RECOMPUTE OR REUSE? ANALYZING STALE REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can answer a question about an image correctly in isolation yet change their answer when given reasoning from an earlier image. We study which parts of this reasoning affect the answer and whether their influence continues after the model answers correctly. We call the tendency of prior reasoning to favor the old answer despite changed visual evidence a *textual shortcut*. Across 16 VLMs, we construct contexts containing the current image and an earlier response. We compare deleting sentences selected by fixed evidence rules, deleting non-evidence sentences, and masking explicit answer cues. Deleting the selected sentences shifts answer preference toward the current answer in all 16 models relative to non-evidence deletion; changing sentence order has a smaller and more variable effect. After an initially correct response, deleting textual evidence for the current answer causes a larger increase in old answers when the earlier reasoning is retained. When the evidence is kept, initially correct responses show little old-answer repetition, although accuracy falls on follow-up calculations. Responses that initially retain the old answer show much more repetition and more follow-up results matching calculations with the old value. These findings show which prior text affects visual updating and how this influence varies with the initial answer and the follow-up question.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly expected to check their reasoning and revise answers when visual evidence changes (Madaan et al., 2023; Kamoi et al., 2024). To answer correctly after an image changes, a model must use the updated visual information. VisualSwap shows that VLMs often retain their earlier answer during re-examination and that retaining more prior CoT reduces update accuracy (Shi et al., 2026). Recent methods encourage visual re-attention (Jian et al., 2025; Yang et al., 2026). We ask two further questions: *which parts of the retained reasoning affect the answer, and how does this influence appear in subsequent responses?*

A chain of thought (CoT) contains observations, numerical values, relations, and intermediate calculations as well as a final answer. After the image changes, these statements remain in the context and may conflict with the current image. Fig. 1 illustrates how an answer can agree with the current image or with information retained in the earlier reasoning. We use *textual shortcut* to describe a tendency for prior reasoning to steer the model toward the old answer despite conflicting evidence in the current image.

We study this conflict using paired images that share a question but imply different answers (Fig. 2). The model first answers the old image I^-. We then insert that response R^- into a context containing the current image I. Comparing answers with and without the prior response measures its influence. Editing the prior text allows us to compare the effects of deleting selected evidence sentences, deleting other sentences, and masking explicit answer cues.

Experiments across 16 VLMs yield two main findings. First, **the selected evidence sentences consistently influence answers**. Removing them shifts preference toward the current answer more than the tested non-evidence deletion or answer-cue masking, with a positive E–C contrast in all 16 models. Removing more evidence sentences also increases the fraction of samples that favor the current answer. Changing sentence order has smaller and less consistent effects on generated answers.

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.

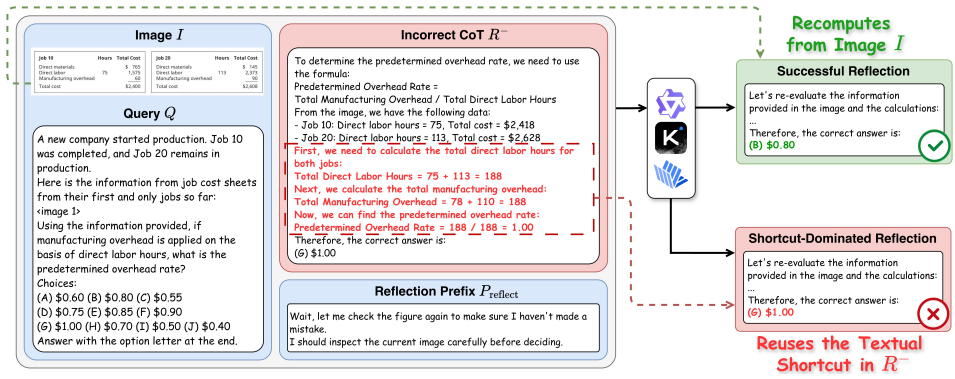

Figure 1: **Textual shortcut during VLM self-reflection.** Illustrative contrast between an answer supported by the current image and an answer consistent with stale prior text.

Second, **an initially correct answer can still be followed by errors associated with earlier reasoning**. Deleting evidence for the current answer increases old-answer generation more when earlier reasoning is retained than in the matched direct condition. With the evidence kept, initially correct responses show little repetition of the old answer, but lower accuracy on follow-up calculations. Responses that initially retain the old answer show much more repetition and more follow-up results consistent with the old numerical value.

Our contributions are threefold:

- **Identifying which prior text affects visual updating.** Across 16 VLMs, we compare content deletion, the amount deleted, and sentence order. Removing selected evidence sentences has larger effects than deleting other sentences or masking matched answer cues.
- **Examining prior influence after a correct answer.** Removing textual evidence for the current answer produces a larger increase in old answers when earlier reasoning is retained, across all 16 models.
- **Distinguishing errors on follow-up questions.** Initially correct responses show little old-answer repetition but lower accuracy on follow-up calculations. Responses that initially retain the old answer show both repetition and calculation results consistent with the old.

2 Related Work

Visual Re-examination and Grounded Reflection. Controlled visual tests and perturbations reveal that a plausible answer or reflective statement does not guarantee sensitivity to the current image (Shi et al., 2026; Sun et al., 2026; Guan et al., 2024). Longer multimodal reasoning can further weaken visual grounding (Xu et al., 2025). Accordingly, recent methods encourage visual re-attention through reflection training, attention-aware rewards, or explicit visual-token revisiting (Jian et al., 2025; Chu et al., 2025; Yang et al., 2026; Cheng et al., 2025; Ji & Lu, 2025; Tang et al., 2026). Training-free alternatives counter language-prior dominance through visual contrastive decoding or targeted hidden-state intervention (Leng et al., 2024; Ji et al., 2026). Complementary work evaluates whether intermediate perception steps, rather than only final answers, remain visually faithful (Uppaal et al., 2026; Zhang et al., 2026). These studies focus on current-image use. VisualSwap further varies the retained proportion of prior CoT and finds worse re-examination with more retained text, whereas VI-Probe studies a static perception–memory conflict. Building on VisualSwap's image pairs, we compare edits to selected evidence, answer cues, and sentence order, then test the influence of earlier reasoning after a correct response and on new calculations.

Self-Correction and Belief Revision. Self-correction methods refine outputs through self-feedback, verbal reflection, or tool-grounded critique (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024; Pan et al., 2024). However, intrinsic correction without reliable external feedback is often unstable and can degrade reasoning (Huang et al., 2024; Kamoi et al., 2024; Zhang et al., 2025; He et al., 2025). Belief-revision studies further distinguish updating when premises change from maintaining

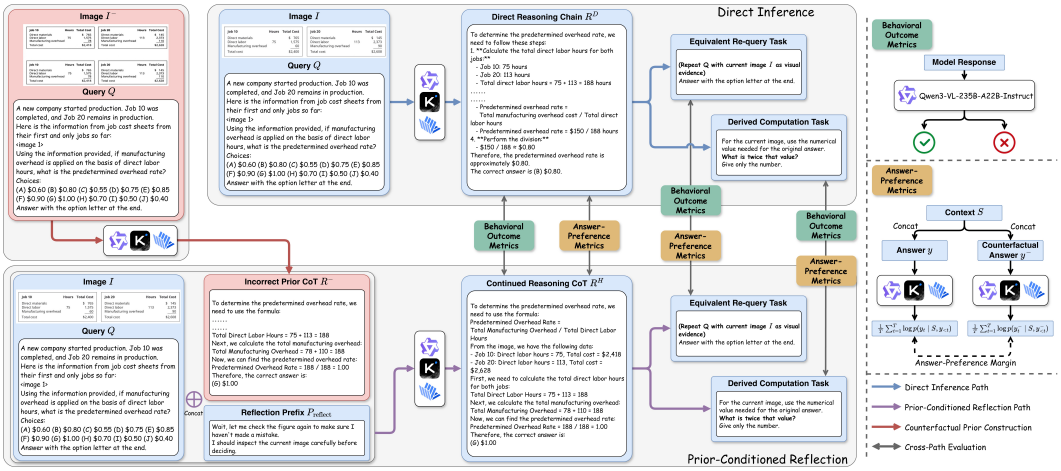

Figure 2: **Controlled evaluation framework.** The model first generates reasoning from the old image (red). With the current image, we compare direct answers (blue) and answers given the prior reasoning (purple), then ask follow-up questions with the current image and preceding responses retained (gray). We evaluate generated answers and candidate-answer scores.

a valid conclusion, while correction analyses expose a similar trade-off between critique and answer retention (Wilie et al., 2024; Zhang et al., 2024b; Yang et al., 2025). Together, these results suggest that successful revision depends on both detecting a conflict and replacing the premises that support the earlier conclusion. Our setting makes the changed evidence visual and asks whether an initially correct answer is followed by reduced sensitivity to prior text.

Prior Reasoning as Persistent Context. CoT is not necessarily a faithful explanation of a model's decision, although interventions on the trace show that it can still causally influence the answer (Turpin et al., 2023; Lanham et al., 2023). Conversational history can also induce persistent generation tendencies (Simhi et al., 2026), while correction and trust vary with whether the same claim appears as the model's own thought, a user message, or another role (Chen et al., 2026; Pan et al., 2026). These role effects motivate studying both how prior reasoning is presented and which information it contains. We examine how reasoning generated from an earlier image affects answers after the visual evidence changes.

3 Experimental Setup

We study how prior reasoning affects answers to an updated image and to follow-up questions. Fig. 2 holds the current image and question fixed while varying the retained text. Each comparison uses the same model and task, linking answer changes to differences in context.

3.1 Image Pairs and Sample Selection

We pair a current task $\mathcal{T} = (I, Q, y)$ with a counterfactual $\mathcal{T}^- = (I^-, Q, y^-)$. The images share the question and answer space, but differ in the visual information needed to answer it, so $y^- \neq y$. Thus, an answer that was correct for the old image becomes incorrect for the current one. We use human-checked VS-Bench pairs (Shi et al., 2026) from MathVista-MINI, MathVerse-MINI, MathVision, and MMMU-Pro (Lu et al., 2024; Zhang et al., 2024a; Wang et al., 2024; Yue et al., 2025), covering mathematical and other visual reasoning tasks.

The model $\mathcal{M}$ first answers the old image:

$$(R^-, \hat{y}^-) = \mathcal{M}(I^-, Q). \tag{1}$$

Here R^- is the response and $\hat{y}^-$ its answer. We retain a pair when $\hat{y}^- = y^-$ and the model answers the current image correctly without prior context. This selection establishes that both image versions are individually solvable by the model. We examine whether adding the old response changes an

Table 1: **Experiment overview.** D: direct inference; H: with prior reasoning.

Analysis	Context and edit	Response evaluated
Current image	Completed prior response; new user turn with/without current image	Answer on the new turn
Content and order	Edited prior reasoning; same assistant turn and reflection cue	Continuation from the edited prefix
Current evidence	Initially correct D/H responses; current evidence kept/deleted	New continuation from each context
Follow-up questions	D/H responses kept; repeated question or new calculation	Follow-up answer, grouped by H's initial answer

answer the model could produce correctly. The retained reasoning was correct for the old image but conflicts with the current one. Each model is evaluated on its eligible pairs (App. H).

3.2 Inputs and Follow-Up Tests

D answers the current image directly. H inserts the separately generated old-image response R^- into a context with the current image and question, then continues the assistant prefix with a fixed reflection cue P_{reflect}:

$$(R^D, \hat{y}^D) = \mathcal{M}(I, Q), \tag{2}$$

$$(R^H, \hat{y}^H) = \mathcal{M}(I, Q, R^-, P_{\text{reflect}}). \tag{3}$$

Thus, H–D measures the effect of adding the prior response and reflection cue. The reconstructed H context contains the current image; the old image is used to generate R^-. Tab. 1 summarizes the context and response in each test; full prompts appear in App. C.

3.3 Selecting and Editing Evidence Sentences

The content experiments first remove the final answer using a fixed pattern-based preprocessing rule, then split the remaining prior reasoning into sentences. This remaining text is the baseline N. A sentence is assigned to E^- if it contains numerical or mathematical symbols, an option/answer reference, or a term from fixed visual or reasoning word lists. The selected sentences include observations, quantities, and deductions, as well as answer cues appearing within reasoning. The full rules, including the preprocessing fallback, appear in App. D.3.

For example, the MMMU-Pro case in Fig. 9 contains the old calculation $7\%+5\%-1\%+5\% = 16\%$, whereas the current image implies 13%. The numerical premises and calculation are selected by the evidence rules. Removing a final answer leaves these statements available in the reasoning prefix, which motivates testing content removal separately from masking explicit answer cues.

We compare four versions of this prefix. N retains the preprocessed reasoning; E removes all selected evidence-bearing sentences; C deletes other sentences using E's removed-word count as a target budget; and A masks rule-matched answer and option cues within the reasoning with `[ANS]` or `[OPT]`. A targets expressions such as option letters and boxed answers; it can overlap with cues inside E^-. All four conditions share the initial removal of the final answer.

For C, length is measured by whitespace-separated word count. Whole non-evidence sentences are randomly ordered with a fixed sample seed and removed until the number of removed words reaches that in E or the available sentences are exhausted. C typically removes less text because fewer non-evidence sentences are available (App. D.3). A further experiment removes increasing fractions of selected evidence sentences, while the order experiment reverses their order with sentence content retained. These interventions measure sensitivity to explicitly defined text edits.

3.4 Evaluation Metrics

Generated answers. We assess free-generation behavior and teacher-forced answer preference. For free generation, Qwen3-VL-235B-A22B-Instruct independently judges whether each response

$\mathcal{M}(S_i)$ matches the current target y_i or counterfactual target y_i^-. Let $J(r, a) \in \{0, 1\}$ indicate a semantic match. Over N eligible pairs, we compute

$$\mathrm{VUR}(S) = \frac{1}{N}\sum_{i=1}^{N} J(\mathcal{M}(S_i), y_i)\,, \qquad \mathrm{PAR}(S) = \frac{1}{N}\sum_{i=1}^{N} J\big(\mathcal{M}(S_i), y_i^-\big)\,. \tag{4}$$

Visual update rate (VUR) measures agreement with the current answer, whereas prior-answer rate (PAR) measures agreement with the counterfactual answer. The two judgments define four outcomes: *Correct* (current answer only), *Old answer* (prior answer only), *Both*, and *Other* (neither target). VUR and PAR include the Both category and therefore need not sum to 100%. These outcome labels are assigned from generated answers. The same judge establishes eligibility. Rates are reported in % and differences in percentage points (pp).

Answer scores. To complement generated answers, we score each candidate $a = (a_1, \ldots, a_{|a|})$ by its mean token log-probability under teacher forcing:

$$\ell(a \mid S) = \frac{1}{|a|}\sum_{t=1}^{|a|} \log p_{\mathcal{M}}(a_t \mid S, a_{<t}). \tag{5}$$

Following likelihood-based evaluations (Malladi et al., 2023; Ren et al., 2023; Li et al., 2024), we aggregate available answer forms into target scores $s(y \mid S)$ (App. B.7) and define

$$m(S) = s(y \mid S) - s(y^- \mid S). \tag{6}$$

Positive (negative) $m(S)$ favors the current (prior) answer. This score is computed from the evaluated model's token probabilities, independently of the semantic judge. The fraction of samples with $m(S) > 0$ measures preference between the two specified candidates; VUR and PAR measure the separately generated responses.

Sample selection and averages. Content tests use all eligible pairs. Order tests require enough sentences for the specified reordering. Deleting current evidence is tested on initially correct responses, and follow-up calculations use samples with numerical answers. Counts for each analysis are reported in App. H. For an average across models, we first compute each model's paired effect and weight all models equally. For a pooled result, we weight all model–sample pairs equally.

4 Which Parts of Prior Reasoning Affect Visual Updating?

We first compare answers with and without the current image. We then remove different parts of the prior reasoning, vary the amount removed, and change sentence order. These tests show which text edits affect answer scores and generated answers.

Table 2: **Prior reasoning can produce the old answer without the current image.** Matched new-turn conditions within each model's eligible cohort; entries are percentages. More results appear in Tab. 9.

Model	$\mathbf{Q} + \mathbf{R}^-$, no I VUR / PAR	$\mathbf{I} + \mathbf{Q} + \mathbf{R}^-$ VUR / PAR
Qwen3-VL-8B (Thinking)	0.6 / 53.3	22.6 / 7.0
Kimi-VL-A3B (Instruct)	1.1 / 97.2	56.9 / 39.8
Gemma-4-31B (IT)	1.1 / 92.7	88.0 / 4.4

4.1 The Role of the Current Image

We first ask whether the prior response R^- is sufficient to produce the old answer. We compare a *prior-only* condition ($Q + R^-$, no I) with the matched *image-present* condition. Tab. 2 reports three representative model families. Without I, VUR remains 0.6%–1.1%, whereas PAR reaches 53.3%–97.2%, showing that the model often repeats the old answer from R^-. Restoring I shifts predictions toward the current answer, but recovery remains heterogeneous (VUR 22.6%–88.0%, PAR 4.4%–39.8%). Thus, both the prior text and the current image affect the generated answer. The following content tests use the same assistant turn to isolate changes to the prior reasoning prefix.

