PREMISE OVER PIXELS: DIAGNOSING VISUAL SELF-CORRECTION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-step reasoning requires a model to overturn its own earlier errors before they propagate. In vision-language models (VLMs), the same image remains available even after the model has stated an incorrect visual premise. We introduce *false-premise replay* to test whether VLMs can use that image to revise their own errors. From a failed reasoning trace, we extract one answer-relevant visual claim contradicted by the image, discard the remaining reasoning and final answer, and replay the claim with the original image and question. Across three model-reviewed evaluation sets, replay lowers accuracy relative to a fresh query by 25.8–31.0 percentage points, while substituting a corrected premise raises it. Image interventions show continued visual influence, but models do not reliably identify the image-supported premise. On InternVL3.5-8B, a false premise also reduces how often sampled answers include a correct one and impairs selection from fixed candidate pools. We consolidate 217 premise pairs into a shared evaluation set and compare correction baselines on five distinct models. On this set, false-premise replay lowers accuracy by 16.6 points even for a 975B-parameter model with reasoning enabled. On all five models, same-model visual-verification feedback improves accuracy over self-reflection with false premises, but lowers it when the earlier premise is correct. The findings distinguish obtaining corrective information from using it, and show why measuring repair under false-premise replay alone can overstate correction quality.

1 INTRODUCTION

Prior work addresses errors in large language model (LLM) reasoning in different ways. ReAct interleaves reasoning with actions and observations from an external environment (Yao et al., 2023), while Reflexion uses feedback from previous trials to guide subsequent attempts by language agents (Shinn et al., 2023). Process reward models evaluate intermediate reasoning steps and use these evaluations to select from a pool of sampled solutions (Lightman et al., 2024). Other approaches explicitly revise an earlier response using feedback obtained either from external tools or from the model itself (Gou et al., 2024; Madaan et al., 2023). Within this revision setting, the source of feedback matters. Correction is effective when reliable external feedback is available, whereas intrinsic self-correction based only on the model's own feedback remains unreliable in general reasoning tasks (Huang et al., 2024; Kamoi et al., 2024). Such feedback is text, such as a critique, tool output, or test result.

Vision-language models introduce a different correction setting. A model can state a reading that disagrees with an image while that image remains available for subsequent reasoning. We call a textual statement describing image content a *visual claim*. When the claim is carried forward as a basis for reasoning, we call it a *visual premise*, such as a stated bar value used in a calculation. Can the model reconsider that reading against the same image? In Figure 1, the model answers correctly on its own but repeats the wrong answer when its earlier statement appears in a prior assistant turn.

Prior work diagnoses several components of this problem. One line of work studies the *source* of visual errors. VLMs can hallucinate objects that are not present in the image (Li et al., 2023), while HallusionBench distinguishes errors arising from visual misperception from those driven by incorrect language-side priors (Guan et al., 2024). These results show that keeping the image available does not ensure that the model's textual interpretation is determined by the image.

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.

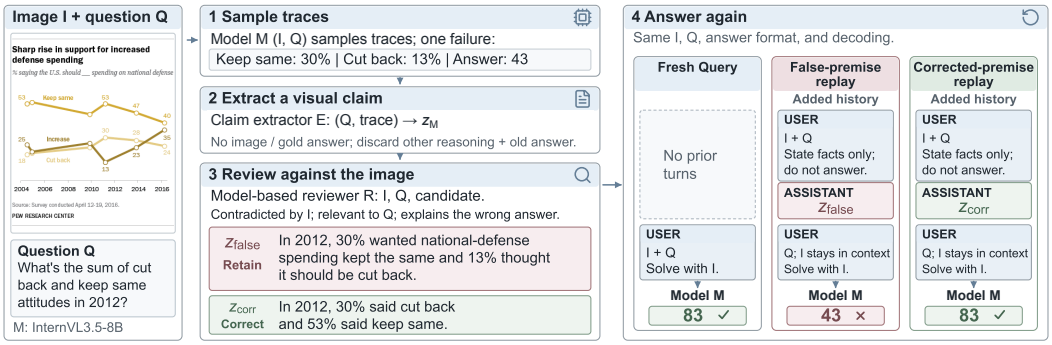


Figure 1: **False-premise construction and replay.** A claim extractor that sees only the question and trace isolates a visual claim. A model-based reviewer checks it against the image and supplies a corrected premise. Only the selected premise enters an earlier assistant turn. With the image, question, and decoding fixed, fresh, false-premise, and corrected-premise replay give 83, 43, and 83.

A second line of work studies competition between image and language-side priors during generation. Visual contrastive decoding contrasts predictions under the original and distorted images to reduce over-reliance on language priors embedded in the LLM decoder (Leng et al., 2024). M3ID instead increases the influence of the image relative to the language prior during decoding (Favero et al., 2024). Counterfactual image experiments further show that model predictions initially reflect memorized priors but shift towards the image in mid-to-late layers (Golovanevsky et al., 2025).

A third setting arises once visual information has entered the textual reasoning history. The model can then reason from its own earlier interpretation rather than directly from the image. MMHalSnowball shows that prior generated hallucinations can mislead later image-based responses even when the relevant visual information remains available (Zhong et al., 2024). VisualSwap retains earlier reasoning while changing answer-relevant visual content and shows that VLMs can fail to update their responses (Shi et al., 2026). Keeping the current image fixed while ablating components of the prior reasoning trace further shows that evidence-bearing spans can carry the previous answer (Fan et al., 2026). These studies show prior-text influence under different image and trace interventions.

Like the ReAct and Reflexion agents above, a VLM that reasons in steps or retries a task conditions on its earlier outputs, so a wrong visual premise can remain in context while the image that contradicts it is still available. A wrong final answer, or feedback that flags it, does not reveal whether the model misread the image or failed to reconsider an earlier misreading, nor which visual statement should be checked. Our *false-premise replay* evaluates one such statement directly. It inserts an answer-relevant visual claim from the model's failed trace into an earlier assistant turn. The image and question are unchanged, and the remaining reasoning and old answer are removed. Review marks the claim false and supplies its correction, so unlike the trace and image interventions above, each false premise has a matched corrected premise under the same image. The paired answers show whether the model overturns an error and whether it keeps a correct statement.

A *claim extractor* reads only the question and a failed, sampled trace, without the image or dataset answer, and extracts one visual claim. A *model-based reviewer* checks the claim's truth, relevance, and relation to the wrong answer against the image, and supplies a corrected version of the same claim. During replay, the evaluated model receives neither the review's *correctness label*, which states whether a premise is correct, nor the other premise of the pair.

Corrected-premise replay shows whether the answer is reachable once the premise is corrected. Testing only false premises would favor any method that simply ignores earlier statements, since ignoring a false premise looks like correcting it. Such a method would also lose the corrected premise. A fresh query, with no earlier turn, and a neutral replay, which inserts a generic statement instead of the premise, test whether the accuracy change is specific to the false premise.

Relative to a fresh query, false-premise replay lowers accuracy by 25.8–31.0 percentage points across three model-reviewed sets. The effect persists on pairs confirmed by human review, and the false–corrected gap persists with short original reasoning prefixes. Explicitly asking the model

to re-check the image leaves primary-set accuracy at 10.5%, whereas substituting the correction raises it to 71.0%. The image still influences the answer, but the model does not reliably select the image-supported premise. On 82 InternVL extension pairs, replacing length-matched neutral text with the false premise lowers the number of pairs with a correct answer among four samples from 41 to 18, and impairs selection among fixed candidates.

On a shared set of 217 premise pairs, false-premise replay lowers accuracy for all six evaluated models, including a 975B-parameter model with reasoning enabled. On the five models with complete correction comparisons, visual-verification feedback improves accuracy over self-reflection by 7.83–22.12 points with false premises, but lowers it when the earlier premise is correct. A separate study on InternVL with constructed premises shows that supplying corrected facts helps more than identifying the error, even while the false premise remains in context. Our contributions are:

- **A diagnostic protocol and a shared evaluation set of 217 premise pairs** that isolates one model-generated visual error, pairs it with a corrected version under the original image, and supports cross-model evaluation.
- **A behavioral diagnosis** that separates continued image influence from premise selection and, on InternVL, missing correct answers among samples from failure to select them.
- **A paired evaluation of correction baselines** that measures repair and retention across both premise conditions, including same-model feedback, self-reflection, voting, and history reset.

2 False-Premise Replay

2.1 Constructing and Replaying a Premise Pair

From a failed trace to one premise. A source item contains an image I, question Q, and dataset answer a^*. The evaluated VLM M samples a reasoning trace τ ending in an answer. For an incorrect trace, a claim extractor E identifies a minimal visual claim:

$$\tau \sim M(\cdot \mid I, Q), \qquad a_{\text{false}} = \text{Ans}(\tau) \not\equiv a^*, \qquad z_M = E(Q, \tau).$$

Here $\not\equiv$ denotes semantic non-equivalence. The extractor sees neither the image nor the dataset answer. Both the extractor and the reviewer are prompted gpt-5.6-terra models (Appendix A). A model-based reviewer R retains z_M as a *false premise* z_{false} only if the image contradicts it, it is relevant to the question, and it explains the wrong answer. R also provides z_{corr}, a corrected version of the same claim that does not state the final answer. We keep only the selected claim and discard the remaining reasoning and old answer. Each accepted model–image–question *episode* contains one premise pair, fixed before replay or correction evaluation.

Replay and comparisons. Review judges that z_{false} contradicts the image. Replay asks a different question: whether the model, with the image still present, overturns this claim or carries it forward. The model is not told that the claim is false. We prepare the three input turns in Table 1, placing z_{false} in the assistant turn, then ask M to generate a new answer $\hat{a}_{\text{false}}$. This answer may match or differ from the old wrong answer a_{false}, and the copy rate defined below measures how often they match.

Fresh query asks the question with the image and no earlier turn. It uses the construction image and question, but deterministic decoding instead of stochastic trace sampling. Inclusion depends only on the failed sampled trace and its review. *Corrected-premise replay* replaces only z_{false} with z_{corr}. To test whether the accuracy change is specific to the false premise, we replace z_{false} in the replay turn with alternative content and query the VLM again. A *neutral replay* inserts a generic statement that provides no information relevant to the question. A *task-irrelevant replay* instead inserts text matched approximately in form and length to the false premise but unrelated to the information needed to answer the question. A *re-check instruction* retains the false premise and asks the model to revisit the image. The image, question, answer format, and deterministic decoding are fixed across comparisons.

2.2 Evaluation Sets

We draw chart and document questions from ChartQA (Masry et al., 2022) and InfographicVQA (Mathew et al., 2022). Geometry and natural-image questions come from Geometry3K (Lu

Table 1: **Replay input and the new answer.** Rows 1–3 are supplied together before the model answers. Row 2 is inserted from an earlier sampled trace. Exact prompts for the shared evaluation set are in Appendix H.

Turn	False-premise replay
1. User	Image I, question Q, and a request for relevant visual facts without calculating or answering.
2. Assistant	z_{false}, taken from the failed trace and accepted by review. No other reasoning or old answer.
3. User	Solve Q using the current image and the specified answer format.
4. Output	M generates $\hat{a}_{\text{false}}$ after receiving rows 1–3.
Fresh query	Only I and the row-3 request, with no earlier turns. M generates $\hat{a}_{\text{fresh}}$.
Corrected replay	Rows 1–3 with z_{corr} in row 2. M generates $\hat{a}_{\text{corr}}$.
Neutral replay	Rows 1–3 with a generic statement in row 2 that provides no information relevant to the question.

et al., 2021) and GQA (Hudson & Manning, 2019). InternVL3.5-8B-HF (Wang et al., 2025) and Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) each sample four traces for each of 1,070 questions, giving 2,140 model–question combinations and 8,560 traces. Extraction and review retain 72 InternVL and 52 Qwen-30B episodes, which form the primary evaluation set of 124 episodes over 101 images.

