DO VISION-LANGUAGE MODELS USE THE REGIONS THEY CITE? A STEP-SPECIFIC INTERVENTION TEST

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models increasingly attach a bounding box to each reasoning step as the visual evidence for that step. Localization metrics measure whether such boxes match reference regions; they do not test whether the step is selectively sensitive to the region it cites. We introduce a step-specific intervention test that requires no region annotations: perturb each cited region in turn, record the change in every fixed step's mean per-token log-likelihood, and compare a step's sensitivity to its own region with its sensitivity to distinct regions cited by other steps of the same chain. Because every pair of distinct regions is compared in both directions, the expected own-region ordering probability is at most 50% under an additive step-and-region sensitivity model with row-exchangeable residuals, so a confidence interval entirely above 50% is evidence against that joint null. For Qwen3-VL-4B the own-region ordering probability is about 61% in each of three image-disjoint GQA cohorts and 72% on A-OKVQA. On the largest GQA cohort, every preregistered change of scoring context and intervention operator (removing the preceding steps, replacing blur with within-box patch permutation, or both) keeps the interval above 50% (lowest point estimate 58%). On a held-out cohort, under a protocol specified in advance, we regenerate each current step with the preceding steps held fixed. The fixed-text margin predicts the relative content change caused by blurring the step's own rather than a sibling step's region (Spearman's $\rho = 0.504$, 95% CI $[0.402, 0.600]$). The advantage is strong for GLM-4.6V-Flash on two image-disjoint cohorts, weak and cohort-variable for InternVL3.5-4B, and partial throughout: about 40% of Qwen3-VL-4B's comparisons in the three fixed-text GQA cohorts favor another step's region. A cited region is a testable claim about evidence use, not proof of it.

1 INTRODUCTION

Vision-language models increasingly reason in explicit steps and attach a bounding box to each step (Shao et al., 2024; Wu et al., 2025; Fan et al., 2025; Zhang et al., 2026). A step such as "identify the animals in front of the trees" arrives with coordinates, so a grounded rationale promises what text-only chain-of-thought (Wei et al., 2022) cannot: a reader can see where the model says it looked.

A cited region is a claim about evidence, and localization cannot tell whether the claim holds. A box can sit exactly on the right object while the step does not depend on it, and an imprecise box can still cover what the step depends on. Work on text-only reasoning already shows that stated rationales need not describe what drove a prediction (Turpin et al., 2023; Lanham et al., 2023). Current evaluations of grounded reasoning score overlap with annotated referents (Kazemzadeh et al., 2014), answer–box consistency (Wu et al., 2025), judged evidence quality (Uppaal et al., 2026) or localization jointly with answer correctness (Li et al., 2026). They describe where the model pointed and what it answered; they do not test whether the step depends on the content of the region it cited.

We test that dependence with an intervention. For each chain we perturb every cited region in turn and record the change in the teacher-forced mean per-token log-likelihood of each step, which gives a matrix $J[s_t, r_k]$ (Figure 1). Our operational target is relative intervention sensitivity: is a step

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.

more sensitive to the region it cited than to the regions its sibling steps cited? Relative sensitivity of this kind does not by itself establish semantic evidence use. Because the controls come from the chain itself, each comparison stays within one image, question and generation, and the assay needs no region annotation. Comparing a cited region with an unclaimed region, or an intervention with no intervention, would not suffice: a region that matters to every step, or a step that reacts to any perturbation, passes such a test without any step-specific link.

Two properties of the construction do the work. Each pair of distinct cited regions contributes both directional comparisons, so a region that matters to every step enters on both sides. Additive step-level effects, such as those of a longer step or a lower baseline likelihood, cancel within each difference. Proposition 1 states the consequence: if sensitivity has an additive mean in a step effect and a region effect and the residuals are row-exchangeable, the expected ordering probability is at most 0.5. An interval above 0.5 is therefore evidence against that joint null; it does not locate the departure or identify a mechanism (Sections 2 and 6).

Qwen3-VL-4B shows a reproducible own-region ordering advantage of about 0.61 in three GQA cohorts that share no images, the last two preregistered. On the largest of them, the advantage persists when the model's preceding steps are removed from the scoring context, when blur is replaced by a within-box patch permutation, and when both change at once; it also appears on an A-OKVQA cohort. On a held-out cohort, the margin predicts how much more the regenerated current step changes, with preceding steps held fixed, when its own region rather than a sibling's is blurred ($\rho_s = 0.504$). Across checkpoints the picture is uneven: GLM-4.6V-Flash, from a distinct model family, shows a strong advantage on two cohorts, while InternVL3.5-4B is weak and cohort-variable. Even where the coupling is clearest it is partial: about four in ten of Qwen's comparisons in the three fixed-text GQA cohorts favor another step's region (Table 1). Our contributions are:

- **An annotation-free own-versus-sibling assay with a stated null.** It applies to any model that emits step-level regions and exposes step-token log-likelihoods; Proposition 1 states the nuisance models it excludes, and counterexamples mark its limits (Appendix C).
- **A replicated but partial own-region advantage.** In Qwen3-VL-4B the ordering probability is about 0.61 in three image-disjoint GQA cohorts, stays above 0.5 under a preregistered 2×2 change of scoring context and operator, and appears on A-OKVQA; post-hoc checks locate it in non-nested, low-overlap region pairs and find it under equal chain weighting.
- **Behavioral predictive validity under a fixed-prefix protocol.** On a held-out cohort the margin predicts relative content change of the regenerated current step ($\rho_s = 0.504\ [0.402, 0.600]$); the prespecified sensitivity analyses keep the interval above zero, and a blinded audit by one author agrees with the rule-based endpoint on 86% of 120 comparisons.
- **Checkpoint heterogeneity and cross-family replication.** Under the same assay the advantage is strong for GLM-4.6V-Flash on two image-disjoint cohorts and weak and cohort-variable for InternVL3.5-4B.

2 A Step-Specific Intervention Assay

Chains and claimed regions. Given an image x and a question q, the model produces a chain of steps $s_1, \ldots, s_T$, each with a claimed box r_t; steps without a valid box are dropped. Box coordinates are removed from the scoring context and from the scored target, so no sensitivity can arise from the model reproducing its own coordinates. We keep three objects apart: the claimed region r_t (where the model says its evidence is), its localization against annotated objects (best IoU), and the interventional sensitivity $J[s_t, r]$ defined below. The assay tests the region as a claim about the step, not whether perturbation affects the model at all (the paired comparison removes that additive effect) or whether the answer is correct.

Two intervention operators. The primary operator I_r replaces the pixels inside box r with a Gaussian-blurred copy whose radius scales with the box, using a feathered edge and no box expansion. Blur suppresses local detail while keeping coarse color and layout; it was chosen over constant fill, noise fill and patch replacement before any sensitivity was computed. The second operator is a deterministic *within-box patch permutation*: it rearranges equal-sized cells of the box, preserving its

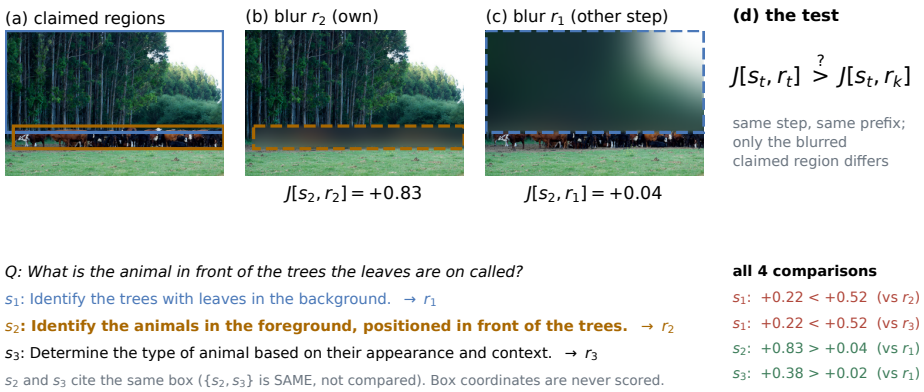


Figure 1: **The assay on one chain.** Qwen3-VL-4B answers a GQA question in three steps. Steps s_2 and s_3 cite the same box, so the pair $\{s_2, s_3\}$ is SAME and excluded; the DISTINCT pairs $\{s_1, s_2\}$ and $\{s_1, s_3\}$ give four ordered comparisons. Blurring the region cited by s_2 lowers its mean per-token log-likelihood by 0.83, whereas blurring the region cited by s_1 lowers it by 0.04; for s_1 the comparison runs the other way (0.22 for its own region against 0.52 for s_2's). The tree region cited by s_1 has best-object IoU 0.72, yet s_1 is more sensitive to the sibling region. Two of the four comparisons favor the own region, a partial effect within one chain. The chain was selected by a rule fixed before any sensitivity was computed; the displayed comparison was chosen while drafting (Appendix G). Image: Visual Genome, CC BY 4.0.

pixel multiset (before edge feathering) while destroying spatial layout. The two operators have different layout and detail profiles, but neither excludes perturbation artifacts they share (Appendix A).

Interventional sensitivity and two scoring contexts. For a scoring context c,

$$J^c[s_t, r] \;=\; \bar{\ell}(s_t \mid c, x) \;-\; \bar{\ell}(s_t \mid c, I_r(x)), \tag{1}$$

where $\bar{\ell}$ is the step's mean per-token log-likelihood under teacher forcing. The default *prefix* context $c = (q, s_{<t})$ conditions on the question and the model's own preceding steps; the *question-only* context $c = (q)$ removes the generated steps and scores the same step text, which tests whether the ordering requires the verbalized prefix. The scored tokens are the step text together with its JSON key, quotes and delimiter (Appendix A). $J > 0$ means the intervention lowered the step's likelihood, and $J < 0$, an *intervention-facilitating* effect, means it raised it; values are kept signed and never clipped. The default prefix context with blur is called cell A below. We use teacher forcing because regenerating the chain under each intervention would change the text of later steps, and a change in wording could not be attributed to any one step. With every step's text fixed, J measures how much the image supports that exact step in the scoring context, not what the model would freely generate; Section 4.3 tests the link to generation separately.