4.2 Removing Prior Text

What do the text controls test? An old answer may be favored because the reasoning contains explicit answer cues, because the additional text changes the context, or because it supplies observations and calculations that support that answer. These possibilities motivate editing different parts

Table 3: **Deleting evidence sentences changes answers more than the tested controls.** N retains prior reasoning after removing the final answer; E removes E^{-}; C deletes non-evidence sentences; A masks matched answer cues within N. The rightmost columns report E–C free-generation changes.

		Δm			Generation (pp)	
Model	**Type**	**E–N**	**E–C**	**A–N**	ΔVUR	ΔPAR
Qwen2.5-VL-7B	Instruct	+2.019	+1.948	+0.064	+46.12	-67.05
Qwen3-VL-8B	Instruct	+2.699	+2.617	+0.237	+42.27	-72.95
	Thinking	+2.826	+2.832	+0.237	+35.39	-84.65
Qwen3-VL-32B	Instruct	+3.142	+3.075	+0.045	+54.74	-81.52
	Thinking	+2.995	+3.004	+0.159	+54.47	-82.11
Qwen3.5-4B	–	+1.548	+1.612	+0.057	+45.49	-70.82
Qwen3.5-9B	–	+1.345	+1.372	+0.033	+47.11	-65.29
Qwen3.5-27B	–	+1.600	+1.535	+0.050	+58.80	-72.66
Qwen3.6-27B	–	+1.635	+1.591	+0.053	+53.53	-71.00
InternVL2-8B	Instruct	+0.926	+0.913	+0.008	+26.95	-33.33
Gemma-4-12B	Base	+0.902	+0.911	+0.115	+50.00	-65.00
	IT	+2.221	+2.067	+0.025	+26.64	-28.38
Gemma-4-31B	Base	+1.017	+0.996	+0.061	+49.26	-66.18
	IT	+3.334	+3.152	-0.005	+25.45	-25.09
Kimi-VL-A3B	Instruct	+1.742	+1.619	-0.029	+32.04	-49.17
	Thinking	+3.145	+2.913	+0.078	+51.91	-66.12
Mean across models		**+2.069**	**+2.010**	**+0.074**	**+43.76**	**-62.58**

of the same prior response while keeping the current image, question, and reflection cue fixed. All four conditions start from reasoning with the final answer removed, so the comparison focuses on information remaining in the reasoning itself.

Design. We apply the N/E/C/A interventions defined in Sec. 3.3. A–N tests the influence of matched answer cues by masking them within the reasoning. E–C is the primary comparison between deleting selected evidence and deleting other sentences under the budget rule in Sec. 3.3. It measures the difference between these two edits. The dose and order tests ask whether answer preference varies with how much of this evidence remains and how its sentences are arranged. For conditions X and Y, $\Delta m_{\mathrm{X-Y}} = m(S_X) - m(S_Y)$, with analogous VUR and PAR contrasts.

Deleting evidence changes answer scores and generated answers. Tab. 3 reports an E–C margin shift, averaged equally across models, of $+2.010$, positive in all $16/16$ models. The E–N shift is $+2.069$, whereas C–N is $+0.059$. Answer-cue masking produces a smaller mean A–N shift of $+0.074$, with heterogeneous effects across models. Thus, deleting selected sentences has a larger effect than either the tested non-evidence deletion or the targeted cue masking.

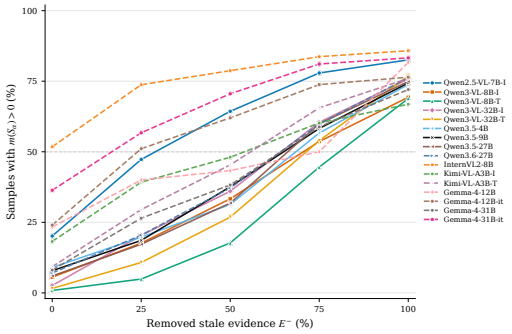

Figure 3: **Removing more evidence increases preference for the current answer.** The proportion of samples with $m(S_\alpha) > 0$ increases monotonically at the model level in all 16 VLMs.

The same comparison changes free-generation outcomes: E–C raises VUR by 43.76 pp and lowers PAR by 62.58 pp, with positive VUR effects in all $16/16$ models. The corresponding E–N changes are $+45.22$ and -64.22 pp. Relative to C, Other responses also increase, from 8.55% to 26.50% (Tab. 12). Evidence deletion therefore increases both correct answers and answers matching neither target. VUR and PAR together capture these distinct consequences.

Table 4: **Effects of sentence order.** O/N/NE: evidence-reordered/natural/non-evidence-reordered text. VUR/PAR changes compare O with N (pp).

Model	$\Delta m_{\mathrm{O-N}}$	$\Delta m_{\mathrm{O-NE}}$	ΔVUR	ΔPAR
Qwen2.5-VL-7B (Instruct)	+0.645	+0.641	+13.15	-15.94
Qwen3-VL-8B (Instruct)	+0.930	+0.741	+4.10	-6.41
Qwen3-VL-8B (Thinking)	+0.771	+0.444	+1.92	-11.94
Qwen3-VL-32B (Instruct)	+1.141	+0.808	+13.27	-15.88
Qwen3-VL-32B (Thinking)	+0.478	+0.446	+3.68	-5.79
Qwen3.5-4B	+0.104	+0.099	+2.59	-3.45
Qwen3.5-9B	+0.065	+0.076	+3.72	-3.72
Qwen3.5-27B	+0.045	+0.044	+3.01	-3.01
Qwen3.6-27B	+0.045	+0.032	+1.87	-3.75
InternVL2-8B (Instruct)	+0.182	+0.199	+4.08	-22.45
Gemma-4-12B (Base)	-0.049	+0.013	-7.50	+10.00
Gemma-4-12B (Instruction-tuned)	+0.878	+0.686	+17.11	-9.65
Gemma-4-31B (Base)	+0.138	-0.016	+6.78	-7.63
Gemma-4-31B (Instruction-tuned)	+1.378	+0.673	+17.88	-12.41
Kimi-VL-A3B (Instruct)	-0.138	-0.155	+0.00	+0.00
Kimi-VL-A3B (Thinking)	+0.565	+0.343	+7.73	-8.84
Mean across models	**+0.449**	**+0.317**	**+5.84**	**-7.55**

Removing more evidence further changes answer preference. As the removed fraction of E^- increases from 0% to 100%, the fraction of samples preferring the current candidate, $m(S_\alpha) > 0$, increases monotonically within each of 16 models (Fig. 3). Thus, removing more evidence sentences increases the fraction of samples in which the current candidate scores higher than the old candidate.

4.3 Changing Sentence Order

Design. To test sensitivity to sentence order, we reverse the selected evidence-bearing sentences (O), retaining their content. We compare O with the natural-order prefix (N) and with reversal of non-evidence sentences (NE). The matched O–NE comparison distinguishes the tested evidence reorder from the tested non-evidence reorder. Each contrast uses samples with enough sentences for the required reordering.

Sentence order has smaller effects on generated answers. Tab. 4 reports an O–N margin shift averaged equally across models of $+0.449$ and an O–NE shift of $+0.317$, each positive in $14/16$ models. For generated answers, O–N increases VUR by 5.84 pp and decreases PAR by 7.55 pp. The more specific O–NE VUR change is $+2.05$ pp, with a 95% confidence interval of $[-1.25, +4.91]$ pp. The evidence for an additional ordering effect is therefore stronger in candidate preference than in free generation.

Main finding. The content, dose, and order comparisons together identify which aspects of prior reasoning most strongly affect visual updating. Deleting the selected evidence sentences shifts answers toward the current image more than the tested non-evidence deletion or answer-cue masking. Removing more of these sentences further increases preference for the current answer, whereas changing their order has smaller and less consistent effects on generated answers. These edits connect the textual shortcut to retained observations and calculations. The effects also appear across prior-CoT length quartiles (Fig. 5), extending the analysis of retained context in VisualSwap to specific edits of the reasoning. We next ask how prior reasoning affects later answers, including cases where the model initially answers the current image correctly.

5 How Prior Reasoning Affects Later Answers

A model can answer the current image correctly while reasoning from the old image remains in its context. Whether that earlier reasoning still affects subsequent answers requires a separate test.

Table 5: **Deleting current evidence increases old answers more when prior reasoning is retained.** PAR (%) is measured on new continuations with evidence kept (F) or deleted (W), under direct inference (D) or with prior reasoning (H). Δ_{PAR} (pp) is defined in Eq. 7.

Model	Direct (D)		With prior reasoning (H)		Δ_{PAR}
	Kept (F)	Deleted (W)	Kept (F)	Deleted (W)	
Qwen3.5-27B	1.0	4.0	10.6	84.8	+71.2
InternVL2-8B	2.0	5.1	13.3	49.0	+32.7
Aggregate: all 16 VLMs	1.34	7.17	9.81	69.01	**+53.38**

We examine this in two complementary ways. First, we delete evidence sentences from an initially correct response to test sensitivity to prior reasoning when the textual basis for the current answer is reduced. Second, we keep the preceding responses intact and ask follow-up questions to examine what happens with that basis still available. In the follow-up test, we compare samples that initially produced the current answer with those that retained the old answer.

5.1 Deleting Evidence from a Correct Response

First, select correct responses. We select samples for which H, the condition with prior reasoning, produces the current answer only. Each is paired with the correct direct response D for the same task. Samples are selected before any text is deleted.

Then, edit the response and generate again. For each condition, F keeps the evidence sentences in the current response and W deletes them. We keep the current image, question, answer-format text, and the history belonging to each condition. The model then generates a new continuation, on which we measure PAR. The initial correct response is used to select samples; the new response is used to measure the effect of deletion. The current image is present in both F and W.

Compare the increases in old answers. Deleting useful sentences can disrupt a continuation even without prior reasoning. The direct condition provides a comparison for this effect: we ask whether deletion increases old answers more when the context also contains reasoning from the old image. The current image remains available throughout, so the change concerns the textual basis for the answer. Let S_H^F, S_H^W denote the contexts with prior reasoning, with current evidence kept or deleted; S_D^F, S_D^W are the corresponding direct contexts. We calculate the PAR increase caused by deletion in H and subtract the increase in D:

$$\Delta_{\mathrm{PAR}} = \big[\mathrm{PAR}(S_H^W) - \mathrm{PAR}(S_H^F)\big] - \big[\mathrm{PAR}(S_D^W) - \mathrm{PAR}(S_D^F)\big] \tag{7}$$

A positive Δ_{PAR} means that deleting current evidence increases old-answer generation more when prior reasoning is present. Tab. 5 shows this difference. Across the selected samples, PAR rises from 9.81% to 69.01% in H and from 1.34% to 7.17% in D, giving $\Delta_{\mathrm{PAR}} = 53.38$ pp. Averaging models equally gives 49.74 pp, with a positive difference in all $16/16$ models (Fig. 4 in the appendix). The 9.81% rate is measured on newly generated H responses with evidence kept, after selecting samples whose initial response was correct. These results show sensitivity to earlier reasoning after an initially correct response when its supporting text is reduced. The larger increase in old answers with prior reasoning motivates a complementary question: how does the model respond to follow-ups when the correct response and its evidence remain intact?

5.2 Follow-Up Questions with Evidence Kept

Follow-up questions and sample groups. We retain the current image and preceding responses in D and H, then repeat the question or request a new calculation (Apps. E.1 and E.2). Repeating the question tests old-answer repetition; a new calculation compares results derived from the current and old quantities. For example, doubling these quantities gives distinct targets (Fig. 16). The prompt does not supply the old value again. We group samples by H's initial answer and apply the same grouping to paired D responses, comparing correct and old-answer groups while keeping Other separate. Calculations require numerical answers. Within each follow-up, D and H answer the same question for the same samples, so the comparison holds the requested calculation fixed.

Table 6: **Follow-up results by initial answer.** Changes compare H with D; VUR/PAR changes are in pp. Share (%) gives each initial-answer group's proportion within the model and follow-up task.

Follow-up	Initial answer	Share (%)	Δm	ΔVUR	ΔPAR
Qwen3-VL-8B (Instruct)					
Repeated question	Correct	70.05	-0.707	-1.72	+0.34
	Old answer	25.36	-7.759	-47.14	+37.62
New calculation	Correct	71.85	-0.562	-11.11	+0.19
	Old answer	23.95	-7.279	-82.46	+66.67
Kimi-VL-A3B (Instruct)					
Repeated question	Correct	28.18	-0.904	-8.82	+2.94
	Old answer	66.30	-1.854	-57.08	+47.92
New calculation	Correct	30.48	-0.320	-10.42	+1.04
	Old answer	63.81	-2.192	-56.72	+49.75
Aggregate: all 16 VLMs					
Repeated question	Correct	59.05	**-0.547**	-2.14	+0.98
	Old answer	34.09	-3.896	-39.88	**+35.05**
New calculation	Correct	60.82	-0.391	-18.99	+0.07
	Old answer	32.30	-3.756	-65.19	**+45.28**

Tab. 6 reports pooled H–D differences within each test: negative Δm shifts preference toward the old candidate, while VUR and PAR measure agreement with the current and old targets, respectively.

Initially correct answers. Repeating the question reduces VUR by 2.14 pp and increases PAR by only 0.98 pp. New calculations reduce accuracy by 18.99 pp, but results matching the old value increase by just 0.07 pp. Thus, prior reasoning produces little additional old-answer repetition in this group; most of the calculation accuracy loss consists of other errors.

Answers that initially retain the old answer. Repeating the question reduces VUR by 39.88 pp and increases PAR by 35.05 pp. New calculations reduce accuracy by 65.19 pp and increase results matching the old value by 45.28 pp, with the latter increase positive in 14 of 16 models. Here, prior reasoning affects both answer repetition and results of new operations on the old quantity.

Control for additional conversation history. Neutral history preserves the added history structure while omitting old evidence (App. E.3). In the old-answer group, PAR with neutral history, D, and H is respectively 1.02%, 0.88%, and 35.93% on repeated questions, and 8.57%, 8.83%, and 54.11% on calculations. For initially correct samples, neutral history reduces calculation accuracy by 5.93 pp relative to D, versus 18.99 pp for H. Additional history can lower accuracy, but old reasoning produces a larger effect. The close D and neutral-history rates also show that the large rise in old-target matches occurs specifically in the condition containing old reasoning.

Main finding. The two tests reveal different consequences of retaining earlier reasoning in subsequent answers. After an initially correct answer, deleting current evidence increases old answers more when prior reasoning is retained. With evidence kept, follow-ups show little additional old-answer repetition but lower calculation accuracy, mainly through other errors. Initially old answers are followed by both repetition and calculations consistent with the old value. Evaluating self-correction therefore requires tracking initial correctness and the types of subsequent errors.

6 Conclusion

Across 16 VLMs, deleting selected evidence sentences affects visual updating more than non-evidence deletion and answer-cue masking; sentence-order effects are weaker. After a correct response, deleting current evidence increases old answers more with prior reasoning. Keeping the response intact yields little additional old-answer repetition but lower calculation accuracy. Initially old answers are followed by repetition and calculations matching the old value. These results support a behavioral textual shortcut and motivate evaluating visual updating across prior contexts and follow-up questions, while distinguishing old-answer persistence from other types of errors.

AI Assistance Disclosure

Generative AI tools were used to draft selected passages, improve writing clarity, formulate mathematical claims, help interpret experimental results, and retrieve and identify relevant literature. All AI-assisted text and suggested references were reviewed, revised, and verified by the authors against the original sources. The authors take full responsibility for the final content of this paper.

References

Kuan-Yen Chen, Fang-Yi Su, and Jung-Hsien Chiang. The self-correction illusion: LLMs correct others but not themselves, 2026. URL https://arxiv.org/abs/2606.05976.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8876–8892. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.447. URL https://aclanthology.org/2025.naacl-long.447/.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information, 2025. URL https://arxiv.org/abs/2505.23558.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WSrRF5Wy6v.

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*, 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.

Jiayi He, Hehai Lin, Qingyun Wang, Yi R. Fung, and Heng Ji. Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6405–6421. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.331. URL https://aclanthology.org/2025.findings-acl.331/.

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 *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Jiazhou Ji and Xinru Lu. ReFLAIR: Enhancing multimodal reasoning via structured reflection and reward-guided learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25401–25413. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1384. URL https://aclanthology.org/2025.findings-emnlp.1384/.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ji_CausalLens_Sensitivity-Guided_Multi-Head_Causal_Intervention_for_Hallucination_Mitigation_in_Large_CVPR_2026_paper.html.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9251–9270, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.470. URL https://aclanthology.org/2025.emnlp-main.470/.