The *false-premise selection rate* is the fraction of candidate questions that yield at least one accepted premise for a given model. It is 72/1,070 (6.7%) for InternVL and 52/1,070 (4.9%) for Qwen-30B. These rates measure the yield of the four-trace construction procedure. The subsequent replay effect instead measures persistence conditional on accepting a premise. Two additional model-specific sets contain 35 Qwen-30B and 58 InternVL episodes. Neither overlaps the primary set in accepted images, but the two share 18 source questions. Appendix A details their construction and model review.

A shared evaluation set for comparing correction procedures. We consolidate 217 pairs over 176 images and evaluate the same pairs on six models (Section 4). The set contains 130 InternVL-origin and 87 Qwen-30B-origin pairs, spanning 27 chart, 28 document, 97 geometry, and 65 natural-image pairs (Appendix F). Every evaluated model answers the same pairs, including pairs built from another model's errors. The set uses a standardized final-line answer format, so its results are reported separately from Figure 2. Reference answers, alternative premises, and correctness labels stay outside the model inputs, and new models can be evaluated without constructing new premises.

Human review. Two reviewers assessed premise truth for 120 pairs over 97 images, stratified by domain and construction set (Appendix A.1, with examples in Appendix B.2). The reviewers agreed on 203/240 statements (84.58%). Both supported 79 pairs, agreed on mismatches in 7, and disagreed in 34. For the 34 disagreements, a third reviewer checked AI-drafted judgments one by one. Two policies for combining this follow-up with the original reviews define additional sensitivity subsets (Appendix A.2).

2.3 Paired Evaluation Metrics

For condition h and N episodes, let $\hat{a}_{i,h}$ be the parsed answer and $c_{i,h} = \mathbf{1}[\hat{a}_{i,h} \equiv a_i^\star]$ indicate correctness. Unparsed answers count as incorrect. We measure accuracy A_h, the fresh-minus-false *replay gap*, and the *false-answer copy rate*:

$$A_h = \frac{1}{N}\sum_i c_{i,h}, \qquad \Delta_{\text{replay}} = A_{\text{fresh}} - A_{\text{false}}, \qquad C_{\text{false}} = \frac{1}{N}\sum_i \mathbf{1}[\hat{a}_{i,\text{false}} \equiv a_{\text{false},i}].$$

The copy rate compares the new answer with the old wrong answer, whereas accuracy compares it with the dataset reference. Corrected-premise accuracy A_{corr} measures performance with a corrected premise on all N episodes.

For a correction procedure S and a reference procedure S_0, a *repair* turns an incorrect answer of S_0 into a correct one, and a *harm* turns a correct answer of S_0 into an incorrect one. *Retention* is the share

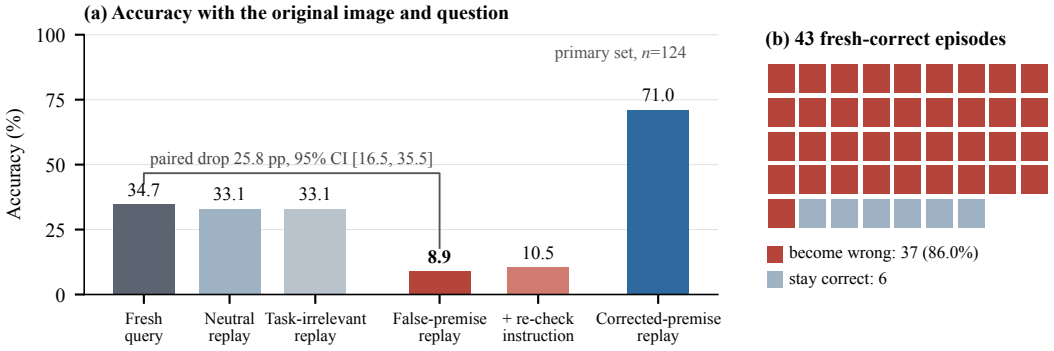


Figure 2: **False-premise replay degrades accuracy on the primary evaluation set.** (a) With the image, question, answer format, and decoding procedure held fixed, neutral and task-irrelevant replay leave accuracy close to fresh query, whereas replaying the false premise reduces accuracy from 34.7% to 8.9%. An explicit instruction to re-check the image provides little recovery, while replacing the false premise with its reviewed correction raises accuracy to 71.0%. (b) Among the 43 episodes answered correctly under fresh query, 37 become incorrect under false-premise replay. The primary set pools InternVL3.5-8B and Qwen3-VL-30B-A3B episodes on each model's own errors.

of correct answers of S_0 that remain correct under S, so every harm lowers retention. False-premise replay mainly tests repair, because the earlier premise supports a wrong answer. Corrected-premise replay mainly tests retention, because the earlier premise supports the right one. We report both accuracies and their equally weighted mean, *balanced accuracy*:

$$B(S) = \tfrac{1}{2}\big(A_{\text{false}}(S) + A_{\text{corr}}(S)\big), \qquad B(S) - B(S_0) = \frac{N_{\text{repair}} - N_{\text{harm}}}{2N},$$

where repairs and harms are counted across both replay conditions.

Paired bootstrap resampling keeps the compared conditions together and clusters shared episodes by source image. The notation $\pm$ denotes two bootstrap standard errors, distinct from percentile intervals labeled 95% CI. Later studies use the evaluation sets specified in Appendix A.

3 What Persists After a Visual Error?

3.1 A Single Premise Can Overturn a Correct Answer

False-premise replay lowers primary-set accuracy from 34.7% to 8.9% (Figure 2a). The two additional model-specific sets show comparable drops of 28.6$\pm$17.3 and 31.0$\pm$13.8 percentage points (Appendix Table 7). Of the 43 primary episodes answered correctly under fresh query, 37 become incorrect after replay (Figure 2b). Model review found that the image contradicts each premise, so a model that checks the premise against the image would keep its fresh answer. These models follow the premise instead. Corrected-premise replay reaches 71.0%, 62.9%, and 67.2% on the three sets.

The content of the assistant turn matters (Figure 2a). Neutral and task-irrelevant replay stay close to fresh query, so adding an earlier assistant turn alone does not explain the loss. An explicit instruction to re-check the image recovers little, far less than supplying the corrected premise (Appendix B). Replaying the full failed trace gives 1.1% accuracy on the 93 additional episodes, with 98.9% of answers repeating the old one, so it cannot show which statement the model failed to reconsider.

The premise contrast persists with short preceding reasoning. On 54 human-supported pairs over 44 images, we keep the original reasoning that preceded the selected claim, 8–71 words long, and follow it with either the false or the corrected premise. Corrected-premise accuracy still exceeds false-premise accuracy by 46.30 points on InternVL (95% CI [30.19, 62.26]) and 31.48 on Qwen3-VL-8B ([14.89, 47.46]). Intervals for the change from the single-premise gap include zero (Appendix J).

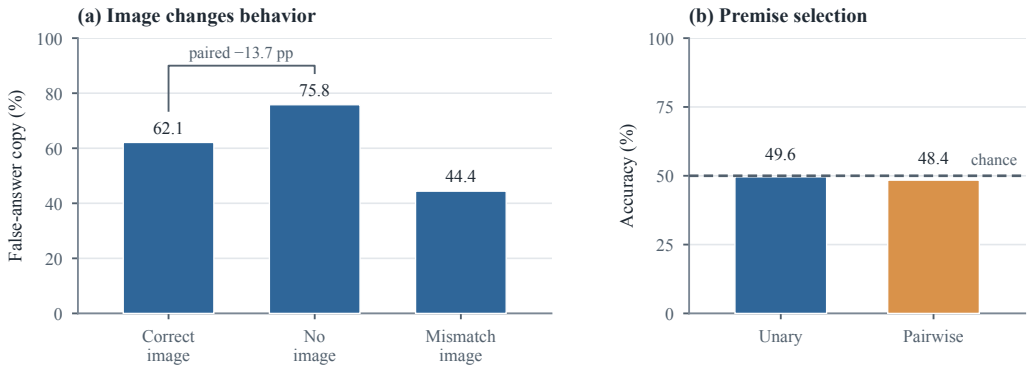


Figure 3: **Image dependence and premise selection measure different properties.** (a) False-answer copying changes when the image is removed or mismatched. (b) Forced premise-selection accuracy remains near chance despite high confidence normalized over the two candidates. Both panels use the primary set, which pools InternVL3.5-8B and Qwen3-VL-30B-A3B episodes.

3.2 Visual Influence Does Not Ensure Correct Premise Selection

The accuracy drop does not by itself show whether the model continues to use the image. If replay caused the model to depend only on the earlier premise, removing or replacing the image should have little effect. Figure 3a keeps the false premise fixed while varying the image. Without an image, the model reproduces the old wrong answer in 75.8% of primary episodes. Restoring the original image lowers this rate to 62.1%, and replacing it with a mismatched image lowers it further to 44.4%. The mismatched image may redirect the model toward another wrong answer. Reduced copying therefore shows visual influence.

We next ask whether this visual influence supports a reliable choice between the false premise and its correction. A unary probe compares the likelihoods of `Supported` and `Contradicted` for each premise separately. A pairwise probe compares the two premises with their positions exchanged. Unlike the image interventions, these probes ask which premise the image supports.

Unary and pairwise accuracy are 49.6% and 48.4%, despite mean two-candidate confidence of 91.9% and 88.3%. Relative to no image, the original image raises the `Supported` margin of both premise wordings by nearly the same amount, 4.134 log-probability units, so this shift cannot identify the correct premise. On separate mirrored visual contrasts, the same test formats reach 90.3–98.4% with images, so they can detect image-supported facts in clean contrasts (Appendix B).

The earlier premise also changes what the model states without supplied premise options. Across all 217 episodes, replay raises the share of responses containing the false premise from 52.5% to 82.5% (95% CI for the change [+22.37, +37.62]). The rate of responses stating only the corrected premise falls, and a re-check instruction lowers it further (Appendix B.1).

3.3 Prior Premises Affect Candidate Generation and Selection

We sample four answers per condition, then hold the candidate text fixed to distinguish missing correct answers from failure to select them. The comparison uses the 82 InternVL extension pairs of the shared evaluation set, held out from the 48 pairs used to develop the analysis. Length-matched neutral text has the premise's token count but no task information. With both user turns and the image fixed, replacing the false premise with this text raises coverage from 18/82 to 41/82, a 28.05-point difference (95% CI [17.07, 39.02]). Of 72 incorrect votes under false-premise replay, 64 have no correct parsed candidate and 8 fail to select one that is present. Voting cannot recover an answer absent from the sampled pool.

We then fix the four answers generated with length-matched neutral text and change only the premise in the assistant turn that precedes the selection prompt. The false premise lowers selection accuracy from 33.54% to 22.56%, a 10.98-point drop (95% CI [3.66, 18.90]), whereas the corrected premise raises it to 40.85%. First-token option scoring, which needs no well-formed response, preserves this

direction. For reference, majority voting over the same pool reaches 37.80%. Appendix C gives order controls, parsing, confidence changes, and the Qwen3-VL-30B-A3B results. On 44 constructed InternVL episodes, masking attention to the earlier turn in middle layers repairs false-premise answers but loses corrected ones (Appendix D), the tradeoff we measure next.

