QUESTION-SPECIFIC TABLE VALUES REDUCE SENSITIVITY TO THE CHART IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a model receives both a chart and an extracted table, it can in principle check the table against the chart. We test whether restating the table entries as the two quantities a question compares makes correct answers less likely when these sources disagree. From 96 synthetic specifications, we render opposite-answer image pairs and cross them with tables that match only one image. Three 8B–12B models receive an experimenter-supplied assistant report that either repeats four table entries or states the two quantities the question compares; both versions come from the same table and are padded to equal token length. Under direct answering, question-specific values lower conflict accuracy from 59–68% to 0–13% while raising agreement accuracy from 84–90% to 96–100% across JSON and prose. In a separate study of model-extracted values on 96 ChartQA-derived images, adding a calculation report to each model's own work record improves average accuracy by 9–17 percentage points across five models, including additional Qwen and InternVL 4B checkpoints. Among the 26 conditions where the extracted values reverse the comparison, however, correct final answers fall from 13 to zero. A chart-directed reasoning request attenuates or reverses the controlled representation effect in seven of eight model–format combinations on four 2B–4B checkpoints (24 reused specifications). On a new panel with the main models, cutting report attention in the middle third of the decoder raises accuracy with question-specific values by 26–33 percentage points, and jointly cutting image attention removes that gain. The raw/values contrast differs significantly between report and equally long text-control cuts for Qwen and InternVL, but not Gemma. These results show that chart sensitivity depends on the form of a table-derived report and on the answering protocol.

1 INTRODUCTION

Chart question answering requires reading quantities from an image and operating on them (Masry et al., 2022). DePlot separates chart reading from subsequent language reasoning by translating plots into tables (Liu et al., 2023a); for numerical reasoning, Program of Thoughts and PAL execute generated programs in an external interpreter (Chen et al., 2023; Gao et al., 2023). Correct arithmetic, however, cannot repair an extraction error. If the error reverses the comparison, the answer computed from the table is wrong for the chart. We ask whether an answering model overrides such a table when it still has the chart and is warned that the supplied numbers may disagree with it.

We study the form of a report inserted between transcription and answering. For the question in Figure 1, the model must compare K34 + K77 with K26 + K30. The report can repeat the four table entries $(70, 25, 15, 17)$ or supply the two sums, 95 and 32, labeled by answer option. Both versions use only the table, but the second presents precomputed quantities aligned with the answer options. We test whether replacing the first report with the second changes the answer when the chart contradicts the table.

We evaluate each context separately with the raw and values reports, holding the image, question and table fixed for this paired comparison. A second comparison changes the chart to reverse the

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.

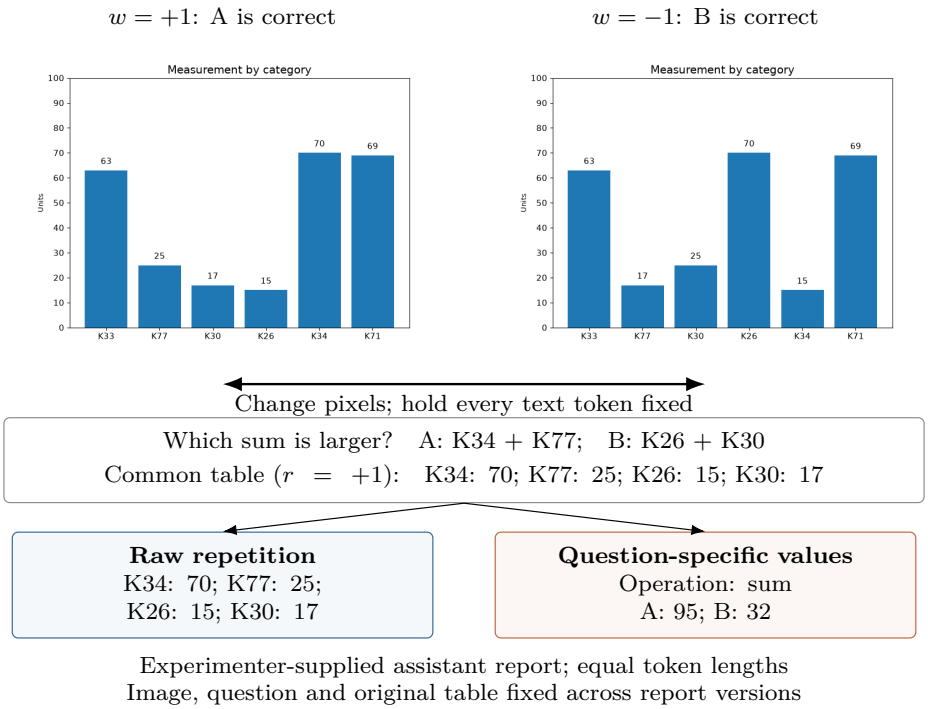


Figure 1: Controlled comparison of report representations. Opposite-answer chart images share identical question and table text; the table agrees with only one image. The assistant report repeats four entries (raw) or gives two question-specific quantities (values). The charts are actual experimental inputs; the reports are abbreviated, omitting padding used to equalize token counts.

correct answer while preserving every text token. Crossing image and table worlds independently then lets us distinguish response patterns that follow the chart from those that follow the table.

Under direct answering, question-specific values yield a lower signed chart response than raw repetition in all six main model–format comparisons (Section 4). This also occurs on read questions, where the report selects and labels entries without new arithmetic. Reports computed from the models' own extracted values on ChartQA-derived charts improve average accuracy in all five models, but correct answers are lost in the subgroup where extraction reverses the comparison (Section 5).

We next examine when the contrast changes. A chart-directed reasoning request often reverses its sign on four smaller checkpoints (Section 6). On the main models, middle-band report cuts raise values-arm accuracy; the signed raw/values contrast differs significantly from the matched text-control cut for Qwen and InternVL, but not Gemma (Section 7).

Our contribution is a paired evaluation of table-derived reports that distinguishes agreement gains from conflict errors, measures chart sensitivity through image changes with text held fixed, and tests how the report effect depends on the answering protocol.

Code and saved results: anonymous.4open.science/r/chart-table-artifact-71A6/.

2 Related Work

Chart QA benchmarks include questions over generated charts in DVQA and PlotQA (Kafle et al., 2018; Methani et al., 2020) and visual and logical reasoning over real-world charts in ChartQA (Masry et al., 2022). MatCha and UniChart connect chart structure and data extraction with numerical reasoning through chart-specific pretraining (Liu et al., 2023b; Masry et al., 2023). For inference, DePlot translates plots into tables for subsequent language reasoning (Liu et al., 2023a).

In our experiments, the answering model retains the chart even when the supplied table contradicts it.

IsoBench and MathVerse evaluate how the representation and information supplied to a model affect its answers. IsoBench compares isomorphic visual, textual and mathematical inputs, including joint presentation through IsoCombination (Fu et al., 2024); MathVerse varies the visual and textual information in mathematical problems to assess diagram use (Zhang et al., 2024). When image and text disagree, source reliance depends on the model, task and prompt (Deng et al., 2025; Pezeshkpour et al., 2025). In the chart setting, Ghosh et al. (2026) study this reliance by gradually degrading either the image or the text while keeping the other source unchanged. Our comparison instead changes a table-derived report at matched token length, leaving the chart and supplied table fixed.

Multimodal hallucination snowballing describes how a hallucinated statement in a conversation can mislead a VLM's later answers (Zhong et al., 2024). We also place a report before the final answer: we supply it ourselves in the controlled experiment and append a calculation to the model's own work record in the natural-extraction study. Chain-of-thought prompting can improve accuracy (Wei et al., 2022), but Sánchez Villegas et al. (2026) show that a trace can appear visually grounded while following misleading text. In our protocol comparison, we score final answers under a request that permits written reasoning and directs the model to the chart. We do not use the trace as evidence of visual reliance.

Attention knockouts and activation interventions test computational dependencies (Geva et al., 2023; Kaduri et al., 2025); their interpretation depends on the intervention, control and readout (Zhang & Nanda, 2024). For conflicts between visual attributes and linguistic priors, Nooralahzadeh et al. (2026) distinguish attribute decodability from the final answer using probes and activation patching. Ortu et al. (2026) use logit inspection and attention-head interventions to study conflicts between parametric knowledge and counterfactual images. We examine externally supplied tables and reports through attention cuts in fixed decoder bands. For descriptive layer readouts, we apply the model's existing output head (the raw logit lens), without fitting the intermediate decoders learned by a tuned lens (Belrose et al., 2023). Appendix G explains how the choice of readout limits interpretation.

3 Controlled representation experiment

Inputs and treatment. A frozen model receives a chart I, a question q, a label–value table $\mathcal{T}$, and a report $C_a = g_a(\mathcal{T}, q)$, with $a \in \{\text{raw}, \text{values}\}$:

$$\widehat{y}_a = f_\theta(I, q, \mathcal{T}, g_a(\mathcal{T}, q); \pi). \tag{1}$$

The protocol π specifies the message sequence, final request, decoding and parser. Raw repetition lists four category–value entries. Question-specific values give the two selected entries, sums or absolute within-pair differences that the question compares. The quantities are labeled A and B but no winner is declared, so the report stops one comparison short of stating the answer. The supplied table stays in the context in both report conditions (arms), and both reports say that they use the reference measurements and have not checked the image. We add deterministic nonnumeric padding to both reports to give them exactly the same token length within each checkpoint and format. Replacing raw with values jointly changes which quantities are displayed, their alignment with answer options, explicit operation cues and, for sum and difference tasks, their precomputation. We estimate the effect of this complete replacement; matching token length does not isolate these components or equalize the salience of the relevant quantities.

Crossed worlds. VQA v2 pairs the same question with similar natural images requiring different answers (Goyal et al., 2017). We use procedurally controlled chart pairs and independently cross the supplied tables. We generate 96 base specifications, equally divided among vertical bars, horizontal bars and dots. Each contains six categories and four queried values. For each task, two rendered worlds reverse the correct canonical alternative $w \in \{-1, +1\}$ while leaving category positions and labels unchanged. Independently, two supplied tables imply alternatives $r \in \{-1, +1\}$. Agreement means $w = r$; conflict means $w \neq r$. We cross three tasks, two option orders, annotated or unannotated marks, JSON or prose report format and both representations. Figure 1 shows the first base's sum task with one of the two table worlds. All charts are rendered procedurally from the saved numerical specifications.

Models and answers. The three main models are Qwen3-VL-8B-Instruct, InternVL3.5-8B-HF and Gemma-3-12B-IT (Bai et al., 2025; Wang et al., 2025; Gemma Team et al., 2025). We use free-running greedy generation, called *native* in the recorded results, capped at 32 tokens. There is no answer prefill and no request for written reasoning. After stripping outer whitespace, a strict parser accepts either bare A/B or exactly one complete answer tag, with no other text. Invalid outputs count as wrong for accuracy and stay in every denominator. Each model–format–report cell contains 4,608 answers, half in conflict. For statistical inference, the 96 bases are the units: all renderings and answers derived from a base are grouped together. Appendix A gives prompts, generation details and checkpoint revisions.

Signed chart response. Order $o = 0$ assigns A to the canonical $+1$ alternative (Figure 1); $o = 1$ reverses the labels. Set $\sigma_o = (-1)^o$ and $u_a = \sigma_o(\mathbf{1}\{\widehat{y}_a = A\} - \mathbf{1}\{\widehat{y}_a = B\})$ for parsed $\widehat{y}_a \in \{A, B, \perp\}$; invalids map to zero. At fixed r and identical text, define

$$\Gamma_a = \mathbb{E}\left[\frac{u_a(+1, r) - u_a(-1, r)}{2}\right], \qquad \Delta\Gamma = \Gamma_{\text{values}} - \Gamma_{\text{raw}}. \tag{2}$$

The expectation is taken over the balanced finite design. Answering both images correctly contributes $+1$, a constant valid answer contributes zero, and two incorrect valid answers contribute -1. Thus Γ_a measures the signed, truth-aligned response: reversing an answer against the chart's truth lowers the score. With accuracy A_a and valid-answer coverage V_a,

$$\Gamma_a = \mathbb{E}[wu_a] = 2A_a - V_a. \tag{3}$$

Equation (3) lets us interpret $\Delta\Gamma$ as twice the accuracy change minus the coverage change. Complete four-world response patterns are reported in Appendix C. We test the six primary model–format contrasts with paired t tests over bases and Holm correction (Holm, 1979). Intervals resample whole bases (Efron, 1979) at the nominal 95% level, without multiplicity adjustment. We also bound how far a contrast can move when invalid answers are replaced with A/B (Section 6) or when responses flagged by repeat/backend discrepancies are allowed to vary over $[-1, 1]$ (Section 7). These bounds concern changes to outputs on the observed panel. The confidence intervals describe sampling uncertainty over bases.

4 Representation effects on the 96-base panel

Figure 2 shows a shift toward the table in every model–format cell. With the values report, conflict accuracy is 0.2–12.5%, down from 58.8–67.9% with raw repetition. Agreement accuracy rises from 83.6–90.3% to 95.9–100.0%. The six primary contrasts range from $\Delta\Gamma = -0.521$ to -0.420, all significant under the six-test Holm correction (Table 1). For raw repetition, Γ_{raw} ranges from 0.423 to 0.582; for values, Γ_{values} is 0.002–0.112. On this signed measure, restatement reduces how much answers track the correct alternative when the image changes, although the text remains identical within each pair.

Table 1: Primary representation contrasts in the 96-base experiment. $\Delta\Gamma = \Gamma_{\text{values}} - \Gamma_{\text{raw}}$; V_r/V_v denotes raw/values coverage over 4,608 answers per arm. p values use paired base-level tests and Holm correction over the six tests.