The sensitivity matrix and eligible comparisons. For each chain we compute the full matrix $J_{t,k} = J[s_t, r_k]$. A comparison of r_t with r_k is meaningful only if the two boxes are genuinely different, so each unordered pair $\{t, k\}$ is classified as SAME, DISTINCT or AMBIGUOUS by a fixed geometric rule on normalized center distance, area ratio, intersection-over-smaller and IoU (Appendix B); only DISTINCT pairs are compared, and the pair set is fixed from boxes alone, before any sensitivity exists. Each pair contributes both ordered comparisons $\Delta_{t,k} = J_{t,t} - J_{t,k}$ and $\Delta_{k,t} = J_{k,k} - J_{k,t}$. The primary endpoint is the own-region ordering probability $\pi_{\text{own}} = P(\Delta_{t,k} > 0)$, a probability of superiority over comparisons: 1.0 means every eligible comparison favors the own region, and 0.5 is the reference under the null of Proposition 1, not a generic chance level. π_{own} describes comparisons; it is not the share of steps that use their evidence, and a single non-win does not mark a step as unfaithful. The median of Δ (null 0) measures the size of the advantage on the J scale; a small median with $\pi_{\text{own}} \approx 0.6$ means the own region wins often, typically by a modest margin. Comparisons are pooled, so chains with more DISTINCT pairs weigh more in the point estimate (equal chain weighting: Section 4.2); uncertainty comes from a bootstrap that resamples chains, never individual comparisons (2000 resamples).

Why the diagonal. Two facts need no distributional assumption. Any additive step-level quantity, such as a step's length, baseline log-likelihood or difficulty, cancels from every $\Delta_{t,k}$; non-additive or variance effects of such quantities do not. Additive step and region effects both cancel from the symmetric pair score $S_{t,k} = \frac{1}{2}(\Delta_{t,k} + \Delta_{k,t})$. The ordering endpoint needs one further condition.

Proposition 1 (Paired-design symmetry)**.** *Suppose that within a chain $J_{t,k} = \alpha_t + \beta_k + \varepsilon_{t,k}$ with arbitrary fixed step effects α and region effects β, and that, conditional on these effects and the fixed DISTINCT pair set, the residuals are row-exchangeable for every DISTINCT pair $\{t, k\}$: $(\varepsilon_{t,t}, \varepsilon_{t,k}) \overset{d}{=} (\varepsilon_{k,t}, \varepsilon_{k,k})$. Then, conditional on α, β and the pair set, $P(\Delta_{t,k} > 0) + P(\Delta_{k,t} > 0) = 1 - P(\Delta_{t,k} = 0)$; hence the expected ordering probability is at most $1/2$, with equality when ties have probability zero, and $\Delta_{k,t} \overset{d}{=} -\Delta_{t,k}$.*

Proof. Write $x = \beta_t - \beta_k$, $U = \varepsilon_{t,t} - \varepsilon_{t,k}$ and $V = \varepsilon_{k,k} - \varepsilon_{k,t}$. Then $\Delta_{t,k} = x + U$ and $\Delta_{k,t} = -x + V$, and row-exchangeability gives $V \overset{d}{=} -U$. Hence $P(U > -x) + P(-U > x) = 1 - P(U = -x)$, and $\Delta_{k,t} \overset{d}{=} -(x + U)$. Since the pair set is fixed before J, linearity carries the identity to the pooled ordering probability. □

The condition allows arbitrary within-chain dependence subject to this pairwise row-exchangeability, residual scale that varies across regions, and asymmetric residuals; ties can only lower the expectation. Because the null is the conjunction of an additive mean and the residual condition, an interval above 0.5 is evidence against the conjunction, not against additivity alone. Three alternatives can produce diagonal dominance without step-specific matching; each has a targeted but partial check. *Step-dependent residual scale* breaks row-exchangeability while the mean stays additive (with residual standard deviations 0.5 and 2 it reaches 0.642 per pair; Appendix C), and it is plausible because scored steps differ in length and content. The symmetric pair score removes an additive mean algebraically and is calibrated when the residual contrast is distributed like its negation, which allows step-dependent scale (Appendix D); it does not cover every residual alternative. A *multiplicative mean*, in which the more image-sensitive step also cites the more influential region, is targeted by the anchor-weighted score of Section 4.7. An *order effect transmitted through the prefix* is removed by question-only scoring (Section 4.2).

3 Experimental Design

Checkpoints. We evaluate three open off-the-shelf checkpoints that met a fixed interface, one box per step under a single prompt, and fixed eligibility criteria applied before any sensitivity was computed: Qwen3-VL-4B-Instruct (Bai et al., 2025) (4.4 B, no thinking trace), InternVL3.5-4B (Wang et al., 2025) (4.7 B, Qwen3-initialized language model, thinking off under this interface) and GLM-4.6V-Flash (Z.ai, 2025) (10.3 B, distinct GLM-V family, thinking trace stripped before parsing; Appendix H). They are general-purpose checkpoints prompted to cite regions rather than systems built for step-level grounded reasoning. Seven further candidate families failed the interface criteria or required an incompatible software stack; GLM was added later under a selection rule written before download and passed the same criteria on the same 120 discovery images.

Data and prompt. Questions come from GQA (Hudson & Manning, 2019), whose Visual Genome images (Krishna et al., 2017) carry scene graphs with annotated objects, and from A-OKVQA (Schwenk et al., 2022) over COCO images (Lin et al., 2014). We take one question per image by a fixed longest-question rule, which avoids analyst selection (for A-OKVQA only the question text is used). A single prompt asks the model to reason step by step and to return a JSON list in which each step has one bounding box; decoding is greedy.

Cohorts. We use four mutually disjoint fixed-text cohorts drawn with fixed seeds: 120-image GQA *discovery*, 200-image GQA *confirmation*, 300-image GQA *robustness*, and 300-image A-OKVQA cohorts; a fifth, held-out cohort serves the behavioral study of Section 4.3. Later cohorts exclude every image used earlier, and the A-OKVQA pool was de-duplicated against GQA by COCO identifiers and perceptual hashing, which agreed on all 49 excluded images (Appendix B).

Model / cohort	Context · operator	Chains / comp.	$\pi_{\text{own}} = P(\text{own} > \text{other})$ [95% CI]	median Δ [95% CI]
Qwen3-VL-4B, GQA — three independent cohorts of 120, 200 and 300 images				
discovery	prefix · blur	78 / 444	**0.610** [0.542, 0.684]	+0.017 [+0.006, +0.022]
confirmation	prefix · blur	116 / 548	**0.608** [0.546, 0.663]	+0.028 [+0.012, +0.042]
robustness	prefix · blur	182 / 916	**0.607** [0.568, 0.647]	+0.022 [+0.013, +0.035]
Qwen3-VL-4B, robustness cohort — same chains and pairs				
H1	question · blur	182 / 916	**0.610** [0.577, 0.644]	+0.049 [+0.034, +0.070]
H2	prefix · perm.	182 / 916	**0.581** [0.542, 0.621]	+0.014 [+0.006, +0.021]
H3	question · perm.	182 / 916	**0.587** [0.544, 0.632]	+0.026 [+0.010, +0.042]
Second dataset and other checkpoints (prefix · blur)				
H4 Qwen, A-OKVQA	prefix · blur	112 / 582	**0.715** [0.668, 0.758]	+0.160 [+0.101, +0.216]
GLM† robustness	prefix · blur	93 / 398	**0.799** [0.761, 0.837]	+0.100 [+0.078, +0.157]
GLM† confirmation	prefix · blur	51 / 238	**0.828** [0.782, 0.877]	+0.188 [+0.132, +0.277]
InternVL3.5-4B‡	prefix · blur	163 / 1036	**0.551** [0.505, 0.599]	+0.0030 [+0.0004, +0.0052]

Table 1: **Own-region ordering under the intervention assay.** Chains: chains with at least one DISTINCT region pair; comp.: own-versus-other comparisons, both orders of each pair; perm.: within-box patch permutation. Comparisons are pooled across chains; intervals resample chains; bold marks a pointwise interval that excludes 0.5, which is a confirmatory test only for the preregistered rows. † GLM-4.6V-Flash, supplementary; the confirmation row is a replication specified after the robustness result was known. ‡ descriptive, robustness cohort.

The 2×2 robustness design. On the robustness cohort the same chains, claimed regions and DISTINCT pairs are scored in four cells, {prefix, question-only} × {blur, patch permutation}: cell A is the within-cohort replication reference, and the other three are the preregistered hypotheses H1 (question-only × blur), H2 (prefix × permutation) and H3 (question-only × permutation).

Preregistration. For the confirmation, robustness and A-OKVQA cohorts, hypotheses, endpoints, thresholds and decision rules were committed before any model output existed (for the last two, also the second operator, the box-expansion levels and the analysis code), and the chains and pairs entering each test were frozen before any sensitivity was computed. A hypothesis passes if the 95% chain-clustered interval of π_{own} lies entirely above 0.5; the joint claim is an intersection-union test, individual claims use Holm correction, and stop rules required at least 100 eligible Qwen chains on GQA and 60 on A-OKVQA. The median of Δ was co-primary in the discovery and confirmation decisions and supportive in the robustness program. All criteria were met (Appendix J).

4 Results

4.1 The own-region ordering advantage replicates across three independent GQA cohorts

For Qwen3-VL-4B, a step was more sensitive to its own claimed region than to another step's in 271 of 444, 333 of 548 and 556 of 916 comparisons in the discovery, confirmation and robustness cohorts ($\pi_{\text{own}} = 0.610$, 0.608, 0.607; Table 1), with medians $+0.017$, $+0.028$ and $+0.022$, each interval above zero. The intervals are wide enough that only the direction and rough magnitude replicate, not the third decimal, and the coupling is partial: about four in ten comparisons favored the other step's region.

4.2 The effect survives a change of scoring context, of operator, and of both

On the robustness cohort the same 182 chains and 458 region pairs were scored in all four cells. Removing the model's own preceding steps from the scoring context gave $\pi_{\text{own}} = 0.610$ $[0.577, 0.644]$ (H1), with a median interval above zero; the within-box patch permutation gave 0.581 $[0.542, 0.621]$ (H2); both changes together gave 0.587 $[0.544, 0.632]$ (H3). All three intervals exclude 0.5, Holm correction rejects each individual null, and all four cells stay above 0.5 at the wider central 97.5% bootstrap intervals (Appendix J). Two paired contrasts on identical comparisons with shared resamples are small and include zero: under prefix scoring, blur minus permutation is $+0.026$

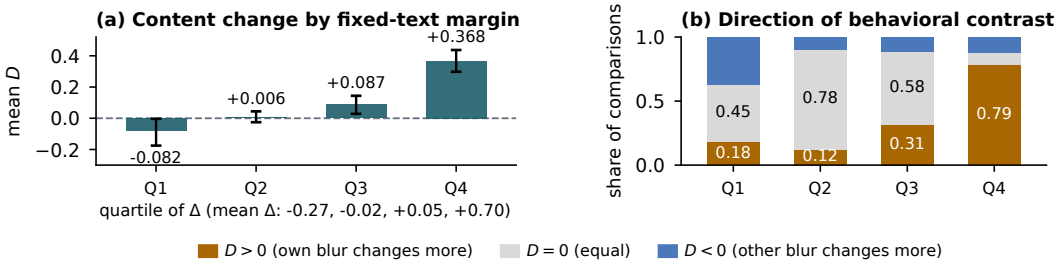


Figure 2: **Behavioral validation, Qwen3-VL-4B, held-out cohort.** The current step is regenerated with preceding steps held fixed. (a) Mean contrast D by quartile of the fixed-text margin Δ, 95% chain-bootstrap intervals; primary statistic $\rho_s(\Delta, D) = 0.504$ $[0.402, 0.600]$ (548 comparisons, 100 chains). (b) Share of comparisons in which own-region blur changed the regenerated step more ($D > 0$), equally ($D = 0$) or less ($D < 0$) than other-region blur.