4 Can Standard Correction Preserve Useful Prior Facts?

Direct replay on the same inputs. We evaluate all 217 pairs on six models from four developers (Table 2), including Inkling-Small and Inkling from Thinking Machines Lab and Kimi-K2.6 from Moonshot AI, all hosted through Tinker (Thinking Machines Lab, 2026), and we also run both Inkling models with reasoning (Appendix F). For every model and setting and under both parsers, false-premise replay lowers accuracy relative to fresh query and corrected-premise replay raises it. With reasoning, false-premise replay still lowers accuracy by 16.6 points for each Inkling model, with 95% intervals excluding zero. The drop appears whether or not the model produced the premise. False-premise replay costs InternVL 31.0 points on pairs built from Qwen-30B errors and 20.8 points on its own, and no hosted model produced any pair (Appendix Table 13). For InternVL and both Qwen models, the same directions hold on the 79 pairs that both reviewers supported and on the 99 pairs supported under both follow-up policies, with 95% intervals excluding zero under the primary exact score (Appendix A.2).

Table 2: **Direct replay on identical premise pairs.** Correct answers out of 217 under a common answer format and the primary parser. Parentheses give the whitespace-tolerant count where it differs. Unparsed counts cover all 868 requests. †Hosted interface without reasoning. ‡Reasoning effort 0.9 with a 16,384-token cap.

	Without a visual premise		With a visual premise		
Model	Fresh	Neutral	False	Corrected	Unparsed
InternVL3.5-8B	84	82	30	143	2
Qwen3-VL-8B	71 (89)	67 (91)	27 (31)	112 (143)	179 (19)
Qwen3-VL-30B-A3B	83 (92)	96 (100)	24 (26)	132 (167)	127 (28)
Inkling-Small†	86 (100)	88 (90)	28 (32)	125 (136)	71 (2)
Inkling†	74 (78)	72 (80)	30 (31)	125 (145)	69 (13)
Inkling-Small‡	115 (118)	122 (123)	79	139 (142)	41 (27)
Inkling‡	123 (130)	125 (130)	87	160 (167)	63 (24)
Kimi-K2.6†	103	102	69	131	25

Correction baselines. We compare six procedures on the shared evaluation set. *Unchanged answer* retains the replay answer. *Self-reflection* appends it to the conversation and asks the same model to check and revise it against the image. *Visual-verification feedback* adapts VISCO's LookBack (Wu et al., 2025). The model asks up to three atomic questions about the premise and answers each from the image in a new conversation. It then critiques its original answer using these observations and an answer generated without the earlier turn, treating both as evidence that may contain errors. The critique is appended to the original conversation for correction. Every call uses only the evaluated model, with no correctness label or reviewed correction, so each procedure must find the error itself.

We also compare four-answer voting with the earlier premise retained, *history reset*, and reset followed by four-answer voting. History reset discards the earlier turns and uses the fresh-query answer in both premise conditions. Appendices F and H give call counts, decoding, parsing, and prompts.

Feedback trades repair against retention. On the full model-reviewed set, feedback raises false-premise replay accuracy over self-reflection by 8.76, 14.75, 22.12, 14.75, and 7.83 points on InternVL, Qwen-8B, Inkling-Small, Inkling, and Kimi-K2.6. Corrected-premise replay accuracy falls by 8.76, 9.68, 8.76, 8.76, and 5.99 points. The paired 95% intervals exclude zero for both directions on all five models and under both parsers (Appendix Table 15). Human-supported subsets, analyzed for the 8B models, keep these directions with wider intervals (Appendix A.2).

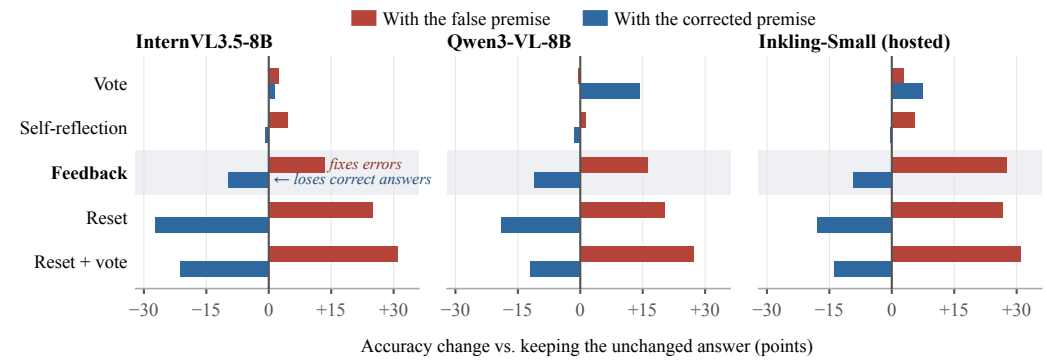


Figure 4: **Feedback and history reset trade repair for retention on the 217 shared pairs.** Each row compares one procedure with keeping the unchanged replay answer. Red bars show the accuracy change with the false premise, and blue bars show it with the corrected premise. Bars pointing in opposite directions mean that a procedure fixes errors but loses correct answers. Every procedure uses only the evaluated model. Primary parser, with counts in Appendix Table 14.

Relative to self-reflection, balanced accuracy changes by 0.00, +2.53, +3.00, and +0.92 points on InternVL, Qwen-8B, Inkling, and Kimi-K2.6, with intervals that include zero, and rises by 6.68 points on Inkling-Small (Appendix Table 15). False-premise replay alone would overstate the benefit on all five models (Figure 4, Appendix Table 15).

Where the losses occur. We split the pairs by whether each model answers correctly under fresh query (Table 16). On pairs a model answers correctly alone, false-premise replay turns 72–79% of those answers wrong on five of the six models, and 39–41% on Kimi-K2.6 and on both Inkling models with reasoning. Relative to the unchanged answer, feedback loses 1–6 corrected-premise answers on these pairs and 11–22 on the remaining pairs, where the fresh answer given to its critique is wrong. Without that fresh answer, feedback on both Inkling models still beats self-reflection under false premises and trails it under corrected ones, with 95% intervals excluding zero (Appendix F).

5 What Corrective Information Can the Model Use?

The procedures in Section 4 must find the error themselves. We ask whether naming the wrong statement suffices or the model also needs corrected facts. On 500 separate sources with constructed rather than model-generated premises, we compare identifying an error with supplying its correction. Feedback gives a *correctness label* for the earlier visual statement rather than for the final answer: `incorrect` for the false premise or `correct` for its correction. We supply this label alone, with task-irrelevant text, or with corrected visual facts, while the image and earlier premise remain present. Both the label and the facts come from premise construction, so this study measures how the model uses correct information that it did not produce.

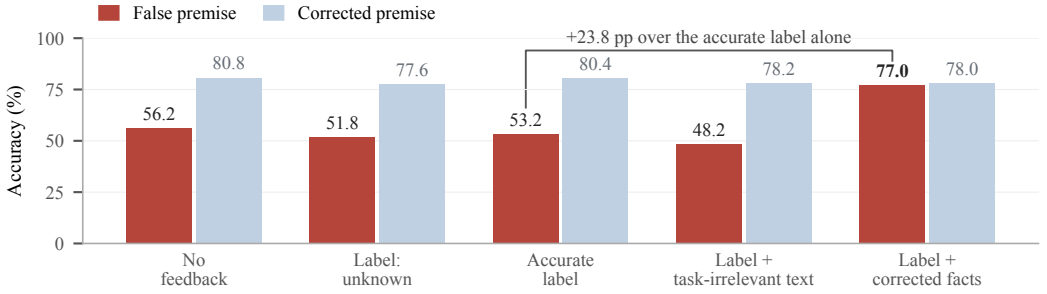


Figure 5: **Feedback content on 500 constructed sources.** InternVL retains the image and earlier premise. Feedback supplies a correctness label alone or with task-irrelevant text or corrected facts. The `unknown` label withholds a correctness judgment, and no feedback leaves the replay unchanged.

An accurate label alone changes false-premise replay accuracy by 1.4 points relative to the `unknown` label (Figure 5, 95% CI [−1.0, 3.8]). Adding corrected facts gains 23.8 points over the accurate label alone and 28.8 over task-irrelevant text (95% CI [24.4, 33.2]). Unlike corrected-premise replay, this feedback does not remove the false premise. The model can use the added facts despite that earlier error.

Without images, an accurate label helps Qwen-30B and corrected facts still help InternVL, so the supplied information alone can produce these gains (Appendix E). On the 82 extension pairs, notes that Qwen3-VL-30B-A3B writes from the image and question alone raise InternVL's balanced accuracy by 14.02 points, whereas InternVL's own notes leave it unchanged (Appendix G).

6 Related Work

Visual verification and rereading. VISCO finds that perceptual errors are particularly difficult for VLMs to identify in their own reasoning. It scores correction by both the errors fixed and the correct answers changed to incorrect ones, and its LookBack method asks the same model atomic questions about the image without the prior reasoning (Wu et al., 2025). Visual process reward models evaluate visual premises in intermediate steps (Wang et al., 2026). ReGround trains selective re-examination of grounding failures (Peng et al., 2026). Studies of prior text and visual decoding are discussed in Section 1. Language models over-commit to their own early mistakes (Zhang et al., 2024) and change correct answers to agree with a user's false claim (Sharma et al., 2024). Deceptive prompts and erroneous user corrections similarly mislead VLMs (Qian et al., 2024; Li et al., 2025). Language models often fail to find their reasoning errors yet correct them once told where the error is (Tyen et al., 2024), and Volcano trains a VLM to revise its answer from its own feedback (Lee et al., 2024). Like VISCO, we count both repairs and harms, but on paired episodes that change only whether the earlier premise is true, separating repair from retention.

7 Limitations

Premise labels were assigned by model review, and human review covers 120 of the 217 shared pairs and checks premise truth only. Under our sampling and review procedure, 4.9–6.7% of candidate questions for each model yield a qualifying false premise. Our results therefore characterize persistence conditional on finding such a premise rather than how often these false premises arise naturally. Checking a premise is costly, since visual-verification feedback adds up to six model calls per replay, compared with one for self-reflection (Appendix F).

8 Conclusion

False-premise replay reveals a conditional failure of visual self-correction. A single false visual premise can persist even when the original image and question remain the same, while replaying its reviewed correction often restores the answer. The effect also holds on human-supported pairs, and the false–corrected gap persists with short original reasoning prefixes. Image interventions show that the model continues to use the visual input, but the unary and pairwise tests show that it does not reliably identify which premise is supported by the image. Thus, access to the image alone is not sufficient to overturn a previously stated visual error. On InternVL, a false premise also reduces correct answers among sampled candidates, and supplied corrected facts help more than a correctness label alone. Reliable visual self-correction requires using the image to verify and replace the false premise. On the shared set, the procedures with the largest false-premise gains also lose accuracy when the earlier premise is correct, so correction should be evaluated under both premise conditions.

Reproducibility Statement

Section 2 and Appendix A describe premise construction, model review, and human review. Appendix F gives decoding, parsing, scoring, and resampling settings, and Appendix H lists the deployed prompts. The supplementary material contains the 217 shared premise pairs with reference answers, the request and scoring code, the model-run scripts, and every recorded response behind the shared-set results. Source images keep their upstream licenses and are identified by SHA-256.

