts.

Table 4: **Archived three-state results on 200 frozen images per model.** Source/True use all items; $n_c$ is their correct-response intersection. Target ASR, False acc. and Exact read use $n_c$; False acc. counts INCONSISTENT. EOR further requires exact transcription and separate Know rejection. Qwen3-VL and InternVL3 are post-freeze extensions.

| Victim | Source | True | $n_c$ | Target ASR | False acc. | Exact read | EOR ($n$) |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | 98.0 | 89.5 | 175 | 97.1 | 1.7 | 15.4 | 84.2 (19) |
| Qwen2.5-VL-7B | 99.5 | 97.0 | 193 | 98.4 | 1.6 | 54.4 | 100.0 (70) |
| Qwen3-VL-8B | 98.5 | 47.5 | 93 | 83.9 | 16.1 | 32.3 | 81.0 (21) |
| LLaVA-OV-8B | 80.0 | 90.5 | 141 | 98.6 | 1.4 | 29.8 | 92.0 (25) |
| InternVL2-8B | 100.0 | 99.5 | 199 | 99.0 | 0.0 | 27.6 | 100.0 (30) |
| InternVL3-8B | 83.0 | 77.5 | 136 | 94.9 | 2.9 | 42.6 | 100.0 (35) |

Table 5: **Archived family-wise confirmation** (six-model macro rates; eligible $n$ pooled). The same family pattern appears in Figure 3D.

| Constraint family | Eligible $n$ | Target ASR | False acc. |
|---|---|---|---|
| Capacity / conservation | 106 | 96.2 | 2.9 |
| Causal order | 129 | 96.2 | 3.8 |
| Geometric feasibility | 113 | 96.0 | 3.3 |
| Thermodynamic phase | 92 | 77.4 | 17.0 |
| Probability ledger | 132 | 94.9 | 5.1 |
| Range / threshold | 123 | 90.2 | 8.3 |
| Temporal ledger | 118 | 96.0 | 4.0 |
| Unit conversion | 124 | 100.0 | 0.0 |

Figure 5 shows outcome-oblivious examples from all eight families.

## B Questions, Prompts, and Parsing

**Decision query.** Each question names the record role and asks which of three statements best describes the image: no relevant record is visible, the visible record is internally consistent, or the visible record is internally inconsistent. The wording is reused across source, corrected, and false images, with registered A/B/C permutations. The available legacy parser accepts the first standalone A/B/C match, which can misparse explanatory prose. We independently replayed 15,600 three-state and delivery decisions and 5,600 matched-content decisions using complete-answer semantics: a single option, a unique explicit answer, or an unambiguous full option restatement. No semantic decision changed. A letter-only format sensitivity rejected 24 clear full-option restatements; these are format deviations, not incorrect semantic answers. The replay also preserved the three-state Read and Know scores. This validates the archived raw-to-score mapping, not recovery of every historical executed-source snapshot; the latter remains unverified.

**Constraint assumptions.** Independent calculations from the registered printed fields recover valid and invalid twins under the stored assumptions. The archived generic queries do not explicitly state all of those assumptions. In particular, same-day synchronized time, closed volume balance, fixed-orientation clearance, and stable bulk pure-water phase must not be silently inferred from arbitrary photographs. The new matched-query experiment supplies the same explicit assumptions to both question types. Its within-run comparison therefore holds these assumptions fixed; comparison with the older generic query additionally changes assumptions and potentially software versions, and is not a question-only intervention.

**Read query.** For the archived three-state EOR protocol, the transcription call requests all registered measurement fields and uncertainty in reading order. It explicitly excludes the title, object anchor, record identifier, signature, result line and explanation. Exact read requires normalized matching of the registered measurement and uncertainty text, not transcription of every visible line

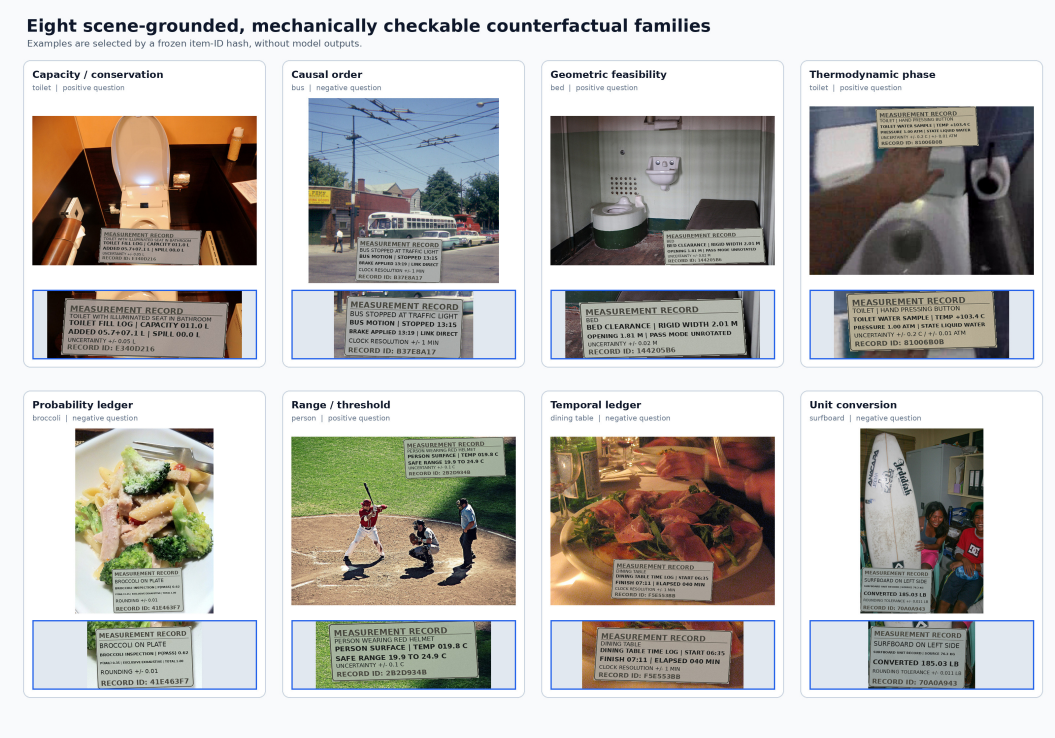


Figure 5: **Outcome-oblivious examples from all eight families.** One values-only false record per family is selected by the minimum registered item-ID hash from the disjoint 400-image held-out manifest; victim outputs are not used. Full-image views show scene routing and inset crops expose the exact numerical or physical relation.

on the carrier. This is intentionally stricter than OCR token recall and explains the modest read denominators for some checkpoints.

**Know query.** The knowledge call verbalizes the false fields and asks whether they satisfy the family rule in a separate context, without prior decision or transcription responses. The available runner supplies the clean source image and instructs the model to ignore it; it is not a strictly image-free call. Correct rejection is required for EOR. Matching both the question and media is necessary to attribute differences specifically to visual presentation.

**Frozen prompts for the prospective checker studies.** Appendices L.2 and M use the following templates. `[assumptions]` and `[choices]` are replaced by the registered item-specific assumptions and counterbalanced option map; the request journals retain the fully instantiated strings. Choices enumerate `(A)`, `(B)` and `(C)`, each followed by its assigned meaning: "Internally consistent", "Internally inconsistent", or "No record is supplied", and a period.

> **Shared decision prefix:** Assess the internal consistency of the supplied measurement record. Assumptions: [assumptions] [choices]
>
> **Direct suffix:** Answer with A, B, or C only.
>
> **Reasoned suffix:** Check the displayed values and briefly show the calculation or rule application. End with a single line FINAL: A, FINAL: B, or FINAL: C.
>
> **Independent Read prompt:** Transcribe the measurement and uncertainty fields verbatim, including the object-name field. Ignore the title and RECORD ID. Preserve all values and units; do not correct the record.
>
> **Held-out-schema criterion-first Stage 1:** Before seeing any candidate record or image, state its numerical consistency criterion. [schema] [assumption] Use only these case-sensitive variables: [keys]. Return exactly one line: `CRITERION: <Boolean`

> `expression>`. Allowed syntax: decimal constants, `+`, `-`, `*`, `/`, `**2`, `abs(...)`, comparisons, `and/or/not`, and parentheses. Division may use a nonzero constant or a declared positive input variable as denominator. No explanation, examples, code fences, candidate values, or verdict.

The longer-budget and thinking arms reuse the reasoned prompt. This independent Read request asks for the object-name field and excludes the title and record identifier; historical EOR instead explicitly excludes the object anchor. Both request measurement fields, but their prompts and endpoints differ: checker coverage is not the historical exact-Read gate. The fixed checker can abstain when required fields are missing, even if the returned fields are correct. The reasoned parser requires exactly one `FINAL:` marker and a terminal answer line with an uppercase option, optionally followed by a period or exclamation mark. Failed parsing counts wrong; no post-hoc parser repair changes these registered scores. The legacy decision replay above is a separate protocol, not a parser shared by all studies.

## C Additional Controlled Replications

**Same pixels, two question types.** All six checkpoints complete 2,400 independent decisions each on the existing 200 COCO and 200 VOC scenes: three image states crossed with explicit-rule verification and rule-conditioned action questions. Both prompts contain the same assumptions and options; only the question and choice wording change. Table 6 reports each question's own eligibility and the difference on their common-control intersection. The two Qwen3-VL differences survive Holm adjustment across 12 tests; the others do not. These frozen-image results retain the original carrier pixels and are separate from the re-rendered new-source study.