Model	Format	$\Delta\Gamma$	Nominal 95% CI	V_r/V_v (%)	Adj. p
Qwen 8B	JSON	-0.521	$[-0.566, -0.475]$	100.0/100.0	$< 10^{-6}$
Qwen 8B	prose	-0.471	$[-0.517, -0.426]$	100.0/100.0	$< 10^{-6}$
InternVL 8B	JSON	-0.470	$[-0.502, -0.437]$	100.0/100.0	$< 10^{-6}$
InternVL 8B	prose	-0.515	$[-0.548, -0.481]$	99.5/100.0	$< 10^{-6}$
Gemma 12B	JSON	-0.422	$[-0.461, -0.384]$	100.0/100.0	$< 10^{-6}$
Gemma 12B	prose	-0.420	$[-0.466, -0.375]$	100.0/100.0	$< 10^{-6}$

The four-world diagnostic (Appendix C) separates table following from a constant option preference. For Qwen/JSON, the number of table-consistent quartets rises from 110 to 1,146 out of 1,152 under the values report. None of its values quartets makes a constant canonical choice.

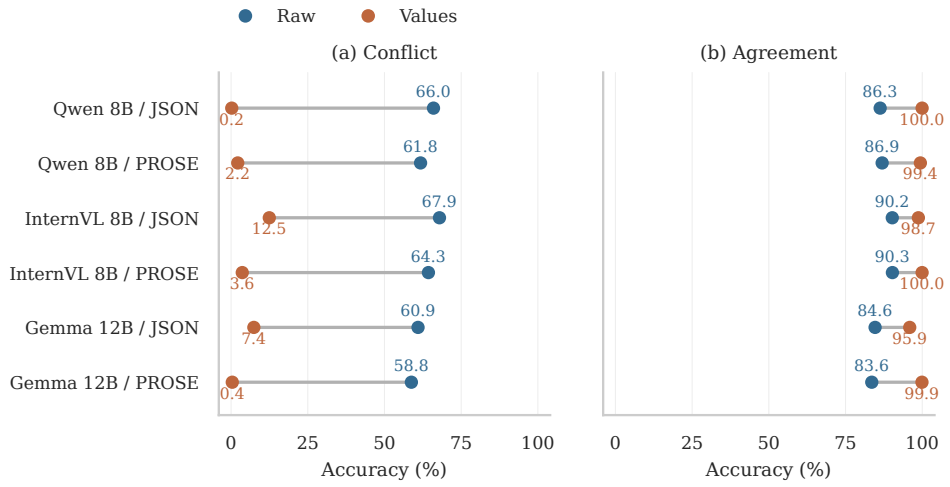


Figure 2: Question-specific values lower conflict accuracy and raise agreement accuracy across the three main models and both report formats. Each point summarizes 2,304 answers from 96 base specifications; invalid answers count as wrong.

The shift also occurs without new arithmetic. In the read task, the values report selects two of the four entries and labels them A and B, while the original table stays in both contexts. The six $\Delta\Gamma$ estimates range from -0.344 to -0.561 (Figure 7, which also separates geometry and annotation). Because selection and option alignment change together, we cannot separate their contributions in this comparison.

5 Calculation reports on ChartQA-derived charts

To examine the pattern with model-extracted values, we use 96 ChartQA-derived images (Masry et al., 2022) with readable annotations: 82 bar, 9 line and 5 pie charts. Read and sum questions in both option orders give 384 conditions per model. Each of five models (Qwen 4B/8B, InternVL 4B/8B and Gemma 12B) is asked to produce a work record containing four values and a claimed winner.

Each saved work record is reused in two final-answer branches, both of which retain the chart. In the first, the model answers directly after the record. In the second, it first receives a deterministic calculation report listing the two quantities, computed winner, claimed winner, internal-agreement field and an explicit note that image truth was not checked. When parsing fails, this branch receives a standard incomplete-record notification before generating its answer. Unlike the equal-length replacement in the main experiment, appending this report lengthens the context.

Adding the report improves average accuracy in every model, by 8.6 percentage points for Qwen 4B to 16.9 for InternVL 8B over the 384 conditions (Figure 3; exact counts in Table 12). Among 1,920 records, the parsed comparison is correct in 1,833, reversed in 26, tied in one and incomplete in 60. We classify records by the comparison implied by their extracted values, so an incorrect individual value can still occur in the correct-comparison group.

The average gain masks a loss in the reversal subgroup. Among its 26 conditions from 11 charts, 13 were answered correctly before the report was added and none afterwards. All 26 came from Qwen 4B, Qwen 8B and Gemma; neither InternVL model produced a reversal in this sample (Appendix D). Besides the two quantities, this report supplies the computed winner, the claimed winner and whether they agree. We therefore measure the effect of the whole report, rather than the values alone.

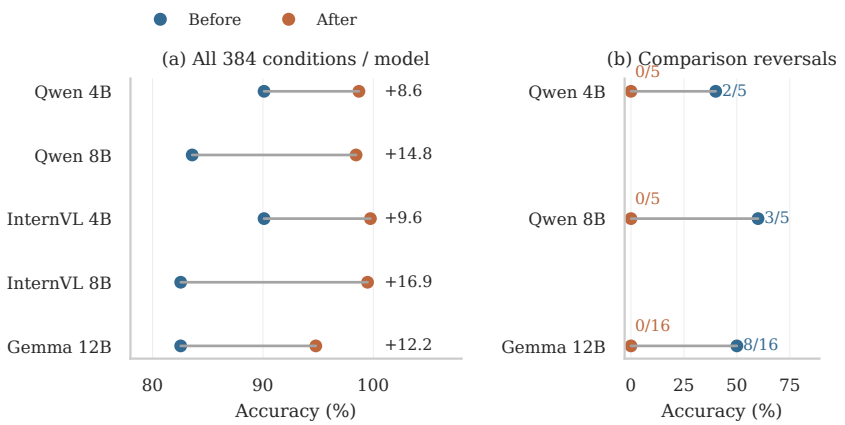


Figure 3: Adding a calculation report raises average accuracy but eliminates correct answers in the observed reversal subgroup. Left: accuracy over 384 conditions per model from 96 ChartQA-derived charts. Right: correct answers fall from 13 to zero among 26 comparison-reversing extraction conditions from 11 charts.

6 Effects of the answering protocol

The protocol study evaluates Qwen3-VL 2B/4B and InternVL3.5 2B/4B on the first 24 bases of the main experiment, keeping the images, formats and annotation states unchanged. We compare direct answering with a 32- or 512-token cap and chart-directed reasoning with a 512-token cap. The direct-answer conditions share a prefix and differ only in the cap; the reasoning condition allows written reasoning and asks the model to work from the chart. All three use the same terminal-answer parser, which accepts preceding reasoning.

The 32-token direct-answer condition yields seven negative contrasts out of eight (Figure 4), each significant under the eight-test Holm correction. Their negative signs are robust to replacing each invalid answer with either A or B, in any combination (Proposition 1). The exception is InternVL 2B/prose ($\Delta\Gamma = +0.059$, adjusted $p = 0.0125$), where only 16.2% of raw answers parse as valid, compared with 72.1% of values answers. In a post hoc check, we remove at most one trailing period from invalid completions and re-parse all outputs. The contrast falls to $+0.003$, with an interval spanning zero, so the positive result is sensitive to answer formatting. All main tests retain the primary parser; Appendix E gives coverage, absolute accuracies and the re-parsing details.

Raising the cap alone changes $\Delta\Gamma$ by -0.012 to $+0.003$ on this panel. At the same 512-token cap, adding the chart-directed reasoning request gives positive estimates of $\Omega = \Delta\Gamma_{\text{reasoning}} - \Delta\Gamma_{\text{direct}}$ in all eight cells. Seven are significant after Holm correction over the fixed 16-test cap/instruction family. Qwen 4B changes from approximately -0.51 to $+0.44/+0.47$ across formats, and InternVL 4B from $-0.71/-0.68$ to $+0.13/+0.34$. For Qwen 4B/JSON, the shift involves fewer correct raw answers, from 875 to 813, and more correct values answers, from 576 to 1,064, out of 1,152 in each arm. Because the request combines a chart directive with permission for written reasoning (Appendix A), we measure their joint effect.

7 Effects of report and image attention cuts

A new panel on the main models. We generated 24 new specifications shared by Qwen 8B, InternVL 8B and Gemma 12B. All 72 base–task units per model were completed using JSON assistant reports, unannotated charts and direct answers capped at 32 tokens. Before generating these evaluation outputs, we fixed the middle-band hypotheses based on an earlier attention-cut study with two 4B models on a separate 24-base panel (Appendix F.5).

Interventions and controls. Following attention-knockout analyses (Geva et al., 2023; Kaduri et al., 2025), we restrict access to specified token spans. In a selected decoder band b, we mask

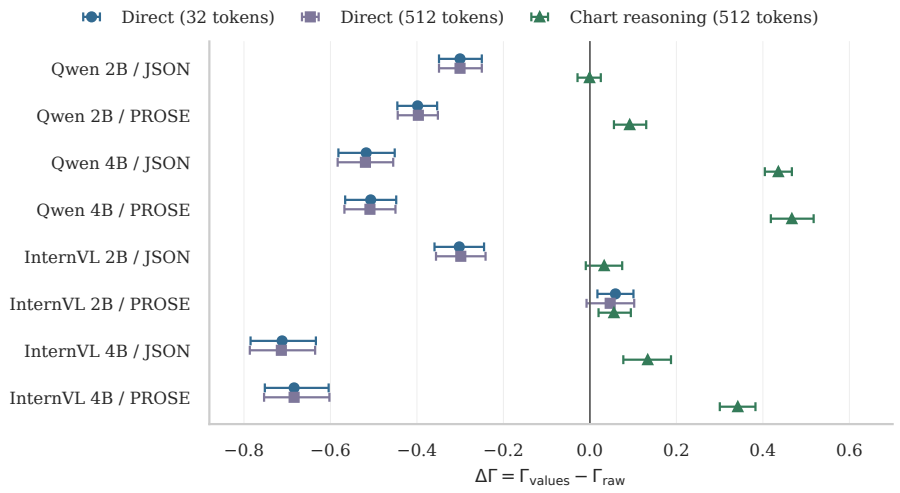


Figure 4: Representation contrasts under three answering protocols on 24 reused bases. The two direct-answer settings differ only in their generation cap; the third permits written reasoning and directs the model to the chart. Points show $\Delta\Gamma$ (values minus raw) with nominal 95% base-bootstrap intervals. Negative values indicate a smaller signed chart response with the values report.

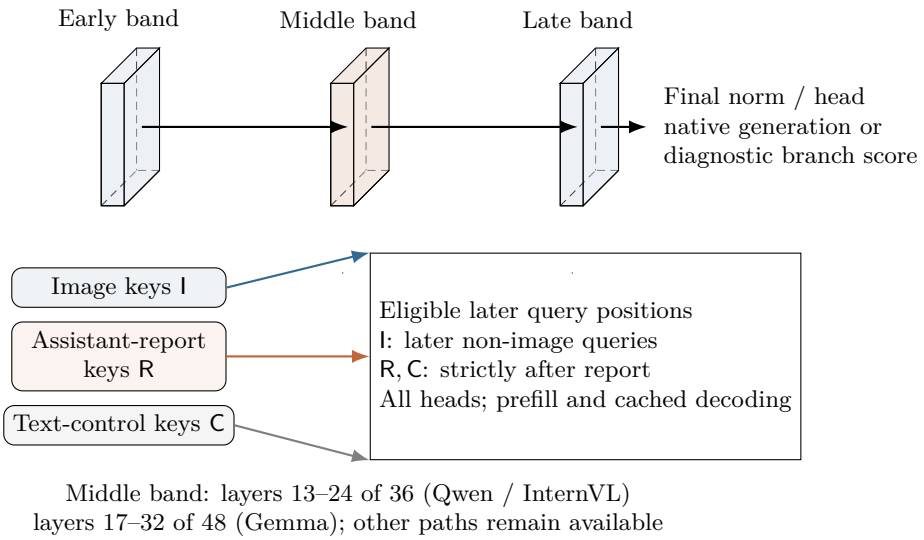


Figure 5: Attention-edge interventions within fixed decoder thirds. Report (R), matched preceding text (C), image (I) and joint (IR) cuts remove the indicated allowed key-to-query edges in all heads during prefill and cached decoding. Indirect paths remain open; the text control matches report length and query destinations.

allowed attention edges in all heads during prefill and every cached decoding step. R_b removes attention to report keys from queries strictly after the report. C_b uses the same query destinations and blocks attention to keys in an equally long text span immediately preceding the report. The control matches the report length and query destinations, but its text can include parts of the question or table. I_b removes image-key access from eligible later non-image queries; IR_b combines image and report cuts. We compare these cuts with the uncut condition none using the same explicit-mask backend. Early, middle and late bands partition decoder depth into thirds; middle means layers 13–24 of 36 for Qwen/InternVL and 17–32 of 48 for Gemma. Figure 5 and Appendix F specify the edge sets.

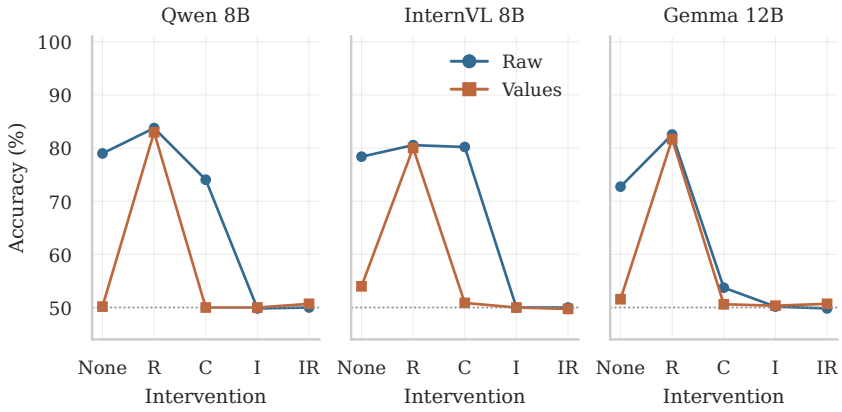


Figure 6: Middle-band report cuts improve values-arm accuracy, while joint image/report cuts yield approximately chance accuracy. Image cuts alone also yield approximately chance accuracy in both arms (Table 21). The study uses 24 new bases shared by the main models, with 1,152 generated answers per report and intervention condition, including two repetitions. The uncut condition uses the same explicit-mask backend.