REFERENCES

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

Wenxiao Fan, Jingling Fu, Fang Li, Luohang Liu, Yu He, Lichen Ma, Zhiyang Yu, Weishan Bi, Junshi Huang, Yan Li, Gu Simiu, and Kan Li. Recompute or Reuse? Diagnosing and Mitigating Textual Shortcuts in VLM Self-Reflection, 2026. URL https://arxiv.org/abs/2608.01930.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-Modal Hallucination Control by Visual Information Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels Versus Priors: Controlling Knowledge Priors in Vision-Language Models through Visual Counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations*, volume 2024, pp. 32808–32824, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8b4add8b0aa8749d80a34ca5d941c355-Paper-Conference.pdf.

Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. *Transactions of the*

Association for Computational Linguistics, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713. URL https://aclanthology.org/2024.tacl-1.78/.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.23. URL https://aclanthology.org/2024.naacl-long.23/.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Shuo Li, Tao Ji, Xiaoran Fan, Linsheng Lu, Leyi Yang, Yuming Yang, Zhiheng Xi, Rui Zheng, Yuran Wang, Xiaohui Zhao, Tao Gui, Qi Zhang, and Xuanjing Huang. Have the VLMs Lost Confidence? A Study of Sycophancy in VLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/07be1a0850e58ca29e2b6ce31fc0c791-Paper-Conference.pdf.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Lei Peng, Shuai Lv, and Wei Hu. ReGround: Restoring Visual Grounding in Multi-Step Reasoning through Self-Diagnosis and Visual Re-Examination, 2026. URL https://arxiv.org/abs/2608.04385. Accepted to ACM Multimedia 2026.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How Easy is It to Fool Your Multimodal LLMs? An Empirical Analysis on Deceptive Prompt. In *Safe Generative AI Workshop @ NeurIPS 2024*, 2024. URL https://neurips.cc/virtual/2024/106314.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs Seeing or Just Saying? Uncovering the Illusion of Visual Re-examination. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65441. Oral presentation.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Thinking Machines Lab. Tinker models and pricing. https://tinker-docs.thinkingmachines.ai/tinker/models/, 2026. Accessed 2026-09-25.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.826. URL https://aclanthology.org/2024.findings-acl.826/.

Junxin Wang, Dai Guan, Weijie Qiu, Zhihang Li, Yongbo Gai, Zhengyi Yang, Mengyu Zhou, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Grounding the Score: Explicit Visual Premise Verification for Reliable Vision-Language Process Reward Models, 2026. URL https://arxiv.org/abs/2603.16253.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9527–9537, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_VISCO_Benchmarking_Fine-Grained_Critique_and_Correction_Towards_Self-Improvement_in_Visual_CVPR_2025_paper.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/oral/12647.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How Language Model Hallucinations Can Snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ay.html.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.648. URL https://aclanthology.org/2024.acl-long.648/.

A Evaluation Sets and Construction

Table 3: **Evaluation sets used for different questions.** Counts are source questions unless the row specifies episodes or pairs. Later studies reuse previously studied sources.

Study	Size	Role
Primary replay set	124 episodes, 101 images	Self-generated errors from 1,070 candidate questions
Additional replay sets	35 Qwen-30B, 58 InternVL episodes	Self-generated errors, with 18 shared source questions between these sets
Premise production	217 episodes	Primary and additional replay episodes
Shared evaluation set	217 premise pairs, 176 images	Same pairs on every evaluated model, including own- and other-model errors
Candidate studies	48 initial, 82 extension InternVL pairs	Initial domain-balanced study and its source-separated extension
Attention masking	44 InternVL sources	Reliability-checked episodes from a separate constructed pool
Feedback content	500 sources	Constructed false and corrected premises
Generated-note replay	48 initial, 82 extension InternVL pairs	External-reader and self-written notes with the premise retained

The primary set pools three source-disjoint construction batches: 395 questions yielding 38 episodes over 30 images, 575 yielding 57 over 46, and 100 yielding 29 over 25. They are pooled into one primary evaluation set. Four traces per question for each of two models produce 2,140 model–question opportunities and 8,560 traces. Extraction and image review retain 72 InternVL and 52 Qwen-30B episodes. The selection rate and the replay effect answer different questions: the former describes the construction procedure, while the latter conditions on an accepted premise.

Annotation and decoding. Trace generation requests four samples of at most 768 new tokens with temperature 0.8, top-p 0.95, top-k 0, and seed 20260729 under a stable model–prompt seed rule. Claim extraction and the full-image primary review use `gpt-5.6-terra` at `xhigh` reasoning effort. Extraction receives only the question and sampled trace. Review checks the image and declared crop, rejects terminal or ambiguous statements, requires an answer-relevant contradiction, and requires an explanation of a semantically distinct wrong answer. The remaining rationale and final answer are not replayed. The review also supplies the corrected premise. These are different information-access stages of one model-based pipeline. Provider-side build and sampling metadata for the hosted annotation calls are unavailable.

The construction records contain 18 secondary and one tie-break review for false premises, and 27 secondary and five tie-break reviews for the answer-free reference facts, also using the same hosted model. A separate backend-alignment pilot agrees on 14 of 15 completed decisions. This checks only workflow agreement. The human review below evaluates a subset separately, without changing these construction records.

Uncertainty. Primary-set bootstraps sample source images and retain paired conditions and both models' episodes when they share an image. Later single-model studies sample sources. We use 10,000 draws. Intervals for the reader comparison resample each source together with its false and corrected presentations. They capture only source variability conditional on the recorded generations.

A.1 Human Review of Premise Truth

We sampled 120 pairs over 97 images from the 217-pair shared evaluation set, selecting by source image and stratifying by domain and construction set without using model outcomes. Each pair received two original human annotations, yielding 240 item annotations and 480 binary statement judgments. Reviewers saw the original image, question, and both premise strings as A and B, without their false/corrected labels or the model's replay answers. The A/B order was reversed between the

two reviewers. Each statement was judged separately, so reviewers could mark both true or both false. This review tests premise truth only, leaving relevance, answer leakage, and sufficiency to explain a wrong answer to model review.

Table 4: **Human premise-truth review by source benchmark.** The first three outcome columns partition the original double review. "Both confirm" requires both reviewers to support both recorded labels. "Mismatch" means they agree but reject at least one label. The last column applies statement-wise majority after the follow-up.

		Original double review			Follow-up
Source	Pairs	Both confirm	Mismatch	Disagree	Majority support
ChartQA	15	10	2	3	12
InfographicVQA	16	11	1	4	12
Geometry3K	54	41	2	11	51
GQA	35	17	2	16	25
Total	120	79	7	34	100

The original reviewers agree on 203/240 statements (84.58%). Agreement between reviewers is distinct from agreement with the construction model's labels. They jointly confirm both labels in 79 pairs, agree on a label mismatch in 7, and disagree on at least one statement in 34 (Table 4). Thus some sampled pairs may remain invalid after the review. Disagreement is particularly frequent on GQA, rather than uniform across domains.

Human-confirmed, AI-assisted follow-up. A third annotator individually checked AI-drafted judgments for all 34 disputed pairs. We record this as human-confirmed, AI-assisted review, which is weaker evidence than an independent blind third annotation. The follow-up supports the original pairing in 21 disputed cases under statement-wise majority, giving 100/120 supported pairs overall and 20 mismatches. Giving priority to the follow-up on the disputed cases also retains 100 pairs, but changes membership for 2 items. The two policies share 99 supported pairs, and neither policy overwrites the original labels.

Item identities, A/B mappings, displayed images, and premise strings match the distributed packs and the original shared set. We preserve the original annotations and AI-assisted follow-up separately, including their item-level disagreements and confidence notes.

A.2 Sensitivity to Human-Supported Premise Selection

We recompute summaries from the unchanged scored responses, selecting pairs only by review outcomes. The *unanimous* subset contains the 79 pairs supported by both original reviewers (62 images). The *common* subset contains the 99 pairs supported under both follow-up policies (80 images). Additional variants use each policy's 100 pairs separately and exclude one flagged chart example from the common subset (98 pairs). These post-hoc robustness checks reuse existing outputs on previously studied pairs and complement the full-set results. We retain the original primary parser, whitespace-tolerant parser, exact score, and numeric-tolerance score. Paired percentile 95% intervals use the same 10,000 image-cluster bootstrap draws and are not adjusted for multiple comparisons.

Table 5: **Direct replay after review-based selection.** Correct-answer counts use the primary parser and exact scoring. Changes are percentage points relative to fresh query, with paired image-cluster 95% intervals. The three open-weight models receive identical inputs within each subset.

Subset (N)	Model	Fresh	False	Corr.	False − fresh	Corr. − fresh
Unanimous (79)	InternVL	27	14	49	−16.46 [−29.27, −4.00]	+27.85 [14.08, 41.46]
Unanimous (79)	Qwen-8B	20	9	39	−13.92 [−25.33, −3.53]	+24.05 [9.09, 39.02]
Unanimous (79)	Qwen-30B	32	7	50	−31.65 [−44.74, −19.74]	+22.78 [5.80, 39.24]
Common (99)	InternVL	34	17	62	−17.17 [−28.26, −6.60]	+28.28 [16.00, 40.21]
Common (99)	Qwen-8B	27	10	45	−17.17 [−28.09, −7.22]	+18.18 [5.10, 31.31]
Common (99)	Qwen-30B	37	8	56	−29.29 [−40.21, −19.19]	+19.19 [5.21, 33.01]

The principal replay contrast persists. Under primary exact scoring, false-minus-fresh is negative and corrected-minus-fresh positive, with intervals excluding zero on all three open-weight models in each of the five supported variants. Table 5 shows the unanimous and common subsets. The common subset also preserves both directions with intervals excluding zero under both parsers and both scores. The conclusion does not hold uniformly for every reference, scoring rule, or domain. On the unanimous subset, Qwen-8B's numeric-tolerance false-minus-fresh contrast is −10.13 [-21.79, 1.23] points. On the common subset, Qwen-30B's exact corrected-minus-neutral contrast is +11.11 [-2.06, 23.76] points. Both intervals include zero. Domain-level directions also vary. The supported claim is the aggregate paired replay contrast.

Table 6: **Feedback minus self-reflection on human-supported pairs.** Exact-score changes in percentage points with paired 95% intervals. WS denotes whitespace-tolerant parsing, and "Both" denotes identical results under both parsers. Balanced accuracy averages false- and corrected-premise replay.

Subset (N)	Model	Parser	False premise	Corrected premise	Balanced
Unanimous (79)	InternVL	Both	+6.33 [−3.80, 16.46]	−15.19 [−28.05, −2.56]	−4.43 [−13.69, 4.79]
Unanimous (79)	Qwen-8B	Primary	+3.80 [−3.75, 11.25]	−7.59 [−20.73, 5.13]	−1.90 [−9.62, 5.88]
Unanimous (79)	Qwen-8B	WS	+11.39 [2.33, 21.52]	−8.86 [−22.22, 3.90]	+1.27 [−7.14, 9.64]
Common (99)	InternVL	Both	+6.06 [−3.23, 15.69]	−12.12 [−23.53, −1.02]	−3.03 [−11.11, 5.05]
Common (99)	Qwen-8B	Primary	+10.10 [2.04, 18.45]	−6.06 [−16.67, 4.35]	+2.02 [−5.05, 9.14]
Common (99)	Qwen-8B	WS	+16.16 [6.93, 26.13]	−8.08 [−19.05, 3.00]	+4.04 [−3.47, 11.50]

Repair and retention remain distinct. Feedback's false-premise replay gain and corrected-premise replay loss retain their point-estimate directions in these subsets, but their condition-specific intervals no longer uniformly exclude zero (Table 6). The balanced feedback-minus-reflection interval includes zero for both 8B models in all five supported variants, under both parsers and both scores. Baseline rankings also change with the subset and parser. These results support measuring repair and retention together. They leave both a net advantage for feedback and the equivalence of the procedures unresolved. In this analysis, Qwen-30B contributes direct replay only, and the hosted runs are excluded.