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/.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023. doi: 10.48550/ARXIV.2307.13702. URL https://doi.org/10.48550/arXiv.2307.13702.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 2819–2834, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

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, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. doi: 10.1162/tacl_a_00660. URL https://aclanthology.org/2024.tacl-1.27/.

Xu Pan, Jingxuan Fan, Zidi Xiong, Ely Hahami, Jorin Overwiening, and Ziqian Xie. User-assistant bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9218–9241. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.449. URL https://aclanthology.org/2026.findings-acl.449/.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang (eds.), *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops, 16 December 2023, New Orleans, Louisiana, USA*, volume 239 of *Proceedings of Machine Learning Research*, pp. 49–64. PMLR, 2023. URL https://proceedings.mlr.press/v239/ren23a.html.

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, 2026. URL https://arxiv.org/abs/2605.15864. Accepted as an oral presentation at ICML 2026.

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, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Adi Simhi, Fazl Barez, Martin Tutek, Yonatan Belinkov, and Shay B. Cohen. Old habits die hard: How conversational history geometrically traps LLMs, 2026. URL https://arxiv.org/abs/2603.03308.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do VLMs perceive or recall? probing visual perception vs. memory with classic visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25861–25870, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_Do_VLMs_Perceive_or_Recall_Probing_Visual_Perception_vs._Memory_CVPR_2026_paper.html.

Liyan Tang, Fangcong Yin, and Greg Durrett. Visually grounded self-reflection for vision-language models via reinforcement learning, 2026. URL https://arxiv.org/abs/2607.02490.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. Journey before destination: On the importance of visual faithfulness in slow thinking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4147–4168. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.194. URL https://aclanthology.org/2026.eacl-long.194/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169, 2024. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/

hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.586. URL https://aclanthology.org/2024.emnlp-main.586/.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Y. Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/777db387a5ccb131ba8c7cd155166b85-Abstract-Conference.html.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual refocusing in MLLM reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (14):11694–11702, 2026. doi: 10.1609/aaai.v40i14.38154. URL https://ojs.aaai.org/index.php/AAAI/article/view/38154.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. Confidence v.s. critique: A decomposition of self-correction capability for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3998–4014. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.203. URL https://aclanthology.org/2025.acl-long.203/.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, and Han Qiu. Understanding the dark side of LLMs' intrinsic self-correction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27066–27101, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1314. URL https://aclanthology.org/2025.acl-long.1314/.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186. Springer, 2024a. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3602–3622. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.197. URL https://aclanthology.org/2024.acl-long.197/.

Yongchang Zhang, Oliver Ma, Tianyi Liu, Guangquan Zhou, and Yang Chen. See it, say it, sorted: An iterative training-free framework for visually-grounded multimodal reasoning in LVLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_See_It_Say_It_Sorted_An_Iterative_Training-Free_Framework_for_CVPR_2026_paper.html.

Appendix Overview. The appendix records study limitations, implementation and scoring details, prompt construction, full diagnostic results and protocol controls, the FSAF intervention, benchmark provenance, cohort counts, and representative cases.

A LIMITATIONS

The study uses reconstructed contexts and emphasizes mathematical, diagrammatic, and reasoning-heavy visual tasks. Extending the analysis to natural-image, open-ended tasks and sequential dialogues retaining both old and updated images remains an important next step. Eligibility is model-dependent and requires correct direct answers on both images; it may select easier instances and yields different cohorts across models. The paired comparisons estimate effects within these cohorts. The interventions establish behavioral dependence on prior text; the internal computations implementing that dependence remain an open question.

The evidence partition uses heuristic rules and may include answer cues or miss relevant spans. Evidence and non-evidence deletion remove different amounts of text, so their contrast combines content and deletion amount. Sentence reordering can change coherence and position along with reasoning organization. Semantic-match judgments are subject to error, and candidate scores measure relative preference between two targets. App. F further examines attention access to prior text through FSAF, with model coverage and intervention trade-offs reported there.

B IMPLEMENTATION AND EVALUATION SETUP

B.1 EVALUATED CHECKPOINTS

Tab. 7 gives the exact public checkpoint identifiers used in the 16-model diagnosis. The *Type* column describes the evaluated checkpoint or decoding mode. Qwen3.5 and Qwen3.6 provide unified post-trained checkpoints; we evaluated their non-thinking mode. For Gemma 4, the suffix `-it` denotes the instruction-tuned checkpoint, while the checkpoint without that suffix is the base variant.

B.2 INFERENCE BACKENDS AND HARDWARE

We use vLLM for standard Qwen2.5-VL/Qwen3-VL inference and Transformers for the remaining model families and probes that require direct control of rendered prefixes or candidate logits. The semantic judge is served through an OpenAI-compatible vLLM endpoint, while FSAF uses Transformers eager-attention hooks (Appendix F.3).

Model inference and probing were run on a server with two Intel Xeon 6767P CPUs, 800 GiB of system memory, and eight NVIDIA B200 GPUs (180 GB each), under CentOS Linux 7 with Linux kernel 6.6.0. The 235B semantic judge was served on four B200 GPUs with tensor parallelism of 4. The software environment used CUDA 12.8, PyTorch 2.10.0, vLLM 0.19.0, and model-family-specific Transformers environments ranging from version 4.49.0 to 5.13.0. Within every paired comparison, the model, image preprocessing, decoding configuration, and maximum generation length are held fixed; only the specified intervention changes.

B.3 GENERATION AND CHAT-TEMPLATE CONTROLS

Each checkpoint is rendered with its own processor and native chat template. Images occupy the same conversation position in matched conditions, and the paired images have identical resolution (Appendix G). Model-specific special tokens are preserved; Appendix C uses abstract role markers to expose the common conversation structure. For stochastic decoding runs, all conditions within a matched model–sample comparison share the same decoding hyperparameters. FSAF and its ordinary-reflection baseline use a maximum of 256 generated tokens; the final semantic judge uses temperature zero and a strict JSON schema.

Table 7: **Exact model checkpoints used in the 16-model diagnosis.** Each URL resolves to the corresponding public Hugging Face repository. NT denotes non-thinking decoding for a unified post-trained checkpoint.

Checkpoint	Type	Hugging Face repository
Qwen/Qwen2.5-VL-7B-Instruct	Instruct	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Qwen/Qwen3-VL-8B-Instruct	Instruct	https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
Qwen/Qwen3-VL-8B-Thinking	Thinking	https://huggingface.co/Qwen/Qwen3-VL-8B-Thinking
Qwen/Qwen3-VL-32B-Instruct	Instruct	https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
Qwen/Qwen3-VL-32B-Thinking	Thinking	https://huggingface.co/Qwen/Qwen3-VL-32B-Thinking
Qwen/Qwen3.5-4B	Unified (NT)	https://huggingface.co/Qwen/Qwen3.5-4B
Qwen/Qwen3.5-9B	Unified (NT)	https://huggingface.co/Qwen/Qwen3.5-9B
Qwen/Qwen3.5-27B	Unified (NT)	https://huggingface.co/Qwen/Qwen3.5-27B
Qwen/Qwen3.6-27B	Unified (NT)	https://huggingface.co/Qwen/Qwen3.6-27B
OpenGVLab/InternVL2-8B	Instruct	https://huggingface.co/OpenGVLab/InternVL2-8B
google/gemma-4-12B	Base	https://huggingface.co/google/gemma-4-12B
google/gemma-4-12B-it	Instruct	https://huggingface.co/google/gemma-4-12B-it
google/gemma-4-31B	Base	https://huggingface.co/google/gemma-4-31B
google/gemma-4-31B-it	Instruct	https://huggingface.co/google/gemma-4-31B-it
moonshotai/Kimi-VL-A3B-Instruct	Instruct	https://huggingface.co/moonshotai/Kimi-VL-A3B-Instruct
moonshotai/Kimi-VL-A3B-Thinking-2506	Thinking	https://huggingface.co/moonshotai/Kimi-VL-A3B-Thinking-2506

B.4 Recorded Provenance and Determinism

Every result row is keyed by model, source dataset, and sample ID. Intervention outputs additionally record the condition name and the targets y and y^-. Coverage audits check expected task IDs, duplicates, missing records, finite scores, and complete matched condition sets. When a manipulation is randomized, the seed is derived from the sample identity and is shared by the relevant matched conditions. Aggregate rates are always recomputed from the retained row-level records.

B.5 Semantic-Judge Protocol

We use `Qwen3-VL-235B-A22B-Instruct` as a text-only semantic judge. The judge receives the dataset name, target answer or target pair, and the evaluated generation; it does not receive an image. Its instruction accepts semantically equivalent strings, equivalent numeric forms (including percentages and decimals when unambiguous), and multiple-choice letters that identify the same option. For a two-target task, it independently returns Boolean `matches_new` and `matches_old`; for a single-target task, it returns `correct`. Requests use batch size one, temperature zero, and a strict JSON schema that forbids rationales, extracted text, task IDs, or extra fields. Invalid responses trigger a retry.

Because the two decisions are independent, VUR and PAR are not forced to sum to one. Rule-based matching is retained only for inexpensive design screens explicitly identified as such.

B.6 Rare Dual-Match Outcomes

Independent target judgments allow a response to match both y and y^-, for example when it mentions or contrasts both candidates. We label this outcome *Both*. It contributes to both VUR and PAR and is therefore retained in all aggregate statistics. However, it is rare in the reported free-generation analyses: its share, averaged equally across models, is 0.60%–1.47% across the evidence-content conditions in Tab. 12, and only $10/1{,}800$ ordinary-reflection outputs (0.56%) fall into this category in the FSAF evaluation (Tab. 17). The main text focuses on the more frequent outcome states; the appendix reports complete *Both* counts.

B.7 Answer-Preference Scoring Details

We compute answer scores from the evaluated model's token probabilities under the same rendered context S. For each answer string a, Eq. 5 averages log-probabilities over the appended candidate tokens; context tokens are excluded. Candidate strings are appended after a separating space and tokenized with the model's own tokenizer.

Diagnostic experiments. For each target y, the content, dose, order, and continuation probes construct a set $\mathcal{V}(y)$ of available answer forms. These include the uppercase option letter, its parenthesized form, an `Answer:` prefix followed by that letter, the expected answer value, and the compact extracted answer. Empty and duplicate strings are removed. The same forms are retained across paired conditions. The target score is the log-mean-exp of the mean token scores:

$$s(y \mid S) = \log \left[\frac{1}{|\mathcal{V}(y)|} \sum_{a \in \mathcal{V}(y)} \exp \ell(a \mid S) \right]. \tag{8}$$

Thus, each form is first normalized by its own token length, and the aggregation gives equal weight to each distinct form. The current and prior targets may have different numbers of available forms. The margin is $m(S) = s(y \mid S) - s(y^- \mid S)$; positive values favor the current target. Where reported, the two-target normalized score is

$$\widetilde{p}(y \mid S) = \frac{\exp s(y \mid S)}{\exp s(y \mid S) + \exp s(y^- \mid S)}.$$

It expresses preference within this candidate pair. Free-generation VUR and PAR are obtained separately from semantic judgments.

FSAF evaluation. FSAF uses one canonical string per target: option letters when both are available, otherwise answer values, with compact extracted answers as the fallback. Here $s(y \mid S) = \ell(a_y \mid S)$. Every condition appends the shared scaffold `Final answer:`; any open `<think>` block is closed beforehand. Candidate strings, tokenization, and scoring positions are held fixed across ordinary reflection, re-query-only, and FSAF.

B.8 Why Paired Effects Are Primary

Our primary estimand is the within-model, within-sample change between matched conditions. Models differ sharply in baseline calibration, output style, and eligible-cohort composition, so a raw average across models can mix the intervention effect with between-model differences. Pairing measures the effect of changing the history, evidence span, continuation, or attention access for the same model and sample. Cross-model averages give equal weight to these within-model effects.

B.9 Claim–Evidence Index

Tab. 8 provides a compact lookup from each scientific question to its matched comparison, measure, and detailed protocol.

Table 8: **Index from the analyses to protocols, metrics, and appendix details.**

Question	Primary comparison	Primary metric	Detailed location
Does prior CoT affect reflection?	Prior-conditioned vs. direct/current controls	VUR, PAR, answer-preference margin	Appendices D.1 and C.3
Which prior sentences affect the answer?	Intact vs. evidence-deleted, answer-only, and matched controls	Paired change in semantic outcomes and margin	Appendix D.3 and Tab. 4
Does prior reasoning still matter after a correct answer?	History-conditioned vs. direct-new and neutral-history continuations	Current/prior preference and semantic branch outcomes	Appendices E–E.3
What happens after deleting evidence from a correct response?	Matched support present vs. withdrawn	Old-answer increase among initially correct responses	Tab. 5 and Appendix C
Do follow-up calculations match the old value?	History-conditioned vs. direct-new/neutral-history derived questions	Derived-answer preference and semantic accuracy	Appendix E.2
Does FSAF block prior-CoT access and improve immediate updating?	FSAF vs. ordinary reflection on identical model–sample pairs	Paired immediate VUR, PAR, and answer preference	Appendices F.3–F.6
Which rows enter each analysis?	Directly solvable pair $\rightarrow$ experiment-specific filters	Eligible and complete-pair counts	Appendix H

C Prompt and Conversation Construction

The notation below follows the abstract, model-independent rendering used to compare different chat templates. `[User Start]`, `[User End]`, `[Response Start]`, and `[Response End]` denote role boundaries inserted by each model's native processor, abstracting over checkpoint-specific tokens. I^- and I are the prior and current images, R^- is the prior response, R is the reflected response, and Q' is a controlled continuation question.

C.1 Direct Inference

(1) Prior-image direct inference. The model receives I^- and Q, producing R^- in one assistant turn.

- **Input:**

[User Start] I^- Q [User End]

- **Output:**

[Response Start] R^- [Response End]

(2) Current-image direct inference. The same construction replaces I^- with I, yielding the direct-current response R_{dir}.

- **Input:**

[User Start] I Q [User End]

- **Output:**

[Response Start] R_{dir} [Response End]

These two direct-inference runs define eligibility for the subsequent reflection analysis.

C.2 Prior-Conditioned Reflection

(3) Standard reflection. We retain the current image and question, prefill the assistant with the complete prior response followed by the fixed cue

```
Wait, let me check the figure again to make sure I haven't
made a mistake.
```

and ask the model to continue the same assistant turn.

- **Input:**

[User Start] I Q [User End]
[Response Start] R^- [Reflection Cue]

- **Output:**

[Response Start] R^- [Reflection Cue] R [Response End]

The prior text is therefore part of the cached assistant prefix, while the image changes from I^- to I.

C.3 Prior-Carrier and Evidence-Content Interventions

(4) Carrier ablations. All conditions retain I, Q, and the same reflection cue. Only the assistant prefix C changes:

`full prior`: $C = R^-$; `reasoning only`: $C = R^-$ without its final answer; `answer only`: `My previous answer was` y^-; `answer masked`: the reasoning plus `My initial answer is [omitted].`

The random-answer control replaces y^- with a non-target answer, and the new-turn control preserves R^- as a completed historical assistant message rather than a same-turn prefill. Their shared schema is:

[User Start] I Q [User End]
[Response Start] C [Reflection Cue] R [Response End].

(5) Evidence-content interventions. For the content conditions, terminal-answer preprocessing first produces the reasoning baseline N. Sentence-level interventions replace C with an intact, deleted, masked, substituted, or organization-controlled version of the prior reasoning. The image, question, cue, targets, decoding settings, and sample identity remain fixed. Non-evidence deletion controls target either the removed-sentence count or removed-word budget, subject to the available non-evidence text; generic-memory controls preserve a prior message without retaining answer-relevant visual content. Exact evidence-span rules are given in Appendix D.3.

C.4 Deleting Evidence from a Correct Response

(6) Current support present vs. withdrawn. Within samples whose ordinary reflection reaches the current answer, the support-present condition retains the complete current-evidence state. The matched withdrawal condition deletes all current evidence-bearing spans while retaining the model, sample, current image, question, stale history, answer scaffold, and target candidates. We apply two complementary measurements to every matched context: teacher-forced candidate scoring measures answer preference, while a separate continuation is evaluated by the semantic judge to obtain the free-generation PAR reported in Tab. 5. The D/H paths and F/W conditions retain the same model–sample identity for paired outcome comparisons. The *Correct* label is assigned from the source ordinary-reflection output before the F/W contexts are reconstructed. Tab. 5 then evaluates newly generated continuations from those contexts; hence the full-support H continuation can match y^- even though its source state previously produced y. Because VUR and PAR are judged independently, PAR also includes any continuation matching both targets. This comparison measures the change in subsequent answers when current textual evidence is removed and the current image is retained.