Table 2: Main-model middle-band contrasts from generated answers on 24 new bases: report versus matched text cut (T) and interaction with the image cut (J). Adjusted p values use Holm correction over the full nine-test primary family.

Model	Contrast	Estimate	Nominal 95% CI	Adj. p
Qwen 8B	T	+0.465	$[+0.382, +0.552]$	9.00×10^{-5}
Qwen 8B	J	+0.550	$[+0.474, +0.628]$	9.00×10^{-5}
InternVL 8B	T	+0.576	$[+0.514, +0.635]$	9.00×10^{-5}
InternVL 8B	J	+0.483	$[+0.425, +0.540]$	9.00×10^{-5}
Gemma 12B	T	+0.045	$[-0.010, +0.099]$	0.144
Gemma 12B	J	+0.392	$[+0.293, +0.495]$	9.00×10^{-5}

Let $D(k) = \Gamma_{\text{values}}(k) - \Gamma_{\text{raw}}(k)$ for intervention k. The two within-band contrasts are

$$T_b = D(\mathsf{R}_b) - D(\mathsf{C}_b), \tag{4}$$

$$J_b = \{D(\mathsf{R}_b) - D(\mathsf{none})\} - \{D(\mathsf{IR}_b) - D(\mathsf{I}_b)\}. \tag{5}$$

T_b compares the signed representation contrasts under report and matched text-control cuts: $T_b > 0$ means $D(\mathsf{R}_b) > D(\mathsf{C}_b)$. J_b compares the report-cut effect on D without and with the image cut: $D(\mathsf{R}_b) - D(\mathsf{none})$ versus $D(\mathsf{IR}_b) - D(\mathsf{I}_b)$. The primary family contains nine tests, corrected jointly with Holm's procedure: for each model, D from the standard inference backend and middle-band T/J from generated answers. Each context is run twice, and repetitions are averaged within base before testing. All recorded generated answers in this panel are valid ($V_a(k) = 1$).

Middle-band results. Figure 6 shows both report arms on a common accuracy scale. Middle report cuts raise values accuracy from 50.2% to 83.0% in Qwen, 54.0% to 80.0% in InternVL, and 51.6% to 81.7% in Gemma. Middle T is significant in Qwen and InternVL. Gemma's estimate is $T = 0.045$, with nominal interval $[-0.010, 0.099]$ and adjusted $p = 0.144$. Middle J is positive and significant in all three models after correction over the nine primary tests. These six within-band contrasts (Table 2) and the three uncut D tests (Table 19) form the nine-test primary family.

Image cuts alone, however, yield approximately 50% accuracy in both report arms for every model (Table 21). Joint image/report cuts also yield approximately 50%, leaving $D(\mathsf{IR}_{\text{middle}}) - D(\mathsf{I}_{\text{middle}})$ close to zero. Consequently, J_{middle} largely reflects $D(\mathsf{R}_{\text{middle}}) - D(\mathsf{none})$. Its significance alone therefore does not identify a specific report–image pathway.

Gemma's nonsignificant T needs to be read alongside both accuracies. The text cut narrows the raw/values gap by lowering raw accuracy from 72.7% to 53.7%, while values accuracy changes

from 51.6% to 50.6%. The report cut also narrows the gap, but raises both accuracies, to 82.6% and 81.7%. Small gaps can therefore occur at very different accuracy levels; nonsignificance does not establish equivalence between the cuts. Across the three models, middle report cuts trade values-arm conflict gains of 58.3–75.7 percentage points for agreement losses of 6.3–10.2 points. If these conditional accuracies are held fixed, the break-even conflict frequencies are 9.68–12.69% (Appendix F); the net benefit depends on how often chart and table disagree.

Some generated answers differ across repetitions or backends. In a post hoc sensitivity analysis, we let every generated response at a flagged context vary over $[-1, 1]$, holding unflagged outputs fixed. The signs of all primary contrasts are robust across the resulting finite-panel ranges. Full numerical checks, all layer bands and descriptive raw-lens curves appear in Appendices F–G.

8 Discussion and conclusion

The report is a deterministic function of the same table and question, yet restating the table entries changes the model's response to the chart. One possible explanation is that the values report reduces the work needed for a table-based answer, although we have not measured the computational cost of answering from either report. The shift also occurs on read questions, so precomputing sums or differences is not necessary for the effect.

The answering protocol can change both the size and direction of the report effect on the smaller checkpoints. The reasoning request directs the model to the chart and permits written reasoning. Because we did not vary these instructions separately, the comparison measures their combined effect.

The positive J contrasts do not establish a specific report–image pathway, because image cuts alone already bring both report arms close to chance.

These interventions concern specific attention edges and controls (Zhang & Nanda, 2024). The control span can contain question or table tokens, and masking renormalizes attention while leaving indirect paths open. For these reasons, the cuts do not identify a natural mediation path or a unique failure layer. The layer readouts in Appendix G instead apply the checkpoint's final normalization and output head to intermediate states. We measure the option preferences induced by this fixed readout, separately from the answers produced by free-running generation. These preferences do not determine whether visual information has been erased. We applied every attention intervention during direct answering, so these experiments leave open how the chart-directed reasoning request changes the report effect.

Middle-band report cuts improve values accuracy on inputs where chart and table disagree but reduce it where they agree (Section 7). The break-even calculation in Appendix F varies the proportion of conflicting inputs while holding the measured agreement and conflict accuracies fixed. It asks when the report cut would improve accuracy under that reweighting. Neither the balanced synthetic design nor the selected natural sample estimates conflict frequency in use. Those accuracies may also change in another setting, so the calculation does not establish a deployment threshold.

The experiments cover English binary comparisons from fixed templates and a small checkpoint set; Qwen and InternVL share language-backbone lineage. In the selected natural-image sample, the added report bundles a winner and consistency fields with computed values. That treatment differs from the equal-length replacement in the main experiment. Smaller and larger checkpoints were also tested on different panels or under different answering protocols in the follow-up studies. Consequently, differences between their results cannot be attributed to model size alone: the evaluation conditions change as well. Appendix H records earlier recovery attempts and incomplete studies.

Under direct answering on the three main models, replacing raw repetition with question-specific values raises agreement accuracy to 96–100% and lowers conflict accuracy to 0–13%. The raw/values contrast changes with a chart-directed reasoning request on the smaller checkpoints and with the specified attention cuts on the main models. In the separate natural-extraction study, adding a calculation report improves average accuracy for every model, yet correct answers fall from 13 to zero on the 26 observed comparison reversals. Even with the chart still available, restating the same table values can change which source the answer follows.

AI USE STATEMENT

GPT-6 Astra was used primarily for experimental implementation. It also supported refinement of hypotheses, formulations and experimental designs, assisted with mathematical proofs and helped review analyses and interpretations. Further uses included literature search, procedural chart generation and screening, translation of author-written text, and editing and polishing the author-written manuscript. The authors originated the core research idea and initial conceptual and mathematical formulations, and led hypothesis development and the analysis and interpretation of results.

Figures were produced programmatically from saved numerical results and chart specifications, without generative image synthesis.

Verification included code tests, replay of saved outputs, and checks against original references. The authors take responsibility for the final manuscript and artifacts.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, June 2025.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KZd1EErRJ1.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Rodela Ghosh, Aviral Gupta, and Guangjing Wang. Which source wins? task-dependent reliance in vision-language models. *arXiv preprint arXiv:2608.17205*, 2026. URL https://arxiv.org/abs/2608.17205.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kaduri_Whats_in_the_Image_A_Deep-Dive_into_the_Vision_of_CVPR_2025_paper.html.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_paper.html.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.714. URL https://aclanthology.org/2023.acl-long.714/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14662–14684. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.906. URL https://aclanthology.org/2023.emnlp-main.906/.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026. URL https://arxiv.org/abs/2604.09364.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.642. URL https://aclanthology.org/2026.acl-long.642/.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. Mixed signals: Decoding VLMs' reasoning and underlying bias in vision-language conflict. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24833–24848. Association for Computational Linguistics, 2025. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1351. URL https://aclanthology.org/2025.findings-emnlp.1351/.

Danae Sánchez Villegas, Samuel Lewis-Lim, Nikolaos Aletras, and Desmond Elliott. Reasoning dynamics and the limits of monitoring modality reliance in vision-language models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=YV2d7n8uMm.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

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 multimodal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1270_ECCV_2024_paper.php.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.648. URL https://aclanthology.org/2024.acl-long.648/.

A Design, prompts and execution details

A.1 Study names and artifact identifiers

Table 3 maps the study names and answering conditions used in the paper to identifiers in the saved records. The identifiers are used only in the appendices and artifact.

Table 3: Correspondence between descriptive names and identifiers in the artifact. The answering-condition identifiers specify the request and generation cap; parser details are given below.

Identifier	Study or answering condition
R01	Report placement and answering protocols on the main models; incomplete panel.
R02	Answering protocols on smaller checkpoints (Section 6).
R03	Activation-swap interventions on two 4B models; incomplete panel with failed controls.
R04	Attention cuts on two 4B models, using a separate supporting panel.
R05	Attention cuts on the main models (Section 7).
P02	Candidate re-scoring against numerical-reading controls.
P03	Geometric verification and its perception gate.
P04	Local predicate repair against the cached six-read median.
P06	Neutral-table calibration of answer-branch scores.
P07	Assignment of candidate values to chart labels.
`direct32`	Direct answering, capped at 32 generated tokens.
`direct512`	Direct answering, capped at 512 generated tokens.
`reason512`	Chart-directed reasoning, capped at 512 generated tokens.

A.2 Panel structure and experimental history

Table 4 lists the completed panels and the comparisons they support. Because tasks, images, option orders and repetitions can share a base chart, statistical comparisons group these observations by base. The main panel and both attention-cut panels draw from the same procedural distribution, although the latter use new seeds to produce new images. The natural-image panel is instead a separate, selected sample of ChartQA-derived images.

Table 4: Completed experimental panels and their dependence structure. Tasks, option orders, image variants and repetitions share base charts. R02 reuses main-experiment bases; R04 and R05 use separate fresh procedural panels. The natural panel uses chart-derived comparisons rather than the original ChartQA questions.

Study	Models	Base images	Evaluated comparison
Main	Qwen 8B, InternVL 8B, Gemma 12B	96 synthetic specifications	Crossed image/table worlds; raw/values; three tasks; JSON/prose; both annotation states.
Natural	Qwen 4B/8B, InternVL 4B/8B, Gemma 12B	96 ChartQA-derived images	Shared generated work record, before/after calculation-report addition; read/sum and both orders.
R02	Qwen 2B/4B, InternVL 2B/4B	First 24 main bases, reused	Raw/values under direct32, direct512 and reason512; both formats and annotation states.
R04	Qwen 4B, InternVL 4B	24 fresh specifications	Fixed-band attention cuts; JSON, assistant report, direct32, no mark annotations.
R05	Qwen 8B, InternVL 8B, Gemma 12B	Another 24 fresh specifications	Same intervention framework; fixed original-model middle-band hypotheses.

We designed R02 after the main results were known. R04 preceded R05, and its results informed the R05 hypotheses, which we fixed before generating the R05 evaluation outputs.

Since the two attention-cut studies use different image seeds, their panels do not provide a controlled comparison of model size. We report earlier incomplete studies and recovery attempts separately in Appendix H; their observations are not pooled with the completed panels.

A.3 PROCEDURAL CHART CONSTRUCTION

Each main base has six fixed category labels and four target values sampled as distinct integers from 12 through 88. For the four targets (v_0, v_1, v_2, v_3), the canonical $+1$ alternative is $v_0 > v_1$ for read, $v_0 + v_1 > v_2 + v_3$ for sum, and $|v_0 - v_1| > |v_2 - v_3|$ for absolute difference; these alternatives are fixed before assigning displayed A/B labels. Minimum comparison gaps are 10 for read, 16 for sum and 12 for absolute difference. To reverse the canonical comparison, the image transformation exchanges the relevant values while preserving category positions and labels. The table world is selected independently of this image world. Because both answer orders are evaluated, we map displayed A/B labels back to the canonical alternatives before computing signed responses.

Table 5 lists the procedural seeds. The saved main dataset contains 1,152 task-indexed image paths but only 576 distinct rendered charts, because different tasks can share an image.

Table 5: Procedural seeds and numbers of evaluation bases. R04 and R05 use separate engineering and evaluation seeds; the main evaluation dataset uses seed 9173102.

Panel	Engineering seed	Dataset/evaluation seed	Evaluation bases
Main	–	9173102	96
R04	9242601	9242602	24
R05	9252601	9252602	24

Each attention-cut evaluation panel has eight bases per geometry. Crossing its 24 bases with three tasks, two option orders, two table worlds and two image worlds gives 576 input contexts for each report and intervention condition. We evaluate every context twice, so each cell contains 1,152 recorded native answers, equally divided between agreement and conflict.

For each model, three task units from three separate bases were used for technical checks of the inference and attention-mask implementation (the engineering phase). These observations are excluded from evaluation estimates.

A.4 MAIN MESSAGE SEQUENCE AND REPORT CONSTRUCTION

The primary system message is reproduced verbatim:

```
Answer the question about the supplied chart. Reference measurements, if
    supplied, may differ from the chart. A computation using those
    measurements is not an independent check of the image. When no image
    is provided, answer as best you can from the supplied information.
    For the final choice return only <answer>A</answer> or <answer>B</
    answer>.
```

The first user turn contains the image, comparison question, reference measurements and the following shell instruction:

```
A computed or repeated reference report follows. It may not match the
    chart. A later turn requests the final answer.
```

An experimenter-supplied assistant turn contains the raw or values report. Both use the source string:

```
Uses reference measurements only; image truth has not been checked.
```

Table 6 lists the fields of the two report versions in JSON.

Table 6: Fields of the assistant report in JSON. Both versions carry the same `source` string shown above.

Field	Raw repetition	Question-specific values
`reference_values`	four label/value objects	–
`operation`	–	the task operation
`question_values`	–	the A and B entries
`source`	the source string	the source string

In prose, the raw report lists the supplied category values, and the values report states the operation and the quantities associated with options A and B. JSON and prose name the report formats; the required final-answer format is the same in both cases.

Token lengths are matched by a deterministic padding procedure.