Flagged examples and scope. The follow-up marks 6 cases as medium confidence. One is a chart whose plotted year alignment makes the question's year uncertain despite the identified values. Excluding this case leaves 98 common-supported pairs and preserves the aggregate primary exact contrasts on all three open-weight models. Another case contains a football/frisbee wording tension and is policy-dependent, so it is absent from the common subset. The human review does not cover the separate attention, candidate-pool, or 500-source feedback sets. Recomputing the unfiltered set reproduces the existing full-set aggregates and paired comparisons, and the original labels and responses remain unchanged.

B Replay Comparisons and Visual Probes

Table 7 gives fresh, false-premise, and corrected-premise accuracy for all three model-reviewed evaluation sets, and Figure 2 in the main text reports the primary-set conditions.

Table 7: **Replay across evaluation sets.** Accuracy in percent. Drop is fresh minus false-premise replay in percentage points, with ±2 SE.

Evaluation set	N	Fresh	False	Corrected	Drop
Primary	124	34.7	8.9	71.0	25.8±9.6
Additional Qwen-30B	35	34.3	5.7	62.9	28.6±17.3
Additional InternVL	58	48.3	17.2	67.2	31.0±13.8

Neutral replay controls for an additional assistant turn without supplying visual evidence. Task-irrelevant replay uses an unrelated model-generated premise approximately matched in form and length. Corrected replay replaces the false premise, while re-checking leaves it in place and changes the final instruction. Full-trace replay is a separate comparison: it retains the wrong rationale and answer, so it cannot isolate the selected premise. On the 93 episodes pooled from the additional sets, full-trace replay yields 1.1% accuracy and 98.9% false-answer copying.

Unary probes compare forced `Supported` and `Contradicted` completions for each premise. Pairwise probes score a choice between the false and corrected premises in both orders. Confidence is normalized over those two candidate completions only. The primary unary score has AUC 0.516. A mirrored control uses 62 images in 31 pairs, with the same question and premise text appearing once with each truth label. Probe accuracy ranges from 90.3% to 98.4% with images and from 47.6% to 50.8% without them. The mirrored examples differ in difficulty and construction from the self-generated errors.

B.1 Generating Visual Facts Without Supplied Options

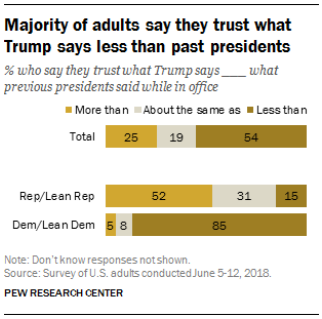
Each model is asked to state the answer-relevant visual fact without answering the question. We compare a fresh query, false-premise replay, and false-premise replay with a re-check instruction. A deterministic matcher and a blinded text-only judge compare each response with the false and corrected premise strings. The judge does not see the image, condition, or model identity. It labels corrected-only, false-only, both, or neither. Ambiguity flags remain in the denominator. This measures only what a response states relative to the reviewed premises.

Table 8: **Generated visual facts on all 217 episodes of InternVL3.5-8B and Qwen3-VL-30B-A3B, pooled.** "False present" includes false-only and both-premise responses.

Earlier assistant text	Corrected only (%)	False present (%)
None	23.5	52.5
False premise	8.8	82.5
False premise + re-check	5.5	93.5

B.2 Recorded Premise Examples

Figure 6 gives one recorded example from each source benchmark. Each panel pairs the original question with the extracted false premise and its model-reviewed correction. Only the selected premise is inserted into the assistant history. The remaining source trace and its final answer are discarded.

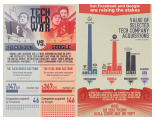


(a) ChartQA: legend–value binding.
Error source: InternVL3.5-8B.

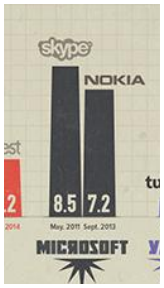
Question. How many Grey bars have a value less than 25?

False premise. The grey Less than bars are 54 for Total, 15 for Rep/Lean Rep, and 85 for Dem/Lean Dem.

Corrected premise. The gray 'About the same as' bars are 19 for Total, 31 for Rep/Lean Rep, and 8 for Dem/Lean Dem.



Full image



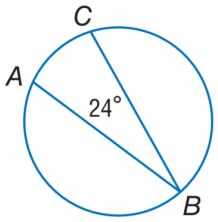
Detail

(b) InfographicVQA: entity–value binding.
Error source: Qwen3-VL-30B-A3B.

Question. Which is Microsoft's biggest acquisition to date?

False premise. Nokia is labeled $8.5 billion in the Microsoft acquisition bar.

Corrected premise. The Microsoft acquisition bars list Skype at 8.5 and Nokia at 7.2, in billions of U.S. dollars.



(c) Geometry3K: angle identification.
Error source: Qwen3-VL-30B-A3B.

Question. Find $m\widehat{AC}$.

False premise. A, B, and C are on a circle, and the angle identified as central angle ACB is given as 24 degrees.

Corrected premise. The inscribed angle ABC is labeled 24 degrees.



(d) GQA: object–attribute binding.
Error source: InternVL3.5-8B.

Question. Does the large car look empty and green?

False premise. The large car is a pickup truck parked next to another vehicle; it is green and appears empty, with no visible people or items inside.

Corrected premise. The large foreground pickup is beige/tan rather than green.

Figure 6: **Recorded premises across all four source benchmarks.** The errors associate chart values with the wrong legend entry, attach an acquisition value to the wrong company, misidentify an inscribed angle, and assign the wrong color to a vehicle. Premise strings are reproduced from the recorded model-reviewed pairs. Panel (b) enlarges a detail for legibility alongside the full image supplied to the model. These are illustrative construction examples.

C Candidate Generation and Fixed-Pool Selection

Generation. The initial 48 sources contain 12 questions from each of the four domains and use InternVL-generated premises. The InternVL extension set contains 82 own-error pairs: two chart, two infographic, 48 geometry, and 30 natural-image pairs. We compare the false premise, neutral text matched to its token count, and a short neutral statement, retaining both user turns and the image. Token counts use the model's tokenizer. The length-matched control also preserves the final user turn's token position, including image-token expansion. It removes the complete premise, including any true subfacts, rather than editing only one incorrect binding.

Each condition generates four answers with temperature 0.8, top-p 0.95, and a 2,048-token cap. InternVL uses batch size four and top-k zero. There is one realized candidate pool per condition. Unparsed outputs remain incorrect and are not discarded or selectively rerun. Voting normalizes numeric and textual answers, ignores unparsed candidates when a parsed answer exists, and breaks ties by the earliest sampled winning answer. A covered question has at least one correct parsed candidate. For this vote, the number of wrong outputs equals the number of uncovered questions plus the number of covered questions whose vote is wrong.

Table 9: **InternVL candidate generation.** Coverage, correct votes, and selection losses count sources. Missing answers count individual generations.

Sources	Earlier text	Coverage	Correct votes	Selection losses	Missing
Initial 48	False	8	5	3	0/192
	Matched neutral	26	18	8	3/192
	Short neutral	24	17	7	5/192
Extension 82	False	18	10	8	0/328
	Matched neutral	41	31	10	11/328
	Short neutral	45	33	12	5/328

On the initial sources, Qwen3-VL-30B-A3B receives InternVL's false premise, making it a cross-model comparison. Its false and matched-neutral coverage is 10/48 and 31/48, but 54/192 and 51/192 responses lack parsed answers. Allowing correct answers in missing outputs gives a coverage-difference range of $[-8.33, 68.75]$ points. The Qwen3-VL-30B-A3B extension study instead uses 73 of its own-error sources. Coverage is 9/73 versus 35/73, with 55/292 versus 90/292 unparsed outputs and a missing-only difference range of $[-1.37, 68.49]$ points. Both ranges include zero.

Fixed-pool selection. Four cached answers appear as options A–D, including duplicates and explicit missing-answer placeholders. No correct candidate is added using the reference label. A deterministic source-specific permutation and its reverse are used. The selector sees the original image, question, earlier assistant statement, and a request for one option letter. Strict parsing requires the complete trimmed output to be A, B, C, or D. We average orders within each source before computing paired uncertainty. On the extension pairs, first-token option argmax reaches 27.44% under a false premise and 42.68% under a corrected premise. Argmax with neutral text reaches 37.20%, preserving the neutral-over-false contrast without requiring a well-formed generated response.

Confidence and candidate reordering. Normalizing the four first-token option probabilities gives a candidate-position distribution. Both false and corrected premises raise its maximum probability relative to neutral text, but they change probability assigned to correct candidates in opposite directions. The analysis counts duplicate answer positions in their total probability mass.

Table 10: **First-token probability changes relative to neutral text.** Changes are percentage points. Reversals count sources with different valid normalized answers and a log-probability gap above 0.1 nats in both conditions in at least one option order.

Sources	Earlier text	Δ max probability	Δ correct mass	Reversals
Initial 48	False	9.23	-6.04	25
	Corrected	9.85	13.72	21
Extension 82	False	5.10	-6.79	38
	Corrected	7.70	6.64	35

A positive power transformation $p'_j \propto p_j^a$ with $a > 0$ preserves strict rankings. Observed reversals therefore cannot all be explained by that rank-preserving sharpening. The analysis reuses the selection generations. For the extension pairs, the 95% interval for the correct-mass change under false-premise replay is [-11.70, -2.68] points, compared with [3.09, 10.79] under a corrected premise.

D Attention to Earlier Assistant Content

The intervention masks attention from downstream query positions to earlier assistant-content tokens in language layers 12–23 of InternVL3.5-8B. Framing tokens and other attention links remain available. The 44 reliability-checked episodes come from a separate 500-source pool, distinct from an earlier 16-episode pilot run. Baseline and masked runs share the same runtime. Masking all 36 language layers provides a magnitude reference, raising false-premise replay correctness from 6/44 to 23/44. Masking layers 12–23 raises false-premise replay correctness from 6/44 to 18/44 but lowers corrected-premise replay correctness from 35/44 to 26/44, and neutral replay changes from 20/44 to 22/44 (Table 11).

Table 11: **Attention masking helps false-premise replay but harms corrected-premise replay.** Correct-answer counts and paired percentage-point changes on 44 InternVL episodes.

Premise	Unmasked	Layers 12–23 masked	Change (pp, 95% CI)
False	6/44	18/44	$+27.3$ [13.6, 40.9]
Corrected	35/44	26/44	-20.5 [-34.1, -6.8]
Neutral	20/44	22/44	$+4.5$ [-4.5, 13.6]

The neutral-replay interval includes zero. This intervention changes downstream access to prior text and leaves the vision encoder unchanged.