$[-0.009, +0.063]$; under blur, prefix minus question-only is -0.003 $[-0.039, +0.033]$. These are conditional contrasts, not factorial main effects, and intervals that include zero do not establish that operators or contexts are equivalent. The ordering therefore does not require the verbalized prefix and is not specific to one operator, although it could still reflect artifacts that both operators share.

Post-hoc checks on the same cohort address region geometry and chain weighting (Table 5). The advantage is carried by the 375 non-nested, low-overlap pairs (0.593–0.639 across the four cells, every interval above 0.5); in the 83 nested pairs, where perturbing the larger box also removes most of the smaller one, it sits near 0.5. It also appears in the 52 pairs whose boxes have similar areas (0.625–0.731, every interval above 0.5), and weighting chains equally gives 0.572–0.629, again with every interval above 0.5. These strata were chosen after the primary results and are unadjusted for multiplicity; they narrow geometric and weighting explanations without exhausting them.

4.3 Behavioral validation of the fixed-text assay

To test whether the fixed-text margin $\Delta_{t,k}$ predicts a change in what the model generates, we ran a study whose cohort, generation protocol, endpoint, analysis plan and sample size were frozen before any behavioral output existed (Appendix I). A held-out GQA cohort was processed in a frozen order until 100 eligible Qwen3-VL-4B chains were obtained (159 images; 548 ordered comparisons). For each comparison the current step was regenerated with the preceding steps held fixed (prefix context, greedy decoding) and only the image changed: unaltered, own region blurred, or other region blurred. A fixed rule-based extractor maps each regenerated step to typed content units (numbers, spatial and relational terms, other content words, and the final assertion); d is the Jaccard distance between unit sets and $D_{t,k} = d(\text{base}, \text{own}) - d(\text{base}, \text{other})$. D is a proxy for content change, not a judgment of semantic equivalence, and $\Delta_{t,k}$ is the cell-A margin of the same comparison.

The prespecified primary statistic, the comparison-weighted Spearman correlation between Δ and D with a chain-resampled interval, is $\rho_s = 0.504$ $[0.402, 0.600]$ (Figure 2). The prespecified sensitivity analyses keep the interval above zero: 0.602 $[0.463, 0.713]$ on the 369 comparisons whose regenerated baseline reproduced the stored step's content units, 0.390 $[0.276, 0.497]$ after removing the ten chains most favorable to the association, and 0.548 $[0.437, 0.650]$ with equal chain weights. A fourth, treating invalid continuations as missing, coincides with the primary analysis because no continuation was invalid. A secondary edit-distance endpoint gives 0.521 $[0.425, 0.606]$. Overall $P(D > 0) = 0.352$, $P(D = 0) = 0.473$ and $P(D < 0) = 0.175$. A tie means equal change, not necessarily none: both distances are zero in 219 comparisons (40.0%) and equal but positive in 40 (7.3%; Appendix I).

Blinded manual audit. One author judged 120 comparisons (30 per Δ quartile), each shown as the baseline step and the two intervened steps in random order with condition, Δ, D, boxes and identifiers withheld. The judgments matched the extractor's direction on 103 of 120 (0.858; Cohen's $\kappa = 0.771$) and on 49 of the 54 with a strict direction on both sides (0.907). This checks the extractor against one reader who is also an author; it is not an independent evaluation. The margin therefore predicts relative content change of the regenerated current step under this fixed-prefix protocol. A post-hoc component check supports sibling controls: own-region sensitivity alone correlates more weakly with D ($\rho_s = 0.282$), while the margin retains an association after conditioning on it (partial $\rho_s = 0.444$; Appendix I). For context, the held-out behavioral cohort has a descriptive fixed-text

ordering probability of 0.557 $[0.498, 0.616]$; this was not a prespecified replication of the fixed-text endpoint (Appendix I).

4.4 The own-region ordering advantage appears on a second dataset

On A-OKVQA, whose human-written questions require commonsense and world knowledge rather than GQA's template-generated queries, 416 of 582 comparisons favored the own region: $\pi_{\text{own}} = 0.715$ $[0.668, 0.758]$, median $+0.160$ $[+0.101, +0.216]$. The larger point estimate could reflect A-OKVQA's larger claimed regions (median 11.3% of the image against 6.4%), its different questions, or both; the design does not separate these, and we compare only whether the interval lies above 0.5, not magnitudes across datasets.

4.5 A second model family, and model size within the Qwen3-VL family

In cell A, GLM-4.6V-Flash showed a strong own-region ordering advantage on the robustness cohort: 318 of 398 comparisons, $\pi_{\text{own}} = 0.799$ $[0.761, 0.837]$. A replication on the image-disjoint confirmation cohort, specified after that result was known with the same checkpoint revision, prompt, budget, parser and assay, gave 197 of 238 comparisons, 0.828 $[0.782, 0.877]$, on 51 eligible chains. An exploratory 2×2 on the robustness chains keeps all four cells between 0.676 and 0.799, each interval above 0.5, so for this checkpoint too the advantage is not tied to the default context or operator (Appendix E). Both extensions lie outside the original preregistration.

Within the Qwen3-VL family, the 2B and 8B Instruct checkpoints were run on the robustness cohort under a protocol frozen after the 4B result was known (same prompt, budget, parser and rules; larger checkpoints exceed the memory of our single 40 GB GPU; Appendix H). Eligibility differs by size (91, 182 and 104 of 300 chains; 8B often emits fewer than two valid regions), and π_{own} is 0.541 $[0.486, 0.594]$, an interval that includes 0.5, then 0.607 $[0.568, 0.647]$ and 0.696 $[0.644, 0.745]$. The point estimates increase with size, but this is an exploratory same-family comparison, not a scaling law: because eligibility is model-dependent, the ordering may partly reflect differences in the analyzed chain populations rather than size alone.

4.6 Boundary: weaker and cohort-variable coupling in InternVL3.5-4B

InternVL3.5-4B produces step-specific region claims in most multi-step chains but cites much smaller regions (median 1.4% of the image against 6.4% for Qwen). Its results are cohort-variable: 0.515 $[0.432, 0.601]$ and 0.543 $[0.497, 0.590]$ in the first two cohorts, with intervals including 0.5, and 0.551 $[0.505, 0.599]$ in the robustness cohort, where both the ordering and the median interval narrowly exclude their nulls. Its own-region sensitivities have a small signed mean (0.011 on the robustness cohort) and about half are negative, so they partly cancel (mean absolute value 0.028). We therefore read this checkpoint as showing weak, cohort-variable coupling rather than no coupling. A preregistered diagnostic asked whether small claimed regions explain the weakness: expanding every box about its center to $2\times$, $4\times$ and $8\times$ its area raised the mean own-region sensitivity from 0.011 to 0.036 with no evidence of an increase in the ordering probability (0.531, 0.549, 0.546), whereas Qwen stayed above 0.5 at every level (Appendix F); the preregistered rule classifies this outcome as mixed, neither footprint-limited nor a persistent non-detection. A post-hoc regeneration with a 2048-token budget did not materially change the ordering (0.544 $[0.499, 0.589]$; Appendix H).

4.7 Secondary analyses

Scale-interaction diagnostic (preregistered, secondary). Two statistics on the same pairs each remove a different nuisance mean. The symmetric pair score S removes an additive step-plus-region mean and gives $P(S > 0) = 0.655$ $[0.597, 0.715]$. The anchor-weighted score $X_{t,k} = \Delta_{t,k} G_k + \Delta_{k,t} G_t$, where G_t is step t's sensitivity to blurring the whole image, removes a proportional step$\times$region mean for either sign of the step factor when the anchors are noise-free and share the assumed factorization; it gives $P(X > 0) = 0.616$ $[0.564, 0.669]$. Both intervals exclude 0.5, so the joint rule passes. Removing a mean is algebraic; each sign test is calibrated only under a further symmetry condition on the residual contrast (for X also on the anchors), and the bootstrap intervals are asymptotic (Appendix D). The pre-execution design study calibrated an earlier, anchor-normalized form of the rule on pairs with two positive anchors. A post-hoc calibration of the rule

applied here (signed anchors, all pairs, the paper's bootstrap) passed in at most 0.017 of synthetic cohorts under additive nulls and 0.040 under proportional nulls with noisy anchors, but in every cohort under two synthetic no-matching stress structures (Appendix D). The diagnostic narrows the scale-interaction explanation; it does not identify step-specific use.

Scored tokens, temporal direction, signed effects and localization. Re-scoring the same teacher-forced sequences over step-content tokens alone, without the JSON structure tokens, raises the ordering probability in all four Qwen cells (0.615–0.654, every interval above 0.5) but changes 13–22% of individual win/non-win outcomes (post-hoc; Appendix E). The aggregate advantage is robust to this scoring choice; individual comparisons are not. Splitting comparisons by whether the competing region belongs to an earlier or a later step gives 0.590–0.631 on the robustness cohort under either context and 0.668 (earlier) and 0.547 (later) on the confirmation cohort. These splits are descriptive: each keeps only one direction per pair, so Proposition 1 does not calibrate a 0.5 null within either split, and the question-only cell of Section 4.2 is the direct check on transmission through the prefix. A stable minority of own-region sensitivities are negative, meaning that blur makes the step more likely (under blur on GQA: Qwen 27–32%, InternVL 43–50%); the sign identifies the direction of the effect, not an absence of image dependence. Within Qwen the advantage shows no monotonic attenuation as the cited region's localization worsens, although the low-IoU strata are small. Further post-hoc diagnostics (effect location, step length, answer correctness) are in Appendix E.

5 Related Work

Grounded reasoning. Referring-expression benchmarks score whether a box matches a described object (Kazemzadeh et al., 2014), and grounding-capable VLMs emit coordinates as text (Chen et al., 2023; Peng et al., 2023). Grounded-reasoning methods attach regions to intermediate steps, supervising them with annotated boxes (Shao et al., 2024), rewarding interleaved box-and-text chains (Fan et al., 2025; Zhang et al., 2026) or scoring answer–grounding consistency (Wu et al., 2025), following earlier work that paired justifications with pointed evidence (Park et al., 2018). These measure where regions land and whether answers agree with them. GRIT comes closest to an intervention: it removes box tokens and reports lower visual attention, an internal readout rather than the likelihood of the step that cited the box.