1. A formatting-padding line and `[DONE]` are appended to each report.
2. The padding units `~`, `|`, `.` and an unspaced `~` are tried in that order, with padding counts from 0 through 128.
3. The first unit that permits a common token length at least eight tokens above the longer unpadded completed report is selected.
4. For that unit, the shortest admissible common length is taken.

If no unit satisfies these conditions, input construction stops. The exact padding strings and counts are saved with each input.

The final user request is:

```
Give the final choice now. Return only <answer>A</answer> or <answer>B</
    answer>.
```

Native greedy generation stops at `</answer>` or EOS, subject to the 32-token cap. The two parser rules are compared below.

R02 uses the same assistant report construction. Its reason512 final request is:

```
Work through the comparison using the chart, then give the final choice.
    You may write reasoning before one final <answer>A</answer> or <
    answer>B</answer> tag.
```

Answer parsing.

Main-panel parser (strict). After stripping outer whitespace, this parser accepts bare A/B or exactly one complete answer tag, with no other text. Preceding reasoning makes the output invalid. Invalid responses count as zero in signed metrics and as incorrect in accuracy; no repair is attempted.

R02 parser (terminal). This parser accepts bare A/B or exactly one complete terminal answer tag. Reasoning before that tag is allowed. All R02 primary arms use the same parser.

The validity counts therefore use different acceptance rules in the main panel and R02. We did not rescore the main panel with the R02 parser.

A.5 Checkpoints and recorded runtime

The checkpoints are pinned to the following revisions:

```
Qwen/Qwen3-VL-8B-Instruct
  0c351dd01ed87e9c1b53cbc748cba10e6187ff3b
OpenGVLab/InternVL3_5-8B-HF
  741a7d03020411e666c6109218ab71e08151ef86
```

```
google/gemma-3-12b-it
  96b6f1eccf38110c56df3a15bffe176da04bfd80
Qwen/Qwen3-VL-4B-Instruct
  ebb281ec70b05090aa6165b016eac8ec08e71b17
OpenGVLab/InternVL3_5-4B-HF
  6bd4487402110ef9889ba50eb7aefeb302526fed
Qwen/Qwen3-VL-2B-Instruct
  89644892e4d85e24eaac8bacfd4f463576704203
OpenGVLab/InternVL3_5-2B-HF
  3f301ffcf3dcbb47893afae6650ea3e78d96fb6d
```

The 2B checkpoint revisions were resolved and locked before R02 prediction, and they are recorded with the other revisions in the paper package. InternVL uses the official final HF checkpoints following Cascade RL. Comparisons across model sizes are observational because training and architecture also differ. Moreover, Qwen and InternVL share language-backbone lineage, which limits their independence as model families.

The paired-image runs use NVIDIA GH200 nodes with 96 GB HBM3 GPU memory and 120 GB LPDDR5X CPU memory.[1] The recorded software is PyTorch `2.11.0+cu130`, CUDA 13.0 and Transformers `5.12.1`; R04/R05 use bfloat16. Numerical results may differ under other software or hardware.

For R05 evaluation, recorded job durations sum to 9.72 hours for Qwen (18 model loads), 13.58 hours for InternVL (33 loads) and 17.91 hours for Gemma (36 loads). These totals also include other job overhead. All evaluation jobs completed successfully. These are sums of per-job durations, not billed GPU-hours. When jobs run in parallel, their summed durations also differ from elapsed time to completion.

B Estimands and inference

B.1 Accuracy, coverage and signed response

For binary truth w and prediction $u \in \{-1, 0, +1\}$,

$$wu = 2\mathbf{1}\{u = w\} - \mathbf{1}\{u \neq 0\}. \tag{6}$$

Averaging this equality over the design gives Equation (3). If we instead average uniformly over the two image worlds while holding the other factors fixed, we obtain $\mathbb{E}_w[wu] = [u(+1, r) - u(-1, r)]/2$, which establishes Equation (2). Consequently, $\Delta\Gamma = 2\Delta A - \Delta V$: a change in the signed contrast can reflect changes in either accuracy or coverage. Both quantities are therefore needed to interpret the contrast, as the InternVL 2B/prose results illustrate (Section 6).

For reference, define table response as $\Lambda_a = \mathbb{E}[ru_a]$. A constant prediction has zero table response on the balanced design, even though it does not respond to the chart. Low table response alone is therefore insufficient evidence of correct chart use.

Proposition 1 (Finite-panel completion bound)**.** *Let $q_a = 1 - V_a$ be the invalid fraction in arm a. Replace each invalid prediction by an arbitrary binary response while preserving valid answers. The resulting representation contrast belongs to*

$$[\Delta\Gamma - (q_{\text{values}} + q_{\text{raw}}), \quad \Delta\Gamma + (q_{\text{values}} + q_{\text{raw}})]. \tag{7}$$

Proof. Each replaced zero changes wu by at most one in absolute value. Summing over the invalid fractions bounds each arm's change by q_a, and applying the triangle inequality to the arm difference gives the result. □

This bound concerns completions of the observed panel, rather than sampling uncertainty. For each of the seven negative R02 direct32 contrasts, even its upper endpoint is negative. Thus none of these signs can be reversed by assigning A or B to the invalid answers while preserving the valid answers.

[1] NVIDIA GH200 Grace Hopper Superchip datasheet, p. 4.

B.2 LINEAR BOUNDS FOR FLAGGED NATIVE EXECUTIONS

Write a native contrast as $\theta = N^{-1}\sum_j c_j v_j$, with $v_j \in [-1, 1]$. Let F contain terms treated as unknown and let $S = \sum_{j \notin F} c_j v_j$. Conditional on all unflagged outputs,

$$\theta \in \left[\frac{S - \sum_{j \in F} |c_j|}{N},\ \frac{S + \sum_{j \in F} |c_j|}{N}\right]. \tag{8}$$

Because v_j is restricted to $[-1, 1]$, the extreme contribution for coefficient c_j is $\pm|c_j|$, which proves the bound.

In R05, a context is flagged if its generated answer differs in any of three comparisons:

- between the two repeated executions;
- between the `original` backend and the explicit-mask `none` execution (the bridge comparison);
- between the `none` execution and the mask-cloning `identity` control.

Once a context is flagged, we treat its answers under all native conditions and both repetitions as unknown. All unflagged outputs retain their recorded values.

The bounds and tests answer different questions. The post hoc bounds quantify how far a contrast can move when responses at flagged contexts vary, with all unflagged outputs held fixed. The base-level tests instead use variation across bases to assess evidence against their null hypotheses under the stated assumptions. A bound can therefore exclude zero even when the corresponding test does not reject, as occurs for Gemma's middle-band T. The bounds are not confidence intervals. Because they condition on the recorded unflagged outputs, they also give no guarantee for a new GPU run.

B.3 INDEPENDENT UNITS AND FIXED FAMILIES

For every representation, protocol and intervention contrast, we first average the paired differences within each base and then give each base equal weight. Since generations from the same base share chart structure, we do not resample them as independent observations.

Table 7 summarizes the tests and bootstrap settings. The Holm families are listed separately below, because their definitions depend on both the endpoint and the layer band.

Table 7: Tests and bootstrap settings for the main, protocol and attention-cut studies. Bases are the independent units; bootstrap intervals are nominal 95% intervals for individual contrasts.

Study	Test	Bootstrap intervals
Main	Two-sided paired base-level t test, 95 degrees of freedom.	30,000 base resamples; seed 9173110.
R02	Paired base-level t test, 23 degrees of freedom.	20,000 base resamples; seed 2026092211.
R04 and R05	100,000 two-sided base sign flips.	20,000 base-bootstrap draws.

Multiplicity correction.

Main. The six primary model–format tests form one Holm family.

R02. The eight direct32 tests form one family. The eight cap-only and eight reasoning-instruction contrasts form a separate 16-test family.

R04. The 14-test family contains two standard-backend D tests from generated answers and 12 score-based T/J tests across three bands. The four middle-band T/J tests from generated answers form a separate family.

R05. The nine-test primary family contains standard-backend D and middle-band T/J from generated answers, with three tests per model. The secondary family contains 30 tests: 12 early/late generated-answer tests and 18 score tests across all bands.

Holm correction is applied separately to each stated family. The bootstrap intervals do not provide simultaneous coverage across a family.

Assumptions and scope. For R04 and R05, the sign-flip test assumes that base contrasts are independent and symmetric under the null. We adopt this modeling assumption for the paired design, which evaluates every condition on each base rather than randomly assigning conditions to bases. The intervention tests compare conditions within each band; they do not directly compare one band with another. The stratified results, raw-lens curves and parser diagnostic are descriptive analyses.

C Additional results for the main 96-base experiment

Table 8 gives the integer counts behind the percentages in the main text (2,304 answers per agreement/conflict state, 4,608 per report cell, invalid outputs included in the denominators). Figure 7 breaks $\Delta\Gamma$ down by task, annotation and geometry.

Table 8: Exact correct and valid answer counts in the primary 96-base cells. Every arrow denotes raw $\rightarrow$ values. Conflict and agreement each contain 2,304 answers per arm; validity uses all 4,608. Invalid outputs are included in the accuracy denominators.

Model	Format	Conflict /2,304	Agreement /2,304	Valid /4,608
Qwen 8B	JSON	1520 $\rightarrow$ 5	1988 $\rightarrow$ 2303	4608 $\rightarrow$ 4608
Qwen 8B	prose	1423 $\rightarrow$ 50	2003 $\rightarrow$ 2290	4608 $\rightarrow$ 4608
InternVL 8B	JSON	1565 $\rightarrow$ 287	2079 $\rightarrow$ 2275	4608 $\rightarrow$ 4608
InternVL 8B	prose	1482 $\rightarrow$ 84	2080 $\rightarrow$ 2303	4585 $\rightarrow$ 4608
Gemma 12B	JSON	1404 $\rightarrow$ 171	1950 $\rightarrow$ 2210	4608 $\rightarrow$ 4608
Gemma 12B	prose	1354 $\rightarrow$ 9	1925 $\rightarrow$ 2302	4608 $\rightarrow$ 4608

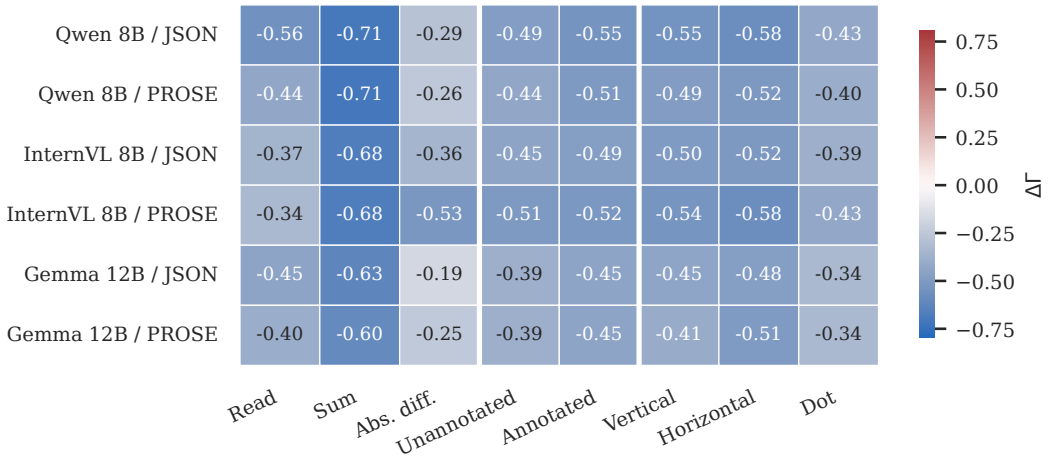


Figure 7: Descriptive representation effects $\Delta\Gamma$ (values minus raw) in the main 96-base experiment, stratified by task, annotation and chart geometry. Negative effects also occur for read, which introduces no new arithmetic.

Four-world responses. Holding the base, task, option order, annotation, report format and report arm fixed, we collect the four outputs at $(w, r) \in \{-1, +1\}^2$. Changing the table world also changes its derived report. Both the table and report remain fixed when only the image world changes.

Each quartet is assigned to exactly one category, tested in the following order.

1. *Invalid in quartet:* at least one of the four outputs is invalid.
2. *Table-consistent:* all four outputs are valid and $u(w, r) = r$ in every cell.

3. *Image-consistent:* all four outputs are valid and $u(w, r) = w$ in every cell.
4. *Constant choice:* all four outputs are valid and have the same canonical sign.
5. *Other valid:* every remaining fully valid quartet.

Because the categories are tested in this order, they are mutually exclusive. Figure 8 includes all 1,152 quartets for each report arm, format and model.

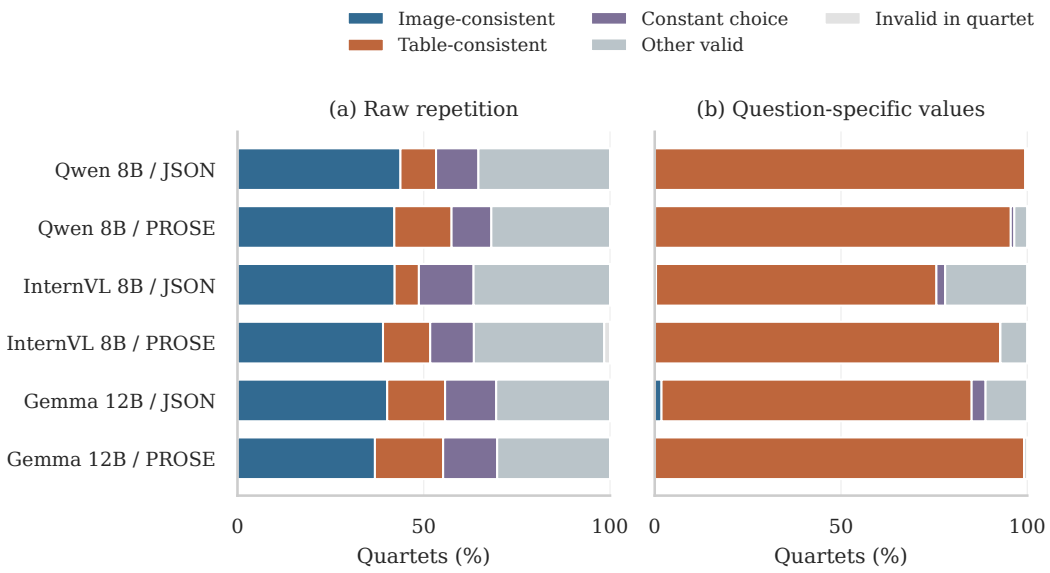


Figure 8: Complete four-world response patterns, over 1,152 quartets per model, format and report arm. A quartet crosses two image worlds with two table worlds. Table-consistent, image-consistent, constant-choice and other-valid patterns require four valid answers; any invalid answer assigns the entire quartet to the invalid category.