E Feedback Content Controls

The 500-source set contains 419 single-fact items and 81 controls with constructed premises, separate from the model-generated errors. A feedback correctness label refers to the earlier visual statement, not the final answer. The label is `unknown`, `correct`, or `incorrect`. The `unknown` label withholds a judgment and does not mean that the statement's annotated truth is unknown. Accurate feedback says `incorrect` for the false premise and `correct` for the corrected premise. We also reverse these labels to test misleading feedback.

Detailed feedback adds either reviewed corrected facts or task-irrelevant text to the accurate label. The two additions are matched in the model tokenizer's token count. Both retain the original assistant premise. This is distinct from neutral replay, which replaces that premise with non-informative text, and from corrected-premise replay, which replaces it with the reviewed correction. The candidate-generation study's length-matched neutral text is another replacement control (Appendix C). The shared evaluation set instead uses the short, unmatched statement quoted in Appendix H.

Table 12: **Supplied correctness labels and image-absent controls.** InternVL accuracies use all 500 sources. A label can be accurate or misleading depending on the premise. Detailed feedback always includes the accurate label.

Feedback / image	False premise (%)	Corrected premise (%)
"Correct" / image	49.4	80.4
"Incorrect" / image	53.2	77.8
Correct facts / no image	72.6	76.0
Task-irrelevant text / no image	19.0	76.6

The image-absent detail conditions were added after the image-present run. Image blocks alone were removed, retaining textual references to the image. The image-absent outputs come from later runs on different hardware. High no-image accuracy shows that the supplied text can support answers under that answer format.

Qwen3-VL-30B-A3B was also evaluated with correctness labels without images. Under false-premise replay, `unknown`, `correct`, and `incorrect` yield 92/500, 93/500, and 187/500 correct answers. Accurate feedback therefore helps even without pixels. Of 29 sources where switching from a

misleading to an accurate label repairs the answer in both premise conditions, 28 are binary questions. On such questions, rejecting one proposition can identify an alternative without rereading the image.

F Baselines and Scoring on the Shared Evaluation Set

The shared evaluation set contains the same 217 accepted premise pairs used in the model-specific studies, now evaluated on every model. A source image can contribute more than one pair. The set contains 27 chart pairs over 24 images, 28 document pairs over 21 images, 97 geometry pairs over 76 images, and 65 natural-image pairs over 55 images. The initial 48 and extension 82 pairs originate from InternVL, while the remaining 87 originate from Qwen3-VL-30B-A3B. No pair is selected according to a baseline's success. Human truth review covers 120 pairs (Appendix A.1). Full-set scores retain model-reviewed labels.

Table 2 reports the complete direct conditions for six models from four developers, with the two Inkling models also run with reasoning. InternVL3.5-8B initializes its language model from Qwen3-8B (Wang et al., 2025), so we count developers rather than language-model lineages. The two 8B models and Inkling-Small have the full correction comparison reported in Figure 4, and Inkling and Kimi-K2.6 have the same comparison in the tables below. Qwen-30B appears in the direct table as an answering model and serves as the external reader only in Appendix G. Rankings among these models depend on answer formatting.

Hosted models. Inkling-Small is a mixture-of-experts model with 276B total and 12B active parameters, as reported by its developer (Thinking Machines Lab, 2026). We query it through the Tinker sampling interface with the provider's chat renderer, reasoning effort 0, temperature 0, the 2,048-token cap, and the request-specific seed. It receives the same 868 direct requests, with the image in the same position. The provider does not expose a pinned model revision, so we mark this run with †. Two answers reach the token cap. The primary parser leaves 71 answers unparsed, mostly because the answer marker ends the final sentence, and the whitespace-tolerant parser leaves 2. The run uses 490,489 input and 86,074 output tokens. Its correction baselines use the same frozen requests and dependency order, with sampled votes at temperature 0.8 and top-p 0.95 and no top-k limit. They add 5,642 calls with 3,396,086 input and 470,001 output tokens. False-premise replay lowers its accuracy by 26.7 points (95% CI [18.9, 34.7]). Inkling, the larger model from the same developer, has 975B total and 41B active parameters (Thinking Machines Lab, 2026). It uses the same interface, settings, requests, and correction baselines. Its direct run has 12 capped answers, with 69 primary and 13 whitespace-tolerant unparsed answers, and its baselines add 5,642 calls with 3,463,011 input and 558,799 output tokens. False-premise replay lowers its accuracy by 20.3 points ([13.6, 27.1]). With reasoning, false-premise replay lowers the accuracy of Inkling-Small by 16.6 points ([9.7, 23.8]) and of Inkling by 16.6 points ([9.6, 23.6]). The reasoning runs repeat the direct requests for both models with reasoning effort 0.9, the renderer default, and a 16,384-token cap, since reasoning precedes the answer. Scoring reads only the final answer text. Inkling-Small and Inkling reach the cap on 27 and 24 requests and use 1,524,139 and 1,724,673 output tokens. Kimi-K2.6 from Moonshot AI has 1T total and 32B active parameters (Thinking Machines Lab, 2026). It runs through the same interface with the interface provider's chat renderer and thinking disabled, which adds the model's default system prompt, and otherwise uses the same requests, decoding, and correction baselines. Its direct run has 25 capped and 25 unparsed answers under either parser, and its baselines add 5,642 calls with 5,063,691 input and 1,008,573 output tokens. False-premise replay lowers its accuracy by 15.7 points ([9.6, 22.0]). No hosted model produced any premise pair.

Premise source. Table 13 splits the direct results by the model whose error produced each pair. False-premise replay lowers accuracy on both sources for every model. The larger drop falls on other-model pairs for InternVL and on own-error pairs for Qwen-30B.

Answer format and decoding. All compared procedures request a brief explanation followed by a final line beginning `FINAL ANSWER:`. The primary parser requires that marker on the final nonempty line. If the marker is absent, one unambiguous boxed answer is accepted. Parsing never consults the reference answer. Unparsed outputs remain incorrect. Greedy answer generation has a 2,048-token cap. Atomic questions, their answers, and critiques have a 512-token cap. Stochastic votes use four independent generations with temperature 0.8, top-p 0.95, and top-k 0. A fixed

Table 13: **Direct replay by premise source.** Correct answers under the primary parser, with the whitespace-tolerant count in parentheses where it differs. "Own" marks pairs built from the evaluated model's errors. †Hosted interface. ‡With reasoning.

Model	Premise source	N	Fresh	False	Corrected
InternVL3.5-8B	InternVL (own)	130	44	17	89
	Qwen-30B	87	40	13	54
Qwen3-VL-8B	InternVL	130	39 (54)	21 (23)	71 (89)
	Qwen-30B	87	32 (35)	6 (8)	41 (54)
Qwen3-VL-30B-A3B	InternVL	130	54 (61)	21 (23)	81 (104)
	Qwen-30B (own)	87	29 (31)	3	51 (63)
Inkling-Small†	InternVL	130	53 (61)	18 (19)	78 (83)
	Qwen-30B	87	33 (39)	10 (13)	47 (53)
Inkling†	InternVL	130	46 (48)	16 (17)	76 (88)
	Qwen-30B	87	28 (30)	14	49 (57)
Inkling-Small‡	InternVL	130	73	44	87 (88)
	Qwen-30B	87	42 (45)	35	52 (54)
Inkling‡	InternVL	130	73 (75)	56	95 (101)
	Qwen-30B	87	50 (55)	31	65 (66)
Kimi-K2.6†	InternVL	130	57	41	78
	Qwen-30B	87	46	28	53

request-specific seed is derived from seed 20260920. Native-checkpoint calls use batch size one, BF16, SDPA attention, and the pinned native processor.

What each baseline changes. Self-reflection retains the image, question, premise, and initial answer, then asks the evaluated model to check that answer. Visual-verification feedback adds question generation, separate image-only answers, and a critique before the final revision. The critique also receives the model's cached fresh answer, explicitly marked as untrusted. The feedback procedure receives neither a gold answer nor the alternative premise in the pair. It processes false- and corrected-premise replay identically. If no atomic question is parsed, critique and correction still run. No such extraction failure occurred for the five models with complete comparisons. This adapts LookBack's atomic questioning to one-premise replay and differs from its original evaluation (Wu et al., 2025).

Voting uses normalized parsed answers, excludes unparsed samples from the vote, and chooses the earliest sampled winning answer to break a tie. If all samples fail parsing, the vote fails. Reset reuses the same fresh answer in both replay conditions. Reset plus voting similarly reuses the same fresh four-answer vote. Each identical request is generated once and reused across the two conditions.

Model calls. Unchanged answering and reset each require one answer generation. Voting requires four sampled answers. Self-reflection adds one call after the initial answer. Feedback with $q \leq 3$ extracted questions adds $q + 3$ calls for question generation, q readings, critique, and correction, and uses one fresh answer cached across the two replay conditions. Thus a standalone feedback evaluation uses the initial answer, a fresh answer, and $q + 3$ further calls. These procedures have different computational costs. Across all four direct conditions and all baseline stages, the recorded totals are 6,508 calls for InternVL and 6,492 for Qwen-8B, including 8 and 140 token-capped calls, respectively. Inkling-Small, Inkling, and Kimi-K2.6 each use 6,510 calls without reasoning, including 45, 63, and 278 token-capped calls. Identical cached requests are counted once.

Paired comparison and uncertainty. Every method is evaluated on all 217 pairs, with no model-specific complete-case selection. Percentile intervals use 10,000 image-cluster bootstrap draws with seed 20260920 and retain all questions, premise pairs, and replay conditions of a selected image. Table 15 directly compares feedback with self-reflection, rather than comparing their separate intervals against unchanged replay. In this comparison, a repair is an incorrect self-reflection answer that becomes correct with feedback, and a harm is the reverse. Table 14 instead uses the unchanged

Table 14: **Complete correction results and answer transitions on the shared evaluation set.** False and corrected columns count correct answers out of 217, and balanced accuracy averages the two replay conditions. Repairs and harms use the unchanged replay answer as reference over 434 presentations. Figure 4 plots the first three blocks. No procedure uses an external reader. †Hosted interface.

	Correct answers			Transitions	
Procedure	False	Corrected	Balanced (%)	Repairs	Harms
InternVL3.5-8B					
Unchanged answer	30	143	39.86	0	0
History reset	84	84	38.71	72	77
Four-answer vote	35	146	41.71	22	14
Reset + four-answer vote	97	97	44.70	81	60
Self-reflection	40	141	41.71	24	16
Visual-verification feedback	59	122	41.71	44	36
Qwen3-VL-8B					
Unchanged answer	27	112	32.03	0	0
History reset	71	71	32.72	70	67
Four-answer vote	26	143	38.94	44	14
Reset + four-answer vote	86	86	39.63	92	59
Self-reflection	30	109	32.03	11	11
Visual-verification feedback	62	88	34.56	52	41
Inkling-Small†					
Unchanged answer	28	125	35.25	0	0
History reset	86	86	39.63	79	60
Four-answer vote	34	141	40.32	35	13
Reset + four-answer vote	95	95	43.78	90	53
Self-reflection	40	124	37.79	20	9
Visual-verification feedback	88	105	44.47	72	32
Inkling†					
Unchanged answer	30	125	35.71	0	0
History reset	74	74	34.10	61	68
Four-answer vote	39	150	43.55	46	12
Reset + four-answer vote	93	93	42.86	82	51
Self-reflection	54	130	42.40	47	18
Visual-verification feedback	86	111	45.39	79	37
Kimi-K2.6†					
Unchanged answer	69	131	46.08	0	0
History reset	103	103	47.47	42	36
Four-answer vote	69	126	44.93	9	14
Reset + four-answer vote	108	108	49.77	48	32
Self-reflection	98	123	50.92	33	12
Visual-verification feedback	115	110	51.84	50	25

replay answer as its reference. The accuracy differences compare procedures within each replay condition on the same pairs. Intervals are per contrast and not adjusted for multiple comparisons.