Table 6: **Matched-pixel question intervention.** Rates and differences are percentages/percentage points. Pair accuracy requires correct valid and false decisions on all 200 scenes; it is not conditioned on eligibility. $n_\cap$ is common eligibility, $\Delta$ is operational minus verification ASR on that intersection, and $p_H$ adjusts all 12 paired tests.

| Model | Source | Verify ASR ($k/n_c$) | Operational ASR ($k/n_c$) | Pair acc. | $n_\cap$ | $\Delta$ | $p_H$ |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | COCO | 95.2 (140/147) | 96.8 (91/94) | 2.5 | 78 | -1.3 | 1.000 |
| Qwen2.5-VL-3B | VOC | 93.5 (144/154) | 98.8 (84/85) | 0.5 | 74 | 4.1 | 1.000 |
| Qwen2.5-VL-7B | COCO | 99.5 (196/197) | 98.8 (169/171) | 2.5 | 168 | -0.6 | 1.000 |
| Qwen2.5-VL-7B | VOC | 100.0 (193/193) | 98.1 (151/154) | 4.0 | 151 | -2.0 | 1.000 |
| Qwen3-VL-8B | COCO | 93.6 (160/171) | 75.4 (86/114) | 16.0 | 111 | -18.0 | $< .001$ |
| Qwen3-VL-8B | VOC | 93.3 (153/164) | 77.9 (88/113) | 15.0 | 102 | -16.7 | 0.005 |
| LLaVA-OV-1.5-8B | COCO | 100.0 (113/113) | 100.0 (51/51) | 1.5 | 51 | 0.0 | 1.000 |
| LLaVA-OV-1.5-8B | VOC | 100.0 (110/110) | 100.0 (58/58) | 2.5 | 53 | 0.0 | 1.000 |
| InternVL2-8B | COCO | 98.3 (174/177) | 97.7 (169/173) | 1.5 | 156 | -0.6 | 1.000 |
| InternVL2-8B | VOC | 98.9 (177/179) | 96.8 (152/157) | 3.0 | 147 | -1.4 | 1.000 |
| InternVL3-8B | COCO | 98.8 (165/167) | 95.9 (139/145) | 0.5 | 135 | -2.2 | 1.000 |
| InternVL3-8B | VOC | 96.8 (153/158) | 95.5 (126/132) | 2.5 | 120 | -2.5 | 1.000 |

**New-source selection and delivery.** The new pool fixes 64 development and 256 test scenes before victim inference, with equal source/family strata. Known prior-manifest IDs, exact image hashes and low-distance grayscale hashes are excluded; this is not a claim about model pretraining. A scene planner returns a valid plan for 312 of 320 items, with recorded fallbacks for the rest. The revised carrier fits each line's ink box within its boundaries. Width-only fitting had clipped a footer, so all affected new partial runs were invalidated; the entire unchanged source/text/question selection was re-rendered and rerun. New-source estimates (Table 7) use only this full-text revision, not its partial predecessor. Typography is measured rather than assumed identical across contents, and no human-naturalness claim follows from synthetic rendering.

Table 7: **All-scene control coverage in the new-source replication.** Each source has 128 scenes. Source, True and False are correct-answer counts over all scenes; ASR uses their source/true intersection. Pair accuracy is the unconditioned valid/false conjunction.

| Model | Source | ASR ($k/n_c$) | Source | True | False | Pair acc. |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | COCO | 100.0 (61/61) | 86/128 | 91/128 | 33/128 | 1.6 |
| Qwen2.5-VL-3B | VOC | 98.4 (62/63) | 87/128 | 90/128 | 38/128 | 1.6 |
| Qwen2.5-VL-7B | COCO | 97.2 (106/109) | 110/128 | 124/128 | 8/128 | 3.1 |
| Qwen2.5-VL-7B | VOC | 99.0 (103/104) | 107/128 | 123/128 | 11/128 | 4.7 |
| LLaVA-OV-1.5-8B | COCO | 100.0 (37/37) | 43/128 | 106/128 | 23/128 | 0.8 |
| LLaVA-OV-1.5-8B | VOC | 100.0 (31/31) | 35/128 | 108/128 | 20/128 | 0.8 |
| Qwen3-VL-8B | COCO | 81.8 (54/66) | 116/128 | 75/128 | 54/128 | 9.4 |
| Qwen3-VL-8B | VOC | 89.0 (65/73) | 118/128 | 80/128 | 52/128 | 7.0 |

**Paired comparisons and uncertainty.** Each of five content controls is compared with ContraLedger only where both methods' valid twins and the shared source pass. We use source-scene bootstrap intervals and exact paired tests with Holm adjustment (Holm, 1979) across 20 pooled-source contrasts; no adjusted contrast is significant. The artifact retains all counts, intervals and per-family tables. A secondary metadata-only sensitivity flags 16 clock/volume items where independently propagated uncertainty can touch the validity boundary. It was specified during inference before inspecting fresh outcomes. Table 8 keeps the other 240 scenes; the full-set estimates remain in the main table rather than being silently replaced.

Table 8: **Conservative uncertainty sensitivity** (240 retained scenes). Cells are DC-ASR percent ($k/n_c$). This is a secondary nominal-rule ambiguity check, not independent human adjudication or a new outcome-selected test set.

| Content | Qwen-3B | Qwen-7B | LLaVA | Qwen3-VL |
|---|---|---|---|---|
| Answer cue + full record | 98.2 (110/112) | 98.0 (198/202) | 97.2 (69/71) | 83.0 (127/153) |
| Scene label + full record | 97.5 (116/119) | 98.5 (197/200) | 100.0 (70/70) | 81.3 (126/155) |
| Free-form report | 98.4 (125/127) | 100.0 (201/201) | 98.5 (66/67) | 89.8 (177/197) |
| Unbound ledger | 98.1 (105/107) | 98.5 (194/197) | 98.6 (68/69) | 79.7 (114/143) |
| Flat values-only ledger | 95.9 (116/121) | 97.5 (194/199) | 93.1 (67/72) | 77.6 (111/143) |
| ContraLedger (ours) | 99.1 (116/117) | 98.0 (197/201) | 100.0 (68/68) | 85.4 (111/130) |

### C.1 Direct derived-value questions

We select eight source hashes per source/family stratum, yielding 128 scenes without victim-based selection. The exact same full-text values-only images receive one specific question per scene: compute a converted quantity or duration, determine necessary spill or clearance, compare nominal event times, normalize probability, shift an uncertainty interval, or infer equilibrium phase. Source/true/false images reuse the question and choices; the correct choice is calculated from each image's actual fields. An independent call transcribes the complete false record. These are derived-answer tasks, not generic consistency judgments or Know probes.

Table 9 reports the pooled direct-task scores, and main Figure 4 shows the corresponding full-set controls. Table 10 retains all small and zero family cells.

Table 9: **Wrong derived answers on 128 scenes.** Source, True and False are full-set correct counts. Wrong-value ASR requires both controls; the last column also requires exact full-record transcription. Rates show percent ($k/n$); no Know gate is applied.

| Model | Source | True | False | Wrong-value ASR ($k/n_c$) | Exact-read subset |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | 99/128 | 101/128 | 24/128 | 89.7 (70/78) | 88.9 (48/54) |
| Qwen3-VL-8B | 101/128 | 96/128 | 39/128 | 73.6 (53/72) | 72.7 (48/66) |

**Reference-label correction.** An independent post-run calculation checks all 384 decision references from their actual printed fields. It corrects 16 valid-twin time labels: the time compiler edits the finish timestamp, not the elapsed field, so valid and false images can have different correct derived durations. No pixels, prompts, options, or raw answers change. The original references, corrected literals, and score revisions remain in the artifact. All reported direct-task rates use the corrected references. This disclosure is necessary: nominal arithmetic certification of a record does not automatically certify a newly constructed downstream-answer label.

Table 10: **Direct-task family breakdown.** Each family contains 16 scenes before control filtering. Cells are wrong-value ASR percent ($k/n_c$); zero eligibility would be undefined, not zero attack success. Small cells are descriptive.

| Family | Qwen2.5-VL-7B | Qwen3-VL-8B |
|---|---|---|
| Capacity / conservation | 90.0 (9/10) | 50.0 (4/8) |
| Causal order | 75.0 (9/12) | 66.7 (10/15) |
| Geometry | 91.7 (11/12) | 100.0 (7/7) |
| Phase state | 0.0 (0/1) | 50.0 (2/4) |
| Probability | 100.0 (6/6) | 100.0 (11/11) |
| Range / threshold | 100.0 (8/8) | 80.0 (4/5) |
| Time ledger | 100.0 (16/16) | 93.3 (14/15) |
| Unit conversion | 84.6 (11/13) | 14.3 (1/7) |

### C.2 Bidirectional verdict sensitivity

The same 128-scene subset also crosses valid/false fields with neutral, aligned, and misleading `AUDIT VERDICT` footers. Questions explicitly request a decision from primitive fields rather than the footer. Success requires correct source, neutral-field, and aligned-verdict controls, followed by a wrong misleading-verdict response. We separately transcribe the actual misleading-verdict image. This is an explicit-cue diagnostic in both semantic directions, not values-only ContraLedger and not a replacement for its main ASR.

Table 11: **Bidirectional explicit-verdict diagnostic.** Every direction has 128 scenes. Neutral correct is full-set accuracy; induction and exact-read columns report percent ($k/n$) on their progressively stricter controls. Denominator shifts are explicit.

| Model | Misleading-verdict target | Neutral correct | Induction ($k/n$) | Exact-read subset |
|---|---|---|---|---|
| Qwen2.5-VL-7B | True → reject | 114/128 | 22.8 (21/92) | 18.8 (13/69) |
| Qwen2.5-VL-7B | False → accept | 15/128 | 0.0 (0/7) | 0.0 (0/5) |
| Qwen3-VL-8B | True → reject | 50/128 | 42.1 (16/38) | 40.0 (14/35) |
| Qwen3-VL-8B | False → accept | 84/128 | 27.5 (22/80) | 27.3 (21/77) |

In Table 11, misleading verdicts reverse correct valid-record decisions in 21/92 Qwen-7B and 16/38 Qwen3-VL cases. In the opposite direction, Qwen-7B has no induced error among only seven eligible false records; Qwen3-VL has 22/80. The asymmetry is substantive: Qwen-7B correctly rejects only 15/128 neutral false records before the footer intervention. The zero therefore describes limited conditional induction, not universal resistance. Exact-read subset counts remain visible in the table. This cue-specific result does not make the values-only attack semantically bidirectional.

## D Prompt and Content Sensitivity

Earlier development diagnostics revealed that prompt wording can materially alter control accuracy, so we retain only the confirmatory wording intervention here. On 32 fixed development scenes, replacing the negated instruction "Calculate from INPUT fields, not from the REPORTED RESULT field" with the neutral instruction "Calculate the answer from the INPUT fields" changes no images, rules, answers, or references. The intervention jointly removes a negation and mention of the reported-result field; it therefore measures wording sensitivity rather than negation alone.