Table 9: Counts of three fully valid four-world patterns out of 1,152 quartets per cell. Arrows denote raw → values. Other-valid and invalid quartets, omitted from these three columns, are shown in Figure 8; the displayed columns need not sum to 1,152.

Model	Format	Table-consistent	Image-consistent	Constant
Qwen 8B	JSON	110 → 1146	504 → 0	131 → 0
Qwen 8B	prose	177 → 1100	485 → 1	123 → 11
InternVL 8B	JSON	75 → 867	486 → 4	169 → 27
InternVL 8B	prose	146 → 1068	450 → 0	135 → 1
Gemma 12B	JSON	179 → 959	463 → 21	158 → 43
Gemma 12B	prose	211 → 1142	425 → 0	167 → 1

Direct-answer controls and format limits. Table 10 gives the two direct-answer control conditions.

Table 10: Direct-answer control conditions in the main experiment: correct and valid answers per model, out of 2,304 (chart only) and 4,608 (chart and table without a report). Invalid outputs count as wrong.

	Chart only		Chart and table, no report	
Model	Correct	Valid	Correct	Valid
Qwen 8B	1,964	2,286	3,712	4,608
InternVL 8B	766	770	1,795	2,178
Gemma 12B	744	768	1,255	1,387

Because invalid outputs count as wrong, the low valid counts of InternVL and Gemma limit their strict accuracy under this answer format and token budget. R02 examines a terminal parser and longer caps, with coverage reported alongside accuracy.

Table 11 accounts for all 91,584 saved answers. Of these, 55,296 form the six primary model–format contrasts. The additional conditions include reports without an image and questions without either an image or a table. Their results are included in the artifact but are not analyzed here.

Table 11: Composition of the 91,584 saved main-experiment answers. Each of the three models has the same 30,528 conditions across all 96 bases. The primary report comparison uses 18,432 answers per model; the remaining rows are auxiliary controls.

Condition	Per model	Three models
Chart, table and report (two formats, two reports)	18,432	55,296
Chart and table, without report	4,608	13,824
Chart only	2,304	6,912
Table and report, without image (two formats, two reports)	4,608	13,824
Question only	576	1,728
Total	30,528	91,584

D Natural records, reference provenance and error cases

D.1 Implemented before/after comparison

For each model, chart, task and option order, the extraction input contains the original image and question but no source-table values. The system says that reference values or claimed results may be incorrect, and that internal consistency does not establish image truth. The extraction request is:

```
Read the four requested chart values, in the stated label order, then
    derive the answer to this question. Return one JSON object with "
    values": [x0,x1,x2,x3] and "claimed_winner": "A" or "B". Use null for
     unreadable values or an unresolved winner. Do not output an answer
    tag. End with [DONE].
```

Final answering uses the following work-record warning in the reconstructed image/question shell:

```
A work record may follow. It is not guaranteed to be correct; a later
    turn requests the final answer.
```

The before/after comparison is built in four steps.

1. *Extraction.* Generation is greedy, capped at 400 tokens and stopped at `[DONE]`. The generated text and token IDs are saved; the existing body-ID routine removes the terminal EOS as specified by the model's native interface.
2. *Shell.* For final answering, the image/question shell is reconstructed with the warning above; the extraction instruction is not part of it.
3. *Shared work record.* The saved generated body IDs are appended to this shell, so both branches reuse exactly the same generated work record. The after branch additionally receives the calculation report before the next user turn.
4. *Final request.* Both branches use the same native next-user boundary and final request, with a 32-token cap and no forced answer prefix.

Schematically,

$$\begin{aligned} e &\leftarrow [\text{system}; \text{user}(I, q, \text{extract})], \\ x_{\text{before}} &= [\text{system}; \text{user}(I, q, \text{shell}); \text{assistant}(e); \text{user}(\text{final})], \\ x_{\text{after}} &= [\text{system}; \text{user}(I, q, \text{shell}); \text{assistant}(e \| \text{calculator}(e)); \text{user}(\text{final})]. \end{aligned} \tag{9}$$

The saved native token IDs define what was actually run; the extraction instruction is replaced by the shell before final answering.

A complete parsed record requires four finite, plain, nonnegative numerical values no greater than 10^{12} and a claimed winner A/B. Complete Markdown JSON fences are accepted; incomplete JSON is not repaired.

The calculation report is computed from the complete record and carries the following fields.

option quantities the two quantities the question compares, computed from the extracted values;

`computed_winner` the winner computed from the extracted values, or JSON `null` on a numerical tie;

`claimed_winner` the winner stated in the model's record;

`internal_agreement` a Boolean that is true exactly when `computed_winner` equals `claimed_winner`; a tie makes it false, since a complete record always contains a claimed A/B winner;

`image_truth` the fixed string `NOT_CHECKED`.

`internal_agreement` checks consistency within the generated record, not the image. If a record is incomplete, the after branch receives the incomplete-record notification and generates a new final answer; it does not fall back to the before-branch outcome.

Adding the report changes both the context length and its content, including the winner and consistency fields. The measured effect therefore applies to this complete addition, rather than to arithmetic alone; the comparison is not length-matched.

D.2 All cases and the reversal subgroup

Table 12 includes every final-answer condition and separately reports the subgroup with reversed comparisons. A record belongs to that subgroup when its parsed generated numbers imply the opposite comparison from the chart-derived truth, regardless of either final answer. Thus the subgroup is defined by the extracted numbers rather than by whether adding the report changes the answer. Because the charts are a selected sample, these counts do not estimate how often reversals occur in use.

Table 12: Natural-extraction outcomes before → after calculation-report addition. Each model contributes 384 conditions from 96 charts. Reversals are defined by the extracted numerical comparison, independently of answer changes. The 26 reversal conditions share 11 charts and are not independent chart samples.

Model	Correct /384	Gain (pp)	Reversals	Correct in reversals
Qwen 4B	346 → 379	+8.59	5	2 → 0
Qwen 8B	321 → 378	+14.84	5	3 → 0
InternVL 4B	346 → 383	+9.64	0	–
InternVL 8B	317 → 382	+16.93	0	–
Gemma 12B	317 → 364	+12.24	16	8 → 0

Across all records, the extracted comparison is correct in 1,833 cases, reversed in 26 and tied in one; the remaining 60 records are incomplete. The incomplete group contains one Qwen 8B record, 55 InternVL 4B records and four Gemma records. To relate these groups to the average gain, let s denote the extracted-comparison state. The aggregate effect then decomposes as

$$\Delta A = \sum_{s} \widehat{p}(s)\, \mathbb{E}[\mathbf{1}\{\widehat{y}_{\text{after}} = y\} - \mathbf{1}\{\widehat{y}_{\text{before}} = y\} \mid s]. \tag{10}$$

Both the proportions and conditional means are empirical quantities on the observed panel. The identity describes how the subgroups contribute to the observed gain. However, the synthetic and natural studies differ in treatment and protocol as well as extraction status. The decomposition does not separate those differences, so it cannot attribute the difference between the studies to extraction status alone.

Source-table references and generated values. The source selection records each upstream chart ID, series, four labels and values. Data-row indices are one-based; the column index is zero-based. For order 0, read compares v_0 with v_1, and sum compares $v_0 + v_1$ with $v_2 + v_3$; order 1 swaps the alternatives.

Independent decimal calculations from the 96 source CSVs reproduce all 1,920 truth labels. These labels use the source-table values as their reference. Separately, the authors made a qualitative visual spot-check of label/order correspondence and displayed values in the reversal-case inspection material.

For example, `two_col_22215` has source-table values $(1194, 474, 2222, 470)$. Qwen 8B generated $(194, 474, 2222, 470)$ for read in both orders. Order 0 was wrong before and after the report (B→B), while order 1 changed from correct to wrong (B→A). The generated value omits the leading digit of 1194; the reference values and option mapping are unchanged. The archive includes locators and generated texts for all 26 reversal conditions.

A separate controlled real-image comparison. A supporting sum-only study holds an explicit reference claim fixed within each pair and replaces four operands with their computed sums when the supplied table is wrong. Correct answers fall from 245 to 67 for Qwen 8B and from 242 to 19 for Gemma, each out of 384 conditions. That denominator comes from 96 images crossed with two option orders and two claim states. Because this comparison supplies both the table and an explicit reference claim, its protocol differs from those of the main experiment and the natural-extraction study.

E COMPLETE SMALLER-CHECKPOINT PROTOCOL RESULTS

R02 contains 57,600 evaluation calls across four checkpoints. Each model–format–protocol–report cell has 1,152 answers, including 576 conflict and 576 agreement conditions. Table 13 gives all representation effects and the reasoning moderation contrast. The reason512 intervals describe within-cell effects. A fixed family contains eight cap-only and eight reasoning-instruction contrasts, with Holm correction applied jointly over these 16 tests.

Table 13: R02 representation effects for every protocol and model–format cell on 24 reused bases. Ω is reason512 minus direct512 in $\Delta\Gamma$. Reason512 intervals are nominal 95% intervals; adjusted p values for Ω use Holm correction over the fixed family of eight cap and eight instruction contrasts.

Model	Format	direct32	direct512	reason512	reason512 CI	Ω	Adj. p
Qwen 2B	JSON	-0.300	-0.300	-0.001	$[-0.029, +0.025]$	+0.299	1.37×10^{-9}
Qwen 2B	prose	-0.398	-0.397	+0.092	$[+0.056, +0.130]$	+0.489	9.55×10^{-13}
Qwen 4B	JSON	-0.517	-0.519	+0.436	$[+0.405, +0.467]$	+0.955	8.20×10^{-18}
Qwen 4B	prose	-0.507	-0.509	+0.467	$[+0.418, +0.517]$	+0.976	3.91×10^{-16}
InternVL 2B	JSON	-0.302	-0.299	+0.033	$[-0.010, +0.075]$	+0.332	8.71×10^{-8}
InternVL 2B	prose	+0.059	+0.047	+0.056	$[+0.020, +0.095]$	+0.009	1.000
InternVL 4B	JSON	-0.712	-0.714	+0.134	$[+0.077, +0.188]$	+0.847	3.92×10^{-12}
InternVL 4B	prose	-0.684	-0.684	+0.342	$[+0.300, +0.383]$	+1.026	3.17×10^{-15}

Table 14 gives the primary direct32 intervals and adjusted p values, along with the cap-only contrasts (direct512 minus direct32). All cap-only p values equal 1 after correction over the 16 cap and instruction contrasts.

Figure 9 shows coverage under the primary terminal parser, with invalid responses retained in the denominator. Table 15 gives the counts behind the one positive direct32 contrast.

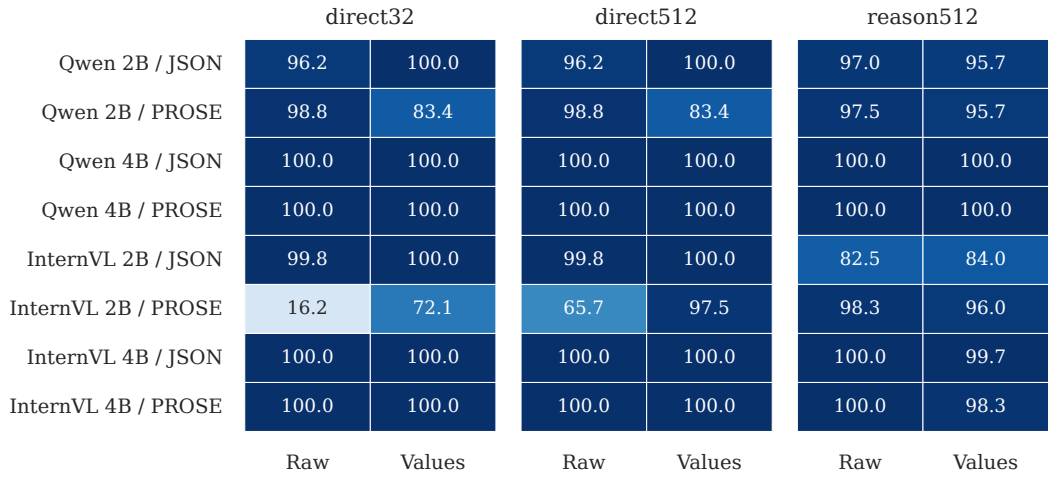


Figure 9: Valid terminal-answer coverage for every R02 model, format, protocol and report arm, with 1,152 answers per cell. Preceding reasoning is allowed by the common primary parser. Low raw coverage in InternVL 2B/prose qualifies its positive direct32 contrast. The post hoc punctuation diagnostic is not substituted here.

Table 14: R02 direct32 primary tests and cap-only contrasts on 24 bases. Holm correction is applied to the eight direct32 tests. Cap contrasts are direct512 minus direct32 and belong to the separate 16-test cap/instruction family, with all adjusted $p = 1$. Displayed intervals are nominal 95% base-bootstrap intervals.