Faithfulness of explanations. Faithfulness is graded and distinct from plausibility (Jacovi & Goldberg, 2020); erasure metrics test textual rationales (DeYoung et al., 2020), and interventions on stated reasoning show that chain-of-thought can misrepresent what drives an answer (Turpin et al., 2023; Lanham et al., 2023). Many such tests measure self-consistency between inputs and outputs rather than faithfulness to the model's internal computation (Parcalabescu & Frank, 2024); our assay is an output-side measurement of this kind. For VLMs, several studies intervene and score the answer: replacing tool-generated zoom-in crops with noise barely changes it (Liu et al., 2025), counterfactual edits of named concepts or removable objects test answer- and explanation-level consistency (Ding et al., 2025; Menzat et al., 2026), and FPVG asks whether answers rely on annotated objects (Reich et al., 2023). Others assess steps or evidence without an intervention contrast (Uppaal et al., 2026; Li et al., 2026), find that claimed visual re-examination often does not reflect renewed processing (Shi et al., 2026), or use masked, automatically localized evidence as a training signal (Xu et al., 2026). Contrastive region masking (Chaturvedi et al., 2025) masks dataset-annotated important regions and compares regenerated reasoning traces with unmasked baselines; random masking is a separate ablation. We instead score the fixed step under masks of its model-cited region versus regions cited by same-chain sibling steps as matched controls.

Perturbation, attribution and attention. Occlusion and perturbation reveal which input regions affect outputs (Zeiler & Fergus, 2014; Fong & Vedaldi, 2017; Petsiuk et al., 2018), with distribution-shift caveats (Hooker et al., 2019), and saliency maps and attention weights need not explain predictions (Adebayo et al., 2018; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Cross-modal ablation and Shapley scores quantify reliance on each modality (Frank et al., 2021; Parcalabescu & Frank, 2023), and contrasting outputs with and without regions can improve grounding (Wan et al., 2024). These methods search for or compute important regions; we test regions the model itself asserted. Mediation and activation interventions locate internal components that carry information (Vig et al.,

2020; Meng et al., 2022); we test whether an explicit output claim is coupled to the model's sensitivity under intervention, not how the model uses the region.

6 Discussion, Limitations and Conclusion

What is established. For Qwen3-VL-4B, the region a step cites predicts which of the chain's cited regions that step is more sensitive to, with the ordering-probability interval above 0.5 in each of three independent GQA cohorts. The advantage does not require the generated prefix, persists under a materially different operator and under both changes at once, and appears on a second dataset. Under a fixed-prefix protocol on a held-out cohort, the margin predicts relative content change of the regenerated current step. A checkpoint from a distinct family shows a strong advantage on two image-disjoint GQA cohorts. The aggregate advantage persists across the tested scoring choices, whereas individual comparisons and effect magnitudes depend on how the assay is implemented.

Implications. Localization and intervention sensitivity measure different properties: a well-localized box can lose its comparison (Figure 1), and the advantage does not weaken monotonically as localization worsens. Systems that present boxes as evidence should therefore not report localization alone. The assay needs no region annotations, but it does need step-level region claims and access to step-token log-likelihoods, which interfaces that return only boxes do not provide. A report should state eligibility, the ordering estimate with chain-clustered uncertainty, and the scoring context and operator. Where generation matters, regenerated steps should also be checked against the margin's prediction. Ordering probabilities of different checkpoints are model-conditional estimates on different eligible populations, not a calibrated faithfulness ranking, and a single non-win does not make a box unfaithful.

Why might coupling differ across checkpoints? InternVL3.5-4B cites much smaller regions (median 1.4% of the image against 6.4%), about half of its own-region sensitivities are negative, and enlarging its boxes increases mean signed sensitivity without a detected increase in ordering. InternVL3.5-4B and Qwen3-VL-4B share a Qwen3-initialized language model yet differ markedly in coupling, so shared language-model lineage alone does not explain the difference; they differ in vision encoder and training (Wang et al., 2025; Bai et al., 2025). GLM-4.6V-Flash, whose point estimates are the largest we observe on its own eligible chains, reasons in a thinking trace before emitting boxes, and reported post-training recipes differ across the three checkpoints (cascaded reinforcement learning for InternVL3.5; reinforcement-learning scaling for GLM-V; Wang et al., 2025; V Team, 2025; Z.ai, 2025). The present experiments do not separate vision encoder, post-training, interface, model size or architecture as explanations; no causal explanation is inferred.

What remains unresolved. Three limits bound these claims. *Identification*: an interval above 0.5 is evidence against an additive mean together with its residual condition; S tests an additive mean under a different residual condition, and X narrows proportional interactions under its anchor assumptions. The evidence supports a departure from these nuisance models in the own-versus-sibling direction, not semantic evidence use, and it does not identify the internal computation behind the departure (Appendix C). *Estimand*: teacher forcing holds each step's text fixed, so the statistic is an intervention-induced change in mean per-token log-likelihood under two off-distribution operators that share perturbation artifacts (Hooker et al., 2019). The behavioral link was measured only for current-step regeneration with preceding steps held fixed; whether it extends to free generation of whole reasoning trajectories is open. *Scope*: three primary off-the-shelf checkpoints (plus two exploratory Qwen sizes) under one prompted interface, the largest with about 10 B parameters, two datasets, two contexts and two operators; specialized grounded-reasoning systems, larger models and other interfaces are untested. The test conditions on chains that assign different regions to different steps (182 of 300 images for Qwen; 93 of 300 and 51 of 200 for GLM), and the coupling is partial (39–42% of Qwen's comparisons reverse across the three fixed-text GQA cohorts and the robustness cells) and heterogeneous for reasons the design does not isolate.

Conclusion. The paired assay finds a reproducible own-region ordering advantage in Qwen3-VL-4B, a replicated one in GLM-4.6V-Flash and weak, cohort-variable results in InternVL3.5-4B; under the fixed-prefix protocol, its margin predicts how the content of regenerated steps changes. A cited region is a testable claim about evidence use, not proof of it.

AI USE STATEMENT

Generative AI tools were used for tasks with required disclosure: Claude Code and Codex assisted in developing the conceptual framework and hypotheses, formulating and proving mathematical claims, designing the methodology and experiments, implementing methods and synthetic calibration studies, and interpreting results; ChatGPT assisted with methodological and statistical review and with checking assumptions and counterexamples. They were also used for tasks with recommended disclosure: literature search, figure preparation, code editing, reproducibility checks and packaging of the supplementary material, and drafting and editing the manuscript. We have reviewed all AI-assisted work: AI-assisted mathematical work was reviewed by the authors, and its simulation and closed-form checks are documented in Appendices C, D and J; empirical results were computed from the checkpoints' outputs by scripts whose implementation checks and independent recomputation are documented in Appendix J, and preregistered and exploratory analyses are distinguished throughout. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A specifies the prompt, scoring contexts and both intervention operators; Appendix B the region rule, coordinate handling and cross-dataset de-duplication; Appendices C and D the proof, counterexamples and synthetic calibration; Appendix I the behavioral protocol, its freeze and its full results; Appendix J cohort construction, protocol records, deviations and commit identifiers. The anonymous supplementary material contains the frozen manifests, signed sensitivity matrices and behavioral records; a CPU-only entry point that recomputes the reported statistics from these frozen outputs and regenerates the figures (Figures 1, 3 and 4 also need two source images); the synthetic simulation code of Appendices C and D, including the historical design study and the post-hoc calibration of the joint rule as applied; and the model-facing generation, intervention and scoring code with its requirements (GPU, checkpoints, source images), which reproduced the released values exactly when re-scoring six chains (Appendix J). Source images are not redistributed.

ETHICS STATEMENT

The study uses publicly available GQA questions with Visual Genome images (CC BY 4.0) and A-OKVQA questions with COCO images. No external human participants were recruited; the behavioral endpoint was additionally checked in a blinded manual audit conducted by one author.

REFERENCES

Julius Adebayo et al. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018. arXiv:1810.03292.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Isha Chaturvedi, Anjana Nair, Yushen Li, Adhitya Rajendra Kumar, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, and Vasu Sharma. Peek-a-boo reasoning: Contrastive region masking in MLLMs. *arXiv preprint arXiv:2512.08976*, 2025.

Keqin Chen et al. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*, pp. 4443–4458, 2020.

Sihao Ding, Santosh Vasa, and Aditi Ramadwar. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. *arXiv preprint arXiv:2510.00047*, 2025.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. GRIT: Teaching MLLMs to think with images. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.15879.

Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of EMNLP*, pp. 9847–9857, 2021.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. arXiv:1806.10758.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. arXiv:1902.09506.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL*, pp. 3543–3556, 2019.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of EMNLP*, pp. 787–798, 2014.

Ranjay Krishna, Yuke Zhu, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. arXiv:1602.07332.

Tamera Lanham et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Xuchen Li, Xuzhao Li, Renjie Pi, Shiyu Hu, Jian Zhao, and Jiahui Gao. Beyond accuracy: Evaluating grounded visual evidence in thinking with images. *arXiv preprint arXiv:2601.11633*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. On the faithfulness of visual thinking: Measurement and enhancement. *arXiv preprint arXiv:2510.23482*, 2025.

Kevin Meng et al. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2202.05262.

Bayar Menzat, Maximilian Süss, Ruizhi Wang, Benno Steinegger, Thomas Lukasiewicz, and Oana-Maria Camburu. Counterfactual tests for measuring chain-of-thought faithfulness in visual language models. *arXiv preprint arXiv:2609.06704*, 2026. Accepted to Findings of EMNLP 2026.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of ACL*, pp. 4032–4059, 2023. doi: 10.18653/v1/2023.acl-long.223.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024. doi: 10.18653/v1/2024.acl-long.329.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8779–8788, 2018.

Zhiliang Peng et al. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of BMVC*, 2018. arXiv:1806.07421.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in VQA. In *Findings of EMNLP*, 2023. arXiv:2305.15015.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. arXiv:2403.16999.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? uncovering the illusion of visual re-examination. *arXiv preprint arXiv:2605.15864*, 2026. ICML 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.04388.

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

V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *Proceedings of ECCV*, 2024. arXiv:2403.02325.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2201.11903.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. In *International Conference on Machine Learning*, 2026. arXiv:2509.20912.

Z.ai. GLM-4.6V-Flash: Official model card. Hugging Face model repository, 2025. URL `https://huggingface.co/zai-org/GLM-4.6V-Flash/tree/411bb4d77144a3f03accbf4b780f5acb8b7cde4e`. Pinned revision 411bb4d7; MIT license; accessed September 17, 2026.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of ECCV*, 2014. arXiv:1311.2901.

Junkai Zhang, Yihe Deng, Kai-Wei Chang, and Wei Wang. Thinking with visual grounding. *arXiv preprint arXiv:2606.16122*, 2026.