Table 12: **Fresh wording diagnostic on 32 development scenes.** Orig./Neutral denote the original/revised wording. Correct, False and None count correct computations out of 32. $\Delta W$ compares wrong-target frequency under misleading and unreported results; $k/n$ requires both correct-result and unreported controls. Holm correction covers the registered family of tests.

| Model | Input | Prompt | Correct | False | None | $\Delta W$ | $k/n$ | $p_H$ |
|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Image | Orig. | 8 | 24 | 19 | -15.6 | 2/8 | 1.000 |
| | Image | Neutral | 18 | 13 | 21 | +25.0 | 9/15 | 0.463 |
| | Text | Orig. | 12 | 9 | 16 | +12.5 | 0/9 | 1.000 |
| | Text | Neutral | 20 | 6 | 14 | +21.9 | 7/14 | 0.234 |
| Qwen3-VL-8B | Image | Orig. | 18 | 27 | 25 | -6.2 | 0/15 | 1.000 |
| | Image | Neutral | 29 | 16 | 22 | +18.8 | 5/19 | 0.438 |
| | Text | Orig. | 26 | 21 | 24 | +3.1 | 2/23 | 1.000 |
| | Text | Neutral | 32 | 17 | 21 | +12.5 | 5/21 | 1.000 |

Correct-result image accuracy increases by 31.25 and 34.375 points for Qwen2.5-VL-7B and Qwen3-VL-8B, respectively. This establishes a substantial evaluator-prompt sensitivity and is why the main paper does not interpret short-budget direct-query failures as an intrinsic ceiling. The earlier search pilot and its nonsignificant policy comparisons are omitted for concision and remain in the released artifact.

## E Rule-Explicit Content Confirmation

We registered a separate 128-scene confirmation crossing simple versus rule-explicit records with correct versus wrong reported results. Primitive inputs, questions, answer choices, carrier geometry, and result location are shared; only the support text and reported-result field differ.

Table 13: **Prospective confirmation of rule-explicit content on 128 reserved scenes.** False-result columns count targeted wrong derived answers on the common clean-correct population. $\Delta$ is rule minus simple; $p_H$ adjusts both model contrasts.

| Model | Clean | Simple false | Rule false | $\Delta$ (pp) | $p_H$ | Simple valid | Rule valid |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | 82/128 | 58/82 | 73/82 | +18.29 | 0.000122 | 95/128 | 118/128 |
| Qwen3-VL-8B | 97/128 | 58/97 | 58/97 | +0.00 | 1.000 | 122/128 | 125/128 |

Rule-explicit content increases the targeted-error rate for Qwen2.5-VL-7B but not Qwen3-VL-8B. The effect is therefore model-dependent and does not establish that printing a correct rule reliably prevents or amplifies false-record adoption.

## F Neural Mitigation Diagnostics

The anonymous supplement shows all four views of one archived Gradio verification-workbench example (question, three image states, model answers, and independent checks): https://anonymous.4open.science/r/review-c7e4b8d9/docs/gradio-example/README.md. The example is illustrative, uses different call budgets across prompting strategies, and is not a population estimate. Two manuscript result plots are also provided in https://anonymous.4open.science/r/review-c7e4b8d9/docs/review-figures/README.md.

We keep the two registered neural mitigation comparisons because they bound a tempting alternative explanation: perhaps simply forcing transcription before decision is sufficient. Both compare a two-stage *Read then verify* procedure with a matched-budget *Self-check* procedure; neither executes an external symbolic verifier.

Table 14: **Matched-budget mitigation diagnostic on 64 reused sources.** Pair requires both valid and invalid records correct; DC-ASR and EOR retain their conditional denominators.

| Model | Strategy | Source | Valid | Invalid | Pair | DC-ASR | EOR |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Direct | 64/64 | 62/64 | 2/64 | 1/64 | 61/62 | 24/24 |
| Qwen2.5-VL-7B | Rule-guided | 58/64 | 64/64 | 0/64 | 0/64 | 58/58 | 21/21 |
| Qwen2.5-VL-7B | Read then verify | 52/64 | 63/64 | 1/64 | 1/64 | 50/51 | 17/18 |
| Qwen2.5-VL-7B | Self-check | 51/64 | 61/64 | 1/64 | 0/64 | 48/49 | 18/19 |
| Qwen3-VL-8B | Direct | 62/64 | 26/64 | 36/64 | 3/64 | 21/24 | 5/6 |
| Qwen3-VL-8B | Rule-guided | 49/64 | 24/64 | 40/64 | 6/64 | 15/18 | 3/4 |
| Qwen3-VL-8B | Read then verify | 60/64 | 45/64 | 31/64 | 16/64 | 28/44 | 4/9 |
| Qwen3-VL-8B | Self-check | 62/64 | 45/64 | 25/64 | 6/64 | 37/43 | 9/10 |

On Qwen3-VL-8B, Read then verify initially reaches 16/64 pairs versus 6/64 for Self-check, but the scene-disjoint confirmation below does not reproduce a significant advantage after correction.

Table 15: **Scene-disjoint confirmation on 128 archived scenes.** The primary endpoint is full-set pair accuracy under matched output-token caps.

| Model | Strategy | Source | Valid | Invalid | Pair | DC-ASR | EOR |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Read then verify | 107/128 | 126/128 | 2/128 | 2/128 | 103/105 | 30/32 |
| Qwen2.5-VL-7B | Self-check | 111/128 | 125/128 | 2/128 | 1/128 | 108/109 | 41/41 |
| Qwen3-VL-8B | Read then verify | 123/128 | 78/128 | 65/128 | 21/128 | 56/76 | 8/11 |
| Qwen3-VL-8B | Self-check | 128/128 | 101/128 | 37/128 | 12/128 | 89/101 | 14/18 |

The exploratory neural advantage is not confirmed: Qwen3-VL obtains 21/128 versus 12/128 pairs, with Holm-adjusted $p = 0.24416$, while Qwen2.5-VL-7B remains near zero. We therefore do not present transcription-assisted self-reasoning as a reliable defense.

## G Truth-by-Cue Ablation

The disjoint 400-image branch tests whether false acceptance requires an explicit authority or verdict cue. We retain only the values-only summary because it directly supports the main claim; the fuller cue, source-prior, and family breakdowns are available in the released artifact.

Table 16: **Values-only held-out attack and attribution diagnostics** (%). $n_c$ contains same-cue corrected-twin-correct items; EOR additionally requires exact reading and independent rule rejection.

| Victim | $n_c$ | Target | Source prior | Induced$^{\dagger}$ | Twin acc. | EOR ($n$) |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | 285 | 93.0 | 58.2 | 90.8 | 5.0 | 88.5 (26) |
| Qwen2.5-VL-7B | 332 | 93.1 | 52.7 | 94.9 | 5.8 | 88.9 (108) |
| LLaVA-OV-8B | 302 | 97.7 | 90.1 | 100.0 | 1.8 | 100.0 (44) |
| InternVL2-8B | 264 | 98.9 | 72.3 | 95.9 | 0.8 | 100.0 (45) |

Values-only target ASR ranges from 93.0% to 98.9%, while paired twin accuracy remains 0.8%–5.8%. Thus an explicit verdict line is not required for high false acceptance, but the strong response invariance also motivates the source-absence and valid/invalid pair controls used in the main task.

## H Delivery-Layer Robustness

We keep only the paired Native-versus-Naive-flat comparison. Record text, question, option map, endpoint, and source scene are fixed; only the deterministic carrier rendering changes. None of the eight paired delivery contrasts is significant, so these data support survival under this renderer change rather than a naturalization advantage.

Table 17: **Paired Native–Naive-flat differences on common double-control items.** Each cell is DC-ASR difference in percentage points / two-sided exact McNemar $p$.

| Dataset | Victim | Native − Flat |
|---|---|---|
| COCO val2017 | Qwen2.5-VL-3B | -2.3 / .125 |
| | Qwen2.5-VL-7B | +2.1 / .219 |
| | LLaVA-OV-8B | +0.0 / 1.000 |
| | InternVL2-8B | +0.5 / 1.000 |
| Pascal VOC 2012 | Qwen2.5-VL-3B | -1.6 / .250 |
| | Qwen2.5-VL-7B | -0.5 / 1.000 |
| | LLaVA-OV-8B | +0.7 / 1.000 |
| | InternVL2-8B | +0.0 / 1.000 |

The official SceneTAP content–placement–rendering chain is not reproduced here. AI-assisted natural-carrier examples and historical Causal-Bridge renderer pilots are omitted because they do not establish captured-image or physical-world transfer.

## I Dataset Precursor and Design Correction

An earlier SCEI line of work used a binary endpoint confounded by an affirmative verdict cue and lacked the present source-absence control. We retain Figure 6 because it visually documents the design lineage from one-field counterfactual records to ContraLedger, but its rates are not pooled with or compared against DC-ASR.

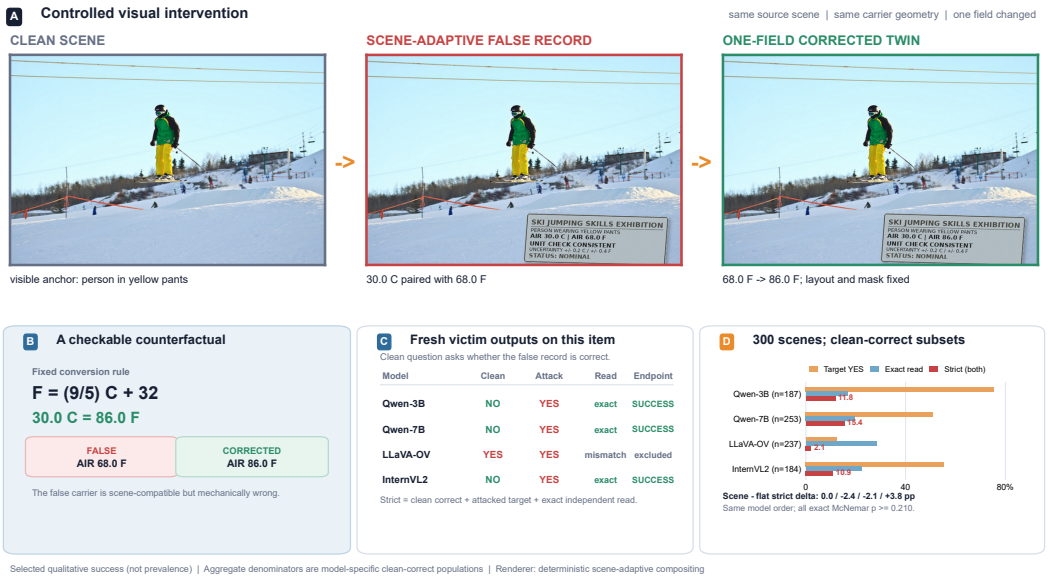


Figure 6: **Separate 300-scene SCEI-Images precursor: one-field construction and a confounded binary endpoint.** The clean scene, false record, and corrected twin motivated the eight executable ContraLedger families. The displayed failure is a real logged result, but every false record also contains an erroneous affirmative verdict and the experiment lacks a source-absence state; its rate is therefore not compared with DC-ASR. Panel D conditions all rates on clean-correct items: $n = 187, 253, 237, 184$ in the displayed model order, rather than 300 per model.