Split by fresh-query correctness. Table 16 splits the shared pairs by whether the model answers correctly under fresh query. On fresh-correct pairs, false-premise replay turns 75.0, 76.1, 75.9, 79.1, 71.6, 38.8% of the answers wrong for InternVL, Qwen-8B, Qwen-30B, Inkling-Small, Inkling, and Kimi-K2.6 (39–77% under the whitespace-tolerant parser). With reasoning, the share is 40.9% for Inkling-Small and 39.0% for Inkling. Relative to the unchanged answer, feedback loses 1, 2, 4, 3, 6 corrected-premise answers on fresh-correct pairs and 20, 22, 16, 11, 15 on fresh-wrong pairs for InternVL, Qwen-8B, Inkling-Small, Inkling, and Kimi-K2.6.

Table 15: **Feedback minus self-reflection.** Paired accuracy changes in percentage points with 95% intervals. Whitespace sensitivity changes only parsing of the recorded outputs. †Hosted interface.

Model	Parser	False premise	Corrected premise	Balanced
InternVL3.5-8B	Primary	+8.76 [2.79, 14.80]	−8.76 [−16.13, −1.43]	+0.00 [−4.98, 5.07]
InternVL3.5-8B	Whitespace-tolerant	+8.76 [2.79, 14.80]	−8.76 [−16.13, −1.43]	+0.00 [−4.98, 5.07]
Qwen3-VL-8B	Primary	+14.75 [8.97, 20.74]	−9.68 [−16.96, −2.70]	+2.53 [−2.26, 7.31]
Qwen3-VL-8B	Whitespace-tolerant	+17.97 [11.54, 24.66]	−13.36 [−21.10, −5.80]	+2.30 [−2.88, 7.60]
Inkling-Small†	Primary	+22.12 [14.83, 29.44]	−8.76 [−15.14, −2.64]	+6.68 [1.64, 11.54]
Inkling-Small†	Whitespace-tolerant	+21.66 [14.41, 28.90]	−11.06 [−18.02, −4.33]	+5.30 [0.22, 10.31]
Inkling†	Primary	+14.75 [7.55, 22.01]	−8.76 [−15.32, −2.31]	+3.00 [−2.27, 8.25]
Inkling†	Whitespace-tolerant	+14.29 [7.11, 21.50]	−13.36 [−20.36, −6.57]	+0.46 [−5.02, 5.92]
Kimi-K2.6†	Primary	+7.83 [2.75, 13.02]	−5.99 [−10.46, −1.90]	+0.92 [−2.18, 3.94]
Kimi-K2.6†	Whitespace-tolerant	+7.83 [2.75, 13.02]	−5.99 [−10.46, −1.90]	+0.92 [−2.18, 3.94]

Table 16: **Correction results split by fresh-query correctness.** Correct answers under the primary parser as false / corrected premise, for pairs the model answers correctly under fresh query and for the remaining pairs. Qwen3-VL-30B-A3B has direct replay only. †Hosted interface.

Model	Pairs	N	Unchanged	Self-reflection	Feedback
InternVL3.5-8B	Fresh correct	84	21 / 75	24 / 72	41 / 74
	Fresh wrong	133	9 / 68	16 / 69	18 / 48
Qwen3-VL-8B	Fresh correct	71	17 / 55	20 / 54	45 / 53
	Fresh wrong	146	10 / 57	10 / 55	17 / 35
Qwen3-VL-30B-A3B	Fresh correct	83	20 / 63		
	Fresh wrong	134	4 / 69		
Inkling-Small†	Fresh correct	86	18 / 75	27 / 72	69 / 71
	Fresh wrong	131	10 / 50	13 / 52	19 / 34
Inkling†	Fresh correct	74	21 / 66	31 / 65	64 / 63
	Fresh wrong	143	9 / 59	23 / 65	22 / 48
Kimi-K2.6†	Fresh correct	103	63 / 101	83 / 98	90 / 95
	Fresh wrong	114	6 / 30	15 / 25	25 / 15

Feedback without the fresh answer. We rerun the critique and correction of both Inkling models without the cached fresh answer, reusing their recorded atomic questions and readings. The critique prompt drops the independent-answer line and its mention, and the correction step is unchanged. For Inkling-Small, feedback then answers 58/217 false-premise and 110/217 corrected-premise replays correctly, against 88/217 and 105/217 with the fresh answer and 40/217 and 124/217 for self-reflection. Relative to self-reflection, accuracy changes by $+8.29$ points (95% CI [2.88, 13.64]) under false premises and -6.45 points ($[-11.89, -0.94]$) under corrected premises. Its corrected-premise losses relative to the unchanged answer split into 7 on fresh-correct and 8 on fresh-wrong pairs. Under the corrected premise, 109 final answers match the fresh answer with it in the critique and 88 without it. For Inkling, the counts are 68/217 and 110/217 without the fresh answer, 86/217 and 111/217 with it, and 54/217 and 130/217 for self-reflection. The changes relative to self-reflection are $+6.45$ ([0.46, 12.72]) and -9.22 ($[-15.77, -2.76]$) points.

Procedure rankings. Under the primary parser, reset plus voting has the highest balanced accuracy for both 8B models, and feedback narrowly has the highest for both Inkling models and Kimi-K2.6. Four-answer voting with the earlier premise retained keeps more correct answers than reset plus voting under corrected-premise replay. For both 8B models the difference in balanced accuracy between feedback and self-reflection includes zero in every human-supported subset, under both parsers and both scoring rules (Appendix A.2).

Formatting and answer matching. The primary score compares normalized text or numerical answers using the existing reference matcher. A secondary chart-only score allows 5% numeric

Table 17: **Format sensitivity on the same outputs.** Missing counts are primary / whitespace-tolerant, across 434 presentations. The last two columns give whitespace-tolerant correct counts out of 217 per replay condition. †Hosted interface.

Model	Procedure	Missing	False	Corrected
InternVL3.5-8B	Unchanged answer	1 / 1	30	143
	History reset	2 / 2	84	84
	Four-answer vote	0 / 0	35	146
	Reset + four-answer vote	0 / 0	97	97
	Self-reflection	1 / 1	40	141
	Visual-verification feedback	1 / 1	59	122
Qwen3-VL-8B	Unchanged answer	93 / 11	31	143
	History reset	84 / 12	89	89
	Four-answer vote	10 / 0	30	148
	Reset + four-answer vote	12 / 0	90	90
	Self-reflection	94 / 12	40	140
	Visual-verification feedback	91 / 7	79	111
Inkling-Small†	Unchanged answer	29 / 0	32	136
	History reset	60 / 4	100	100
	Four-answer vote	1 / 0	34	144
	Reset + four-answer vote	4 / 0	97	97
	Self-reflection	26 / 0	47	135
	Visual-verification feedback	27 / 1	94	111

Table 18: **Domain breakdown on the shared evaluation set.** All three procedures use the same pairs within each row. †Hosted interface.

Model	Domain	N	Unchanged	Self-reflection	Feedback
InternVL3.5-8B	Chart	27	3 / 23	2 / 20	8 / 19
InternVL3.5-8B	Document	28	3 / 22	5 / 22	9 / 14
InternVL3.5-8B	Geometry	97	16 / 41	25 / 46	23 / 37
InternVL3.5-8B	Natural image	65	8 / 57	8 / 53	19 / 52
Qwen3-VL-8B	Chart	27	4 / 19	4 / 19	13 / 17
Qwen3-VL-8B	Document	28	4 / 22	7 / 23	12 / 13
Qwen3-VL-8B	Geometry	97	14 / 34	14 / 33	25 / 29
Qwen3-VL-8B	Natural image	65	5 / 37	5 / 34	12 / 29
Inkling-Small†	Chart	27	8 / 19	9 / 20	16 / 14
Inkling-Small†	Document	28	5 / 24	7 / 24	17 / 20
Inkling-Small†	Geometry	97	10 / 28	15 / 26	24 / 32
Inkling-Small†	Natural image	65	5 / 54	9 / 54	31 / 39

tolerance. A separate post-hoc whitespace sensitivity accepts a unique terminal `FINAL ANSWER:` marker even if it does not begin a new line. It only reparses the recorded outputs, without gold answers or a fallback for capped output. Both Qwen models and both Inkling models often place a valid-looking answer marker on the preceding text's line, making their primary scores sensitive to this distinction. The opposing changes under false- and corrected-premise replay also persist under the chart-tolerant matcher. Table 17 reports missing answers and whitespace-tolerant counts for the first three models, and Table 15 compares both parsers for all five.

Domain breakdown. Table 18 reports primary correct counts under false / corrected premises by domain. Geometry and natural images account for most pairs.

G GENERATED VISUAL EVIDENCE WITH THE EARLIER PREMISE RETAINED

External-reader feedback tests whether generated visual evidence helps while the earlier premise remains in context. Self-written notes test same-checkpoint rereading. Different-image notes control for task relevance. Each comparison includes false- and corrected-premise replay.

Qwen3-VL-30B-A3B writes two notes from the image and question alone, and InternVL answers with these notes and its earlier premise. On the 82 extension pairs of Section 3.3, correct false-premise answers rise from 11/82 to 36/82, whereas notes written for a different image from the same domain give 10/82. Corrected-premise answers change from 51/82 to 49/82. Notes written by InternVL itself give 25/82 with false premises and 37/82 with corrected premises, leaving balanced accuracy unchanged.

Models. The model is InternVL3.5-8B-HF at revision `741a7d03020411e666c6109218ab71e08151ef86`. The independent reader is Qwen3-VL-30B-A3B-Instruct. Two notes per source are used without truth-based filtering. Each note is produced greedily with a 512-token limit. The answering model uses greedy decoding with a 2,048-token limit.

Reader input. Each pass receives the same image and full question, with neither replay history nor a reference answer. The two focuses are fixed:

```
Independently reread the current task image for the full question
as a generic visual self-refinement pass.
Question: {question}
Pass {j} of 2: {focus}
Write one concise answer-relevant evidence note. Do not mention
any earlier answer, do not use a structured status protocol,
and do not give the final answer.
```

The first focus is "Review the overall scene and the directly visible evidence relevant to the question." The second is "Independently verify visible text, quantities, objects, and relationships that could change the answer." These are separate calls with no earlier messages, so the second reader never sees the first note.

Model input. The original image-bearing user turn and assistant premise are unchanged. The following block is inserted before the original final solve instruction:

```
<VISUAL_EVIDENCE_NOTES>
- Note 1: {note_1}
- Note 2: {note_2}
</VISUAL_EVIDENCE_NOTES>
These are untrusted notes about the current task image.
Re-examine the image yourself and ignore unsupported notes.
```

The instruction discourages direct answer disclosure but does not guarantee it, particularly when the requested visual fact itself determines the answer.