A PROMPT, SCORING CONTEXTS AND INTERVENTION OPERATORS

Generation prompt. Identical for every checkpoint, dataset and cohort:

```
Answer the question by reasoning step by step.  For EACH
step, cite the single image region you used, as a bounding
box.  Output ONLY a JSON list, one object per step:
[{"step":  "<one short reasoning sentence>", "bbox_2d":
[x1,y1,x2,y2]}, ...]
The final element must state the answer.
Question:  <question>
```

Decoding is greedy with batch size one and at most 384 new tokens (2048 for GLM-4.6V-Flash, the uniform budget used for every externally sourced candidate, whose thinking traces are stripped before parsing). A chain parses if the output contains a JSON list whose every element has a non-empty `step` string; a step's box is valid if it has four numbers that yield a non-degenerate rectangle after conversion to pixels. All three checkpoints emit coordinates normalized to $[0, 1000]$; we convert with $x \cdot W/1000$, $y \cdot H/1000$ and clamp to the image.

Scoring contexts. For step s_t the model's chat template wraps the image and the prompt, and the assistant turn is teacher-forced as a JSON list containing only step text. In the *prefix* context the list is `[{"step": ` s_1`}, ..., {"step": ` s_{t-1}`}, {"step": ` s_t`}`, where the prefix contains the preceding steps that carry valid regions. In the *question-only* context it is `[{"step": ` s_t`}` for every step. The scored tokens are those following the context in the tokenized full string, including the JSON key, quotes and closing object delimiter (but no closing list bracket); $\bar{\ell}$ is their mean log-probability, and within a context the scored token set is identical for x and every $I_r(x)$. For a chain's first region-bearing step the two contexts coincide; this identity was verified exactly (maximum absolute difference 0.0 on all 182 such rows) as a preregistered implementation check. For GLM-4.6V-Flash the list follows the model's own chat head and an empty thinking block. Baseline likelihoods are computed once per step and context, and each intervened image once per region and operator; every rendered intervention was verified against a SHA-256 hash frozen before scoring.

Operator 1: blur. $I_r(x) = x \odot (1 - M_r) + \mathrm{Blur}_\sigma(x) \odot M_r$, with M_r the box mask smoothed by a Gaussian of radius 3 pixels and $\sigma = \mathrm{clip}(0.15\sqrt{\mathrm{area}(r)}, 3, 40)$ pixels. Rejected before any sensitivity was computed: constant or mean fill (a flat patch is a salient low-level artifact); noise fill (adds energy); patch replacement from elsewhere in the image (substitutes new content).

Operator 2: within-box patch permutation. The box is tiled into $n_x \times n_y$ cells with $n_x = \min(w, \max(2, \mathrm{round}(w/c)))$ for $w \geq 2$ (otherwise $n_x = 1$), analogously for n_y and $c = \mathrm{clip}(\mathrm{round}(0.25\sqrt{wh}), 4, 32)$ pixels; cells of identical shape are cyclically shifted along a random order drawn from a seed hashed from the image identifier and the pixel rectangle, so equal-shaped cells change position, and a cell with a unique shape is rotated by $180°$. Pixel-multiset preservation refers to this rearrangement before feathering. The same 3-pixel feathered mask as blur is used. The operator moves only the box's own pixels. It was chosen before any robustness data existed, from image statistics alone on 238 annotation boxes from images already excluded from every cohort:

Candidate operator	layout kept	detail kept	color change
Gaussian blur (operator 1)	0.833	0.000	1.031
patch permutation (operator 2)	**0.159**	**1.302**	**0.076**
pixelation (block means)	0.896	0.062	0.880
box-mean flat fill	0.628	0.000	1.848
surround-mean flat fill	0.528	0.000	1.875
matched Gaussian noise	0.093	7.272	0.568

Medians over boxes. *Layout kept*: correlation of 8×8 gray thumbnails before and after; *detail kept*: ratio of Laplacian variance; *color change*: L1 distance between color histograms. Blur and permutation sit at opposite corners of the layout/detail plane; pixelation is another low-pass operator; the flat fills and noise reproduce the artifacts rejected for operator 1. Limitations of operator 2: cell seams add high-frequency energy ($+30\%$), texture finer than one cell and the color distribution survive, and on tiny boxes it approaches pixel noise.

B Region measurement and cohorts

Region relation rule. For boxes a, b let nDC be the center distance divided by the mean box diagonal, aRat the smaller-to-larger area ratio, IoS the intersection over the smaller area, and IoU the usual overlap. Then SAME if nDC ≤ 0.15 and aRat ≥ 0.70; otherwise DISTINCT if IoS ≥ 0.80 and aRat ≤ 0.70 (a nested sub-region) or if IoU < 0.30; otherwise AMBIGUOUS. The rule was fixed on a design set before any intervention cohort was drawn and is unchanged across all cohorts; the manifest builder used for the robustness and A-OKVQA cohorts reproduces the confirmation-cohort manifest exactly.

Coordinate validity. Each checkpoint's coordinate convention was validated structurally before any sensitivity: raw values in range, boxes inside the frame, exact round-trip, few degenerate boxes, and the declared convention beating algebraically distinct alternatives by a wide margin (0.697 Qwen, 0.289 InternVL, 0.678 GLM). Average localization quality was not a criterion.

Attrition.

	discovery		confirmation		robustness		A-OKVQA
	Qwen	InternVL	Qwen	InternVL	Qwen	InternVL	Qwen
generated	120	120	200	200	300	300	300
valid chains	112	109	185	186	266	281	233
≥ 2 valid regions	104	99	165	165	249	259	203
eligible (≥ 1 DISTINCT pair)	78	57	116	105	182	163	112
SAME pairs	142	281	195	393	272	732	189
DISTINCT pairs	222	168	274	314	458	518	291
median claimed area (%)	6.0	1.1	5.8	1.0	6.4	1.4	11.3

GLM-4.6V-Flash on the robustness cohort: 300 generated, 279 valid, 156 with ≥ 2 valid regions (its final answer step often carries an empty box), 93 eligible, 224 SAME / 199 DISTINCT pairs. Parse failures are format attrition handled by the unchanged parser (mostly malformed closing brackets for Qwen; truncation at the token limit for InternVL).

Cohort construction and cross-dataset de-duplication. GQA cohorts are drawn from validation images with balanced questions, scene graphs and at least two annotated objects; the robustness cohort (seed 20260916) excludes all 520 GQA images used in any earlier stage: the 40-image measurement pilot, two 80-image development sets used to define claim eligibility (the second was later reused to debug the generation wrapper and for the technical qualification of Appendix I), and the 120 discovery and 200 confirmation images. These sets are pairwise disjoint, so their union is exactly 520 (820 with the robustness cohort). The A-OKVQA cohort (seed 20260917) uses the official v1.0 validation split with original COCO val2017 files; of its 1,122 images, 49 appear in the GQA validation frame according to Visual Genome's COCO-identifier metadata, and perceptual hashing (pHash ≤ 6, pixel correlation ≥ 0.90) identified exactly the same 49 pairs; all were excluded before sampling, leaving 1,073, of which 300 were drawn. Three of the 49 had been used earlier in the project.

Measurement pilot. A 40-image pilot on Qwen3-VL-4B, disjoint from all intervention cohorts, compared claimed regions with area-, aspect- and eccentricity-matched non-claimed rectangles (mean difference in J $+0.168$, 95% CI $[0.110, 0.229]$) and found that an own-versus-other mean difference could be carried by a minority of large effects; this motivated the ordering endpoint. The pilot is not used for inference.

C PROPOSITION 1: PROOF, SCOPE AND COUNTEREXAMPLES

Setting. Pooling chains, $\hat{\pi}_{\text{own}} = (2|\mathcal{D}|)^{-1} \sum_{\{t,k\}\in\mathcal{D}} W_{t,k}$ with $W_{t,k} = \mathbb{1}[\Delta_{t,k} > 0] + \mathbb{1}[\Delta_{k,t} > 0]$, where $\mathcal{D}$ is the set of DISTINCT pairs, fixed by box geometry before any sensitivity is computed. In every model–cohort cell the number of comparisons equals twice the number of pairs.

Proof. Under $J_{t,k} = \alpha_t + \beta_k + \varepsilon_{t,k}$, $\Delta_{t,k} = x + U$ and $\Delta_{k,t} = -x + V$ with $x = \beta_t - \beta_k$, $U = \varepsilon_{t,t} - \varepsilon_{t,k}$ and $V = \varepsilon_{k,k} - \varepsilon_{k,t}$. Row-exchangeability, $(\varepsilon_{t,t}, \varepsilon_{t,k}) \overset{d}{=} (\varepsilon_{k,t}, \varepsilon_{k,k})$, implies $\varepsilon_{t,k} - \varepsilon_{t,t} \overset{d}{=} \varepsilon_{k,k} - \varepsilon_{k,t}$, i.e. $V \overset{d}{=} -U$. Therefore $\mathbb{E}[W_{t,k}] = P(U > -x) + P(-U > x) = P(U > -x) + P(U < -x) = 1 - P(U = -x)$, and $\Delta_{k,t} \overset{d}{=} -\Delta_{t,k}$. Because $|\mathcal{D}|$ is fixed, $\mathbb{E}[\hat{\pi}_{\text{own}}] = \frac{1}{2} - \bar{\tau}/2$ with $\bar{\tau}$ the average tie probability. Each equally weighted pair of directional distributions is symmetric about zero, as is their pooled mixture. Hence zero is a population median of Δ; it need not be the unique median, nor must a finite-sample median equal zero. □

Scope of the assumptions. Residual symmetry is not required. Within-chain dependence is unrestricted subject to row-exchangeability; it changes the variance of $\hat{\pi}_{\text{own}}$, which the chain-level bootstrap addresses, not its expectation. Residual scale may vary across regions. The statement is exact in expectation conditional on the fixed effects, not an identity for each sample. Ties count as non-wins and can only lower $\mathbb{E}[\hat{\pi}_{\text{own}}]$; the original full-token analyses contain no observed tie, which is a property of those samples and not of every distribution the null admits. The post-hoc content-only scoring of Appendix E produces one tie, counted as a non-win under the same rule.

Counterexamples. Table 2 lists the expected per-pair ordering probability $\mathbb{E}[W]/2$ for one DISTINCT pair, from 2×10^6 synthetic draws; values agree with closed forms where available (for the step-dependent scale row, $\frac{1}{2}[\Phi(1/(0.5\sqrt{2})) + 1 - \Phi(1/(2\sqrt{2}))] = 0.642$). The example's step effects and residual scales are its parameters, not a threshold that a violation must reach. The script (seed 0) is in the supplementary material, and rerunning it reproduces every entry. *The positive-scale multiplicative-noise row follows a separate sign argument, not Proposition 1.