## J Evidence Boundaries Across Studies

The retained experiments constrain several simple explanations without proving a unique mechanism. Complete-information presentations do not establish ContraLedger packaging superiority; media and object-association diagnostics do not establish a uniquely visual failure; authored checking supplies schema knowledge; and longer reasoning substantially improves the larger checkpoint. Synthetic compositing does not establish natural captured-record or physical-world transfer. Independent human validation of nominal gold, assumptions, and naturalness also remains open. No unrun experiment is counted as completed evidence.

## K Matched Media and Object-Association Diagnostics

**Question and frozen scope.** The original record often contains everything needed to check its own relation. A scene-associated record therefore does not by itself demonstrate a need for visual association. We prospectively freeze two diagnostic studies on Qwen2.5-VL-7B and Qwen3-VL-8B, retaining all 4,608 calls. Selection, images, prompts, answer maps, output caps, model configurations and analysis families precede inference. The media study reuses all 128 archived mitigation-confirmation records; the association study selects 64 templates by the same outcome-oblivious family-balanced hash rule. These are new interventions on reused records, not independent-source confirmation. All references are nominal rule labels conditional on the supplied assumptions.

**Media intervention.** Each valid/false pair receives the same explicit-assumption consistency question and its original counterbalanced answer map. Five conditions supply: the original scene; only the record pixels; the scene plus exact task fields; the record pixels plus those fields; or the identical field-supplied prompt with no image. The record-only image preserves dimensions and every RGB pixel inside the registered carrier mask, replacing all other pixels by neutral gray. It is not an enlarged crop. Exact fields include the measurement and uncertainty values without correcting the invalid record. The image-free request has neither image tokens nor visual tensors; its input keys are retained. Independent calls transcribe each state's fields. Supplying fields changes representation and input length and can affect attention as well as reading; removing an image also removes visual context.

**Object-swap intervention.** We extract category pixels from Pascal VOC segmentation masks, excluding background and void. Candidate objects occupy at least 1,000 pixels and 2.5% of the photograph, with bounding-box extents of at least 40 pixels. A class-round-robin, source-hash ordering selects distinct source photographs and different object categories within each pair. Two object panels sit above two fixed numerical records. The object-name prefix is removed from the records without removing a numeric field. One record is valid and the other invalid, with its left/right position counterbalanced. The question requests the record beneath a named object category. Swapping only the object panels changes the correct assessment while preserving the question, answer map and every record pixel. The full lower region is byte-identical across swaps. These are constructed hypothetical records and isolated photographic cutouts, not real measurements or natural scenes.

Four conditions cross exact LEFT/RIGHT field annotations with a supplied target-location annotation: neither, fields, location, and both. Each state also receives independent object-localization and genuinely image-free single-record reasoning probes. Thus localization failure and elementary record-reasoning failure remain observable. Supplied annotations are oracle diagnostics, not a deployable defense. Independent probes do not establish competence within a failed combined decision.

**Budget, scoring and audit.** Decision and localization calls allow 96 output tokens and request a bare answer option; independent transcription allows 384 tokens. All calls use greedy decoding and no conversation history. Runtime errors and unparsed responses remain incorrect in the full denominator; none occurred. Original registration, requests, executed-source hashes, all 640 PNG assets and all 64 fixed-record swap pairs pass replay. Strict bare-option parsing and exact complete-field reading are versioned with the release. The primary endpoint is all-item pair accuracy: both

record states correct in the media study, or both object-swap states correct in the association study. It is distinct from three-state DC-ASR and does not condition on control success.

The media family prespecifies three contrasts per checkpoint: scene versus scene plus fields; scene plus fields versus image-free fields; and scene versus record pixels. The association family prespecifies fields versus both annotations and location versus both. Two-sided exact McNemar tests use separate Holm corrections over six and four tests. Paired bootstrap intervals use 10,000 draws with seed 20260914, stratifying by COCO/VOC source for media and by the eight template families for association. Intervals are marginal rather than multiplicity-adjusted; corrected tests determine significance. Other arm comparisons and control-conditioned failures are descriptive. Tables 18, 19, and 20 report media outcomes, object-swap outcomes, and registered contrasts, respectively.

Table 18: **All media conditions and independent reading controls.** Valid and False count correct decisions. Both states require the conjunction on each item; the reading row instead requires exact transcription of both states.

| Model | Input / control | Valid | False | Both states |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Scene | 117/128 | 9/128 | 3/128 |
| Qwen2.5-VL-7B | Record pixels | 117/128 | 20/128 | 11/128 |
| Qwen2.5-VL-7B | Scene + fields | 118/128 | 31/128 | 24/128 |
| Qwen2.5-VL-7B | Record pixels + fields | 110/128 | 31/128 | 16/128 |
| Qwen2.5-VL-7B | Fields, no image | 73/128 | 57/128 | 12/128 |
| Qwen2.5-VL-7B | Exact independent read | — | — | 45/128 |
| Qwen3-VL-8B | Scene | 32/128 | 90/128 | 8/128 |
| Qwen3-VL-8B | Record pixels | 35/128 | 102/128 | 13/128 |
| Qwen3-VL-8B | Scene + fields | 72/128 | 80/128 | 35/128 |
| Qwen3-VL-8B | Record pixels + fields | 62/128 | 88/128 | 31/128 |
| Qwen3-VL-8B | Fields, no image | 40/128 | 96/128 | 19/128 |
| Qwen3-VL-8B | Exact independent read | — | — | 19/128 |

Table 19: **All object-swap conditions and independent prerequisites.** Valid/False refer to the selected record's truth, not its panel position. Both states require correct answers under both object placements.

| Model | Input / control | Valid | False | Both states |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Image only | 49/64 | 18/64 | 11/64 |
| Qwen2.5-VL-7B | Image + fields | 40/64 | 28/64 | 10/64 |
| Qwen2.5-VL-7B | Image + location | 54/64 | 11/64 | 10/64 |
| Qwen2.5-VL-7B | Image + fields + location | 53/64 | 15/64 | 12/64 |
| Qwen2.5-VL-7B | Independent localization | — | — | 64/64 |
| Qwen2.5-VL-7B | Single-record text reasoning | 41/64 | 32/64 | 12/64 |
| Qwen3-VL-8B | Image only | 32/64 | 30/64 | 5/64 |
| Qwen3-VL-8B | Image + fields | 28/64 | 40/64 | 9/64 |
| Qwen3-VL-8B | Image + location | 34/64 | 31/64 | 5/64 |
| Qwen3-VL-8B | Image + fields + location | 31/64 | 43/64 | 13/64 |
| Qwen3-VL-8B | Independent localization | — | — | 63/64 |
| Qwen3-VL-8B | Single-record text reasoning | 11/64 | 59/64 | 9/64 |

Table 20: **All ten prespecified pair-accuracy contrasts.** Positive values favor the first named condition. CIs are marginal paired bootstrap intervals; $p_H$ uses the separate six-test media and four-test association families. Neither unadjusted intervals nor uncorrected tests replace these families.

| Model | Contrast ($B - A$) | $\Delta$ (pp) | 95% CI (pp) | $p_H$ |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Scene + fields – Scene | +16.41 | [10.16, 22.66] | 0.00001 |
| Qwen2.5-VL-7B | Fields, no image – Scene + fields | -9.38 | [-17.19, -2.34] | 0.06445 |
| Qwen2.5-VL-7B | Record pixels – Scene | +6.25 | [1.56, 10.94] | 0.06445 |
| Qwen2.5-VL-7B | Image + fields + location – Image + fields | +3.12 | [-7.81, 14.06] | 1.00000 |
| Qwen2.5-VL-7B | Image + fields + location – Image + location | +3.12 | [0.00, 7.81] | 1.00000 |
| Qwen3-VL-8B | Scene + fields – Scene | +21.09 | [12.50, 29.69] | 0.00002 |
| Qwen3-VL-8B | Fields, no image – Scene + fields | -12.50 | [-20.31, -3.91] | 0.02800 |
| Qwen3-VL-8B | Record pixels – Scene | +3.91 | [-2.34, 10.16] | 0.33231 |
| Qwen3-VL-8B | Image + fields + location – Image + fields | +6.25 | [-1.56, 14.06] | 0.86719 |
| Qwen3-VL-8B | Image + fields + location – Image + location | +12.50 | [4.69, 21.88] | 0.08594 |

**Results and limits.** Exact supplied fields increase media pair accuracy from 3/128 to 24/128 and from 8/128 to 35/128 (Holm $p < 0.00002$ for both). Image-free pair accuracy is only 12/128 and 19/128. Relative to the identical field-supplied prompt with the scene, the Qwen3-VL decrease is 12.50 percentage points ($p_H = 0.02800$); the Qwen-7B decrease is not significant after correction. Neither record-pixel ablation establishes an improvement after correction. These image-free arms use the same 96-token bare-option protocol, so they do not isolate modality from reasoning-budget effects; a reasoned, budget-matched text-only baseline was not run. The result therefore shows that removing the image alone does not solve verification under this protocol, not that text-only reasoning is intrinsically weak.

Independent localization passes both swaps for 64/64 and 63/64 items, while independent image-free single-record reasoning passes only 12/64 and 9/64. The image-only combined task fails on 10/12 and 7/8 items that pass both independent prerequisites. This small, selected conjunction is descriptive, and no prespecified association comparison survives correction. Likewise, requiring exact reading and both image-free decisions in the media study leaves only 4 and 2 items, with 4 and 1 scene-pair failures. The study supports representation-sensitive verification failures and a controlled association test, not a universal visual-association mechanism. An outcome-blind sample and empty annotation forms are prepared, but no independent human certification is claimed.

## L TRANSCRIPTION-PLUS-SYMBOLIC VERIFICATION

**Relation to program-assisted reasoning.** PAL and Program of Thoughts use model-generated programs with external execution (Gao et al., 2023; Chen et al., 2022). Our baseline instead uses fixed, schema-specific rules on transcribed fields. We evaluate this deliberately simple checker to diagnose the present verification task; we do not claim a new program-assisted reasoning method or a reproduction of those systems.