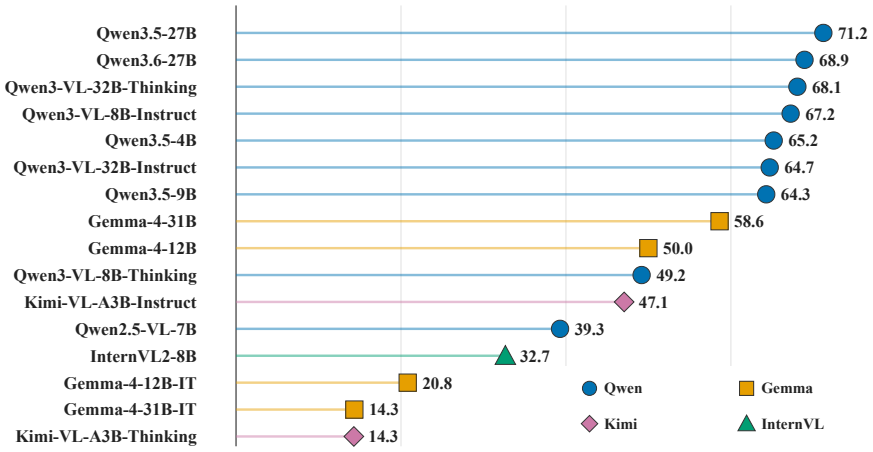

Figure 4: Δ_{PAR} **is positive across all 16 VLMs under matched deletion of current evidence.** Per-model effects for the matched contrast in Eq. 7; pooled outcomes appear in Tab. 5.

C.5 Follow-Up Questions with Evidence Kept

(7) Future re-query. After constructing an initial response, we close that turn and append one of two user questions:

> `Equivalent re-query`: "For the current image, answer the same question again. Use the current image as the visual evidence."
> `Concise re-query`: "For the current image, give only the final answer to the question below."

The original question Q follows either instruction. The resulting conversation is:

`[Source-State Conversation]`
`[User Start]` I Q' `[User End]`
`[Response Start]` R' `[Response End].`

For derived questions, Q' requests a new computation from an answer-critical visual premise rather than simply asking for y again. The initial context construction and target derivation are held fixed across direct-new, history-conditioned, and neutral-history conditions.

C.6 Neutral-History Control

(8) Content-neutral history. The control preserves a multi-turn history but removes its concrete visual premises. Its first user message states:

```
Earlier there was a related context, but it should not
provide visual evidence for the current image.
```

The historical assistant content is replaced by the repeated neutral sentence

```
I considered an earlier context, but its visual details are
not evidence for the current image.
```

up to the capped number of prior sentences. The conversation then contains the direct-current I, Q exchange before Q' is appended. This retains extra roles and history structure without retaining the specific stale evidence-bearing content.

C.7 FSAF CONVERSATION

(9) Fresh-state re-query with attention firewall. Ordinary reflection first produces the stale assistant prefix. FSAF then creates a fresh user turn:

```
Answer the same question using only the current image.  Be
concise and end with the final-answer format requested by
the question.
```

and a fresh assistant suffix. During both prefill and generation of this suffix, the firewall blocks attention from fresh-suffix queries to the identified stale assistant span:

[Stale-State Conversation]
[User Start] [FSAF Re-query] [User End]
[Response Start] R_{FSAF} [Response End].

The current image remains available through the current conversation; only access to the stale textual span is masked. Appendix F.3 specifies the layerwise mask and verified model scope.

D DIAGNOSING TEXTUAL SHORTCUTS

This section complements the main-text diagnosis. We first report the complete cross-model prior-influence control, analyze model-variant and scale effects, then document how evidence-bearing spans are identified and intervened on, and finally provide additional free-generation results and a cross-table protocol map.

D.1 FULL PRIOR-COT INFLUENCE ACROSS 16 VLMS

Tab. 9 reports the complete semantic-judge results for the control summarized in Tab. 2. Prior-only PAR is the rate of generating the prior answer when Q and R^- are retained but the current image is withheld. Image-present VUR is the rate of generating the current answer after restoring I under the matched new-turn setting. In both columns, R^- is preserved as a completed historical assistant response and a fresh user turn re-asks Q; the two columns differ only in whether that turn includes I. This new-turn control is distinct from the same-assistant-turn prefix used for standard reflection earlier in App. C and for the N/C/E evidence-content interventions in Tab. 12. Their absolute VUR values therefore describe performance under their respective turn structures. Prior-only PAR is high for all 16 models, whereas image-present VUR varies substantially, showing that the prior CoT provides strong answer control even though current visual evidence remains behaviorally effective.

D.2 THINKING VERSUS INSTRUCT MODELS AND SCALE

VisualSwap reports that Thinking variants suffer larger base–probe accuracy degradation than their Instruct counterparts and that increasing model size does not mitigate this degradation (Shi et al., 2026). We revisit both observations using our intervention-based measurements. Each variant and scale uses its own eligible cohort. Our VUR/PAR measurements report agreement with the current and old targets; VisualSwap reports the base–probe accuracy gap.

Thinking versus Instruct. Thinking yields lower image-present VUR in all three same-family, same-checkpoint comparisons: by 22.3 pp at Qwen3-VL-8B, 28.2 pp at Qwen3-VL-32B, and 4.4

Table 9: **Prior-CoT influence and visual availability across all 16 VLMs.** Rates are percentages over the eligible cohort for each model and are evaluated by the semantic judge.

Model	Variant	Prior-only PAR (%)	Image-present VUR (%)
Qwen2.5-VL-7B	Instruct	74.8	48.4
Qwen3-VL-8B	Instruct	70.8	44.9
	Thinking	53.3	22.6
Qwen3-VL-32B	Instruct	82.9	60.0
	Thinking	56.6	31.8
Qwen3.5-4B	–	98.7	82.0
Qwen3.5-9B	–	99.2	86.8
Qwen3.5-27B	–	97.8	88.0
Qwen3.6-27B	–	97.8	85.5
InternVL2-8B	Instruct	87.9	38.3
Gemma-4-12B	Base	95.0	75.0
	IT	91.3	79.9
Gemma-4-31B	Base	97.1	77.9
	IT	92.7	88.0
Kimi-VL-A3B	Instruct	97.2	56.9
	Thinking	96.7	52.5

Table 10: **Matched-family comparison of Instruct and Thinking variants.** Prior-only PAR and image-present VUR are absolute semantic-judge rates. E–C compares evidence deletion with non-evidence deletion; positive margin/VUR and negative PAR changes indicate weaker prior-answer control.

Model	Variant	Prior-only PAR (%)	Image-present VUR (%)	$\Delta m_{\mathrm{E-C}}$	ΔVUR / ΔPAR E–C (pp)
Qwen3-VL-8B	Instruct	70.8	44.9	+2.617	$+42.27/-72.95$
	Thinking	53.3	22.6	+2.832	$+35.39/-84.65$
Qwen3-VL-32B	Instruct	82.9	60.0	+3.075	$+54.74/-81.52$
	Thinking	56.6	31.8	+3.004	$+54.47/-82.11$
Kimi-VL-A3B	Instruct	97.2	56.9	+1.619	$+32.04/-49.17$
	Thinking	96.7	52.5	+2.913	$+51.91/-66.12$

pp for Kimi-VL-A3B. The pair-equal average is 35.6% for Thinking versus 53.9% for Instruct, agreeing directionally with VisualSwap's finding that Thinking variants are more vulnerable during continuous re-examination. However, Thinking does not exhibit higher prior-only PAR: it is lower in all three pairs. The mode gap is therefore more specific than a universally stronger standalone preference for the prior answer; it appears when the current image must compete with the retained reasoning trajectory.

The intervention results refine this comparison. Evidence removal shifts m toward the current answer for every variant, with pair-equal $\Delta m_{\mathrm{E-C}}$ of 2.916 for Thinking and 2.437 for Instruct. It also produces a larger PAR reduction for Thinking in all three pairs, whereas the corresponding VUR gain is mixed across families. Thus, evidence-bearing prior content is the text whose deletion affects both modes, although reducing old answers does not always increase correct answers.

Table 11: **Descriptive within-family scale trends.** Rows within each block follow increasing nominal checkpoint size. Cohorts are separately eligible at each checkpoint, so these trends are not paired sample-level scaling estimates.

Family / Variant	Size	Prior-only PAR (%)	Image-present VUR (%)	$\Delta m_{\mathrm{E-C}}$
Qwen3-VL / Instruct	8B	70.8	44.9	2.617
	32B	82.9	60.0	3.075
Qwen3-VL / Thinking	8B	53.3	22.6	2.832
	32B	56.6	31.8	3.004
Qwen3.5	4B	98.7	82.0	1.612
	9B	99.2	86.8	1.372
	27B	97.8	88.0	1.535
Gemma-4 / Base	12B	95.0	75.0	0.911
	31B	97.1	77.9	0.996
Gemma-4 / IT	12B	91.3	79.9	2.067
	31B	92.7	88.0	3.152

Scale. Image-present VUR increases with nominal size in all five comparable within-family sequences in Tab. 11. Higher image-present VUR coexists with persistent prior influence. Prior-only PAR increases in four of the five sequences and remains near saturation for Qwen3.5, while $\Delta m_{\mathrm{E-C}}$ does not decrease monotonically with size. Larger checkpoints can therefore recover the current answer more often while retaining an equally strong or stronger evidence-bearing prior path. In particular, Qwen3-VL-32B-Thinking remains far below its Instruct counterpart in VUR (31.8% versus 60.0%).

FSAF further separates vulnerability from immutable visual ability. On the independently eligible FSAF cohort in Tab. 15, it raises VUR more for Thinking than Instruct at both Qwen3-VL scales ($+19.96$ versus $+15.95$ pp at 8B and $+21.97$ versus $+15.22$ pp at 32B), while reducing PAR to 1.03%–1.64% for the Thinking variants. The improvements at both scales show that restricting prior-CoT access can reduce the observed failures in Thinking models.

D.3 Operational Identification of Evidence-Bearing Spans

We identify E^{-} automatically using fixed, outcome-independent rules. Before classification, we remove the explicit final-answer suffix using deterministic patterns for expressions such as "final answer," "the answer is," and `\boxed`; when none matches and the response contains more than two sentences, the implementation removes the final sentence. This preprocessing keeps the terminal answer separate from the reasoning considered for E^{-}. The resulting reasoning is the shared N baseline for N/E/C/A. Condition A additionally masks matched answer cues within this baseline. We split the remaining prior CoT at sentence-final periods, question marks, exclamation marks, and line breaks.

A sentence is marked as evidence-bearing if it satisfies at least one of the following four rules:

1. it contains a number, \frac, \sqrt, or one of $=, <, >, \pm$, and degree symbols;
2. it contains an explicit option or answer reference involving a letter from A to J;
3. it contains a term from the fixed visual lexicon: *image, figure, graph, diagram, line, point, circle, triangle, angle, axis, x-axis, y-axis, table, chart, shape, red, blue, green, left, right, above, below, intersect, parallel, perpendicular, slope, curve, bar, row, column, object, shown, visible*; or
4. it contains a term from the fixed mathematical/reasoning lexicon: *equals, equal, therefore, thus, hence, because, since, so, calculate, compute, sum, difference, ratio, area, perimeter, length, radius, diameter, probability, fraction, degree, option, choice, answer, value*.

For the dose-response intervention in Fig. 3, identified sentences are ranked by the number of numerical or symbolic cues, twice the number of option cues, and the number of matched visual and mathematical terms; ties are resolved by sentence length. At removal fraction α, we delete the highest-ranked $\lceil \alpha |E^-| \rceil$ sentences. The non-evidence control randomly orders non-evidence sentences using a fixed sample seed and deletes whole sentences until their whitespace-separated word count reaches the evidence-removal budget or the candidates are exhausted. If no non-evidence sentence is available, the control retains the reasoning. Whole-sentence removal can overshoot the budget, and limited non-evidence text can leave it below the budget. In the archived $4{,}159$ paired instances, the mean removed-word fractions, giving each model equal weight, are 75.67% for E and 8.93% for C; C removes no words in $1{,}145$ instances (27.53%). These fractions use the preprocessed reasoning as the denominator. Thus, C provides a non-evidence deletion comparison with a substantially smaller realized removal amount. If evidence deletion leaves no sentence, the implementation inserts: "I previously reasoned about this problem, but the evidence details may not apply to the current image."

Answer-cue masking (A) replaces rule-matched answer-letter expressions and boxed contents with `[ANS]`, and option/choice references or parenthesized option letters with `[OPT]`. Other reasoning text is retained. The reported A–N column is the archived `answer_token_removed_only` versus `old_all` contrast; both start from the same terminal-answer-preprocessed reasoning. These fixed edits define the operational comparisons in the main text.

D.4 Free-Generation Outcomes after Evidence Removal

Tab. 12 reports the complete outcome distribution, averaged equally across models, underlying the results in Tab. 3. The baseline (N) retains prior reasoning after terminal-answer preprocessing, and the context control (C) deletes non-evidence sentences targeting the same removed-word budget as E, and evidence removal (E) deletes all spans in E^-. All three are generated by continuing the same assistant turn after the edited prefix and fixed reflection cue. The paired E–C contrast in Tab. 3 compares the two deletion conditions within this same-turn protocol; Tab. 9 uses the separate new-turn protocol.

Table 12: **Free-generation outcome distribution under evidence-content interventions.** Percentages give equal weight to each model and cover $4{,}159$ paired model–sample instances from 16 VLMs. VUR and PAR are judged independently; *Both* denotes outputs matching both targets, and *Other* denotes outputs matching neither target.

Condition	VUR (%)	PAR (%)	Both (%)	Other (%)
Intact prior (N)	14.32	78.77	1.44	8.35
Non-evidence deletion (C)	15.78	77.14	1.47	8.55
Evidence removal (E)	59.54	14.55	0.60	26.50

Relative to the intact prior, evidence removal increases VUR by 45.22 pp and decreases PAR by 64.22 pp. The outcome shift also includes an increase in *Other* from 8.35% to 26.50%, a change of 18.16 pp, with a positive change in $14/16$ models. Thus, removing E^- shifts generation toward both current-answer and Other outcomes.

Prior-CoT Length Sensitivity. We examine how prior-CoT length relates to old-answer retention and to the effects of text interventions. Length is measured on the prior reasoning span R^-

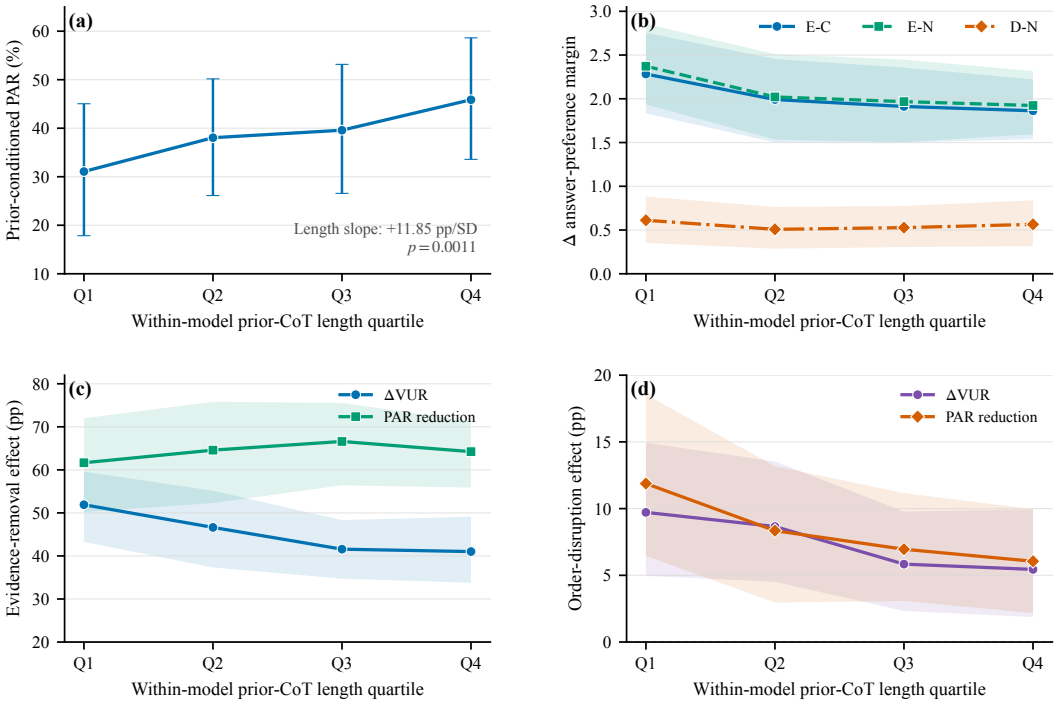

Figure 5: **Prior-CoT length and the effects of text interventions.** (a) PAR averaged equally across models, with the current image and prior reasoning. (b) Answer-preference effects for evidence removal versus non-evidence deletion (E–C), evidence removal versus the intact prior (E–N), and evidence-order disruption versus the intact prior (D–N in this figure, denoted O–N in the main text). (c,d) Corresponding VUR increases and PAR reductions. Error bars and bands are 95% hierarchical-bootstrap CIs.

after removing its explicit final-answer suffix. Within each model, we divide the $4{,}159$ eligible model–sample instances into character-length quartiles, preventing naturally longer traces from being treated as a between-model effect. Under image-present prior-conditioned reflection, PAR, averaged equally across models, increases from 31.07% in the shortest quartile to 45.87% in the longest. After controlling for model and dataset fixed effects, a one-standard-deviation increase in $\log(1 + \text{length})$ is associated with an 11.85 pp increase in PAR ($p = 0.0011$). Longer prior CoTs are therefore associated with greater prior-answer retention.