Table 2: **Counterexamples illustrating the scope of Proposition 1.** Which assumption each structure violates and the resulting expected per-pair own-region ordering probability (2×10^6 synthetic draws; closed forms where available).

Structure	Assumption violated	per-pair π_{own}
additive, i.i.d. residuals	—	0.500
additive, residual scale depends on the region	—	0.500
additive, residual scale depends on the step (sd 0.5 vs 2), $\beta_t \neq \beta_k$	row-exchangeability	0.642
$J_{t,k} = \alpha_t\beta_k + \varepsilon$, $\alpha = (2, 1)$, $\beta = (1, 0)$	additivity	0.581
$J_{t,k} = s_t g_k \eta_{t,k}$, multiplicative noise, $s_t > 0$	sign case*	0.500
$J_{t,k} = s_t g_k \eta_{t,k}$ with step scales of opposite sign	sign condition	0.717
additive plus a prefix term lowering sensitivity to an earlier step's region	separability	0.602

A locally ambiguous sensitivity pattern. Diagonal dominance within a single two-step block is compatible with a signed nuisance model that contains no matching term, for every such block. Write $C = J_{tt} + J_{kk} - J_{tk} - J_{kt}$. If $C = 0$, the block is exactly additive: $J_{tk} = \alpha_t + \beta_k$ with $\alpha = (0, J_{kt} - J_{tt})$ and $\beta = (J_{tt}, J_{tk})$. If $C \neq 0$, then $\det(J + q\mathbf{1}\mathbf{1}^\top) = \det J + q\,C$ vanishes at $q = -\det J/C$, so the block equals a product of rank at most one plus a constant. In both cases the block lies in a signed nuisance class—additive, or proportional with an additive offset—whose factors and interaction coefficient may be negative, and no entry needs to be attributed to matching. For example $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is a constant plus diagonal matching, and is also $\alpha_t + \beta_k + c\,\alpha_t\beta_k$ with $\alpha = (0, 3)$, $\beta = (2, 1)$, $c = -2/3$ and no matching term. This is an ambiguity of the local pattern under a signed nuisance class. It is not a claim that the pooled endpoints are uninformative, and pooling does not by itself resolve the ambiguity: what the endpoints rest on are the distributional conditions stated above. The multiplicative-mean case is the most consequential. Its bias has the

sign of $(\alpha_t - \alpha_k)(\beta_t - \beta_k)$: diagonal dominance arises when the more image-sensitive step of a pair claims the more influential region, without any step-specific matching. A size effect that scales with step sensitivity is an instance of this case, so region size is neutralized only in its additive form. The prefix counterexample is what the question-only scoring context was designed to address empirically; the multiplicative case is what the secondary diagnostic of Appendix D addresses, with the limits stated there.

D SCALE-INTERACTION SENSITIVITY DIAGNOSTIC

Definition. For each DISTINCT pair take $S_{t,k}$ as in Section 2 (the factor $\frac{1}{2}$ does not affect its sign) and let $X_{t,k} = \Delta_{t,k}G_k + \Delta_{k,t}G_t$, where G_t is step t's sensitivity to blurring the whole image (same context and operator family). Under an additive mean $J_{t,r} = \alpha_t + \beta_r + \varepsilon$, $S_{t,k}$ contains only residuals; under a proportional mean $J_{t,r} = a_t b_r + \varepsilon$ with anchors $G_t = a_t b_F$, $X_{t,k} = b_F(a_k U + a_t V)$ contains only residuals, for either sign of a_t. Removing the mean is algebraic; calibrating the sign test is a separate condition. If, conditional on the fixed effects and the pair set, the contrast has the same distribution as its negation, then $P(\cdot > 0) = \frac{1}{2}[1 - P(\cdot = 0)] \leq \frac{1}{2}$, with equality only when ties have probability zero. Independent symmetric row-difference errors are sufficient for this, and they allow the residual scale to differ across steps, but they are not necessary. Symmetry of each row difference separately is not sufficient: for (U, V) uniform on $(-3, -3)$, $(-1, 3)$, $(1, 1)$, $(3, -1)$, both margins are symmetric and $\mathbb{E}[S] = 0$, yet $P(S > 0) = 3/4$, because the joint law is not centrally symmetric. For X the same requirement applies to $a_k U + a_t V$ together with the anchors, which are treated as noise-free and as sharing the factorization $G_t = a_t b_F$. The diagnostic passes if both $P(S > 0)$ and $P(X > 0)$ have chain-clustered 95% intervals above 0.5 (an intersection-union rule over the two null classes). The product form avoids division, so no pair is excluded for a small, zero or negative anchor. It was selected on synthetic data only.

Three synthetic checks, three rules. The false-pass rates below come from three different computations, which we keep apart. (i) The *historical design study* (before the robustness experiments) evaluated an earlier joint rule: S together with the anchor-normalized component $SG_{t,k} = \Delta_{t,k}/G_t + \Delta_{k,t}/G_k$, computed only on pairs whose two anchors are positive, with 1000 multinomial chain resamples per cohort. (ii) A *component-only check* evaluated the signed product X on all pairs, but alone, not jointly with S. (iii) A *post-hoc calibration added in this revision* evaluates the rule applied to the data: S and X on all pairs, signed anchors, and the paper's chain bootstrap. On pairs with two positive anchors $X = G_t G_k \, SG$, so the two components agree in sign there; but the rules analyze different pair sets, their per-chain denominators and bootstrap distributions differ, and for anchors of opposite sign X and the normalized form have opposite signs. On the real Qwen robustness cohort 27.1% of pairs have a non-positive anchor.

(i) Historical design study (synthetic, 170-chain cohorts, 300 cohorts per design). False-pass rate of the normalized positive-anchor joint rule: additive nulls 0.000; proportional nulls with noisy anchors 0.017–0.030 (the ordering endpoint alone: 0.750), with step-dependent residual scale 0.020–0.023. For synthetic no-matching stress structures outside the stated null classes: affine 0.030–0.053; an extra whole-image factor independent of claims 0.090–0.097; conflicting additive and proportional structure, or a whole-image factor anti-correlated with claimed-region influence, 1.000. Power was high under proportional matching and low (0.17–0.20) under additive matching of size 0.2 aligned with claims. Pairs with two positive anchors made up 96–100% of pairs in these designs.

(ii) Component-only check of X. The signed product without exclusion, on its own, gave false-pass rates of 0.010–0.027 under the six proportional nulls of the denominator study (baseline, very noisy anchors, image-insensitive steps, both, half of the steps image-insensitive, and residual scale shrinking with sensitivity); this check says nothing about the joint rule. The scripts and outputs of (i) and (ii) are unchanged in the supplementary material, and rerunning them in this revision reproduced every frozen output exactly (Appendix J).

(iii) Calibration of the rule as applied (post hoc, added in revision). After the real-data results were known, we calibrated the joint rule that the paper applies, under a plan committed before any result (supplementary folder `revision_signed_joint_calibration`). It reuses the frozen data-generating processes of both historical scripts with their published parameters (27 scenarios,

Table 3: **Post-hoc calibration of the joint rule as applied** (added in revision; 300 synthetic cohorts of 170 chains per row, number of cohorts that pass). S and X: each component alone; current joint: S and X on all pairs with signed anchors (the rule applied to the data); historical joint: S and the anchor-normalized component on pairs with two positive anchors, recomputed on the same cohorts and resamples. Last column: share of pairs with a non-positive anchor (real Qwen cohort: 27.1%). Design B rows use the data-generating process of the denominator study.

Scenario	S	X	current joint	historical joint	non-pos. anchor (%)
Additive-mean nulls					
additive	6	0	0	0	0.0
additive, step-dependent scale	6	0	0	0	1.1
additive, anti-aligned	5	299	5	5	0.0
additive (design B)	7	0	0	0	0.0
additive, image-insensitive steps	16	0	0	0	7.7
additive, image-insensitive, anti-aligned	4	219	4	4	7.8
Proportional-mean nulls, noisy anchors					
proportional	300	9	9	7	0.6
proportional, step-dependent scale	300	9	9	11	2.1
proportional, anti-aligned	0	8	0	0	0.1
proportional (design B)	300	4	4	4	0.6
proportional, very noisy anchors	300	12	12	22	6.6
proportional, image-insensitive steps	297	7	7	9	20.0
prop., image-insensitive, very noisy anchors	297	10	10	15	26.4
prop., half the steps image-insensitive	202	7	6	7	37.0
prop., residual scale shrinks with sensitivity	287	6	6	5	1.7
No-matching stress structures (outside the stated null classes)					
affine	300	8	8	12	3.9
conflicting additive and proportional parts	300	300	300	300	0.0
aligned additive and proportional parts	300	0	0	0	0.0
extra whole-image factor, independent	300	26	26	29	1.1
extra whole-image factor, anti-correlated	300	300	300	300	0.0
Matching alternatives (power)					
additive matching, $\delta = 0.5$	300	300	300	300	0.0
additive matching, $\delta = 0.2$	300	47	47	47	0.0
proportional matching, $\delta = 0.5$	300	300	300	300	0.4
proportional matching, $\delta = 0.2$	300	300	300	300	0.5
proportional matching, $\delta = 0.1$	300	250	250	250	0.5
proportional matching, $\delta = 0.2$ (B)	300	300	300	300	0.5
prop. matching, image-insensitive steps, $\delta = 0.2$	300	279	279	291	19.6

300 cohorts of 170 chains each) and computes, on every cohort, the current rule with the frozen analysis code and the historical rule, both with one shared set of 2000 chain resamples (seed 20260826), as for the real data. Because the bootstrap differs from the historical one, the historical rule's rates here differ slightly from those in (i), which stand as published. Table 3 lists every scenario. Under the additive nulls the current rule passed in at most 5 of 300 cohorts, protected by S even where X alone often passes (219 and 299 of 300); under the proportional nulls with noisy anchors in at most 12 of 300 (0.040, Monte Carlo interval $[0.021, 0.069]$, very noisy anchors), including designs with 20–37% of pairs carrying a non-positive anchor, against at most 22 of 300 for the historical rule on the same cohorts. Among the stress structures, the extra independent whole-image factor gave 26 of 300 ($[0.057, 0.124]$), and the conflicting mixture and the anti-correlated whole-image factor gave 300 of 300 under both rules. A count of 0 of 300 has Monte Carlo upper bound 0.012 and is not an error rate of exactly zero. The calibration covers these synthetic designs only; it does not extend the stated null classes, and passing it does not identify step-specific use.