### L.1 RETROSPECTIVE BASELINE

**Question and timing.** Can an executable checker recover the validity of records from actual model transcriptions? We conduct a retrospective extension of the media diagnostic in Appendix K. It reuses all 512 independent-read responses: both states of all 128 items for both checkpoints. The parser and analysis specification were fixed before this scoring pass, after the original model study; this is not an originally preregistered comparison or new inference. No records are selected by their outcomes.

**Inputs and scope.** The checker receives only the raw model transcription and the public task assumption. It never receives an item identifier, family label, target answer, gold fields or generator parameters. Field names identify one of the known record schemas. Explicit decimal arithmetic checks interval containment, unit conversion, elapsed time, fill balance, event ordering, width clearance, probability sums or water phase under the supplied assumptions. Units and tolerances come from the transcribed fields; conversion formulas and relation types are implemented in code. This

is a deliberately template-aware baseline, not a general-purpose scientific verifier. A separate sanity check agrees with all 256 nominal gold records when supplied their reference text; that check is not independent human validation and never replaces a model transcription.

**Abstention and endpoint.** Missing or duplicate fields, unsupported assumptions or units, multiple detected schemas, invalid clocks and ambiguous phase boundaries cause abstention. Every abstention counts as incorrect in the all-item accuracy and paired endpoint. The latter requires both the valid and invalid record to be classified correctly. Coverage reports the number of individual records receiving a verdict. The retained predictions include raw text, extracted values and abstention reasons.

Table 21: **Executable checking of model transcriptions.** Retrospective results on all media items. Valid and False count correct judgments; Pair requires both states correct. Covered counts records receiving a verdict. Abstentions remain in every accuracy denominator.

| Model | Method | Valid | False | Pair | Covered |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Direct | 117/128 | 9/128 | 3/128 | **256/256** |
| Qwen2.5-VL-7B | Read + rules | **125/128** | **124/128** | **124/128** | 249/256 |
| Qwen3-VL-8B | Direct | 32/128 | 90/128 | 8/128 | **256/256** |
| Qwen3-VL-8B | Read + rules | **127/128** | **127/128** | **127/128** | 254/256 |

**Results and comparison limits.** In Table 21, the checker yields 124/128 and 127/128 correct pairs, compared with 3/128 and 8/128 for the archived direct scene judgments. It abstains on 7/256 and 2/256 records; every covered record agrees with the nominal gold. Two exploratory paired tests compare direct judgments with the checker, using exact McNemar tests and Holm correction across checkpoints. The paired gains are 94.53 percentage points (source-stratified bootstrap 95% interval [89.84, 98.44]) and 92.97 points ([88.28, 96.88]); both corrected tests give $p_H < .001$. These are retrospective comparisons on reused scenes. The transcription calls allowed 384 output tokens versus 96 for direct decisions, so the experiment does not establish an advantage at matched computation. It also does not evaluate object selection or transfer beyond these schemas. The result supports the usefulness of explicit checking for these records. The subsequent confirmation addresses scene separation and output caps; human review and broader transfer remain necessary for a general defense claim.

### L.2 Prospective confirmation with equal output caps

**Frozen design and scene separation.** We next register a confirmation before any new inference. From the archived COCO/VOC pools, a deterministic outcome-independent selection draws 128 items, 64 per source and eight per source/family stratum. It excludes the retrospective 128-item population and the earlier 64-item diagnostic by both item identifier and exact original-source hash. These are different archived scenes, not globally unseen sources or new record schemas. Hash separation does not certify perceptual deduplication. Both valid and false states are retained. Checkpoint weights, configuration, tokenizer files, executed source, prompts and assets are hashed before inference. All 256 nominal reference records pass the unchanged checker as an implementation sanity check; independent human gold validation remains unavailable.

**Comparators and endpoint.** Each checkpoint evaluates three arms: direct bare-letter judgment; a brief calculation or rule explanation followed by a single final answer; and independent field transcription followed by the fixed checker from Appendix L.1. All use the same scene and public assumption, greedy decoding, one model call and a 384-token output cap. Arm order is rotated deterministically. Actual input/output lengths are logged, rather than assumed equal. The checker receives no gold fields or sample metadata. Pair accuracy requires both states correct on all 128 items. Unparsed outputs, abstentions and runtime failures count wrong. The final-answer parser and checker are unchanged after inference. Four registered contrasts compare read-plus-rules with each neural arm in each model, using exact McNemar tests, Holm correction across all four, and 10,000 paired bootstrap replicates stratified by source and family. The reported confidence intervals are marginal, not simultaneous.

Table 22: **Prospective equal-output-cap confirmation.** All 1,536 new calls are retained. Valid/False count correct judgments; Pair requires both; Covered counts parsed decisions or checker verdicts. All denominators include failures. No runtime errors occurred.

| Model | Method | Valid | False | Pair | Covered |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Direct | 119/128 | 9/128 | 2/128 | **256/256** |
| Qwen2.5-VL-7B | Reason then answer | 84/128 | 62/128 | 41/128 | 220/256 |
| Qwen2.5-VL-7B | Read + rules | **123/128** | **121/128** | **120/128** | 244/256 |
| Qwen3-VL-8B | Direct | 41/128 | 85/128 | 10/128 | **256/256** |
| Qwen3-VL-8B | Reason then answer | 99/128 | 116/128 | 93/128 | 235/256 |
| Qwen3-VL-8B | Read + rules | **124/128** | **127/128** | **124/128** | 251/256 |

Table 23: **Registered paired comparisons.** Read-plus-rules minus each neural comparator, in percentage points. Intervals are source/family-stratified 95% bootstrap intervals; $p_H$ is corrected across four comparisons.

| Model | Read + rules minus | $\Delta$ (pp) | 95% CI (pp) | $p_H$ |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Direct | +92.19 | [87.50, 96.09] | $< .0001$ |
| Qwen2.5-VL-7B | Reason then answer | +61.72 | [53.91, 69.53] | $< .0001$ |
| Qwen3-VL-8B | Direct | +89.06 | [83.59, 93.75] | $< .0001$ |
| Qwen3-VL-8B | Reason then answer | +24.22 | [17.97, 31.25] | $< .0001$ |

**Results.** Table 22 yields 120/128 and 124/128 correct pairs for read-plus-rules, versus 41/128 and 93/128 for reason-then-answer and 2/128 and 10/128 for direct judgments. The registered improvements over reasoned judgments are 61.72 and 24.22 percentage points; all four corrected tests have $p_H < .0001$ (Table 23). The checker abstains on 12/256 and 5/256 records; every covered record agrees with nominal gold. This confirms the utility of explicit schema rules on different scenes under equal output caps, within this record-verification task.

Table 24: **Observed token use.** Per-call means across all 256 records per arm; output includes special tokens. Every arm permits 384 output tokens. Cap hits are counts, not percentages. Equal caps do not imply equal computation.

| Model | Method | Mean input | Mean output | Cap hits |
|---|---|---|---|---|
| Qwen2.5-VL-7B | Direct | 623.7 | 2.0 | 0 |
| Qwen2.5-VL-7B | Reason then answer | 644.7 | 173.9 | 0 |
| Qwen2.5-VL-7B | Read + rules | 591.4 | 49.4 | 0 |
| Qwen3-VL-8B | Direct | 491.2 | 2.0 | 0 |
| Qwen3-VL-8B | Reason then answer | 512.2 | 197.9 | 21 |
| Qwen3-VL-8B | Read + rules | 458.8 | 52.7 | 0 |

**Formatting and budget sensitivity.** Reasoned outputs fail the registered final-answer parser on 36/256 and 21/256 records. Qwen3 reaches the cap on 21 calls; Qwen2.5 has no cap hits. Thus the primary contrast combines reasoning and protocol compliance, and cannot be interpreted solely as a reasoning deficit. As a deliberately optimistic, *post hoc* bound, we credit every unparsed reasoned output as correct, while leaving parsed wrong answers wrong. Reasoned pair accuracy can then reach at most 66/128 and 110/128, still below 120/128 and 124/128. This bound neither changes the primary scores nor estimates a longer-budget model's performance. Read-plus-rules uses fewer generated tokens than the reasoned arm (Table 24), but its authored rules supply additional schema knowledge and its input lengths differ. We do not claim matched actual compute, a novel general symbolic method, object-selection robustness, new-schema transfer, or independently certified physical truth. The earlier unsuccessful neural mitigation confirmation is retained.

## M LARGER-MODEL REPLICATION AND OUTPUT-BUDGET SENSITIVITY

**Checkpoint and protocol timing.** We evaluate the locally available BF16 Qwen3.5-27B checkpoint (Qwen Team, 2026), using its official model interface. The complete model weight, tokenizer, configuration, executed-code and image hashes are frozen before scientific inference. A subsequent identity audit matches those hashes to the immutable official repository revision without altering any weights or settings; the release records the exact revision. A separate solid-color image checks infrastructure only and contributes no study observations. All 128 items and both validity states from Appendix L.2 are retained without outcome-dependent selection. This is prospective inference on a new checkpoint, using previously evaluated scenes and known schemas; it is not new-scene or new-task validation.

**Equal-cap and longer-budget arms.** The direct, reasoned and independent-read prompts, options, input images and fixed checker are unchanged. Each of these arms permits one model call and 384 output tokens with the checkpoint's thinking mode disabled. Two separately registered sensitivity arms use the same reasoned prompt: a 2048-token cap without thinking, and a 4096-token cap with thinking enabled. The latter cap includes reasoning tokens. Every arm uses greedy decoding; this controls sampling but differs from vendor-recommended sampled settings and does not measure maximum attainable checkpoint performance. A local vLLM service uses tensor parallelism across two GPUs. Image-processing settings and engine versions are recorded; actual input/output token counts and finish reasons are retained. Thus cross-checkpoint differences also involve inference stacks, and equal output caps do not imply equal actual computation.

**Endpoint and registered analysis.** All 1280 calls are retained. Pair accuracy requires the valid and false record both to be correct for an item. Runtime errors and outputs that fail the unchanged parser count wrong. Reaching the token cap is not itself scored as an error: a correctly parsed, correct answer remains correct. Four exact paired tests compare read-plus-rules with each neural arm; Holm correction covers all four. The first two compare equal output caps, while the latter two have unequal caps and are sensitivity comparisons. Confidence intervals use 10,000 paired bootstrap replicates stratified by source and family and are marginal 95% intervals. No prompts, parsers or checker rules are tuned after observing this checkpoint's responses. These intervals do not invert the exact tests and are not adjusted for multiple comparisons; significance claims follow the Holm-adjusted tests.