Model	Format	direct32 CI	Adj. p	Cap contrast	Cap-contrast CI
Qwen 2B	JSON	$[-0.349, -0.250]$	1.35×10^{-10}	+0.0000	$[+0.000, +0.000]$
Qwen 2B	prose	$[-0.445, -0.353]$	2.33×10^{-13}	+0.0017	$[+0.000, +0.005]$
Qwen 4B	JSON	$[-0.582, -0.451]$	7.23×10^{-13}	-0.0017	$[-0.005, +0.000]$
Qwen 4B	prose	$[-0.566, -0.448]$	2.33×10^{-13}	-0.0017	$[-0.005, +0.000]$
InternVL 2B	JSON	$[-0.359, -0.245]$	1.36×10^{-9}	+0.0035	$[+0.000, +0.009]$
InternVL 2B	prose	$[+0.017, +0.101]$	0.0125	-0.0122	$[-0.041, +0.019]$
InternVL 4B	JSON	$[-0.785, -0.634]$	3.85×10^{-14}	-0.0017	$[-0.005, +0.000]$
InternVL 4B	prose	$[-0.752, -0.604]$	4.87×10^{-14}	+0.0000	$[-0.005, +0.005]$

Table 15: InternVL 2B/prose under the two direct-answer caps: correct and valid answers out of 1,152 per arm. Under direct32, 878 raw and 314 values completions reach the generation cap.

Cap	Arm	Correct	Valid
direct32	raw	143	187
direct32	values	499	831
direct512	raw	448	757
direct512	values	658	1,123

Under direct32 the values arm has 356 more correct answers and 644 more valid answers than the raw arm, so $\Delta\Gamma = 2(356/1152) - 644/1152 = 0.05903$. Raising the cap to 512 tokens changes both arms substantially while the contrast changes little. This is why absolute accuracy and coverage need to accompany the representation contrast.

The post hoc punctuation diagnostic is applied uniformly to every output.

1. An answer that already has a valid primary parse is kept unchanged.
2. For an invalid answer, outer whitespace is stripped; if the completion ends in an ASCII period, that one character is removed.

3. The same parser is applied again. Unfinished tags, arbitrary trailing text and multiple answers are still rejected.

The diagnostic accepts 1,298 additional outputs, of which 1,091 are correct. For InternVL 2B/prose under direct32, the contrast becomes $\Delta\Gamma = 0.00347$ with nominal interval $[-0.03819, 0.04427]$. We report this as a formatting diagnostic only and retain the original parser for the primary result.

An earlier secondary strict-validity counter treated parser failure sentinels as valid because it checked only whether the output was non-null. The corrected counter instead requires the parsed output to be A or B. This correction affects only secondary strict coverage: strict accuracy, primary parsing, primary coverage and all fixed tests are unchanged. Figure 9 therefore reports the unchanged primary coverage.

Table 16: R02 correct answers under all three protocols. Arrows denote raw → values, each out of 1,152 answers per model–format–protocol cell. Invalid answers are included. Absolute accuracy and the representation contrast can change differently when coverage changes.

Model	Format	direct32 /1,152	direct512 /1,152	reason512 /1,152
Qwen 2B	JSON	747 → 596	747 → 596	1057 → 1049
Qwen 2B	prose	793 → 475	793 → 476	1021 → 1064
Qwen 4B	JSON	874 → 576	875 → 576	813 → 1064
Qwen 4B	prose	869 → 577	870 → 577	826 → 1095
InternVL 2B	JSON	834 → 661	834 → 663	906 → 934
InternVL 2B	prose	143 → 499	448 → 658	1052 → 1071
InternVL 4B	JSON	995 → 585	996 → 585	962 → 1037
InternVL 4B	prose	972 → 578	972 → 578	934 → 1121

Table 16 reports the correct counts for both report arms under all three protocols. Table 17 gives the chart-only control, which uses reason512 and has 576 answers per model.

Table 17: Chart-only control in R02 (reason512): correct answers out of 576 per model under the primary parser and after the period diagnostic.

Model	Primary parser	After period diagnostic
Qwen 2B	567	573
Qwen 4B	576	576
InternVL 2B	421	576
InternVL 4B	574	576

InternVL 2B's apparent shortfall is therefore sensitive to answer formatting. Because the chart-only control uses reason512, comparing it with direct32 changes both the context and the answering protocol.

F Attention cuts, complete contrasts and numerical checks

F.1 Source positions, destinations and layer bands

Figure 5 uses PlotNeuralNet[2] blocks to illustrate the decoder thirds and the imposed attention cuts. Let $\mathcal{E}_\ell$ be the model's originally allowed query/key edges at layer ℓ, K_I its image positions, and $[s, e)$ the half-open interval occupied by the report. The report and control key sets are $K_R = [s, e)$ and $K_C = [s - (e - s), s)$; an input is rejected if its control interval overlaps or precedes the image region. For layers in band b, each cut deletes the following edges $(t, k) \in \mathcal{E}_\ell$.

Report cut R_b $t \geq e$ and $k \in K_R$.

Text-control cut C_b the same query positions $t \geq e$, and $k \in K_C$.

[2] https://github.com/HarisIqbal88/PlotNeuralNet

Image cut I_b $k \in K_I$, with t a later non-image query.

Joint cut IR_b the union of the image and report edge sets.

We apply each cut in every head during both prefill and cached decoding. Attention is then renormalized over the remaining allowed edges. Because neither the token sequence nor the residual connections change, information may still travel along indirect paths through earlier positions.

The attention-cut records distinguish the following execution conditions.

`original` the standard inference backend, recorded as a separate bridge condition;

`none` an explicit causal-mask scaled dot-product attention (SDPA) execution with no cut, serving as the uncut control;

`identity` the same execution with a cloned mask, as a control;

cuts R_b, C_b, I_b and IR_b in the early, middle and late bands, run with the same explicit-mask backend;

invariance checks report and image cuts applied across all layers, to check the corresponding masking invariances.

Gemma's native sliding window is 1,024 tokens, so it does not bind for the saved prefixes of 549–564 tokens and the short continuations in this panel.

As a deterministic control-span example, the first saved Qwen 8B R05 evaluation input has a 734-token native prefix, image region $[78, 510)$, report $[625, 703)$ and text control $[547, 625)$. The report and control each contain 78 tokens, and their cut destinations begin at query 703. The control token IDs match across both report versions and image worlds in that group. The match is on length and destination only; the control span can contain question, table, shell and chat-delimiter tokens.

F.2 Native answers and branch scores

For a fixed checkpoint, let $i = (n, q, o, r, w)$ index a base, task, option order, table world and image world, with report arm a kept explicit. The multimodal input x_{ia}^{native} is the complete generation prefix described in Appendix A, including the final user request and native assistant generation boundary. Native answers are generated from this prefix without an answer prefill and mapped to u_a using $\sigma_o = (-1)^o$ from Section 3. All 34,560 native evaluation outputs per model in R05 pass the fixed parser: each of 15 conditions has 1,152 outputs per report arm, including the two repetitions. No outputs are excluded to obtain this coverage. Thus $V_a(k) = 1$ and $\Gamma_a(k) = 2A_a(k) - 1$ for every recorded native condition in this panel.

For diagnostic scoring, the tokenizer separately encodes the complete strings `<answer>A</answer>` and `<answer>B</answer>` with `add_special_tokens=False`. Let d be their longest common *token* prefix and s_A, s_B their remaining token sequences. The diagnostic input is

$$x_{ia}^{\text{diag}} = x_{ia}^{\text{native}} \| d, \tag{11}$$

where $\|$ appends token IDs without retokenizing the native prefix or changing its image tensors. "Common" refers to the two candidate answers for a checkpoint; it does not mean that prefixes are identical across models or report arms. Table 18 gives the exact saved sequences for both attention-cut studies. The readout tokens t_A and t_B are the first tokens of s_A and s_B.

For condition k and repetition $\nu \in \{1, 2\}$, the score is the sum of teacher-forced log probabilities over the entire continuation:

$$L_{ia,c}^{(\nu)}(k) = \sum_{j=1}^{|s_c|} \log P_\theta^{(\nu)}(s_{c,j} \mid x_{ia}^{\text{diag}}, s_{c,<j}; k), \quad c \in \{A, B\}, \tag{12}$$

$$Z_{ia}^{(\nu)}(k) = \sigma_o \tanh\left(\frac{L_{ia,A}^{(\nu)}(k) - L_{ia,B}^{(\nu)}(k)}{2}\right), \qquad \Gamma_a^{\text{score}}(k) = \mathbb{E}_{i,\nu}[w_i Z_{ia}^{(\nu)}(k)]. \tag{13}$$

The superscript ν distinguishes repeated numerical executions with the same model parameters. We score each candidate in three steps.

Table 18: Saved diagnostic prefix and complete scored continuations under the pinned tokenizers. Strings show decoded sequences; lists give exact token IDs. The Qwen and InternVL column applies to the 4B and 8B checkpoints. Each continuation includes the closing answer tag and has four tokens.

Sequence	Qwen / InternVL	Gemma 12B
d: text	`<answer`	`<answer>`
d: IDs	`[27,9217]`	`[236820,14433,236813]`
s_A: text	`>A</answer>`	`A</answer>`
s_A: IDs	`[23465,522,9217,29]`	`[236776,954,14433,236813]`
s_B: text	`>B</answer>`	`B</answer>`
s_B: IDs	`[36721,522,9217,29]`	`[236799,954,14433,236813]`

1. Form an input from the diagnostic prefix followed by all tokens of s_c except the last one. This supplies the preceding tokens needed to evaluate every term in its continuation likelihood.
2. Run a separate forward pass with the selected intervention active throughout, then compute the full-vocabulary log-softmax in float32.
3. Sum all four continuation log probabilities, without length normalization or an EOS term. Using only the first token would give a different score.

The aggregation also has three steps.

1. Average the two repetitions within each input.
2. Average equally over the 24 inputs per base, which cross three tasks, two option orders, two table worlds and two image worlds.
3. Give each of the 24 bases equal weight.

The expectation in Equation (13) denotes this empirical average. Although the native and score endpoints use the same definitions of D, T_b and J_b, they measure different outcomes: the bounded score is not a calibrated probability that the generated answer will be correct.

F.3 All fixed R05 results

The six middle-band tests in Table 2 and the three uncut tests in Table 19 form the nine-test primary family. Table 20 lists all 30 secondary tests. Holm correction is applied separately to these two families. Figure 10 displays all three bands for generated-answer and score endpoints, with nominal intervals. Gemma's primary T is nonsignificant, as are all early/late generated-answer T/J contrasts in the secondary family.

Table 19: The three uncut representation contrasts completing the R05 primary family. These standard-backend generated-answer D tests share one nine-test Holm correction with the six middle-band T/J tests in Table 2. Estimates use 24 independent bases, averaging repetitions within base.

Model	Endpoint	Contrast	Band/backend	Estimate	Nominal 95% CI	Adj. p
Qwen 8B	native	D	original	-0.575	$[-0.653, -0.500]$	9.00×10^{-5}
Gemma 12B	native	D	original	-0.424	$[-0.500, -0.344]$	9.00×10^{-5}
InternVL 8B	native	D	original	-0.509	$[-0.571, -0.441]$	9.00×10^{-5}

Table 21 gives native accuracy for both report arms under every condition in the band comparison, with 1,152 answers per arm. We report the original and explicit-mask backends separately because their outputs can differ. For example, InternVL has 618 correct values-arm answers under `original` and 622 under `none`. Report-cut gains are computed relative to `none`, so that the cut and its baseline use the same backend.

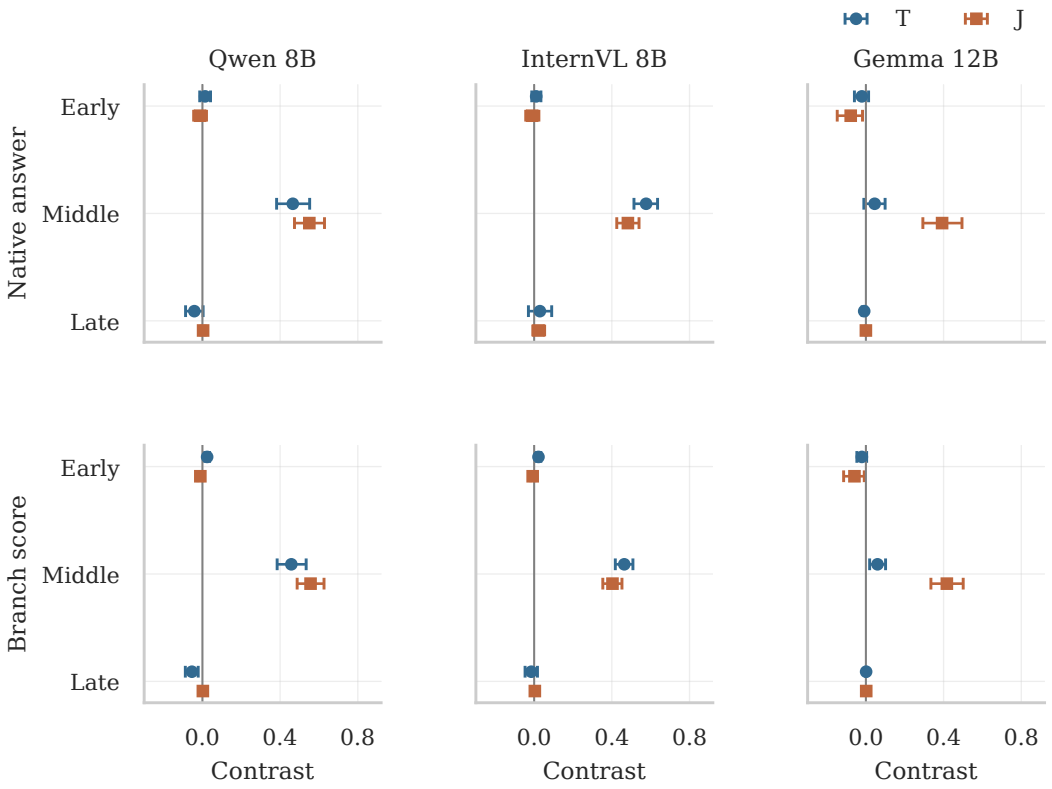


Figure 10: R05 report-versus-text contrasts T_b and report–image cut interactions J_b across all decoder thirds, for native generation and diagnostic branch scores. Bars show nominal 95% intervals over 24 bases. Middle generated-answer contrasts belong to the nine-test primary family; early/late generated-answer and all score contrasts belong to the separate 30-test family. Holm correction is applied within each family.

Table 20: All 30 fixed R05 secondary hypotheses: early/late native T/J and all-band diagnostic-score T/J for three models. Adjusted p values use Holm correction over these 30 tests. Tests compare interventions within each band.