Results. Qwen3-VL-4B, robustness cohort, 458 pairs, none excluded: $P(S > 0) = 0.655$ $[0.597, 0.715]$, $P(X > 0) = 0.616$ $[0.564, 0.669]$ (*diagnostic pass*); 27.1% of pairs have a non-positive anchor, and restricted to pairs with both anchors positive $P(X > 0) = 0.698$ (descriptive). InternVL3.5-4B (descriptive only): $P(S > 0) = 0.581$ $[0.517, 0.642]$, $P(X > 0) = 0.525$ $[0.476, 0.572]$; 61.6% of pairs have a non-positive anchor. The diagnostic is not calibrated against every non-additive null and does not uniquely identify latent step-specific use.

E SECONDARY ESTIMANDS, TEMPORAL DIRECTION AND LOCALIZATION

Symmetric pair score and signed effects. Table 4 is descriptive, on the same chains as Table 1. Earlier cohorts: negative own-region J Qwen 31.7% $[25.5, 37.9]$ and 27.8% $[22.4, 33.1]$; InternVL 46.5% $[40.0, 53.1]$ and 43.1% $[38.1, 47.9]$. Under blur on GQA the shares therefore span 27–32% for Qwen and 43–50% for InternVL; the permutation cells and A-OKVQA differ (Table 4). Negative J identifies the direction of the intervention effect only. Because negative values offset positive ones, a small signed mean does not imply a small effect: on the robustness cohort InternVL's mean own-region J is 0.011 but its mean absolute value is 0.028 (Qwen: 0.228 and 0.264); the absolute values are a descriptive computation added in revision, and raw J is not compared across checkpoints.

Table 4: **Symmetric pair score and signed effects** by model, cohort and cell (descriptive; same chains as Table 1). Median symmetric pair score S with the share of pairs above zero, the share of own-region sensitivities that are intervention-facilitating ($J < 0$), and the mean own-region J.

Cell	median S (share > 0)	own-region $J < 0$	mean own J
Qwen, confirmation, prefix × blur	+0.032 (0.628)	27.8%	+0.280
Qwen, robustness, prefix × blur	+0.035 (0.655)	27.1%	+0.228
Qwen, robustness, question-only × blur	+0.088 (0.712)	32.1%	+0.229
Qwen, robustness, prefix × permutation	+0.022 (0.609)	34.4%	+0.113
Qwen, robustness, question-only × permutation	+0.048 (0.675)	36.6%	+0.101
Qwen, A-OKVQA, prefix × blur	+0.248 (0.797)	9.0%	+0.558
GLM-4.6V-Flash, robustness, prefix × blur	+0.177 (0.910)	23.9%	+0.238
InternVL, robustness, prefix × blur	+0.002 (0.581)	49.8%	+0.011

Temporal direction. Each DISTINCT pair contributes one comparison against an earlier step's region and one against a later step's region, so the two directions have equal counts and the overall P is their average. The confirmation-cohort split was post hoc (with its partition fixed before computation); the robustness-cohort split was predeclared as descriptive. Each split keeps only one direction per pair, so Proposition 1 does not calibrate a 0.5 null within either split; the intervals below describe the splits and are not tests. The direct check on transmission through the prefix is question-only scoring on the full pair set (Section 4.2).

Cohort	Context	competitor from earlier step	competitor from later step
discovery	prefix	0.608 [0.519, 0.692]	0.613 [0.521, 0.698]
confirmation	prefix	0.668 [0.606, 0.729]	0.547 [0.455, 0.634]
robustness	prefix	0.614 [0.564, 0.664]	0.600 [0.542, 0.661]
robustness	question-only	0.631 [0.577, 0.684]	0.590 [0.530, 0.658]

Localization strata (Qwen3-VL-4B). A region is LOW, MEDIUM or HIGH if its best-object IoU is below 0.25, below 0.50, or at least 0.50; a comparison takes the lower stratum of its two regions; strata share chains. Discovery: LOW 0.628 [0.560, 0.711] (19 chains), MEDIUM 0.634 [0.520, 0.731] (18), HIGH 0.592 [0.505, 0.678] (52). Confirmation: LOW 0.602 [0.444, 0.750] (20), MEDIUM 0.694 [0.607, 0.775] (27), HIGH 0.582 [0.509, 0.655] (74). There is no monotonic attenuation as localization worsens; the LOW stratum is small and its confirmation interval includes 0.5, so this is not evidence of independence. The robustness-cohort preregistration listed localization strata as a descriptive analysis, but the frozen analysis code did not implement it; to keep the program closed, no stratum result is reported for that cohort.

Post-hoc diagnostics on the robustness cohort (exploratory, plans committed before computation). *Effect location.* Generic tertiles of the target step's whole-image anchor G_t (cut points 0.190, 0.656): π_{own} 0.552 $[0.485, 0.615]$ (124 chains, 306 comparisons), 0.580 $[0.505, 0.657]$ (121, 305), 0.689 $[0.619, 0.757]$ (135, 305); median Δ $+0.006$, $+0.021$, $+0.142$. $|\Delta|$ quartiles 0.029, 0.091, 0.332 nats; $|\Delta| < 0.01$ in 8%, < 0.05 in 38%, < 0.10 in 52% of comparisons. Tertiles of the target step's own-region J give 0.270, 0.678, 0.875, but this conditions on one term of Δ itself and is not evidence about where a step-specific effect lives. *Step length.* Scored tokens per step: median 17

(deciles 12–28). Spearman of length with $|J[t,t]|$ 0.22 $[0.13, 0.30]$; with the spread of the step's row over its partners 0.10 $[-0.08, 0.28]$ (247 steps with two or more partners); of the pair's length difference with Δ -0.02 $[-0.09, 0.06]$. π_{own} by length-difference tertile: 0.584 $[0.518, 0.649]$, 0.621 $[0.566, 0.684]$, 0.618 $[0.550, 0.688]$; by sign: 0.581, 0.645, 0.623; length-matched pairs ($|\delta L| \leq 2$ tokens, 434 comparisons) 0.613 $[0.559, 0.672]$. *Answer correctness.* Final step contains the normalized GQA gold answer: 47 chains (224 comparisons) π_{own} 0.625 $[0.539, 0.717]$, median $+0.029$; incorrect: 135 chains (692) 0.601 $[0.557, 0.646]$, median $+0.021$; chain accuracy 0.26 under this strict rule. On the behavioral cohort (30 correct, 70 incorrect chains) $\rho_s(\Delta, D)$ is 0.570 $[0.361, 0.729]$ and 0.476 $[0.335, 0.586]$. None of these is a confirmatory result.

Post-hoc geometry and weighting checks (robustness cohort; exploratory, not preregistered). The DISTINCT rule admits pairs that qualify through nesting, where one box lies at least 80% inside the other at an area ratio of at most 0.70, and pairs that qualify through low overlap, IoU < 0.30; 286 of the latter 375 pairs have IoU exactly 0 and their median IoU is 0. Splitting the 458 pairs this way keeps both directions of every pair, so the argument of Proposition 1 applies within a stratum under the added assumption that row-exchangeability holds conditional on the geometry. The geometry was fixed before any sensitivity was computed, but this split was chosen after the primary results were known, so it is exploratory. Counts reconstruct the totals exactly ($83 + 375$ pairs, $166 + 750$ comparisons, no ties). Table 5 gives the strata.

Table 5: **Post-hoc geometry and weighting diagnostics on the Qwen3-VL-4B robustness cohort.** Ordering probability by pair-geometry stratum (number of pairs in parentheses) in each of the four cells; intervals use the frozen chain bootstrap; analyses are exploratory and unadjusted for multiplicity.

Stratum (pairs)	prefix·blur	question·blur	prefix·perm.	question·perm.
nested or near-nested (83)	0.494 [.425,.559]	0.482 [.409,.559]	0.524 [.440,.603]	0.500 [.422,.586]
non-nested, low overlap (375)	0.632 [.585,.680]	0.639 [.602,.677]	0.593 [.548,.639]	0.607 [.558,.655]
similar area, aRat $\geq$ 0.70 (52)	0.683 [.557,.806]	0.625 [.528,.720]	0.731 [.610,.839]	0.644 [.535,.757]
dissimilar area (406)	0.597 [.557,.639]	0.608 [.574,.645]	0.562 [.519,.605]	0.580 [.535,.627]

In nested and near-nested pairs, perturbing the larger region also removes most of the smaller region's content, and the ordering sits near 0.5 in all four cells; this is consistent both with the paired design canceling a geometric effect and with the assay having little sensitivity in that configuration, and the two readings are not separated here. The effect is carried by the non-nested, low-overlap pairs, and it is also present in the small stratum whose two boxes have similar areas, where an additive or proportional size channel has least room to act. A directional split by whether a step's own region is the larger or the smaller of its pair is reported for completeness but is descriptive only: it keeps one direction per pair, so the paired cancellation no longer applies and 0.5 is not a calibrated reference. Its values are 0.696, 0.650, 0.624, 0.604 when the own region is larger and 0.521, 0.571, 0.538, 0.571 when it is smaller, across the four cells. Finally, weighting each chain equally rather than pooling comparisons gives 0.617 $[0.578, 0.657]$, 0.629 $[0.595, 0.664]$, 0.572 $[0.529, 0.616]$ and 0.598 $[0.557, 0.639]$: the ordering advantage is also present when each chain receives equal weight. All intervals here use the frozen chain bootstrap; strata are post hoc and unadjusted for multiplicity.

Content-only scoring (robustness cohort; exploratory, not preregistered). The scored span contains the JSON key, quotes and closing delimiter as well as the step text. Re-running the frozen scorer while retaining per-token log-probabilities reproduces every frozen J (maximum absolute difference 9.2×10^{-7} over 182 chains, no sign disagreement) and allows the same comparisons to be re-scored over the content tokens alone; content is 74% of the scored tokens under prefix scoring and 72% under question-only scoring, and no span had an empty content mask. Table 6 compares the two scorings.

The content-only point estimate is higher in all four cells, and the paired differences, defined as full-token minus content-only on identical comparisons with shared resamples, exclude zero in three of them: -0.045 $[-0.084, -0.004]$, -0.044 $[-0.074, -0.013]$, -0.034 $[-0.074, +0.004]$ and -0.058 $[-0.088, -0.028]$ for the four rows above. The reported effect therefore does not depend on includ-

Table 6: **Content-only re-scoring of the four robustness-cohort cells** (Qwen3-VL-4B; post-hoc). Full-token versus content-only ordering probabilities on the same teacher-forced sequences, and the number of comparisons whose win/non-win indicator changed.