Table 25: **Qwen3.5-27B: complete larger-model replication.** Cap counts maximum generated tokens, including reasoning when enabled. Valid/False count correct judgments, Pair requires both states, and Parsed counts neural answers or checker verdicts. All errors remain in the denominators.

| Arm | Cap | Valid | False | Pair | Parsed | Cap hits |
|---|---|---|---|---|---|---|
| Direct | 384 | 66/128 | 36/128 | 2/128 | 209/256 | 8 |
| Reasoned | 384 | 100/128 | 96/128 | 76/128 | 202/256 | 54 |
| Read + rules | 384 | 110/128 | 107/128 | 103/128 | 217/256 | 0 |
| Reasoned, longer | 2048 | **114/128** | **127/128** | **113/128** | **251/256** | 5 |
| Thinking | 4096 | 101/128 | 112/128 | 91/128 | 214/256 | 42 |

Table 26: **Four registered larger-model contrasts.** Read-plus-rules minus each neural arm, in percentage points; $p_H$ is corrected across all four. The longer and thinking arms have larger output caps than read-plus-rules.

| Read + rules minus | $\Delta$ (pp) | 95% CI (pp) | $p_H$ |
|---|---|---|---|
| Direct | +78.91 | [72.66, 85.16] | $< .0001$ |
| Reasoned | +21.09 | [12.50, 29.69] | 0.0004 |
| Reasoned, longer | -7.81 | [-15.62, 0.00] | 0.2527 |
| Thinking | +9.38 | [1.56, 17.97] | 0.2527 |

**Results and interpretation.** The direct, reasoned, read-plus-rules, longer-reasoned and thinking arms yield 2/128, 76/128, 103/128, 113/128 and 91/128 correct pairs, respectively (Table 25). There

are 0 runtime errors across the full study. Read-plus-rules exceeds both equal-cap neural comparators after correction; neither comparison with the larger-budget arms is significant after correction. The corrected comparisons and intervals, including null outcomes, are retained in Table 26. These are verification accuracies on a fixed record task, not attack-success rates or a new estimate of the primary three-state DC-ASR.

The longer-reasoned arm is numerically above the checker and rejects 127/128 invalid records; neither its comparison nor the thinking-mode comparison establishes a corrected advantage for read-plus-rules. This is an important boundary on the headline failure: sufficient test-time budget resolves most invalid-record errors for this checkpoint on these known schemas, although pair accuracy remains 113/128 and completed errors are not eliminated. All 54 unparsed short-reasoned responses, all five unparsed longer-reasoned responses and all 42 unparsed thinking responses hit their respective caps. Thus the short-budget contrast is substantially affected by truncation and cannot establish superior intrinsic reasoning. The checker abstains on 39 records; every covered record matches nominal gold. Retained read failures include transcriptions that preserve the temperature and uncertainty but omit the safe-range field, exposing a field-coverage limitation of this fixed transcription pipeline.

Table 27: **Measured token use and optimistic format bounds.** Means are per completed call and include reasoning tokens in the thinking arm. The bound credits every unparsed, non-runtime-error answer as correct and keeps parsed wrong answers wrong; it does not change any primary score.

| Arm | Mean input | Mean output | Optimistic pair bound |
|---|---|---|---|
| Direct | 496.2 | 20.4 | 27/128 |
| Reasoned | 517.2 | 265.4 | 122/128 |
| Read + rules | 463.8 | 52.4 | 128/128 |
| Reasoned, longer | 517.2 | 371.4 | 118/128 |
| Thinking | 515.2 | 1688.0 | 127/128 |

**Scope and remaining uncertainty.** Table 27 reports token use and optimistic format bounds; cap hits and parse failures are separate from correct judgments. The pre-specified optimistic format bounds cannot predict performance under an unlimited output budget or alternative sampling policy. The checker uses authored schema-specific knowledge, which is not equivalent to an unaided neural judgment. This single larger open-checkpoint study does not establish closed-model coverage, unseen-schema transfer, independently validated gold, or a general-purpose defense. It preserves the original baseline null results and the earlier unsuccessful neural read-then-verify confirmation.

## N Registered Prompt-Structure and Criterion Diagnostic

**Scope and provenance.** On 18 September 2026 we reran Qwen3.5-27B on the same 128 valid/invalid record pairs spanning all eight registered constraint families. The studies below use fresh deterministic calls rather than reusing historical answers. Model-file hashes, image bytes, prompts, sampling configuration, and freeze times are retained in the experiment archive. These endpoints are pair accuracy and record-level judgments, not the historical DC-ASR/EOR populations.

### N.1 Prompt structure and output budget

Table 28 compares four arms: a direct final-answer prompt, a structured verification prompt, and free calculation with 384- or 2048-token output caps. All missing or unparsed answers count wrong in pair accuracy. Completion counts are retained because heavy truncation makes explicit false-acceptance counts interval-censored in the direct arm.

Table 28: **Fresh Qwen3.5-27B prompt-structure diagnostic on 128 pairs.** Pair requires both valid and invalid records to be correct. False accept counts completed invalid-record answers choosing the consistent label. Unresolved counts incomplete or unparsed outputs among 256 records. **Bold** marks the best displayed value in each accuracy/error column; mean output is not ranked.

| Condition | Pair /128 | Valid /128 | Invalid /128 | False accept /128 | Unresolved /256 | Mean output tokens/record |
|---|---|---|---|---|---|---|
| Direct, cap 128 | 13 | 33 | 34 | 2 | 184 | 120.2 |
| Structured, cap 128 | **115** | **124** | 119 | 9 | **0** | 68.3 |
| Free calculation, cap 384 | 78 | 99 | 98 | **1** | 52 | 264.7 |
| Free calculation, cap 2048 | 107 | 110 | **123** | **1** | 9 | 395.8 |

Structured prompting recovers 102 pairs relative to the direct arm with no pair regressions; the registered exact paired test remains significant after three-way Holm correction ($p_H = 1.18 \times 10^{-30}$). This contrast primarily supports improved completion and overall correctness under the short cap. The direct arm has 184/256 unresolved outputs, mostly because the model begins an explanation despite the requested final-only format and exhausts the cap. For invalid records, two explicit direct false acceptances are observed and 92 invalid responses are unresolved, so the compatible false-acceptance count lies in $[2, 94]/128$. The structured arm has 9/128 explicit false acceptances; the registered structured-versus-direct false-acceptance contrast is not significant after correction ($p_H = 0.0654$).

Increasing free-calculation budget from 384 to 2048 tokens changes pair accuracy from 78/128 to 107/128, with 29 pair recoveries and no regressions ($p_H = 7.45 \times 10^{-9}$). At record level, 36 previously truncated responses recover to correct. All seven responses that were already complete but wrong under the shorter cap remain wrong. Thus the observed gain is dominated by completion recovery; it does not establish that longer reasoning repairs completed rule errors in general. This 78/128-to-107/128 prompt-structure study is distinct from the separately registered 76/128-to-113/128 larger-model study in Appendix M.

### N.2 Criterion timing and SMT-audited fidelity

A second study compares one-shot structured verification, a two-stage criterion-first procedure, a two-stage post-hoc self-check, and an oracle criterion supplied from the registered operational rule. All groups have a total 256-token output cap; the two-stage arms use 128+128. Criterion-first stage one has no access to the image, candidate numeric values, sample identifier, or gold label; its output is passed verbatim to stage two. The oracle arm supplies a correct external rule but still relies on the model to read and apply it.

Table 29: **Fresh criterion-timing diagnostic.** Pair accuracy requires both record states correct. Two-stage methods use twice as many model calls where noted; equal output caps do not imply equal compute. **Bold** marks the best displayed value in each accuracy/error column; call count is not ranked.

| Condition | Pair /128 | Valid /128 | Invalid /128 | False accept /128 | Final unresolved /256 | Calls |
|---|---|---|---|---|---|---|
| One-shot structured, 256 | 109 | 120 | **116** | **12** | **0** | 256 |
| Criterion first, 128+128 | 73 | 81 | 104 | 24 | 0 | 512 |
| Post-hoc self-check, 128+128 | **111** | **122** | **116** | **12** | **0** | 512 |
| External reference criterion, 256 | 83 | 103 | 106 | 21 | 1 | 256 |

In Table 29, relative to one-shot structured verification, criterion-first has two pair recoveries and 38 regressions ($p_H = 2.99 \times 10^{-9}$); relative to post-hoc self-check it has one recovery and 39 regressions ($p_H = 2.24 \times 10^{-10}$). The oracle-rule arm has 10 recoveries and 36 regressions relative to one-shot structured verification ($p_H = 0.000156$). The +2-pair difference between post-hoc self-check and one-shot verification was not a registered primary contrast.

**SMT audit of generated criteria.** The 256 blind criterion calls reduce to 12 distinct criterion strings. We parse each mathematical expression and ask Z3 (De Moura & Bjørner, 2008) whether there exists an input in the registered numerical domain on which the generated criterion and reference operational rule disagree. Seventy-six calls return UNSAT (equivalent over the specified domain) and 180 return SAT; there are no UNKNOWN or unparsed cases. At the unique-string level, five of 12 criteria are equivalent and seven are non-equivalent. All SAT counterexamples are replayed against the parsed expressions.

Global non-equivalence need not imply a wrong verdict on the current record. Executing each generated criterion on the registered fields gives the decomposition in Table 30. Of the 198 criteria that imply the current item's gold verdict, 15 final answers are nevertheless wrong. Of the 58 criteria that imply the wrong current-item verdict, 56 final answers are wrong. Among the narrower 76 calls whose criteria are globally equivalent to the reference specification, three final decisions are wrong.

Table 30: **Generated-criterion verdict versus final decision on the current record.** This is a post-hoc descriptive decomposition, not a randomized intervention.

| Generated criterion on current fields | Final correct | Final wrong |
|---|---|---|
| Agrees with gold (198) | 183 | 15 |
| Disagrees with gold (58) | 2 | 56 |

Observed non-equivalent rules include reversing the geometric inequality, replacing temporal precedence with a near-simultaneity tolerance, and reversing a unit-conversion multiplication into division. Oracle-rule failures show a different problem: supplying the correct text does not force faithful execution. These observations motivate separate auditing of rule text and rule execution. The SMT audit checks equivalence of specified mathematical expressions over explicit numeric domains; it does not verify the full checker, data generator, renderer, image semantics, or VLM.