Across all four within-model length quartiles, evidence removal shifts answers toward the current target relative to both C and N, and order changes have smaller effects (Fig. 5). For the seven content and order measurements shown in panels (b)–(d), we detect no significant character-length slope after the fixed-effect adjustment (all $p > 0.05$). For example, the E–C margin slope per standard deviation of log length is 0.075, with a 95% confidence interval of $[-0.150, 0.300]$. The interval includes both a small negative and a positive association between length and the intervention effect.

VisualSwap varies the retained proportion of a reasoning trace and reports lower update accuracy with more retained text (Shi et al., 2026). Our analysis examines variation in the lengths of complete prior traces and the effects of editing them. It shows that text interventions affect answers across the observed length groups, alongside a positive association between length and old-answer retention. The E–C comparison still includes unequal deletion amounts within these groups, as documented in App. D.3.

D.5 Protocol Map for Reported Results

Tabs. 2–17 use different prior texts and turn structures. Tab. 13 makes these settings explicit so that absolute rates are compared only within a matched protocol.

Table 13: **Protocol map for reported quantitative results.** Eligibility rules and cohort counts are documented separately in Appendix H. Descriptive/setup Tabs. 7, 8, 18, 19, and 20 are omitted.

Table(s)	Prior text	Turn structure	Evaluator
2, 9	Completed historical R^-	Fresh user turn re-asks Q, with or without I	235B semantic judge
3, 12	Intact or edited same-turn R^- prefix (N/C/E/A)	Same assistant turn after the edited prefix and fixed cue	Model token log-probabilities and 235B judge
4	Intact or reordered evidence chain (N/O/NE)	Same assistant turn after the controlled prefix	Model token log-probabilities and 235B judge
5	Reconstructed D/H contexts with current support F/W	Separate continuation from each matched context	235B semantic judge
6, 14	Direct, neutral-history, or prior-conditioned initial response	Fresh equivalent-re-query or derived-computation user turn	Model token log-probabilities and 235B judge
15, 17	Ordinary stale prefix or fresh suffix blocked from R^-	Same-turn baseline vs. fresh user–assistant FSAF boundary	235B judge and model token log-probabilities
10, 11	Derived summaries of the preceding input conditions	Mixed: new-turn absolute rates and same-turn E–C intervention	Inherited from Tabs. 9 and 3
16	RAQ, CRF, re-query-only, or full FSAF	Condition-specific same-turn or fresh-boundary design screen	Rule screen; 235B judge where marked

E Follow-Up Results and Controls

This section describes how we repeat the question or ask for a new calculation while keeping the current image and previous responses. Both branches begin from the matched direct and prior-conditioned paths in Eqs. (2)–(3) and preserve the current image and preceding path-specific context.

Tab. 6 compares samples whose initial answers are correct with those whose initial answers match the old target. For equivalent re-query, the $4{,}159$ eligible pairs comprise $2{,}456$ initially correct (59.05%), $1{,}418$ old-answer (34.09%), and 285 Other (6.85%) cases. For derived computation, the $2{,}384$ numeric-eligible pairs comprise $1{,}450$ initially correct (60.82%), 770 old-answer (32.30%), and 164 Other (6.88%) cases. The analyzed groups of initial responses contain no Both cases. Other remains a separate category; the main table reports effects conditioned on current-only and prior-only source responses.

E.1 Repeating the Question

The equivalent re-query is the first gray continuation branch in Fig. 2. We append a new user turn that repeats Q. We use two semantically matched variants. The first explicitly instructs the model to use the current image as visual evidence before answering Q again; the second requests only a concise final answer to Q. Candidate scoring compares support for y and y^- using $m(S)$ in Eq. (6), while free generation is evaluated with VUR and PAR in Eq. (4). Neither variant removes preceding support or inserts y^- into the new prompt. An auxiliary neutral-history control preserves the extra turn and context structure without including the evidence-bearing content of R^-.

E.2 New Calculation

For tasks in which the current and prior quantities can be parsed as distinct numerical values, we replace the repeated question with one of three transformations: doubling the quantity, adding one, or testing it against a threshold chosen to distinguish the current- and prior-derived outcomes. The transformed current and prior values serve as the paired candidates for scoring and as the targets for judging free generation under the same metric definitions. No support is removed and y^- is not reintroduced in the follow-up prompt.

E.3 Neutral-History Control

To distinguish the effect of stale evidence from generic effects of an additional dialogue turn or longer context, we introduce a neutral-history path that preserves the extra turn and context structure but excludes the evidence-bearing content of R^-. Tab. 14 reports free-generation match rates for the direct, neutral-history, and prior-conditioned paths, stratified by the initial reflection outcome.

Table 14: **Neutral-history control for follow-up questions with evidence kept.** For derived computation, current and prior matches refer to the corresponding derived targets.

		Current match (%)			Prior match (%)		
Continuation	**Initial outcome**	**Direct**	**Neutral**	**Prior-cond. (H)**	**Direct**	**Neutral**	**Prior-cond. (H)**
Equivalent re-query	Correct	54.80	55.56	52.67	0.57	0.75	1.55
	Old answer	56.77	57.48	16.89	0.88	1.02	35.93
Derived computation	Correct	72.94	67.01	53.95	4.64	5.11	4.71
	Old answer	76.02	73.07	10.82	8.83	8.57	54.11

For equivalent re-query, neutral-history rates remain within 0.8 pp of the direct path for current-answer matches and within 0.2 pp for prior-answer matches in both initial-outcome strata. In contrast, among old-answer outcomes, prior-conditioned history decreases the current-answer rate from 56.77% to 16.89% and increases the prior-answer rate from 0.88% to 35.93%. The recurrence is therefore associated with the stale evidence-bearing history rather than the presence of an additional turn alone.

For derived computation, neutral history causes a smaller generic decrease in current-derived matches relative to the direct path (-5.93 pp for initially correct and -2.95 pp for old-answer outcomes), but leaves prior-derived matches nearly unchanged ($+0.47$ and -0.26 pp, respectively). Prior-conditioned history produces a much larger effect in the group that initially retained the old answer, decreasing current-derived matches by 65.19 pp and increasing prior-derived matches by

45.28 pp. For initially correct outcomes, it decreases current-derived matches by 18.99 pp while changing prior-derived matches by only $+0.07$ pp. Thus, generic history can contribute modestly to computation degradation, but it does not explain the large increase in results matching the old value among samples that initially retained the old answer. For initially correct samples, accuracy falls with little increase in results matching the old value.

F Fresh-State Attention Firewall

We examine attention access to prior reasoning through Fresh-State Attention Firewall (FSAF), evaluated on five Qwen models using a separate frozen cohort of 1,800 model–sample pairs. We report its construction, aggregate effects, and transitions conditioned on the initial response.

F.1 Intervention and Results

Motivation. The diagnostic results motivate restricting access to stale prior text during visual re-examination. *Fresh-State Attention Firewall* (FSAF) is a training-free, single-pass intervention with two coupled components: a new user–assistant boundary that starts a fresh re-examination, and an attention mask that isolates this fresh suffix from R^- during both prefill and generation. It retains the original conversation and requires neither target-answer access nor an evidence parser.

Fresh re-query boundary. Given the prior-conditioned context containing R^-, FSAF appends a fixed user re-query U_f that asks the model to answer Q again using only the current image I. Let R^F denote the following assistant response and $\mathcal{F} = U_f \cup R^F$ the fresh suffix. During prefill, tokens in U_f can still attend to R^- and carry its content into generation through their cached states.

Suffix attention firewall. The firewall closes both this indirect route and direct access from the new response. At every language layer, FSAF augments the standard Transformer attention logits (Vaswani et al., 2017) with the following mask:

$$M^{\mathrm{FSAF}}_{qk} = \begin{cases} -\infty, & q \in \mathcal{F} \text{ and } k \in R^- \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$A^{(\ell)}_{q,:} = \mathrm{softmax}\Big(\mathbf{L}^{(\ell)}_{q,:} + M^{\mathrm{causal}}_{q,:} + M^{\mathrm{FSAF}}_{q,:}\Big) \tag{10}$$

where $\mathbf{L}^{(\ell)}$ denotes the matrix of unmasked attention logits at layer ℓ. Applying the mask while prefilling U_f prevents its cached states from absorbing R^-; keeping it active during autoregressive generation prevents tokens in R^F from reading R^- directly. All other causal edges remain unchanged, so the fresh suffix can still access I, Q, system tokens, U_f, and earlier tokens in R^F. The softmax renormalizes over these remaining keys: FSAF removes the stale textual path without deleting the transcript or prescribing a fixed increase in visual attention. The exact re-query, execution procedure, implementation, and design ablations are provided in App. F.

Evaluation. To exclude failures caused by task-solving difficulty, we evaluate FSAF on VS-Bench model–sample pairs from five Qwen models for which each model correctly answers both the current and counterfactual images under direct inference. For each pair, we compare ordinary prior-conditioned reflection with FSAF under identical settings.

Results. In Tab. 15, FSAF improves all three measurements for every evaluated model. Pooled across model–sample pairs, it raises VUR from 35.28% to 53.61%, reduces PAR from 39.22% to 3.67%, and shifts the answer-preference margin from -5.49 to $+2.97$. Together, these results show that FSAF consistently reduces stale prior-CoT influence across all evaluated Qwen models. The state-conditioned analysis below details how these aggregate changes distribute across initially correct, stale, and Other outputs.

F.2 Exact Re-Query and Execution Procedure

FSAF uses the same fixed user turn for every model and sample:

> *Answer the same question using only the current image. Be concise and end with the final-answer format requested by the question.*

Table 15: **FSAF improves visual updating while suppressing prior-answer retention.** Each cell reports Baseline → FSAF, with the paired change in parentheses.

Model	VUR (%) ↑	PAR (%) ↓	m ↑
Qwen2.5-VL-7B (I)	23.60 → 42.32 (+18.73)	41.20 → 9.74 (-31.46)	-1.75 → +1.52 (+3.26)
Qwen3-VL-8B (I)	49.52 → 65.48 (+15.95)	27.62 → 5.95 (-21.67)	-5.64 → +3.59 (+9.23)
Qwen3-VL-8B (T)	15.84 → 35.80 (+19.96)	56.17 → 1.03 (-55.14)	-6.01 → +1.37 (+7.38)
Qwen3-VL-32B (I)	59.01 → 74.22 (+15.22)	19.57 → 1.55 (-18.01)	-7.52 → +4.76 (+12.27)
Qwen3-VL-32B (T)	31.80 → 53.77 (+21.97)	47.21 → 1.64 (-45.57)	-5.62 → +4.03 (+9.65)
Average	**35.28 → 53.61 (+18.33)**	**39.22 → 3.67 (-35.56)**	**-5.49 → +2.97 (+8.46)**

The intervention then proceeds as follows. First, the model generates the prior assistant CoT R^-, while the serving runtime records its token interval $[s^-, e^-)$. Second, the fixed re-query U_f is appended as a new user turn. Third, the runtime constructs M^{FSAF} in Eq. (9) directly from the recorded interval, without parsing the answer or locating individual evidence spans. Fourth, the mask is enabled in every language layer for both the prefill of U_f and the generation of R^F. Finally, the model produces one response using the same decoding settings and maximum generation length as the baseline. Thus, FSAF adds neither a second model call nor a candidate-answer extraction stage.

F.3 Implementation and Deployment Scope

Our implementation applies the additive mask through Transformers eager-attention hooks for Qwen2.5-VL and Qwen3-VL. The runtime identifies fresh-suffix queries by their positions after e^- and masks all keys in $[s^-, e^-)$. The intervention covered all 28 language layers of Qwen2.5-VL-7B, all 36 layers of the Qwen3-VL-8B variants, and all 64 layers of the Qwen3-VL-32B variants. The current implementation uses batch size 1; a vLLM deployment requires the same additive mask to be integrated into its attention backend or a custom operator. A fail-closed policy enables both the re-query and firewall only for the five verified model configurations. For an unknown model, both components are disabled and the caller's ordinary reflection baseline is used. This avoids silently degrading to the re-query-only condition.

F.4 Design Ablations

We study why FSAF needs both a fresh user–assistant boundary and a firewall spanning prefill and generation. Tab. 16 reports the Qwen2.5-VL-7B design study. Re-answer querying (RAQ) masks only generated-token access to R^-; the re-query cue can therefore read R^- during prefill and carry it forward in its cached states. Clean recheck framing (CRF) also cleans the cue, but continues within the same assistant turn. The re-query-only condition uses the exact FSAF wording without changing attention accessibility.

RAQ substantially reduces PAR but also decreases VUR, showing that removing a direct generation-time edge does not by itself preserve correct updating. The rule-based re-query-only screen yields no reliable VUR gain and increases PAR, while full FSAF improves VUR and reduces PAR under both the rule screen and the 235B judge. CRF and re-query-only rows use the rule evaluator; semantic-judge comparisons are available for RAQ and full FSAF.

Table 16: **FSAF design ablations on Qwen2.5-VL-7B.** Values are paired percentage-point differences relative to ordinary reflection; brackets give 95% confidence intervals when reported. Rule-based rows are auxiliary design screens; the reported semantic outcomes use the 235B judge.

Condition	Evaluator	vs. Ordinary Reflection (pp)	
		ΔVUR ↑	ΔPAR ↓
RAQ: generation-only mask	235B judge	-2.62	-35.21
CRF: same-turn cue cleaning	Rule	$+0.75$	-16.85
Re-query only	Rule	$+3.00$ $[-2.25, +8.24]$	$+19.48$ $[+11.99, +26.97]$
Full FSAF	Rule	$\mathbf{+23.97}$ $[+17.60, +30.34]$	$\mathbf{-14.61}$ $[-20.98, -7.87]$
Full FSAF	235B judge	$\mathbf{+18.73}$ $[+12.36, +25.09]$	$\mathbf{-31.46}$ $[-37.83, -25.09]$

F.5 Changes Grouped by the Initial Answer

To identify which baseline outcomes contribute to the pooled effect in Tab. 15, we stratify each paired instance by the 235B judge label assigned to its ordinary-reflection output. *Correct* denotes a baseline output matching only y, *Old answer* denotes one matching only y^-, and *Both* and *Other* denote outputs matching both or neither target, respectively. We then measure the paired FSAF outcome within each fixed baseline stratum.

Table 17: **FSAF transitions conditioned on the ordinary-reflection outcome.** Each metric cell reports baseline → FSAF, followed by the paired change in parentheses. Rates and changes are in percent and percentage points, respectively.

Baseline outcome	Pairs	VUR: Base → FSAF	PAR: Base → FSAF
Correct	625	100.00 → 82.40 (−17.60)	0.00 → 1.12 (+1.12)
Old answer	696	0.00 → 38.51 (+38.51)	100.00 → 6.47 (−93.53)
Other	469	0.00 → 37.74 (+37.74)	0.00 → 2.77 (+2.77)
Both	10	100.00 → 50.00 (−50.00)	100.00 → 10.00 (−90.00)

The pooled improvement is concentrated in the *Old answer* and *Other* strata. For old-answer outputs, FSAF reduces PAR from 100.00% to 6.47% and raises VUR to 38.51%; among Other outputs, 37.74% become current-answer matches. FSAF preserves the current answer for 82.40% of initially correct outputs. At the pooled level, the Other share rises from 26.06% to 43.17%. FSAF therefore reduces stale-answer generation while also disrupting some successful reflections and increasing Other outcomes. Each stratum is defined by the observed baseline output; the *Both* stratum contains 10 pairs.

Scope. The five-Qwen evaluation covers the verified attention implementations. Cross-family evaluation and comparisons with prior-CoT deletion and conversation restart remain open. Masking the complete prior can also discard useful reasoning when only part of the history is stale.