Model	Endpoint	Contrast	Band/backend	Estimate	Nominal 95% CI	Adj. p
Qwen 8B	native	T	early	+0.014	$[-0.014, +0.042]$	1.000
Qwen 8B	native	T	late	-0.042	$[-0.087, +0.003]$	1.000
Qwen 8B	native	J	early	-0.012	$[-0.045, +0.021]$	1.000
Qwen 8B	native	J	late	+0.003	$[-0.007, +0.014]$	1.000
Qwen 8B	score	T	early	+0.024	$[+0.012, +0.037]$	0.027
Qwen 8B	score	T	middle	+0.458	$[+0.385, +0.534]$	3.00×10^{-4}
Qwen 8B	score	T	late	-0.054	$[-0.088, -0.022]$	0.108
Qwen 8B	score	J	early	-0.011	$[-0.024, +0.003]$	1.000
Qwen 8B	score	J	middle	+0.557	$[+0.488, +0.626]$	3.00×10^{-4}
Qwen 8B	score	J	late	+0.002	$[-0.000, +0.005]$	1.000
Gemma 12B	native	T	early	-0.021	$[-0.059, +0.014]$	1.000
Gemma 12B	native	T	late	-0.009	$[-0.019, +0.000]$	1.000
Gemma 12B	native	J	early	-0.078	$[-0.148, -0.017]$	0.639
Gemma 12B	native	J	late	+0.000	$[+0.000, +0.000]$	1.000
Gemma 12B	score	T	early	-0.021	$[-0.046, +0.004]$	1.000
Gemma 12B	score	T	middle	+0.060	$[+0.019, +0.101]$	0.215
Gemma 12B	score	T	late	+0.001	$[-0.002, +0.005]$	1.000
Gemma 12B	score	J	early	-0.059	$[-0.115, -0.008]$	0.773
Gemma 12B	score	J	middle	+0.417	$[+0.335, +0.501]$	3.00×10^{-4}
Gemma 12B	score	J	late	+0.002	$[-0.000, +0.003]$	1.000
InternVL 8B	native	T	early	+0.010	$[-0.012, +0.035]$	1.000
InternVL 8B	native	T	late	+0.030	$[-0.030, +0.090]$	1.000
InternVL 8B	native	J	early	-0.010	$[-0.043, +0.024]$	1.000
InternVL 8B	native	J	late	+0.023	$[-0.007, +0.054]$	1.000
InternVL 8B	score	T	early	+0.022	$[+0.010, +0.034]$	0.031
InternVL 8B	score	T	middle	+0.464	$[+0.418, +0.509]$	3.00×10^{-4}
InternVL 8B	score	T	late	-0.015	$[-0.047, +0.017]$	1.000
InternVL 8B	score	J	early	-0.008	$[-0.019, +0.004]$	1.000
InternVL 8B	score	J	middle	+0.403	$[+0.353, +0.452]$	3.00×10^{-4}
InternVL 8B	score	J	late	+0.004	$[+0.001, +0.007]$	0.641

Table 21: R05 native accuracy (%) for both report arms under every displayed condition, with 1,152 answers per arm. Original denotes the untouched backend; none and all cuts use the explicit-mask backend. Early, middle and late refer to fixed decoder thirds. All recorded answers are valid; coverage is one in every condition.

Condition	Qwen raw	values	InternVL raw	values	Gemma raw	values
original	78.91	50.17	79.08	53.65	72.74	51.56
none	78.99	50.17	78.39	53.99	72.74	51.56
identity	78.99	50.17	78.39	53.99	72.74	51.56
R_early	79.69	50.17	79.25	53.12	71.88	50.87
C_early	80.56	50.35	80.90	54.25	71.35	51.39
I_early	50.00	50.00	54.77	50.69	58.42	50.00
IR_early	50.09	50.00	55.82	50.52	54.60	50.26
R_middle	83.77	82.99	80.56	80.03	82.55	81.68
C_middle	74.05	50.00	80.21	50.87	53.73	50.61
I_middle	49.83	50.00	50.00	50.00	50.17	50.35
IR_middle	50.00	50.69	50.00	49.74	49.83	50.69
R_late	78.99	50.35	78.47	56.25	72.92	51.39
C_late	76.56	50.00	74.74	51.04	72.57	51.48
I_late	78.99	50.17	78.65	53.47	72.74	51.56
IR_late	79.17	50.35	79.17	55.03	72.92	51.39

Agreement/conflict tradeoff. Table 22 gives the values-arm counts behind the trade-off.

Table 22: Values arm under the middle-band report cut: correct answers out of 576 in the conflict and agreement states (explicit-mask `none` → report cut). The break-even conflict weight sets $\Delta A(\rho) = 0$ in Equation (14).

Model	Conflict /576	Agreement /576	Break-even ρ
Qwen 8B	2 → 438	576 → 518	11.74%
InternVL 8B	50 → 386	572 → 536	9.68%
Gemma 12B	36 → 442	558 → 499	12.69%

To examine how this trade-off depends on the mixture of cases, let ρ be the conflict weight. If only this weight changes and the measured conditional accuracies stay fixed, the overall accuracy change is

$$\Delta A(\rho) = \rho \Delta A_{\text{conflict}} + (1 - \rho)\Delta A_{\text{agreement}}. \tag{14}$$

The resulting break-even weights are listed in Table 22. They describe a reweighting of this panel's measured conditional accuracies. They are not deployment thresholds, because both conflict frequency and conditional accuracy may change outside this panel.

F.4 Numerical variation and conditional bounds

All 900 R05 engineering/evaluation groups pass the fixed structural checks. Recorded score-identity, all-layer report-invariance, all-layer image-invariance and score-repeat differences are zero. Some native labels differ across backends or repetitions. Table 23 reports these discrepancies on the evaluation panel, and Table 24 applies Equation (8) to the recorded outputs.

The maximum score gap in Table 23 is

$$\max_{i,a,\nu} \left| Z_{ia}^{(\nu)}(\text{original}) - Z_{ia}^{(\nu)}(\mathsf{none}) \right|.$$

We compare the two backends at the same input, report arm and repetition, then take the maximum over all 2,304 evaluation comparisons per model, before averaging repetitions. Thus the quantity is a difference in the transformed continuation score Z, not in the log likelihood L or the aggregated response Γ^{score}. The saved field uses the A/B sign before canonical remapping; multiplying both scores by the same σ_o leaves their absolute difference unchanged. This backend comparison is separate from the final-head logit check described in Appendix G.

Table 23: Recorded R05 evaluation discrepancies. Max $|\Delta Z|$ is the largest absolute per-input score difference between `original` and `none`, before averaging repetitions. Bridge and identity count native-label disagreements, and repeat compares fixed native repetitions. Flagged counts give the union of affected report-specific contexts. Denominators count saved comparisons or contexts, not independent bases.

Model	Max $\lvert\Delta Z\rvert$	Bridge	Identity	Repeat	Flagged
Qwen 8B	0.244919	3/2,304	0/2,304	6/17,280	7/1,152
InternVL 8B	0.186706	16/2,304	2/2,304	11/17,280	17/1,152
Gemma 12B	0.000000	0/2,304	0/2,304	8/17,280	7/1,152

The ranges in Table 24 hold every unflagged output fixed and let native responses at flagged contexts vary over $[-1, 1]$. Appendix B explains why Gemma's positive lower bound for T is compatible with its nonsignificant base-level test.

F.5 Separate R04 supporting panel

The supporting R04 study completed all 72 evaluation units per model on its separate 24-base panel. As Figure 11 shows, middle report cuts raise values-arm accuracy from 50.17% to 75.69% for Qwen

Table 24: Conditional finite-panel ranges for R05 primary contrasts when every native response at any flagged context can vary in $[-1, 1]$, holding unflagged observations fixed. These are post hoc sensitivity calculations.

Model	D(original)	T_{middle}	J_{middle}
Qwen 8B	$[-.5868, -.5625]$	$[.4410, .4896]$	$[.5035, .6007]$
InternVL 8B	$[-.5243, -.4653]$	$[.5069, .6250]$	$[.3646, .6007]$
Gemma 12B	$[-.4340, -.4097]$	$[.0278, .0764]$	$[.3507, .4479]$

4B and from 50.69% to 86.89% for InternVL 4B. When the specified image edges are also cut, accuracy falls to 50.17% and 50.61%, respectively. Table 25 reports both the fixed 14-test family and the separate four-test family for middle-band contrasts from generated answers.

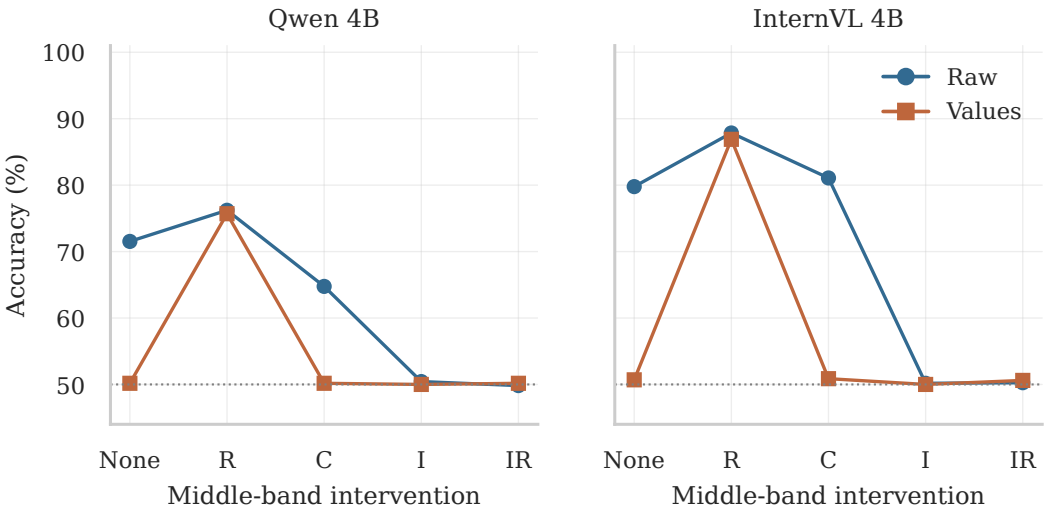


Figure 11: Supporting R04 native-accuracy results for two 4B checkpoints on a separate new 24-base panel. Each report arm has 1,152 answers per condition. Middle report cuts improve values accuracy, and joint image/report cuts remove that gain. These images differ from R05, so the panels are not a controlled model-size comparison.

Under their respective fixed test families, the middle-band score and native T/J contrasts are positive and significant in both models. InternVL's early score T has a positive nominal interval, but its adjusted $p = 0.0533$ does not meet the 0.05 threshold after correction. Both late score T contrasts are negative; therefore, they do not support a report-cut advantage over the text-control cut on this endpoint.

Table 25: All fixed R04 tests on its separate 24-base panel. Holm correction is applied separately to the first 14 rows (standard-backend generated-answer D and all-band score T/J) and the last four (middle-band generated-answer T/J). Intervals are nominal 95% base-bootstrap intervals; repetitions are averaged within base.

Model	Endpoint	Contrast	Band/backend	Estimate	Nominal 95% CI	Adj. p
Qwen 4B	native	D	original	-0.422	$[-0.491, -0.349]$	1.40×10^{-4}
Qwen 4B	score	T	early	-0.001	$[-0.018, +0.014]$	1.000
Qwen 4B	score	T	middle	+0.293	$[+0.239, +0.346]$	1.40×10^{-4}
Qwen 4B	score	T	late	-0.078	$[-0.118, -0.038]$	0.008
Qwen 4B	score	J	early	-0.001	$[-0.020, +0.017]$	1.000
Qwen 4B	score	J	middle	+0.409	$[+0.344, +0.474]$	1.40×10^{-4}
Qwen 4B	score	J	late	+0.001	$[-0.002, +0.004]$	1.000
InternVL 4B	native	D	original	-0.578	$[-0.663, -0.495]$	1.40×10^{-4}
InternVL 4B	score	T	early	+0.027	$[+0.008, +0.044]$	0.053
InternVL 4B	score	T	middle	+0.514	$[+0.431, +0.592]$	1.40×10^{-4}
InternVL 4B	score	T	late	-0.226	$[-0.255, -0.197]$	1.40×10^{-4}
InternVL 4B	score	J	early	+0.012	$[-0.010, +0.032]$	1.000
InternVL 4B	score	J	middle	+0.510	$[+0.446, +0.577]$	1.40×10^{-4}
InternVL 4B	score	J	late	-0.001	$[-0.004, +0.003]$	1.000
Qwen 4B	native	T	middle	+0.281	$[+0.219, +0.340]$	4.00×10^{-5}
Qwen 4B	native	J	middle	+0.401	$[+0.325, +0.472]$	4.00×10^{-5}
InternVL 4B	native	T	middle	+0.585	$[+0.483, +0.682]$	4.00×10^{-5}
InternVL 4B	native	J	middle	+0.552	$[+0.453, +0.653]$	4.00×10^{-5}

G DESCRIPTIVE LAYER READOUTS

The readouts in Figures 12 and 13 use the untouched `original` backend. Thus they contain neither attention cuts nor the explicit-mask `none` execution. We run a separate full-prefix forward pass for each $p \in \{\text{native}, \text{diag}\}$, using the corresponding input defined in Appendix F.2. Each readout is collected once per input, so these curves do not average the two repetitions used for scores and generated answers.

State and token position. Let $h_{\ell,t}(x)$ be the residual vector returned by decoder block ℓ, captured by a forward hook on the whole block. This is after that block's residual updates and before any subsequent additions by the enclosing language model, including Qwen's model-level DeepStack additions. It is not the attention-sublayer output. For each prefix type, we select its last token, $t_p = |x_{ia}^{(p)}| - 1$, using zero-based token positions.

Fixed readout. We apply the checkpoint's actual final normalization and output-embedding module:

$$z_{\ell,ia}^{(p)} = H_\theta(h_{\ell,t_p}(x_{ia}^{(p)})), \qquad H_\theta(h) = \text{lm_head}_\theta(\text{final_norm}_\theta(h)). \tag{15}$$

This readout uses the existing checkpoint modules: we fit no translator and add no rescaling or soft-capping. We cast the resulting logits to float32 before computing margins and the full-vocabulary softmax.