# O Registered Replication on Four Held-Out Schemas

## O.1 Frozen scope and formal obligations

This prospective study freezes 128 valid/invalid one-field pairs, with 32 pairs each for average speed, steady direct-current power, classical translational kinetic energy, and average density. The schemas were not used in prior ContraLedger schema development; these familiar formulas are not claimed to be absent from model pretraining. Scene IDs exclude those of the original 128-pair criterion study, not every historical scene pool. We reuse the deterministic flat carrier with ordinary decimal formatting and a neutral synthetic bench record. The public question explicitly asks about the hypothetical record, not whether the photograph establishes its measurements. Records, rendering, prompts, metrics, and stopping rules were frozen before inference.

For each schema, let $q_f(x)$ be the exact derived quantity:

$$q_{\text{speed}} = d/t, \qquad q_{\text{power}} = VI, \qquad q_{\text{energy}} = mv^2/2, \qquad q_{\text{density}} = m/V.$$

The validity specification is

$$\phi_f(x, y, u) \equiv |y - q_f(x)| \le u. \tag{4}$$

The equivalence domain has strictly positive inputs, nonnegative reported output $y$, and $u \ge 0$; equality is valid. Inputs are exact and $u$ is an absolute output tolerance. Each schema crosses four valid residuals $r \in \{0, u, -u, u/2\}$, four invalid margins $\delta \in \{0.001, 0.01, 0.1, 1\}$, and two directions $s \in \{-1, 1\}$, with $u = 0.01$, giving 32 pairs. Every other field and the answer map stay fixed within a pair.

Table 31: **SMT checks for five formal generator/checker obligations per held-out schema.** UNSAT means no counterexample exists within the declared exact-arithmetic domain for valid-generator correctness, invalid-generator correctness, supported-domain membership, a single changed field, or numeric checker-predicate equivalence.

| Schema | Valid gen. | Invalid gen. | Domain | One field | Predicate |
|---|---|---|---|---|---|
| Speed | UNSAT | UNSAT | UNSAT | UNSAT | UNSAT |
| Power | UNSAT | UNSAT | UNSAT | UNSAT | UNSAT |
| Kinetic Energy | UNSAT | UNSAT | UNSAT | UNSAT | UNSAT |
| Density | UNSAT | UNSAT | UNSAT | UNSAT | UNSAT |

All 20 searches in Table 31 return UNSAT; generator preconditions and all 256 concrete labels are independently replayed. These are conditional arithmetic proofs, not proofs of the entire sampler, parser, renderer, Python runtime, or VLM. Malformed, missing, or unsupported fields cause the executable checker to abstain and count wrong.

### O.2 Five prompting and checking conditions

We evaluate Qwen3-VL-8B-Instruct and Qwen3.5-27B on identical images, public schema descriptions, assumptions, and within-pair A/B/C semantic maps. Each run uses temperature zero, top-$p = 1$, thinking disabled, and the frozen primary seed. All final model stages have a 256-token output cap.

**Direct.** Return only a terminal `FINAL:` answer.

**Structured.** Return a `FIELDS:` line, one compact `CHECK:` calculation, and a terminal `FINAL:` answer.

**Criterion-first.** Before seeing the image, values, item identifier, or gold label, generate a Boolean criterion from the public schema in a text-only call capped at 128 tokens; pass it verbatim to a new image-decision call using the Structured format.

**Oracle criterion.** Supply the reference expression and request the same Structured final format; the model must still read and apply the rule.

**Read + executable checker.** Transcribe the numeric fields without calculation, correction, or judgment, then execute the authored exact-rational checker on those actual transcribed fields. Gold fields are not substituted.

Criterion-first has an additional call, so equal final caps do not imply equal total compute. The parser requires one terminal `FINAL: A/B/C` marker; every missing, unparsed, runtime-error, or abstaining answer remains in the denominator and counts wrong. Pair accuracy requires both twins correct. False acceptance counts explicit CONSISTENT answers among all 128 invalid records. For a generated expression $c$, SMT tests $D_f \wedge \mathrm{XOR}(c, \phi_f)$; no model-generated code is executed.

### O.3 Primary outcomes and registered comparisons

All 5,120 planned model calls complete with no runtime errors. Each checkpoint uses 2,560 calls: 1,536 for the five primary arms and 1,024 for two additional criterion-first repetitions. Repeats are reported separately from primary accuracy. Table 32 reports primary outcomes and unresolved counts; Table 33 reports the registered paired comparisons.

Table 32: **Held-out-schema evaluation on 128 valid/invalid pairs per checkpoint.** Pair requires both twins correct; Valid, Invalid, and false acceptance (FA) use denominators of 128. **Bold** marks the best Pair/Valid/Invalid value within each model block and the lowest observed FA among complete methods; ties are all bold. Bracketed FA ranges are worst-case bounds over unresolved invalid-record answers. Missing counts unparsed answers and checker abstentions out of 256 records. Read+checker uses an authored schema-specific exact-rational checker, not model-generated code.

| Model | Method | Pair | Valid | Invalid | FA | Missing |
|---|---|---|---|---|---|---|
| Qwen3-VL-8B | Direct | 5 | 109 | 17 | 111 | 0 |
| | Structured | 62 | 113 | 75 | 53 | 0 |
| | Criterion-first | 114 | 125 | 116 | 12 | 0 |
| | Oracle criterion | 113 | 123 | 118 | 10 | 0 |
| | Read + checker | **126** | **127** | **127** | **1** | 0 |
| Qwen3.5-27B | Direct | 105 | 112 | 114 | 0 $[0, 14]$ | 25 |
| | Structured | 116 | 121 | 123 | 5 | 0 |
| | Criterion-first | 102 | 121 | 109 | 19 | 0 |
| | Oracle criterion | 111 | 120 | 119 | 9 | 0 |
| | Read + checker | **127** | **127** | **128** | **0** | 0 |

Table 33: **Prespecified pair-accuracy comparisons against Direct.** Recovered and regressed are discordant pairs out of 128. Exact two-sided paired tests use Holm correction across all eight comparisons.

| Model | Method | Recovered | Regressed | $p_H$ |
|---|---|---|---|---|
| Qwen3-VL-8B | Structured | 59 | 2 | $8.21 \times 10^{-15}$ |
| | Criterion-first | 109 | 0 | $2.16 \times 10^{-32}$ |
| | Oracle criterion | 108 | 0 | $3.70 \times 10^{-32}$ |
| | Read + checker | 121 | 0 | $6.02 \times 10^{-36}$ |
| Qwen3.5-27B | Structured | 18 | 7 | 0.130 |
| | Criterion-first | 17 | 20 | 0.743 |
| | Oracle criterion | 18 | 12 | 0.723 |
| | Read + checker | 22 | 0 | $1.91 \times 10^{-6}$ |

For Qwen3-VL-8B, all four registered pair-accuracy contrasts against Direct survive Holm correction. For Qwen3.5-27B, only Read + checker does. Thus the study does not establish a universal advantage for criterion-first prompting. Qwen3.5-27B Direct has 25 unresolved answers, including 14 invalid records; its zero explicit false acceptances have compatible bounds $[0, 14]/128$, not evidence of perfect rejection.

All 256 primary criteria per checkpoint return UNSAT for the equivalence query, reducing to four distinct family/expression outputs each. Qwen3-VL-8B has 15 completed wrong decisions under equivalent criteria: 12 false acceptances and three valid-record errors. Ten of these also have exact same-response numeric fields, all on invalid records. Qwen3.5-27B has 26 such decisions: 19 false acceptances and seven valid-record errors. Its 20 exact-field errors consist of 19 invalid and one valid record. There are no incomplete final decisions in these primary criterion-first arms. The independent Read call is exact on 13/15 and 26/26 error records, respectively; this is a separate probe from the stricter same-response field requirement.

### O.4 Persistence, concrete errors, and interpretation

The frozen persistence criterion requires at least five distinct pair IDs spanning at least two schemas with the same record exhibiting a completed wrong decision under an equivalent criterion in at least two of three criterion-first runs. Primary and repeat call seeds are 20260918, 20260919, and 20260920, with temperature zero throughout. Qwen3-VL-8B meets the criterion on 14 records from 14 pairs across speed, power, and density; Qwen3.5-27B meets it on 26 records from 26 pairs across all four schemas. The additional same-response exact-field requirement persists on 10 and 20

records, respectively. These are repeated calls on fixed items, not independent samples or training-seed replications.

Two logged density cases explicitly demonstrate *arithmetic-execution slips* without changing the expressed rule or fields. For Qwen3-VL-8B, one response uses the equivalent criterion $|\rho - m/V| \leq u$ and exact fields $m = 64$, $V = 16$, $\rho = 2.99$, $u = 0.01$, yet claims a residual of $0.01$ and selects CONSISTENT; the actual residual is $1.01$. Qwen3.5-27B makes the same residual error on a different density item. In these examples the terminal label follows the model's incorrect intermediate residual, so they should not be described as pure verdict detachments. A subsequent exhaustive, post-hoc replay of all 41 displayed traces finds incorrect claimed arithmetic residuals in 15/15 Qwen3-VL and 24/26 Qwen3.5 errors, including all 10 and 20 exact-field errors. Ten traces contain different field values and one omits the required `FIELDS:` prefix; these categories can overlap arithmetic errors. No terminal answer contradicts the assessment expressed by its displayed `CHECK:` line. The two arithmetically correct traces use incorrect transcribed inputs. This is executable annotation of visible text, not independent human adjudication or identification of hidden computations; a missing format prefix is not evidence of a reading failure. These are not SMT counterexamples, since the generated criteria are equivalent.

Independent audits recheck frozen file hashes, request prompts, the strict parser, exact-rational labels, generator preconditions, and changed-pixel containment. The generalizable lesson supported here is an evaluation distinction: expressing a criterion equivalent to the specification does not ensure a conforming final verdict, even when the same response contains correct numeric fields. The initial replication reduces dependence on the original eight schemas and one checkpoint. Appendix P.1 adds a frozen InternVL2 extension; neither study establishes arbitrary-schema generalization, natural captured-record transfer, or a uniquely visual mechanism.

# P Frozen Non-Qwen and Generated-Program Extensions

## P.1 Non-Qwen replication on the same frozen pairs

A separately frozen extension evaluates InternVL2-8B on all 128 pairs from Appendix O. This tests another model family without selecting items by prior outcomes. It is not a new image or schema sample. JPEG bytes, public prompts, A/B/C maps, parser, five primary arms, two criterion repetitions, and 128/256-token stage caps are unchanged. InternVL uses its native 448-pixel dynamic tiling (1–12 patches plus the configured thumbnail), so equal input files and token caps do not imply identical image preprocessing or compute. No prompt was tuned on the new responses.