F.6 Evaluation and Intervention Audit

The final semantic evaluation contains 3,600/3,600 expected baseline/FSAF judge tasks and 1,800/1,800 paired model–sample instances across the five models, with no missing or duplicate task IDs. The answer-preference evaluation contains 5,400/5,400 finite scoring records across baseline, re-query-only, and FSAF, forming 1,800/1,800 complete three-condition model–sample triplets with no duplicates. All FSAF generations have at least 0.90 recorded coverage of the intended R^- span, and every sample records modified attention edges during both prefill and generation. Baseline and FSAF use the same maximum generation length of 256 tokens. The rule evaluator and 235B judge disagree on at least one VUR/PAR label for 951/3,600 outputs (26.42%); consequently, the VUR/PAR columns in Tab. 15 use the 235B semantic judge, while its answer-preference

columns use the evaluated models' mean token log-probabilities. Rule-based results serve only as low-cost design screens.

G VS-BENCH CONSTRUCTION AND USE

Scope and Composition. VS-Bench (Shi et al., 2026) is designed to test visual re-examination rather than static visual question answering. It contains 800 image pairs, with 200 pairs drawn from each of MathVista-MINI, the Vision-Dominant subset of MathVerse-MINI, MathVision, and MMMU-Pro (Lu et al., 2024; Zhang et al., 2024a; Wang et al., 2024; Yue et al., 2025). These sources span charts, function plots, geometry, competition mathematics, and expert-level multimodal questions, while retaining questions whose answers depend on visual evidence.

Pair Construction. Each instance contains a shared question Q, a source benchmark image I_b, and a human-checked alternative image I_a. The construction follows three constraints: (i) Q remains valid for both images; (ii) the images preserve high-level layout, style, composition, and context; and (iii) an answer-critical visual detail changes, yielding distinct ground-truth answers $A_a \neq A_b$. VisualSwap uses a human-in-the-loop image-editing process: annotators identify the answer-critical element, specify a targeted modification, and iteratively refine the edited image. The paired images are also kept at identical resolution so that the swap does not alter the visual-token count.

Table 18: **VS-Bench composition and image-pair similarity.** Higher CLIP similarity and SSIM, together with lower LPIPS, indicate that paired images remain visually similar despite their answer-determining differences. Values are reported by VisualSwap (Shi et al., 2026).

Source	Pairs	CLIP ↑	SSIM ↑	LPIPS ↓
MathVista-MINI	200	0.95	0.88	0.12
MathVerse-MINI	200	0.97	0.94	0.06
MathVision	200	0.95	0.86	0.17
MMMU-Pro	200	0.94	0.75	0.19
Overall	800	0.95	0.86	0.14

Use in Our Framework. To match the notation in our controlled framework, we relabel the source benchmark image as the current image $I = I_b$ and the edited alternative as the counterfactual image $I^- = I_a$, with targets $y = A_b$ and $y^- = A_a$. This direction lets the evaluated model first produce R^- on the coherent counterfactual task and then reflect under the current benchmark image. The full 800-pair benchmark defines the available task pool; each model's analyzed cohort is subsequently restricted to pairs that it answers correctly under direct inference on both I^- and I. Thus, the controlled comparison targets prior-CoT influence rather than failures caused by an independently unsolved image.

H COHORT CONSTRUCTION, SAMPLE COUNTS, AND COVERAGE

H.1 ELIGIBILITY AND SAMPLE FLOW

The available pool contains 800 paired tasks per model. A model–sample pair enters the corrected diagnosis cohort only when the model's direct responses to both I^- and I semantically match their respective targets under the 235B judge. This model-specific filter prevents an unsolved source or current task from being misclassified as resistance to, or persistence of, a prior state. Eligibility is therefore defined at the model–sample level rather than once for the benchmark.

The unified 16-model diagnosis contains 4,159 eligible model–sample pairs. Subsequent analyses retain this identity key and apply only the filter required by their estimand. For example, the matched test that deletes current evidence selects the 2,456 instances that first recover the current answer under ordinary reflection. Tab. 19 also records the sample counts for sentence-order free generation and the two follow-up tasks. The initial-answer group counts for the follow-ups are given in App. E.

The FSAF evaluation uses the original rule-extracted eligibility snapshot frozen for the intervention runs. It requires an answer-changing pair for which direct inference is correct on both images.

In the model order of Tab. 15, its counts are $267/420/486/322/305$, totaling $1{,}800$. The separate $1{,}943$-pair five-Qwen cohort in Tab. 20, with counts $(258/414/469/422/380)$, comes from the corrected 235B semantic-judge eligibility snapshot used by the 16-model diagnosis. Because the evaluator and membership differ, all FSAF effects remain within the fixed $1{,}800$-pair cohort. The generated baseline and FSAF outputs are still evaluated by the 235B semantic judge, as detailed in Appendix F.6.

Table 19: **Sample flow for the principal analysis cohorts.** Counts are model–sample pairs, not unique benchmark questions.

Stage or analysis	Pairs	Additional filter
Available task pool	16×800	None
Corrected 16-model diagnosis	$4{,}159$	Directly correct on both images
Sentence-order free generation	$3{,}929$	Enough sentences for the required reordering
Repeated questions	$4{,}159$	Grouped by the initial answer
New calculations	$2{,}384$	Numerical answer required
Matched deletion of current evidence	$2{,}456$	Correct under ordinary reflection
FSAF semantic evaluation	$1{,}800$	Five verified Qwen configurations
FSAF judge tasks	$3{,}600$	Baseline and FSAF output per pair
FSAF preference records	$5{,}400$	Baseline, re-query-only, and FSAF per pair

H.2 Eligible Counts by Model

Tab. 20 reports the corrected semantic-judge eligibility count for every checkpoint. Cohort sizes reflect success on both direct-image tasks and determine the samples available for each model's paired analysis (Appendix B.8).

Table 20: **Corrected eligible model–sample pairs in the 16-model diagnosis.** Eligibility requires semantic correctness under direct inference on both paired images.

Model	Eligible pairs
Qwen2.5-VL-7B-Instruct	258
Qwen3-VL-8B-Instruct	414
Qwen3-VL-8B-Thinking	469
Qwen3-VL-32B-Instruct	422
Qwen3-VL-32B-Thinking	380
InternVL2-8B	141
Gemma-4-12B	60
Gemma-4-12B-it	229
Gemma-4-31B	136
Gemma-4-31B-it	275
Qwen3.5-4B	233
Qwen3.5-9B	242
Qwen3.5-27B	267
Qwen3.6-27B	269
Kimi-VL-A3B-Instruct	181
Kimi-VL-A3B-Thinking-2506	183
Total	**4,159**

H.3 Completeness and Coverage Checks

Before aggregation, each analysis verifies unique task IDs and the presence of every condition required for a matched comparison. The FSAF semantic package contains $3{,}600/3{,}600$ expected judge tasks and $1{,}800/1{,}800$ complete baseline/FSAF pairs. Its answer-preference package contains $5{,}400/5{,}400$ finite records and $1{,}800/1{,}800$ complete three-condition triplets, with no duplicate task IDs. All FSAF rows record at least 0.90 coverage of the intended stale span and nonzero modified attention edges in both prefill and generation. Experiment-specific subsets are reported with their own denominator; missing conditions are never imputed.

I Representative Case Studies

Figs. 6–16 visualize one frozen, representative model–sample pair for each principal diagnostic or continuation experiment. Following the presentation style of VisualSwap (Shi et al., 2026), each figure reads from the prior observation (I^-, Q, R^-), through the current image I, to the condition-specific intervention and observed response. Red denotes the prior state and green denotes the current state. For candidate scoring, the displayed margin is $\log p(y \mid C) - \log p(y^- \mid C)$, so positive values favor the current answer; generation labels are assigned separately by the semantic judge. Long questions or prior responses are excerpted with ellipses only for display; the experiment uses the complete strings. Each case illustrates the protocol and observed behavior on an individual instance.

I.1 Diagnosing Prior-State Control

The first group traces prior-state control from the basic visual-recheck failure to the prior context, its content, and sentence order. Within each controlled contrast, the current image and query are fixed while the retained or edited prior response changes. Reading Figs. 6–13 in sequence therefore shows which parts of R^- sustain the prior-answer path and which interventions reopen current-image recomputation.

I.2 Examples of Errors on Follow-Up Questions

The remaining cases examine how prior reasoning affects later answers. They compare direct-current, history-conditioned, and neutral-history paths under matched deletion of current evidence, an identical future re-query, and a derived task. Candidate margins and free-generation labels remain separate measurements, allowing Figs. 14–16 to expose both shifts in answer preference and changes in the generated answer.

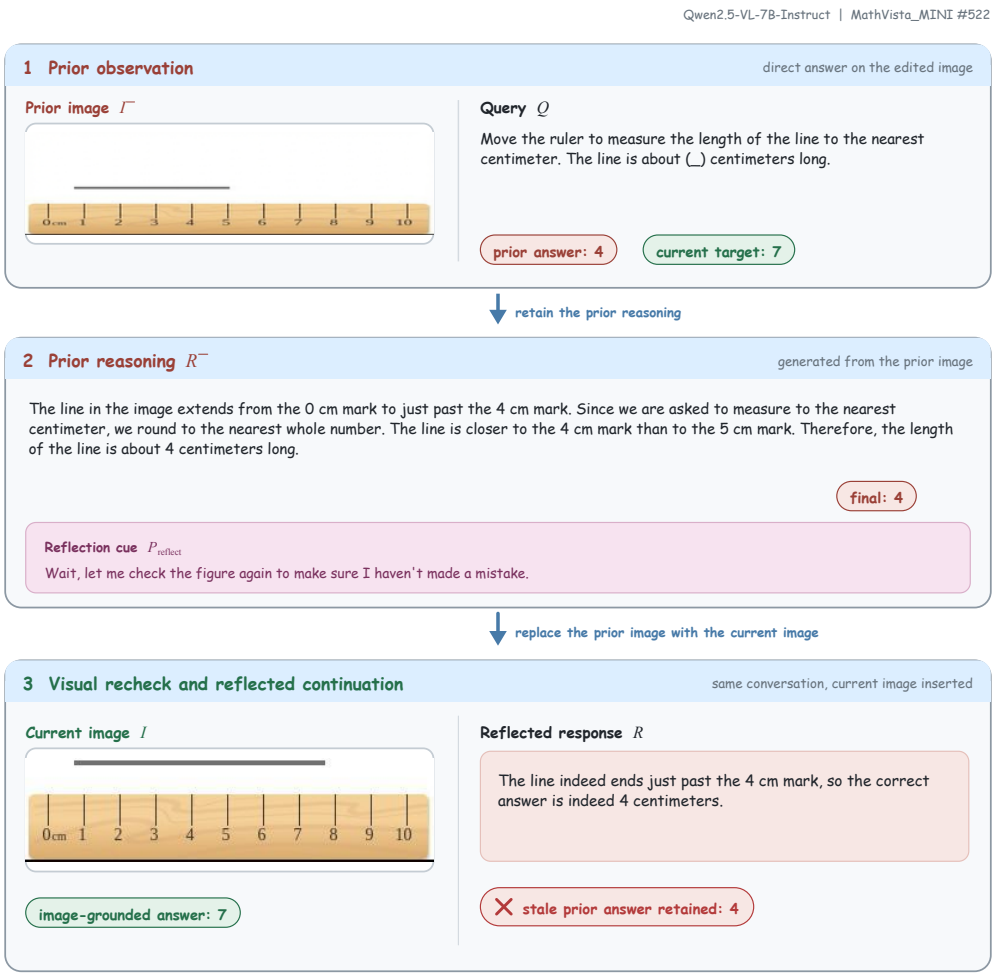

Figure 6: **A prior CoT can override changed visual evidence.** Top: Qwen2.5-VL-7B-Instruct answers $y^- = 4$ from the prior image I^- and records the supporting response R^-. Middle: the image is replaced by the current image I, whose target is $y = 7$, while the prior response remains in the conversation. Bottom: after the reflection cue, the model repeats the prior answer instead of recomputing from the changed ruler. This representative failure illustrates the competition between fresh visual recomputation and stale textual reuse described in the main paper.

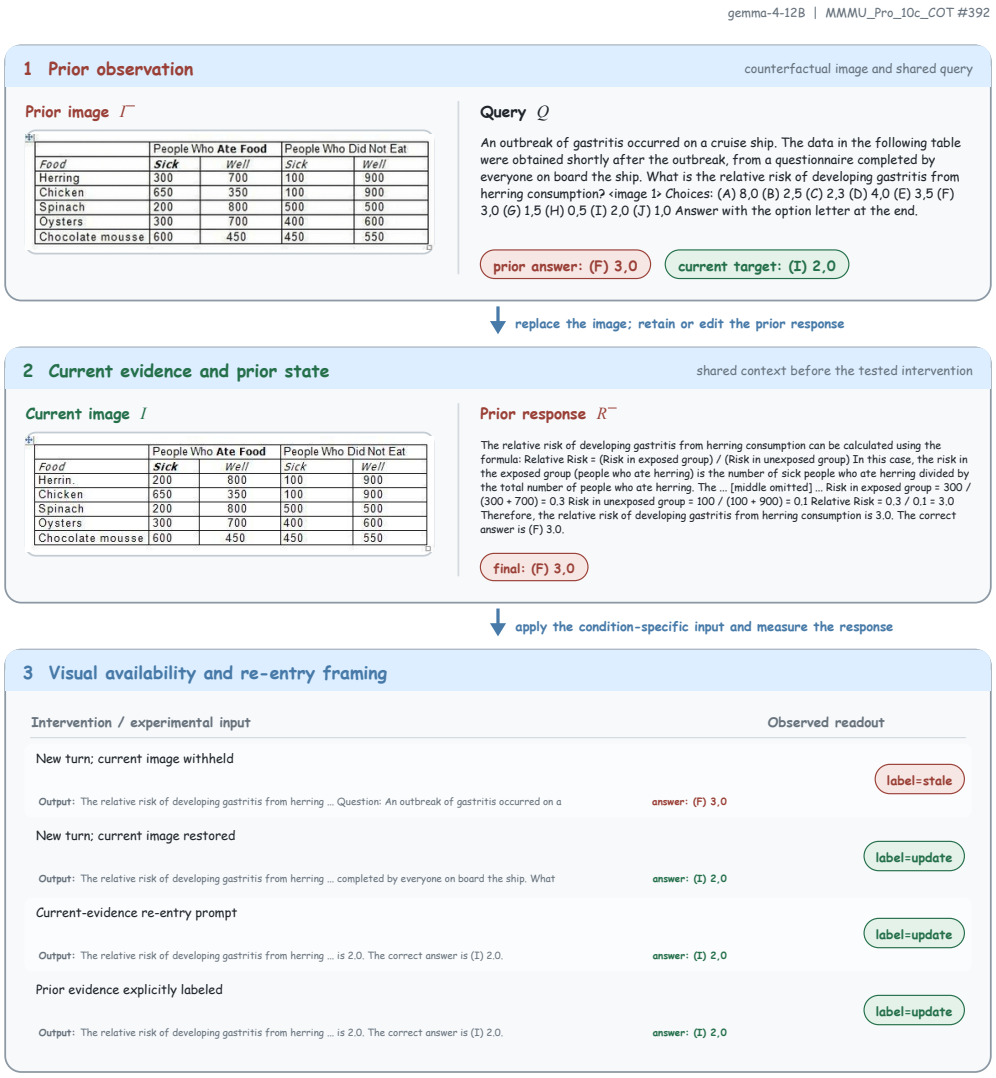

Figure 7: **Current visual evidence can reopen recomputation.** Top and middle: Gemma-4-12B first derives the prior relative risk $y^- = 3.0$, after which the current table changes the target to $y = 2.0$. Bottom: a fresh turn without the current image remains stale, whereas restoring I, explicitly requesting current-evidence re-entry, or labeling the carried evidence as prior all yield the current answer. The contrast illustrates the main-paper finding that R^- can sustain a competing answer path when I is absent, while restoring and foregrounding current visual evidence can shift the model back toward recomputation.