Numerical check. We compare the readout at the final block and last prefix token with the model's returned logits at that token. We take the maximum absolute difference over the full vocabulary, with a fixed tolerance of 0.2 logit units.

The saved R05 evaluation records contain 2,304 such checks per model: 1,152 native prefixes and 1,152 diagnostic prefixes. All pass. Across both prefix types, the observed maximum error is 0 for Qwen 8B and 0.125 for each of InternVL 8B and Gemma 12B. These are readout-versus-returned-logit differences within the original backend, whereas Table 23 compares transformed continuation scores between two backends. Their numerical values are therefore not subject to the same tolerance.

Margins, probability mass and aggregation. The two option IDs are $t_A = 23465, t_B = 36721$ for Qwen/InternVL (tokens `>A`, `>B`), and $t_A = 236776, t_B = 236799$ for Gemma (`A`, `B`); these

are the branch-leading tokens in Table 18. For each input, define the canonical signed margin and probability mass

$$m^{(p)}_{\ell,ia} = \sigma_{o_i}\big(z^{(p)}_{\ell,ia,t_A} - z^{(p)}_{\ell,ia,t_B}\big), \tag{16}$$

$$\mu^{(p)}_{\ell,ia} = \mathrm{softmax}(z^{(p)}_{\ell,ia})_{t_A} + \mathrm{softmax}(z^{(p)}_{\ell,ia})_{t_B}. \tag{17}$$

For a fixed report arm, let $\mathcal{I}_n$ contain the 24 input conditions for base n, crossing the three tasks, two option orders, two table worlds and two image worlds. To obtain the curves, we average within each base and then equally over bases:

$$\mathcal{L}^{(p)}_{\ell,a} = \frac{1}{24}\sum_{n=1}^{24}\frac{1}{24}\sum_{i\in\mathcal{I}_n} w_i m^{(p)}_{\ell,ia}, \tag{18}$$

$$M^{(p)}_{\ell,a} = \frac{1}{24}\sum_{n=1}^{24}\frac{1}{24}\sum_{i\in\mathcal{I}_n} \mu^{(p)}_{\ell,ia}. \tag{19}$$

This averaging gives equal weight to bases, tasks, option orders, table worlds and image worlds. If the other factors are held fixed, multiplying each margin by the image-world sign and averaging is equivalent to taking half the difference between the two image-world margins. In contrast, probability mass is averaged without either the option-order sign or the image-world sign.

The plotted quantities therefore need to be distinguished from the two response endpoints. Both the generated-answer response Γ and the transformed continuation response Γ^{score} lie in $[-1, 1]$, whereas the raw logit-margin contrast $\mathcal{L}$ has no such bound. The probability mass satisfies $M \in [0, 1]$. Figure 12 uses the end of x^{diag}, while Figure 13 uses the end of x^{native}. In both figures, layer indices run from 1 to the number of blocks, and shading marks the fixed middle third.

The prefix distinction matters when interpreting the mass. At Gemma layer 32, the diagnostic A/B mass is 0.0520 for raw and 0.0459 for values; at the final layer, it is approximately one. At the native-prefix end, however, the next token begins the answer tag, and final-layer A/B mass is only approximately 1.68×10^{-5} for raw and 3.16×10^{-6} for values. This small mass measures the probability of immediately producing the specified branch-leading tokens, not the probability of eventually answering A or B. Those tokens are `>A` and `>B` for Qwen/InternVL and `A` and `B` for Gemma; their mass excludes the tokens needed to open the answer tag.

Calibration and peak location. There is no calibration guarantee when the final head is applied to intermediate states. The tuned lens addresses this mismatch by learning layer-specific affine transformations (Belrose et al., 2023), whereas our curves use the fixed raw readout. The location of a peak depends on this readout, so a peak alone does not identify a causal layer.

Margin magnitude and sign. A diagnostic margin can respond to the image while retaining the same sign. For example, if $m(w) = \alpha + \beta w$ with $\beta \neq 0$ and $|\alpha| > |\beta|$, changing the image world changes the margin but not the binary preference induced by its sign. That preference uses only the branch-leading token logits. It is distinct from both the native generated answer and the score based on the complete continuation.

Readouts and interventions. An attention cut can change generated answers through multiple interacting pathways. The lens curves and cuts therefore provide different evidence: the curves measure decodability under a fixed readout, and the cuts measure changes in outputs under specified interventions. A readout contrast alone does not establish the effect of an intervention, nor does an intervention identify everything that can be read from the intermediate states.

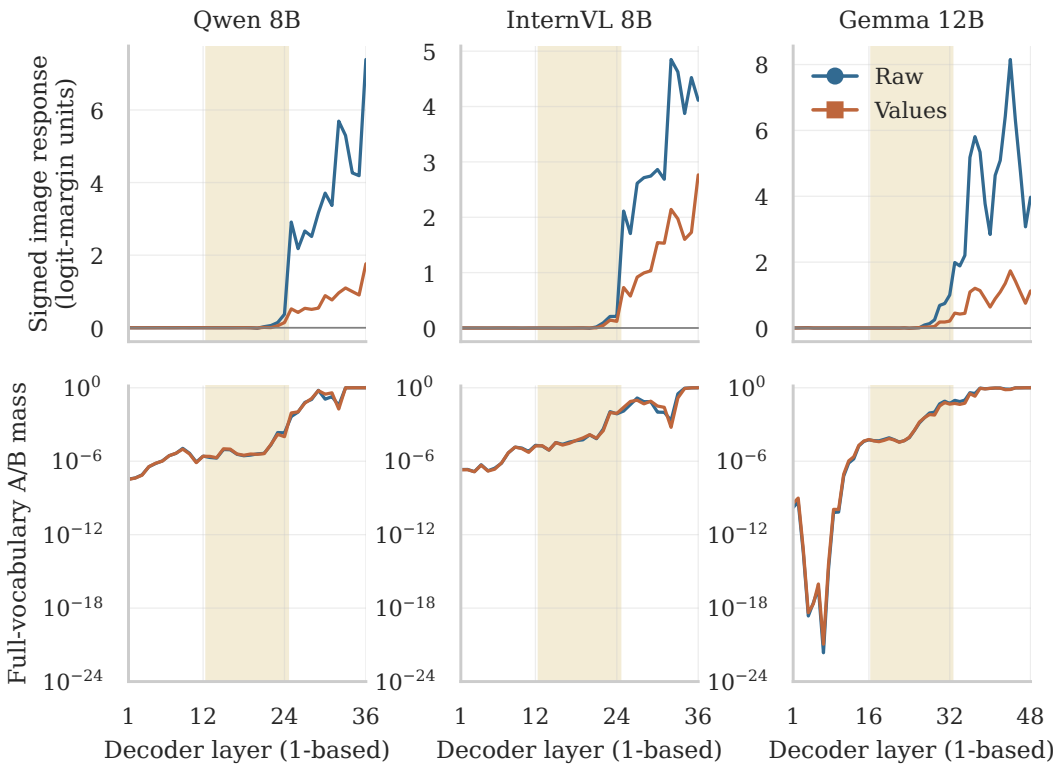


Figure 12: R05 raw logit-lens readouts at the diagnostic-prefix end, across every decoder layer. Signed image contrasts of the A/B branch-token logit margin are accompanied by their probability mass under a full-vocabulary softmax. Shading marks the fixed middle third. Readouts use the untouched backend and the fixed final normalization and output head. Equations (18)–(19) define the averages; they are descriptive, not causal localization.

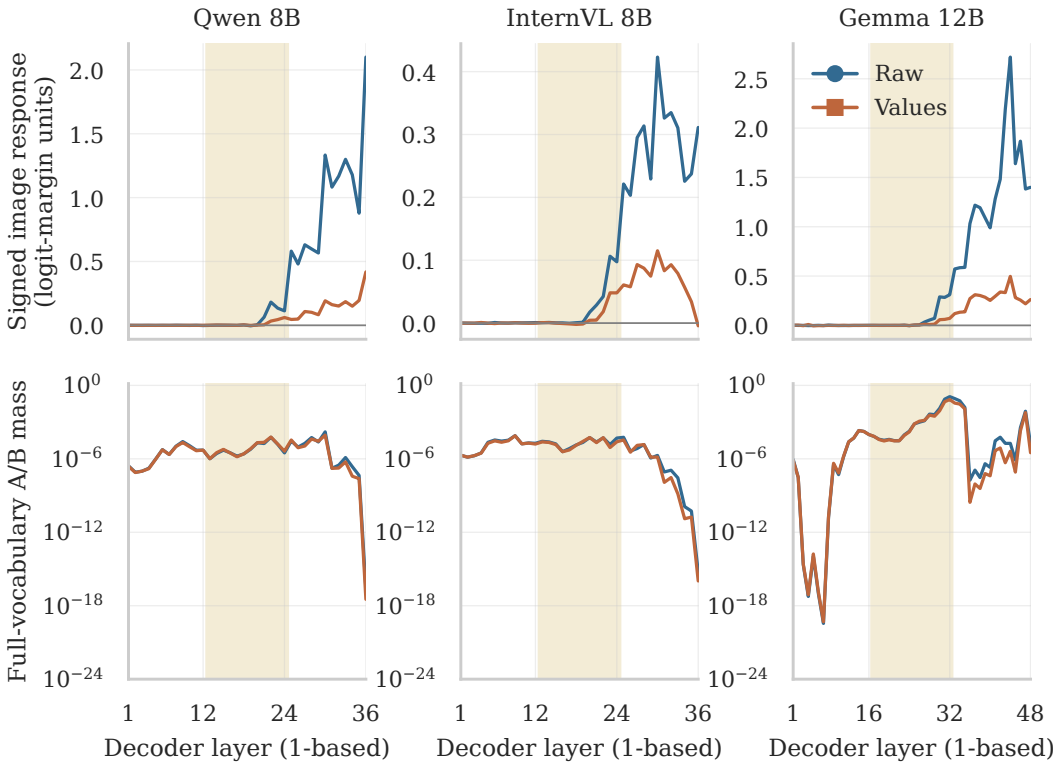


Figure 13: R05 raw logit-lens readouts at the native generation-prefix end, with the same layers and middle-band shading as Figure 12. Mass is assigned to the same branch-leading token IDs as in the diagnostic readout. At this prefix generation may first open an answer tag; the mass is not the probability of the eventual answer or calibrated intermediate-layer confidence.

H DEVELOPMENT RESULTS AND INCOMPLETE STUDIES

Table 26 summarizes earlier recovery attempts and incomplete studies. Their protocols and denominators differ from the completed primary panels. Available summaries and source snapshots for these development studies are provided in the code and results repository linked in the Introduction. Study IDs are retained to locate the corresponding records. Two entries use terms that need definition.

Local repair (P04). *Repairs* count initially wrong answers made correct; *harms* count initially correct answers made wrong. Both are relative to the cached six-read median, and they are separate counts, not a numerator and a denominator.

Neutral-table policy (P06). The policy averages a table with its question-specific value swap, then averages the image-minus-no-image answer-branch log odds over both option orders. Its archived definition and primary controls are in `code/development_source/p06_core.py`.

Table 26: Archived development outcomes and incomplete studies relevant to the paper. These use different protocols and denominators from the completed primary panels. The table includes negative results and incomplete runs, with their original denominators.

Study	Recorded outcome and interpretation
P02: candidate re-scoring	Candidate re-scoring does not beat the numerical median: Qwen and InternVL medians reach 288/288 semantic questions; Gemma P02 and median both score 281/288. JSON-fence rejection is retained as a formatting limitation.
P03: geometric verification	All 64 engineering charts per model finish, but no model passes the perception gate. Geometric checking plus fallback scores 269/384, 292/384 and 334/384, versus numerical medians 380/384, 378/384 and 339/384 (Qwen, InternVL, Gemma). The failed gate prevents a calibrated evaluation of the perception method.
P04: local predicate repair	The primary policy yields 2 repairs and 3 harms for Qwen, 0 repairs and 4 harms for InternVL, and 5 repairs and 16 harms for Gemma. Conditional predicates can be correct while the remaining uncorrected values still imply the wrong task answer.
P06: neutral-table calibration	The fixed 24-base confirmation yields 12 negative primary policy contrasts, 11 Holm-significant. Table-swap invariance does not establish accurate chart response.
P07: value assignment	A fixed assignment policy improves over independent marginalization and summary answering but loses to first-read and median numerical controls. Six of twelve fixed contrasts are positive and six negative, all Holm-significant, without overall superiority over the numerical controls.
R01: report placement and protocol	Only 92 of the 576 units per main model are available: 15 complete bases and part of a sixteenth. The remaining 484 units per model are unobserved, restricting the evidence to this partial panel.
R03: activation swaps	Only 71 of the 72 intervention units per 4B model are complete. Repeated real-model self/identity checks disagree on two remaining contexts, so we do not run the planned complete-panel inference with correction over 16 tests. Donor-state analyses are descriptive.

For R03, baseline outputs are available for 572 of the 576 contexts per report arm for Qwen and all 576 for InternVL. We report baseline availability separately from intervention completion because an available baseline does not establish that its intervention checks passed. The shared fully observed subset contains 22 bases and supports descriptive donor comparisons, but the failed controls prevent the planned complete-panel causal inference.

Activation swaps and accompanying text. Earlier single-model activation-swap studies also changed claims and alternative text. Therefore, any apparent restoration in those studies cannot

be attributed to the activation swap alone, and their broad depth profiles do not isolate a layer. R04 and R05 instead use the edge-cut design to test access to report and image tokens.

Table-swap invariance. Table-swap invariance alone does not show that an answer is correct. For example, a constant prediction is invariant to table swaps but is only 50% correct on balanced opposite-answer worlds.

Local repair. Local repair poses a different problem. If we change only one coordinate of an extracted vector, the remaining coordinates may still be wrong. A comparison can be correct given those remaining coordinates and still disagree with the comparison for the true vector. Answering the local repair question correctly therefore does not guarantee recovery of the task answer.