G.1 RESULTS ON ALL INTERNVL-ORIGIN PAIRS

The comparison evaluates all 130 InternVL-origin pairs under both replay conditions with four conditions: unchanged replay, external-reader feedback, self-written notes, and same-domain different-image notes. All 1,040 open-answer requests are complete. Table 19 keeps the initial 48 and extension 82 pairs separate. These pairs were studied previously.

Table 19: **Generated-evidence diagnostic under both premise conditions.** False and corrected columns count correct answers within each source group. External-reader feedback uses Qwen3-VL-30B-A3B notes with InternVL3.5-8B as the evaluated model. Self-written notes use InternVL throughout. Both retain the assistant premise. Unchanged replay comes from this comparison's own run, so its false-premise count on the 130 pairs differs by one from Table 13.

Sources	Added evidence	False	Corrected	Balanced (%)
Initial 48	Unchanged replay	7	38	46.88
	External-reader feedback	26	31	59.38
	Self-written notes	17	25	43.75
	Different-image notes	2	35	38.54
Extension 82	Unchanged replay	11	51	37.80
	External-reader feedback	36	49	51.83
	Self-written notes	25	37	37.80
	Different-image notes	10	49	35.98

On the extension pairs, external-reader feedback raises balanced accuracy by 14.02 points relative to unchanged replay (95% CI [5.49, 22.56]). Self-written notes improve false-premise replay correctness but lose answers under corrected-premise replay, leaving the balanced point estimate unchanged. The different-image control further separates relevant generated evidence from the presence of extra text. These results show that the evaluated model can use generated evidence despite the retained premise. They do not isolate reading without earlier turns from reader capability or establish that the combined system outperforms the reader itself.

H Workflow and Deployed Correction Prompts

H.1 Construction and Evaluation Have Different Information Access

Table 20 distinguishes construction from correction. The extractor identifies a statement from an already sampled trace without seeing the image. After review, that statement is inserted as an assistant turn in the prepared replay input. The evaluated model then answers the final solve request and does not generate the inserted turn. Corrected-premise replay replaces only that turn. The visual-verification feedback baseline is separate: it must generate its own critique without access to the construction review's correction.

Table 20: **Inputs and outputs by phase.** This table specifies what information each stage can access. The deployed templates for the shared evaluation set follow below.

Phase	Available input	Retained output
Trace sampling	Image and question	Sampled reasoning and answer
Claim extraction	Question and one sampled trace, without image or gold answer	One candidate visual claim
Construction review	Image evidence and the candidate's source context	Accept/reject decision and corrected visual premise
Replay	Image, question, and inserted assistant premise	Newly generated answer
Self-reflection	Replay conversation and initial answer	Revised or retained answer
Visual feedback	Image, claim, question, initial answer, and independently generated observations and fresh answer	Critique, then a revised or retained answer

H.2 Replay on the Shared Evaluation Set

The following text is exported from the deployed prompt functions for the shared evaluation set. Braced fields mark substituted content. The image is attached to the first user message. In replay, that message contains:

```
Inspect the current image for the visual evidence needed by the
    original question. State only the relevant visual fact or facts; do
    not calculate or give the final answer yet.

Question: {question}
```

The next message has role `assistant` and contains only the selected false or corrected premise. The final user message contains:

```
Solve the question using the current image. Give a brief explanation of
    at most six sentences. End with the final answer alone on the last
    line in this exact format: FINAL ANSWER: <answer>. Replace <answer>
    with your answer, without those angle brackets.
Question: {question}
```

Fresh query uses a single image-bearing user message with this final instruction and no earlier turns. Neutral replay substitutes `This is a conversation.` for the assistant premise. This neutral text is not token-length matched. All six procedures on the shared evaluation set use this answer format. Earlier model-specific replay runs instead request a boxed answer and are reported separately.

H.3 Self-Reflection

Append the initial replay answer as an assistant message, then the following user message. The original image-bearing turn and premise remain in the conversation.

```
Check your preceding answer against the current image. Correct any
    error you find; otherwise retain it.
Solve the question using the current image. Give a brief explanation of
    at most six sentences. End with the final answer alone on the last
    line in this exact format: FINAL ANSWER: <answer>. Replace <answer>
    with your answer, without those angle brackets.
Question: {question}
```

H.4 Visual-Verification Feedback

Every phase uses the same evaluated model checkpoint. First, a fresh image-bearing user message asks for atomic questions about the supplied premise:

```
The following is a visual claim from an earlier reasoning step. List up
    to three simple questions that can verify its visual information.
    Each question must be answerable from the image alone, without the
    earlier reasoning. Do not answer the questions. Use one numbered
    question per line, or N/A if no visual information needs
    verification.
Claim: {premise}
```

For each parsed numbered question, use a separate fresh image-bearing user message. It contains no earlier reasoning or other answers:

```
{atomic_question_1}
Answer briefly.
```

Generate a critique in another fresh image-bearing user message. This example shows three extracted questions. If fewer are extracted, only those question–observation pairs are included. If none are extracted, the evidence field contains `No atomic visual questions were extracted.`

```
Check the visual claim and proposed answer against the image. Explain
    any factual error briefly. The observations and independent answer
    below may also be wrong; do not assume they are reliable.
Question: {question}
```

```
Claim: {premise}
Proposed answer: {initial_answer}
Independent answer: {fresh_answer}
Visual questions and observations:
{atomic_question_1}
Observation: {observation_1}
{atomic_question_2}
Observation: {observation_2}
{atomic_question_3}
Observation: {observation_3}
```

Finally, retain the original replay conversation and append its initial answer as an assistant message. Append the following user message containing the generated critique:

```
Here is an unverified critique of the preceding answer:
{generated_critique}
Check it against the image, then retain or correct your answer.
Solve the question using the current image. Give a brief explanation of
    at most six sentences. End with the final answer alone on the last
    line in this exact format: FINAL ANSWER: <answer>. Replace <answer>
    with your answer, without those angle brackets.
Question: {question}
```

The independent answer is the cached fresh answer from the same model, not a reference answer. The supplied claim is whichever premise is already present in that replay condition. This baseline receives neither a correctness label nor a reviewed replacement premise.

H.5 Feedback-Content Diagnostic

The separate 500-source experiment in Appendix E inserts the following review before the final solve request, after its existing image-reference prefix. Unlike the baseline above, it supplies an annotated correctness label and, in one condition, reviewed corrected facts. The original template names the label field `Verdict`, and its allowed values are `correct`, `incorrect`, and `unknown`. For label-only feedback, `details` is `No additional facts are supplied.` Detailed feedback substitutes corrected facts or task-irrelevant text of the same token length. The image, earlier assistant premise, and solve request are retained.

```
<visual_review>
This review concerns the previous assistant's visual statement, not the
    final answer. The verdict reports whether that statement agrees with
    the image; unknown means no verdict is available.
Verdict: {correctness_label}
Details: {details}
</visual_review>
Use the review and the current image to answer the original question.
{original_solve_request}
```

I Data, Provenance, and Generative-Model Use

Results derive from recorded outputs and versioned scoring reports. The artifact retains source identifiers, image hashes, premise strings, messages, settings, and scoring code. Likelihoods, answers, and model-reviewed labels remain distinct records. Source images retain their upstream licenses, and no additional personal data are collected.

Generative models assisted premise extraction and review, code and manuscript editing, and conceptual figures. AI assistants also implemented, ran, and scored experiments, helped design additional runs and controls, assisted in interpreting results, drafted manuscript text, and retrieved and checked citations. The authors reviewed all AI-assisted content and take full responsibility for the paper. The human truth review, human-confirmed AI-assisted follow-up, and review-based sensitivity analysis are versioned separately from the original model-reviewed shared set (Appendix A.1). The sensitivity analysis reuses existing responses without changing their labels or answer scoring.

J REPLAY WITH ORIGINAL PRECEDING REASONING

The single-premise comparison removes the rest of the sampled reasoning. We also test whether the false–corrected contrast persists when the selected premise follows its original preceding reasoning. Both conditions retain the same prefix, image, and question, and differ only in the appended premise.

Source selection and input construction. Starting from the 79 pairs confirmed by both original human reviewers, we recover the original sampled responses and locate the first complete sentence or list item expressing the selected false visual claim. We retain the exact text before that item, remove the item and all subsequent reasoning and the old answer, and append the unchanged reviewed false or corrected premise. An eligible prefix contains at least eight whitespace-delimited words and no explicit final-answer declaration. Text alignment, inclusion, and exclusion are fixed before the new model outputs are generated. This retains 54 pairs over 44 images. Human review covers only the premise pairs.

Prefixes contain 8–71 words (median 22.5). Each model's tokenizer gives 13–104 tokens (median 32). The construction appends the reviewed premise after the retained text. It does not preserve the complete sampled trace, and other errors can remain in the prefix. The experiment therefore tests robustness to short preceding reasoning rather than a controlled effect of trace length.

Matched evaluation. InternVL3.5-8B and Qwen3-VL-8B use the shared evaluation set's checkpoint revisions, image processing, user turns, greedy decoding, and 2,048-token answer cap. The evaluated model receives neither a correctness label nor the alternative premise. All 54 pairs are complete for both models. We compare the new outputs with the existing single-premise outputs on exactly these sources. Images and processed image tensors match between the two evaluations. Unparsed answers remain in the denominator as incorrect. We use the same two parsers and exact and numeric-tolerance scoring rules as Appendix F. Paired intervals resample source-image clusters with 10,000 draws.

Table 21: **False versus corrected premises after an original prefix.** Each row uses 54 matched pairs. F/C are correct-answer counts with the prefix. Gap is corrected-minus-false accuracy. Gap change subtracts the same-source single-premise gap. Effects are percentage points with 95% image-cluster intervals. P denotes the primary parser and WS the whitespace-tolerant sensitivity. InternVL's two parsers agree.

Model	Parser	Score	F	C	Gap	Gap change
InternVL	P	Exact	6	31	+46.30 [30.19, 62.26]	+7.41 [−5.56, 21.57]
InternVL	P	Relaxed	10	32	+40.74 [25.00, 56.60]	+9.26 [−3.28, 23.08]
Qwen-8B	P	Exact	9	26	+31.48 [14.89, 47.46]	−3.70 [−14.81, 7.69]
Qwen-8B	P	Relaxed	14	29	+27.78 [11.54, 43.64]	−5.56 [−17.31, 6.00]
Qwen-8B	WS	Exact	9	28	+35.19 [19.23, 50.98]	−7.41 [−16.07, 0.00]
Qwen-8B	WS	Relaxed	14	31	+31.48 [16.00, 47.17]	−7.41 [−16.07, 0.00]

Premise contrast and context change. With the primary exact score, single-premise false/corrected counts are 10/31 for InternVL and 7/26 for Qwen-8B. Adding the prefix gives 6/31 and 9/26, respectively. Relative to the single-premise answers, InternVL repairs 2 answers and harms 6 under a false premise, and repairs 2 and harms 2 under a corrected premise. Qwen-8B repairs 3 and harms 1 under a false premise, and repairs 3 and harms 3 under a corrected premise. The corrected-minus-false contrast persists on both models. All gap-change intervals include zero, and the two whitespace-tolerant intervals for Qwen-8B end at 0.00.

InternVL's new prefix answers have no parsing failures or capped outputs. Qwen-8B's primary parser leaves 3 false-condition and 5 corrected-condition answers unparsed. The whitespace-tolerant parser leaves 2 and 2, respectively. Qwen-8B reaches the token cap on 2 false-condition and 2 corrected-condition outputs.