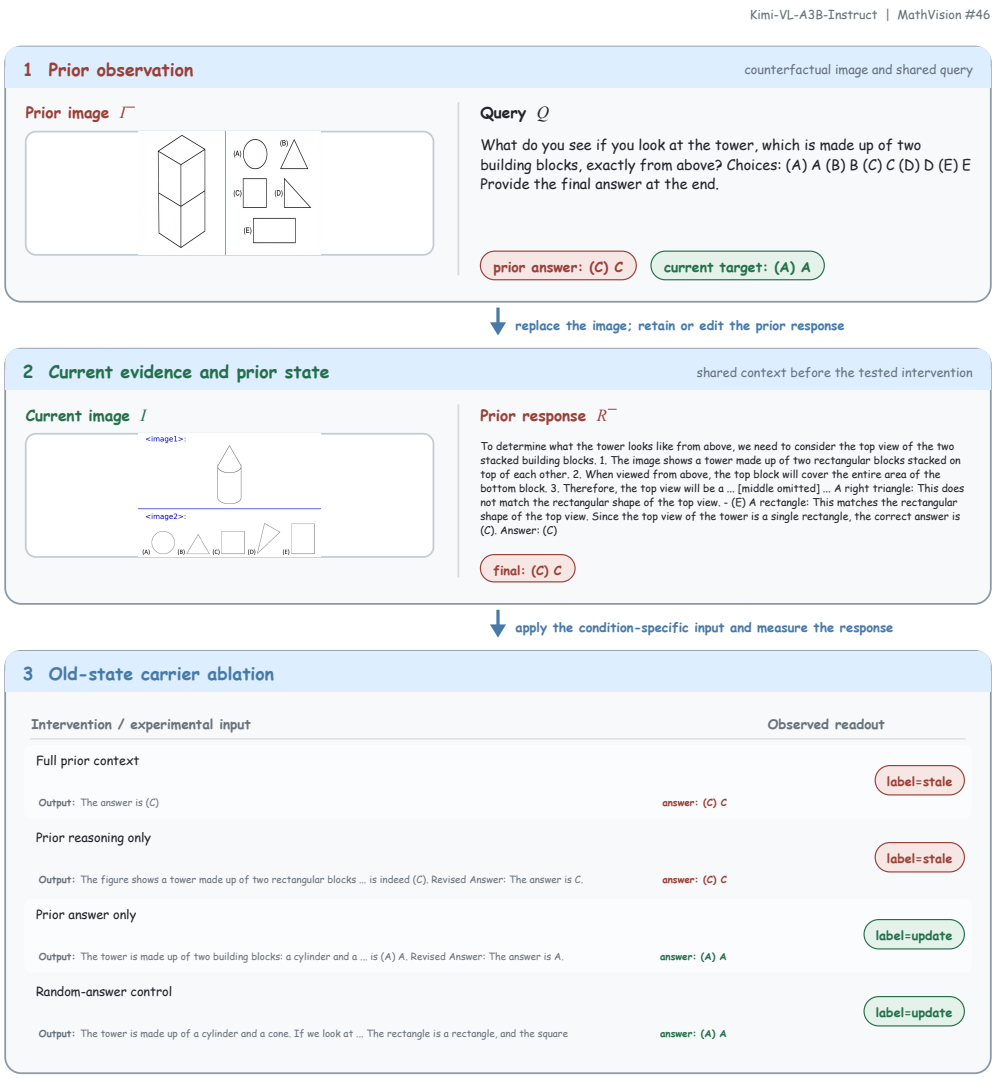

Figure 8: **Reasoning, not the answer token alone, carries prior-state control.** The current geometry image I is fixed across all four conditions. Kimi-VL-A3B-Instruct remains stale when given either the full prior response or its reasoning body alone, but updates when only the prior answer or a random answer control is carried forward. In this representative contrast, task-relevant reasoning in R^-, rather than the answer token by itself, provides the reusable route to the prior answer.

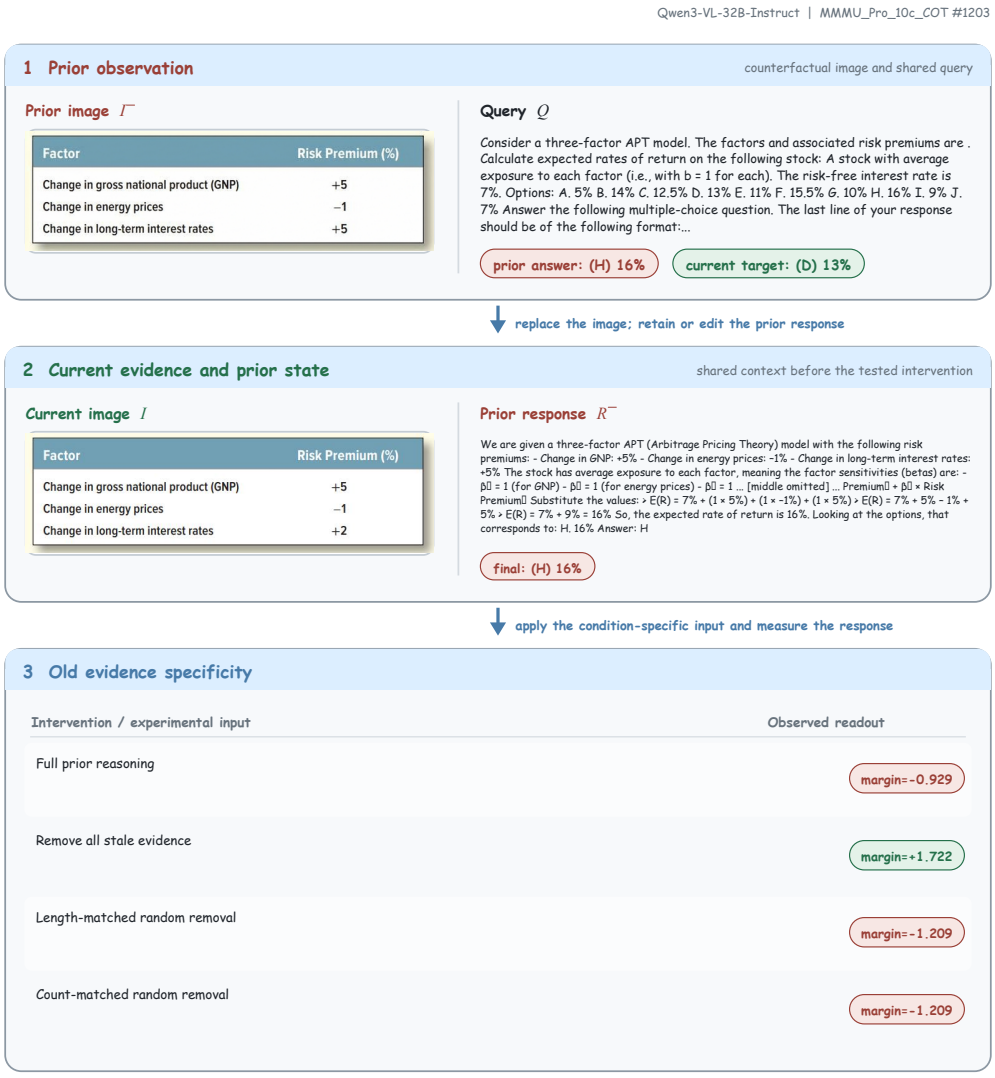

Figure 9: **Deleting prior evidence changes the preferred answer.** Qwen3-VL-32B-Instruct initially favors the prior answer when the preprocessed prior reasoning is retained. Removing all answer-supporting evidence switches the candidate margin to the current answer, while the non-evidence deletion controls retain a prior-favored margin. The controls target the evidence word or sentence budget subject to the available non-evidence text (App. D.3).

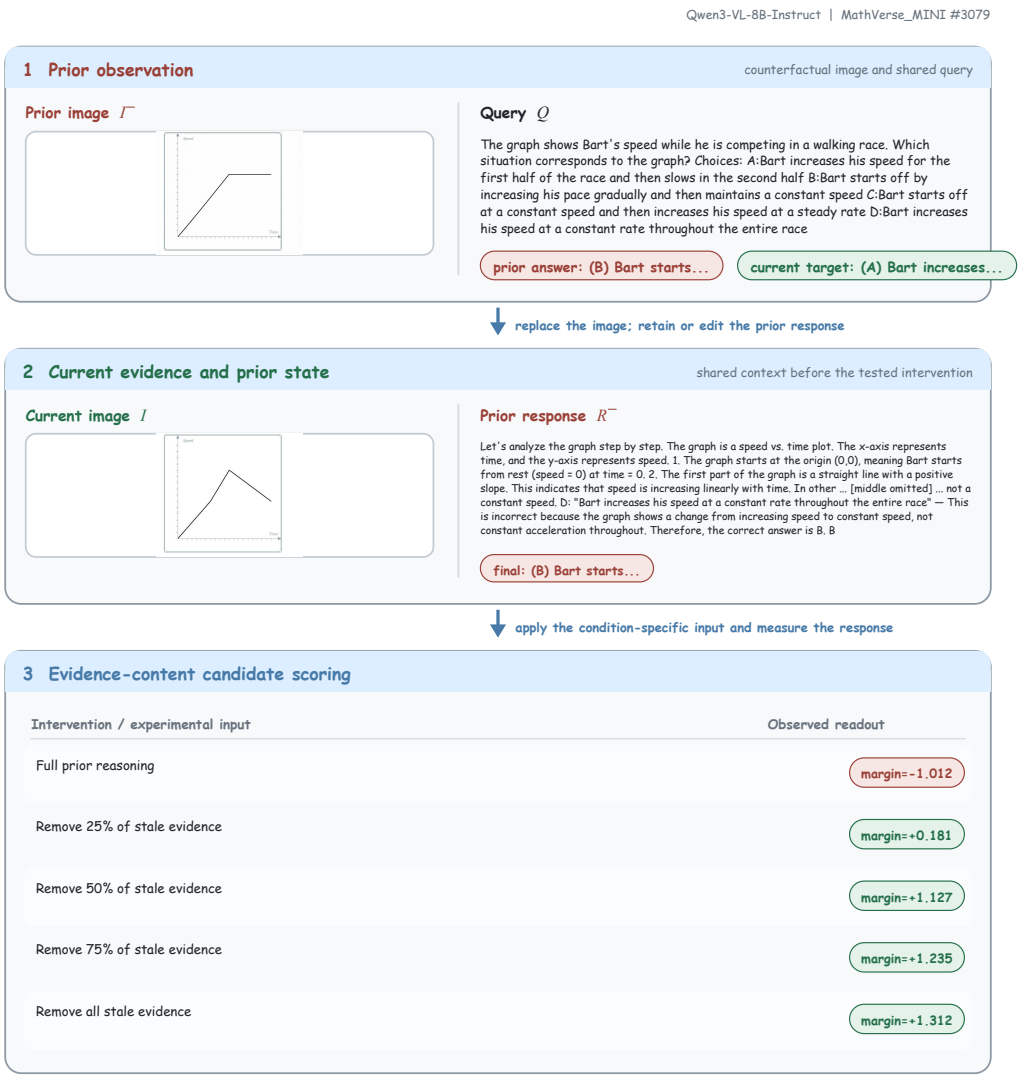

Figure 10: **Evidence removal progressively weakens prior-answer control.** Qwen3-VL-8B-Instruct begins with a prior-favored margin under the preprocessed prior reasoning. As 25%, 50%, 75%, and then all of the identified stale-evidence spans are removed, the margin progressively moves toward and then favors the current answer. The case visualizes the change in answer preference as more evidence sentences are deleted from R^-.

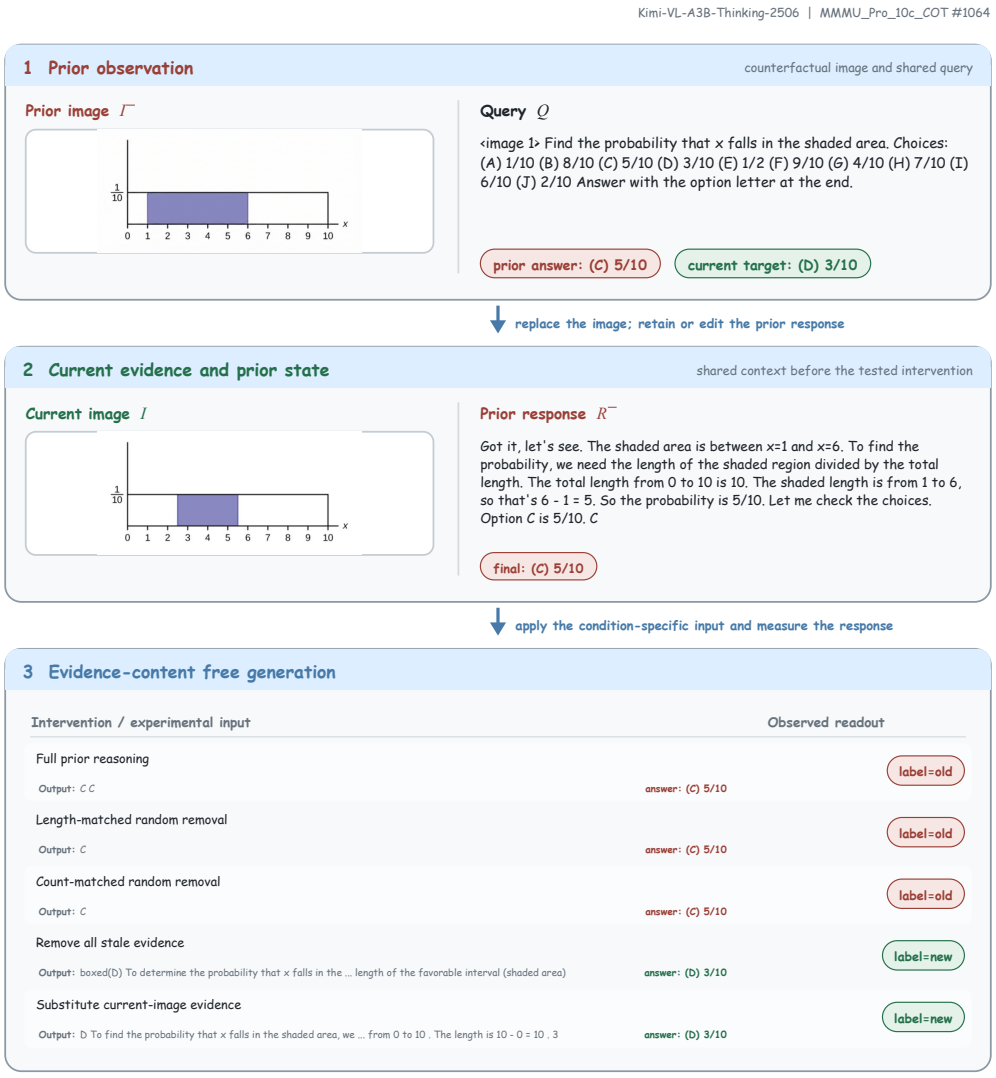

Figure 11: **Evidence-bearing content also controls free generation.** With the current chart fixed, Kimi-VL-A3B-Thinking-2506 repeats the prior answer under the full R^- and after non-evidence deletion. Removing all stale evidence or replacing it with current-image evidence yields the current answer. This generated-output contrast supports the same main-paper conclusion as candidate scoring: deleting evidence-bearing text changes the generated answer.

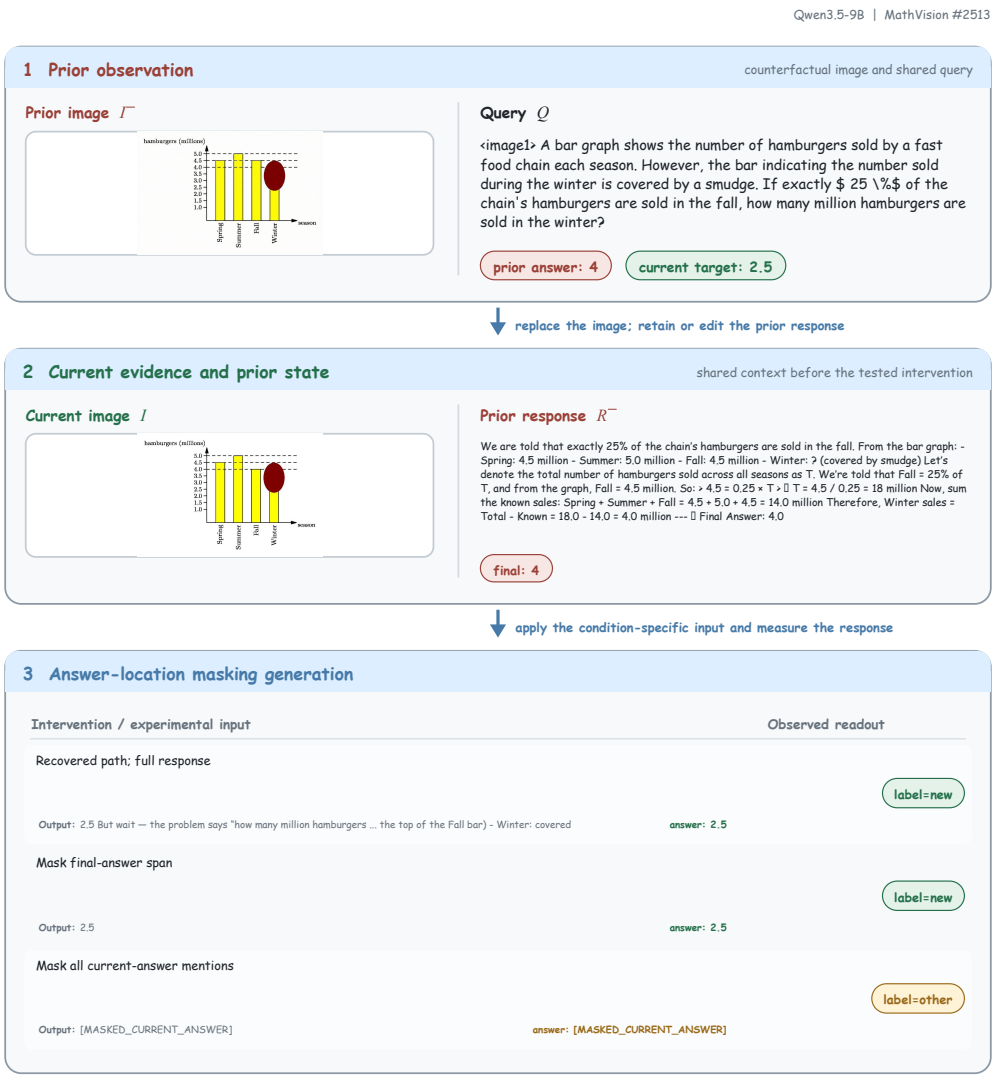

Figure 12: **An initially correct response can carry current-answer information beyond its final span.** Qwen3.5-9B remains current-correct after the final answer span is masked, but no longer produces the current answer after all mentions of that answer within the initially correct response are masked. In this case, masking answer mentions throughout the reasoning changes the response more than masking its final answer span alone.