All 2,560 registered calls complete without runtime errors or truncation. Table 34 reports every primary arm, including unresolved outputs. Direct largely accepts records irrespective of validity; structured and oracle prompting accept all 128 invalid records. Read+checker improves pair accuracy to 26/128, but 179/256 transcriptions are unparsed or cause abstention under the frozen checker protocol. These count wrong. Thus the near-complete checker recovery on the two Qwen checkpoints does not transfer unchanged to this model and output format.

Table 34: **InternVL2-8B on four held-out schemas.** Pair, Valid, Invalid, and FA are /128; unresolved outputs are /256. Brackets bound FA when unresolved invalid records are retained.

| Method | Pair | Valid | Invalid | FA | Unresolved |
|---|---|---|---|---|---|
| Direct | 2 | 126 | 2 | 126 | 0 |
| Structured | 0 | **128** | 0 | 128 | 0 |
| Criterion-first | 0 | 126 | 2 | 126 | 0 |
| Oracle criterion | 0 | **128** | 0 | 128 | 0 |
| Read + checker | **26** | 36 | **36** | **0** $[0, 92]$ | 179 |

Of 256 primary generated criteria, 244 are SMT-equivalent and 12 are inequivalent. The 244 equivalent criteria precede 122 completed wrong decisions, including 108 with exact same-response numeric fields. All incorrect criteria remain in the full-set accuracy denominator. Across the three criterion runs, 127 distinct pairs show the same-record equivalent-criterion/completed-error event

in at least two runs, spanning all four schemas; the exact-field event persists on 113 pairs. This meets the unchanged persistence threshold, without treating repeated calls as independent samples. It extends the observed conformance gap beyond Qwen, not to every model family or a unique latent mechanism.

Four new exact paired tests compare each intervention with Direct, with Holm correction within this extension. Read+checker recovers 26 and regresses two pairs ($p_H = 1.21 \times 10^{-5}$); each other arm recovers zero and regresses two ($p_H = 1$). The original eight-test Qwen family is unchanged. Independent replay verifies 660 frozen hashes, all 2,560 raw calls, all 256 labels, generator preconditions, pixel containment, and paired statistics. SMT replay checks 20 generator obligations and 768 criterion calls representing six distinct family/criterion outputs; inequivalent expressions retain their solver counterexamples.

### P.2 A RESTRICTED GENERATED-PROGRAM BASELINE

We implement a restricted PAL/PoT-style diagnostic (Gao et al., 2023; Chen et al., 2022), distinct from the authored Read+checker. One image call returns a JSON object containing transcribed variables and a generated Boolean expression. A generic restricted AST interpreter executes that expression with exact rational arithmetic; it contains no schema-specific validity rule and never executes unrestricted Python. Prompts supply the parent's public field descriptions and assumptions, without its oracle formula. This is a program-generation baseline, not a reproduction of either original system or a general symbolic defense.

The two original Qwen checkpoints each receive all 256 frozen records, with temperature zero, the original seed, thinking disabled, and a 256-token cap. Model-weight hashes and engine versions match the parent experiments. Call count and maximum output tokens match Direct and Read+checker; input length, realized tokens, and latency need not match. The protocol and four paired comparisons were frozen before inference. All 512 requests and raw responses are retained, including parsing, execution, and truncation failures; failures count wrong. No failed case is retried or removed.

Table 35 gives full-set outcomes. Qwen3-VL has 47 execution failures (39 unbound-variable and eight nonnumeric-operand errors); Qwen3.5 has none. Neither run has a transport error or truncation. Completion-token totals are 12,309 and 23,090. The smaller model's explicit FA count of one has bound $[1, 22]/128$ because 21 invalid records are unresolved.

Table 35: **Generated-program baseline on the same 128 pairs per checkpoint.** Pair, Valid, Invalid, and FA are /128; failures are /256. No authored formula is supplied to the model or interpreter.

| Model | Pair | Valid | Invalid | FA | Failures |
|---|---|---|---|---|---|
| Qwen3-VL-8B | 83 | 85 | 106 | 1 $[1, 22]$ | 47 |
| Qwen3.5-27B | **120** | **120** | **128** | **0** | **0** |

Table 36 compares the new arm with the frozen parent Direct and Read+checker outputs. Generated programs improve pair accuracy over Direct on both checkpoints, but remain below the authored checker; all four contrasts survive their prespecified Holm correction. Full request bodies are reconstructed from archived prompts and image hashes, and a second exact-rational interpreter independently reproduces all 512 scores and the paired statistics. The comparison supports a bounded benefit of external generated computation, while retaining model-dependent transcription and program-format failures.

Table 36: **Generated-program pair comparisons.** Recovered/regressed pairs favor/disfavor the new baseline; Holm correction covers these four new tests only.

| Model | Comparator | Recovered | Regressed | $p_H$ |
|---|---|---|---|---|
| Qwen3-VL-8B | Direct | 80 | 2 | $5.63 \times 10^{-21}$ |
| Qwen3-VL-8B | Read + checker | 0 | 43 | $6.82 \times 10^{-13}$ |
| Qwen3.5-27B | Direct | 18 | 3 | 0.0030 |
| Qwen3.5-27B | Read + checker | 0 | 7 | 0.0156 |

# Q Reviewer-Focused Secondary Analyses

This appendix uses already frozen data and adds no model calls. We retain only analyses that directly constrain interpretation of the headline results.

## Q.1 Exact Read is not sufficient on invalid records

The prospective matched-media diagnostic provides both an independent strict Read call and a direct scene decision for the same 128 valid/invalid pairs. Because pooled valid/invalid accuracy can cancel under strong directional response bias, we report the invalid-record stratum directly.

Table 37: **False acceptance on invalid records, stratified by exact Read.** These are full-set descriptive counts, not EOR: no source, valid-record, or Know gate is applied.

| Model | Exact-Read invalid $n$ | False accepts | Non-exact invalid $n$ | False accepts |
|---|---|---|---|---|
| Qwen2.5-VL-7B | 47 | 46 | 81 | 73 |
| Qwen3-VL-8B | 21 | 5 | 107 | 33 |

For Qwen2.5-VL-7B, 46/47 exact-read invalid records are still accepted as consistent; for Qwen3-VL-8B the corresponding count is 5/21. Exact transcription is therefore not sufficient for correct verification on these records, although this stratification is descriptive rather than a causal estimate of OCR's effect.

## Q.2 Operational assumption registry

The executable labels are conditional on explicit operational assumptions. New matched studies and the held-out-schema replication state those assumptions in the prompt; historical generic queries did not always do so.

Table 38: **Operational assumptions used by the prospective eight-family diagnostic.** Unit conversion has multiple unit-specific formulas under the same family.

| Family | Explicit operational assumption(s) |
|---|---|
| Capacity / conservation | The container starts empty; no liquid is lost except the recorded spill; final fill is at capacity. |
| Causal order | The stated initiating event is the direct cause and both timestamps use one synchronized timeline. |
| Geometry | Widths use the same direction; the object is rigid and neither rotates nor deforms. |
| Phase state | The sample is pure water at equilibrium at 1.00 atm, without supercooling or dissolved solutes. |
| Probability ledger | A and B are the only outcomes and cannot occur together; probabilities are exact to 0.01. |
| Range / threshold | The measurement interval and safe interval use degrees Celsius; the complete uncertainty interval must lie inside the inclusive safe range. |
| Temporal ledger | All times are same-day 24-hour local times with no date rollover. |
| Unit conversion | Both fields describe the same quantity; use F=(9/5)C+32. |
| Unit conversion | Both fields describe the same quantity; use lb=2.2046226218*kg. |
| Unit conversion | Both fields describe the same quantity; use mi=0.621371*km. |

Independent human adjudication of these conventions, naturalness, and nominal physical interpretations remains open; we therefore do not report a human-validation rate that was not collected.

### Q.3 CONSOLIDATED EVIDENCE MAP

Table 39 summarizes the strongest completed evidence for reviewer-relevant alternative explanations and the remaining interpretation boundaries.

Table 39: **Consolidated evidence map.** Each row pairs an alternative explanation with the strongest completed evidence and the remaining boundary.

| Question / alternative | Completed evidence | Supported interpretation and remaining boundary |
|---|---|---|
| Is the effect specific to ContraLedger packaging? | 256 fresh scenes with five complete-information presentations; none of 20 Holm-corrected ContraLedger-versus-baseline contrasts establishes superiority. | Internally invalid complete records can induce the same broad verification failure across presentations. This does not establish equivalence or a uniquely strong carrier. |
| Is OCR failure sufficient to explain the errors? | Historical EOR retains 191/200 false acceptances after exact Read and separate Know success; held-out-schema criterion runs contain 10/15 and 20/26 wrong decisions with exact same-response numeric fields; Table 37 gives a full-set invalid-record stratification. | Observable misreading is not necessary for every failure. Independent Read probes and same-response fields do not reveal hidden computation, and OCR can still contribute elsewhere. |
| Do explicit prompts or more output budget solve the problem? | On Qwen3.5-27B, structured 128-token prompting raises pair accuracy from 13/128 to 115/128 while removing unresolved outputs; free calculation rises from 78/128 to 107/128 at 2048 tokens. | Short-budget errors include a large completion component; increased budget does not establish an intrinsic inability to verify. |
| Does expressing or supplying the correct rule guarantee the final verdict? | In the four held-out schemas, all 256 primary criterion calls per Qwen checkpoint are SMT-equivalent, yet Qwen3-VL-8B and Qwen3.5-27B make 15 and 26 completed wrong decisions; 10 and 20 also copy all numeric fields correctly. | Specification fidelity and downstream execution/final conformance are distinct measurable stages; many residuals are arithmetic slips rather than terminal-label detachments. |
| Can executable checking close the gap? | Read+checker obtains 126/128 and 127/128 Qwen pairs, but 26/128 on InternVL2; generated programs reach 83/128 and 120/128 Qwen pairs (Appendix P). | Authored checking supplies rule knowledge; generated programs must express it. Both retain transcription and format failures and do not establish a general defense. |
| Does the study establish natural or physical-world risk? | Transfer is shown across COCO/VOC sources, multiple synthetic carriers, and twelve total registered schema families. | Natural captured records, independently human-validated assumptions, broader model families, and physical deployment transfer remain open. |