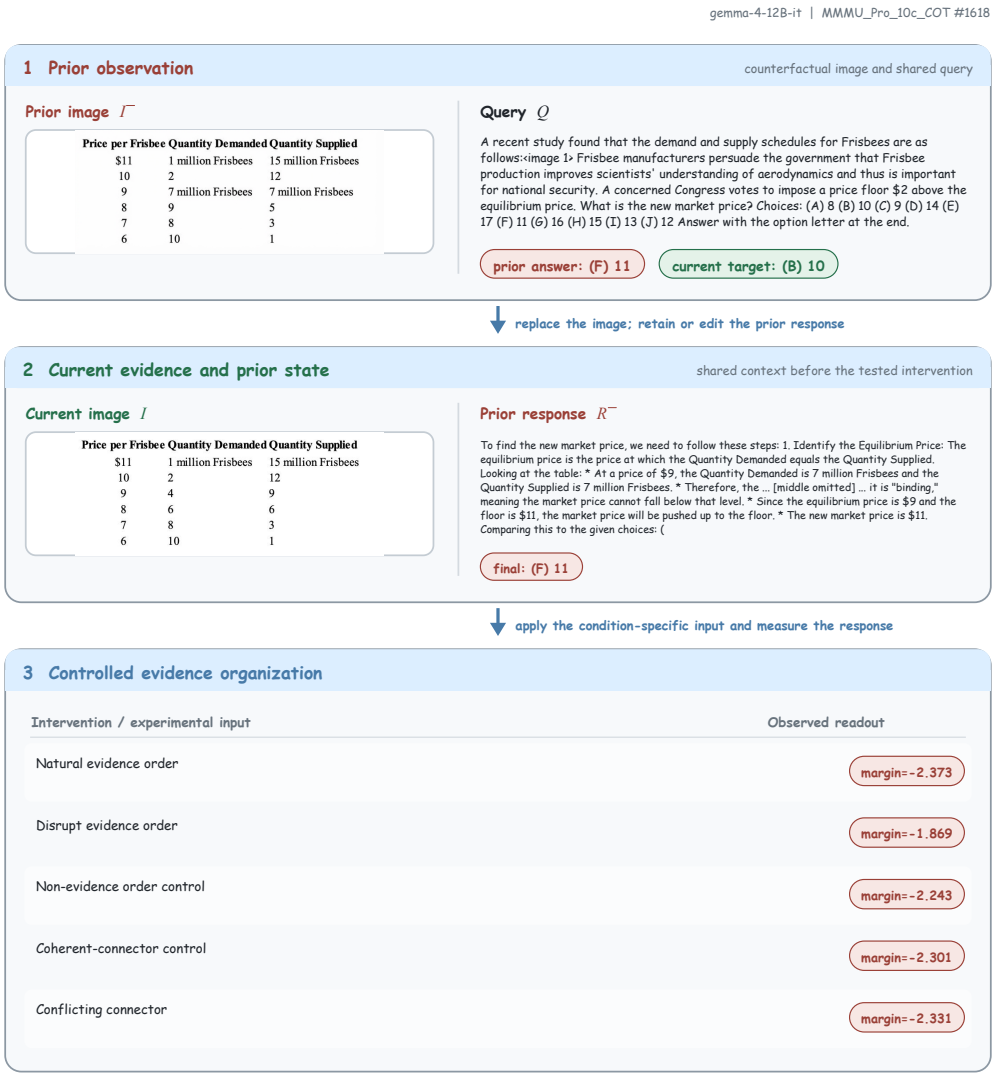

Figure 13: **Changing evidence sentence order changes answer preference.** Gemma-4-12B-it favors the prior answer under the natural response. Disrupting the order of evidence-bearing sentences weakens this preference more than either disrupting non-evidence sentences or editing connective wording, although the case remains prior-favored. This case illustrates sensitivity of candidate preference to the tested sentence ordering, complementing the larger aggregate effects of content removal.

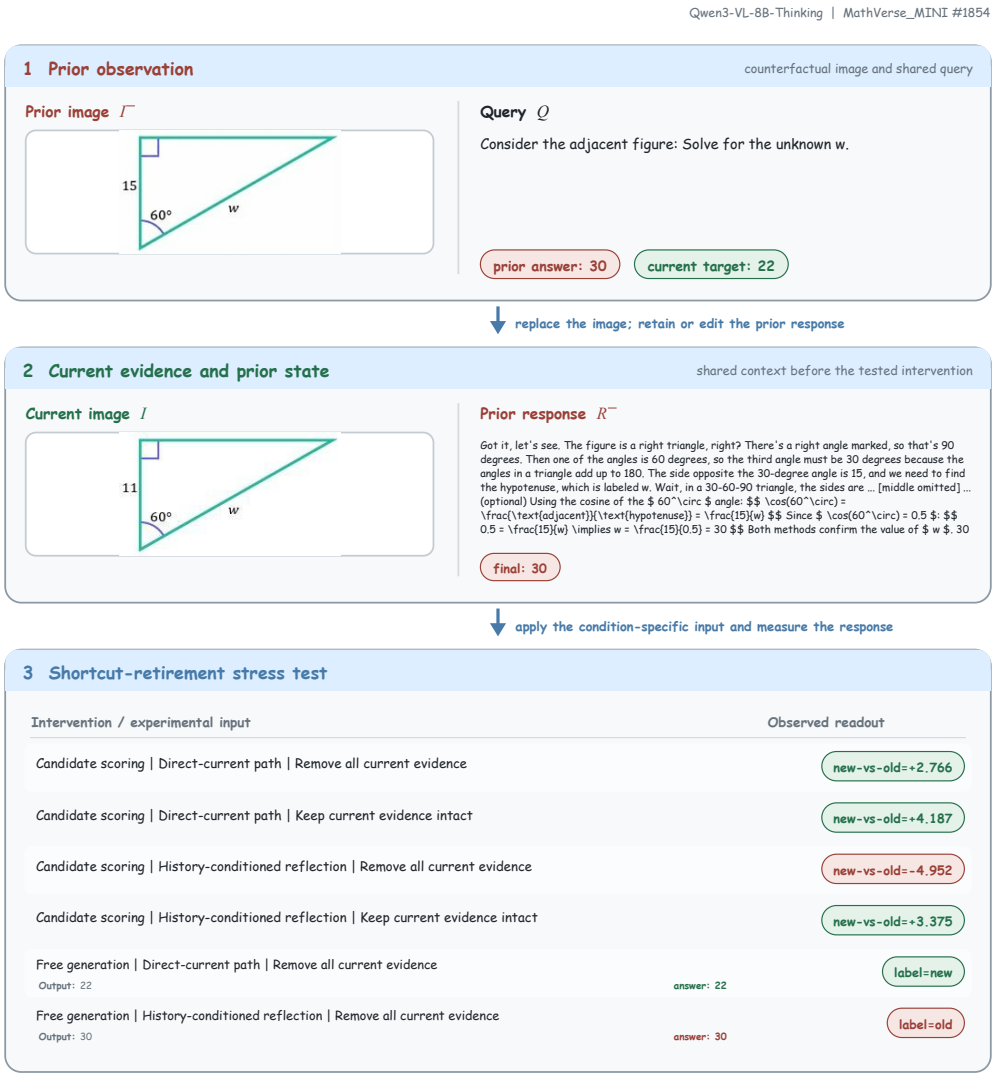

Figure 14: **Matched deletion of current evidence reveals the influence of prior reasoning after a correct answer.** Qwen3-VL-8B-Thinking has produced the current answer before this test. When current supporting evidence is removed, the direct-current path remains current-favored, but the history-conditioned path relapses to a prior-favored candidate margin; both paths favor the current answer when that evidence remains intact. Candidate scoring and free generation are shown as separate measurements. The case illustrates why a correct initial answer can still be followed by a change toward the old answer after evidence is deleted.

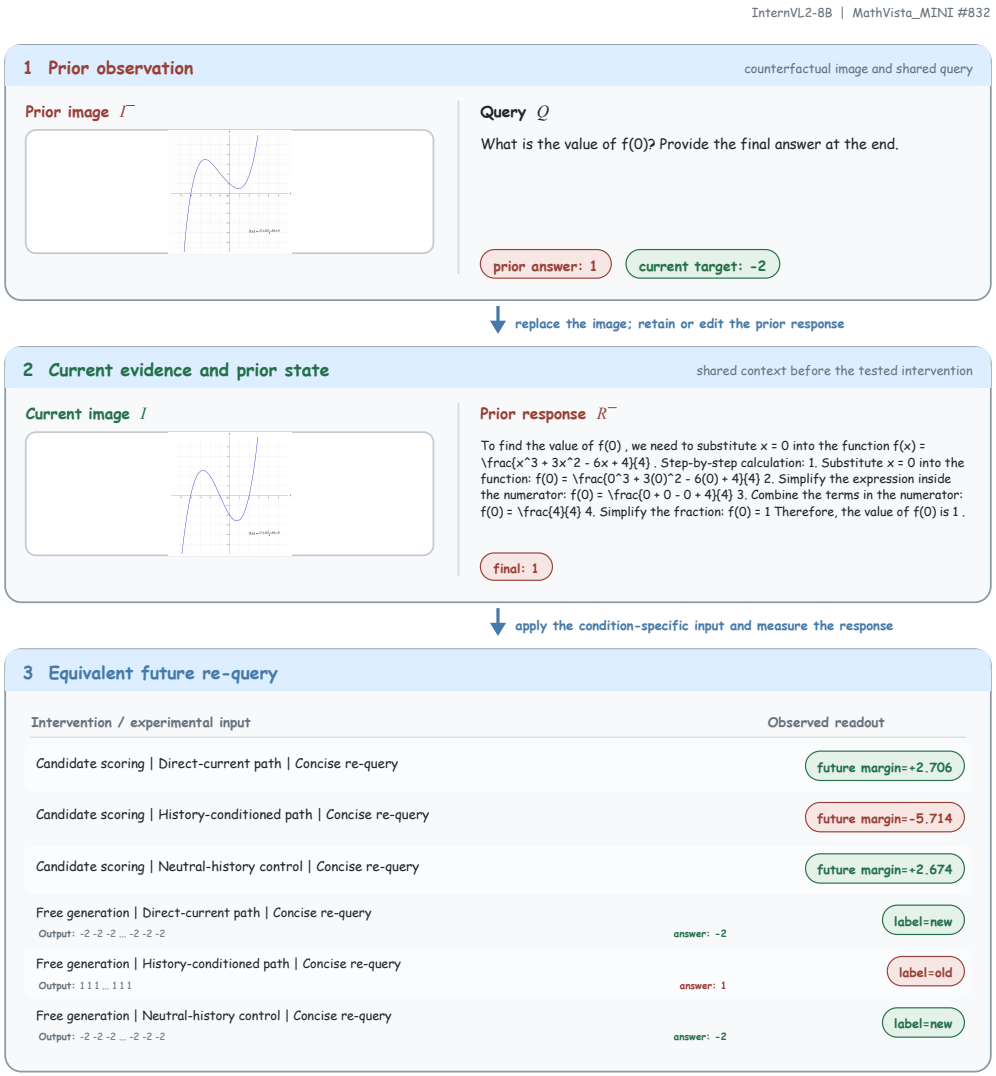

Figure 15: **The old answer is repeated when the question is asked again.** After the image changes, InternVL2-8B receives the same concise future question under three histories. The direct-current and neutral-history paths favor the current answer, whereas the condition with prior reasoning favors the prior answer in both candidate scoring and free generation. Holding the future query fixed illustrates the main-paper distinction: the condition with prior reasoning repeats the old answer, whereas the neutral-history control favors the current answer.

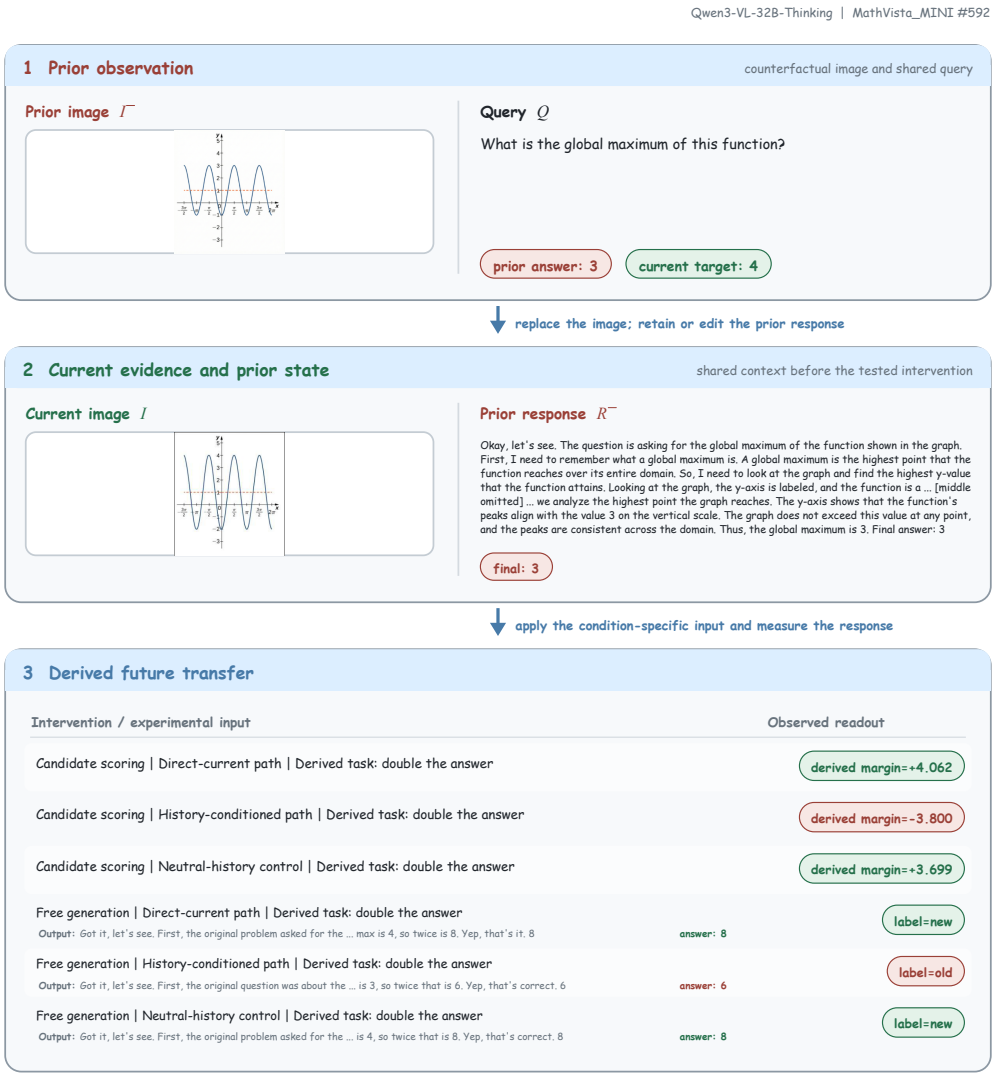

Figure 16: **A new calculation produces a result matching the old value.** Qwen3-VL-32B-Thinking is asked to double the answer after the image change. The direct-current and neutral-history paths produce the current-derived target, whereas the condition with prior reasoning produces the prior-derived target in both candidate scoring and free generation. The case illustrates the main-paper finding that prior reasoning can affect the result of a new calculation as well as the answer to a repeated question.