Cell	full-token P	content-only P	win/non-win indicator changes
prefix · blur	0.607 [.568,.647]	0.652 [.613,.690]	205/916
question-only · blur	0.610 [.577,.644]	0.654 [.621,.686]	126/916
prefix · permutation	0.581 [.542,.621]	0.615 [.578,.651]	193/916
question-only · permutation	0.587 [.544,.632]	0.645 [.601,.689]	121/916

ing the structural tokens in the score; the tokens themselves remain in the teacher-forced sequence and in the context, and only the positions that are averaged change. They are not negligible for an individual comparison: writing $\Delta_{\text{full}} = C + S$ with $C = (L_{\text{content}}/L)\,\Delta_{\text{content}}$, the structural part S carries the larger absolute contribution in 24–41% of comparisons and changes the win/non-win indicator in 205, 126, 193 and 121 of the 916 comparisons. This decomposition is computed from the unrounded J matrices; the per-token log-probabilities released alongside them are rounded to five decimals, which is adequate for inspection but cannot resolve a handful of near-zero comparisons, so it is not used for any statistic. Content-only scoring produces one tie in the prefix · blur cell, counted as a non-win under the unchanged > 0 rule. For GLM-4.6V-Flash the same check gives 0.779, 0.784, 0.744 and 0.681 across the four cells, against full-token 0.799, 0.769, 0.711 and 0.676.

GLM-4.6V-Flash 2×2 **(exploratory extension, not preregistered).** Proposed after the frozen cell-A result was known, with the protocol fixed before the new scoring, on the same 93 chains, regions and DISTINCT pairs; cell A reproduces the frozen values (maximum absolute difference 4.0×10^{-7}, no sign disagreement). Full-token ordering probabilities are 0.799 $[0.761, 0.837]$, 0.769 $[0.725, 0.811]$, 0.711 $[0.659, 0.759]$ and 0.676 $[0.626, 0.723]$ for prefix·blur, question-only·blur, prefix·permutation and question-only·permutation. Between-cell differences on identical comparisons with shared resamples, each interval taken on its own and not corrected for multiplicity: under prefix scoring, blur minus permutation $+0.088$ $[+0.032, +0.147]$; under blur, prefix minus question-only $+0.030$ $[-0.012, +0.073]$. Both operators keep this checkpoint's ordering above 0.5 while the operator contrast under prefix scoring excludes zero. The corresponding Qwen contrasts are $+0.026$ $[-0.009, +0.063]$ and -0.003 $[-0.039, +0.033]$; we report each model separately and run no test of a model-by-operator interaction, so no such interaction is claimed. This is not a replication cohort and not a cross-model ranking.

Post-hoc robustness (confirmation cohort, not preregistered). No single chain contributed more than 18 of 548 comparisons; dropping the 10% of chains most favorable to the effect gave $P = 0.579$ $[0.520, 0.639]$; the leave-one-chain-out range of P was $[0.597, 0.618]$.

F INTERVENTION-FOOTPRINT DIAGNOSTIC

Claimed boxes were expanded about their centers, preserving aspect ratio, to $m \in \{2, 4, 8\}$ times their area and clamped only at image borders; σ was recomputed from the expanded box; scoring used the prefix context; comparisons used the original DISTINCT pairs. Area: median share of the image; DISTINCT: share of original pairs still DISTINCT after expansion; own J: mean own-region sensitivity. The levels were fixed before any robustness-cohort data existed, $8\times$ being added because only it brings InternVL's median footprint to Qwen's scale in earlier geometry. Paired differences use identical comparisons and shared resamples. Table 7 reports the diagnostic. Frozen categories: *footprint-limited* requires, at $4\times$ or $8\times$, an interval above 0.5 together with a paired increase over $1\times$; *persistent non-detection* requires neither at every expanded level; otherwise *mixed*. InternVL's outcome is mixed. Expanded boxes are intervention footprints, not model claims.

Table 7: **Intervention-footprint diagnostic for InternVL3.5-4B and Qwen3-VL-4B** (robustness cohort, cell A; preregistered). Each claimed box is expanded about its center to $2\times$, $4\times$ and $8\times$ its area; paired differences from $1\times$ use identical comparisons and shared chain resamples.

Model	Level	Area	DISTINCT	P [95% CI]	$P - P_{1\times}$ [95% CI]	own J
InternVL3.5-4B	$1\times$	1.4%	100%	0.551 [0.505, 0.599]	—	0.011
	$2\times$	2.7%	97.1%	0.531 [0.487, 0.574]	−0.020 [−0.069, +0.026]	0.020
	$4\times$	5.1%	88.2%	0.549 [0.505, 0.591]	−0.002 [−0.051, +0.048]	0.027
	$8\times$	9.3%	81.9%	0.546 [0.499, 0.594]	−0.005 [−0.054, +0.044]	0.036
Qwen3-VL-4B	$1\times$	6.4%	100%	0.607 [0.568, 0.647]	—	0.228
	$2\times$	11.8%	96.3%	0.609 [0.566, 0.654]	+0.002 [−0.028, +0.033]	0.280
	$4\times$	21.6%	91.3%	0.617 [0.582, 0.650]	+0.010 [−0.032, +0.049]	0.305
	$8\times$	35.4%	83.0%	0.587 [0.548, 0.625]	−0.020 [−0.061, +0.018]	0.341

G REPRESENTATIVE SENSITIVITY MATRICES

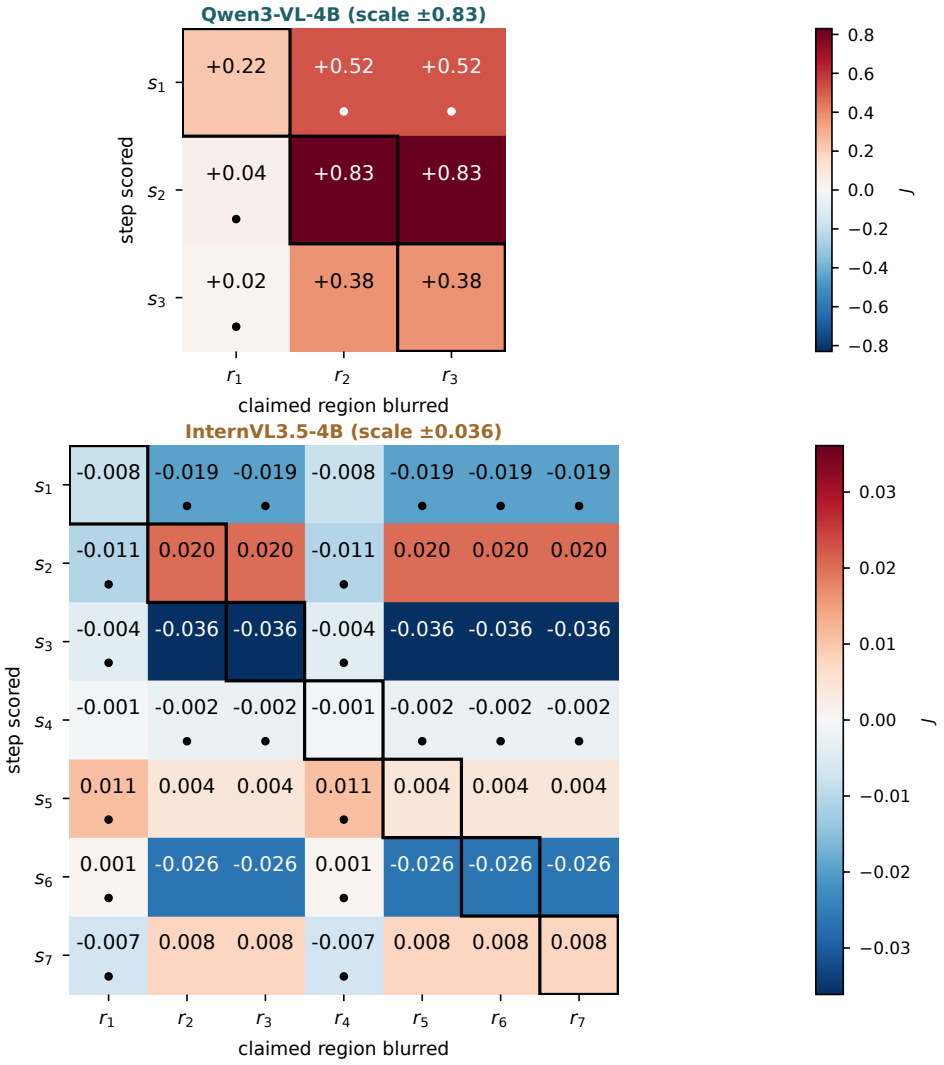


Figure 3: Sensitivity matrices for representative chains. Cells: $J[s_t, r_k]$. Outlined: each step's own region. Dots: the other-step regions that each own-region cell is compared against. Color scales differ between panels because raw J is not comparable across checkpoints.

Figure 3 shows one confirmation-cohort chain per incumbent checkpoint, selected by a rule fixed before any sensitivity was computed: the chain whose own-minus-other median is closest to its model's overall median. Much of the variation runs along columns—some regions affect nearly every step—which is the additive region-level structure the paired comparison is built to cancel. The test reads only whether outlined own-region cells exceed their dotted partners. InternVL's seven-step chain re-uses two boxes, which appear as identical columns and are excluded as SAME pairs. Figure 1 uses the Qwen chain whose frozen identifier begins `8c1ef7e8`; the displayed own-versus-other comparison is s_2 against r_1 ($J[s_2, r_2] = 0.83$ versus $J[s_2, r_1] = 0.04$), with the reverse comparison for s_1 (0.22 versus 0.52) also shown; s_2 is the chain's first step with exactly one DISTINCT partner, the display rule fixed while drafting. The photographs are Visual Genome images 2339240 and 2395089, credited to the Visual Genome dataset (Krishna et al., 2017), distributed under CC BY 4.0 as stated on its official attribution page; overlays and blur are our modifications. Figure 4 shows the corresponding claimed boxes for the representative chains.

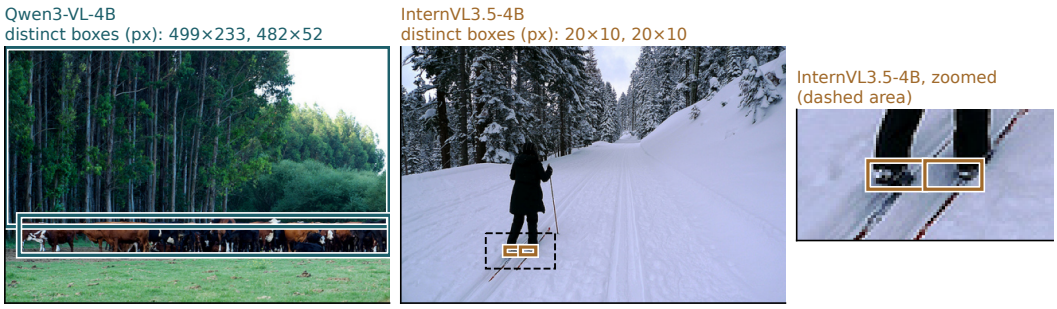


Figure 4: Distinct claimed boxes in the representative confirmation-cohort chains of Figure 3. InternVL's boxes are 20×10 pixels on a 500×333 image, so the blur radius is at its 3-pixel floor. Images: Visual Genome, CC BY 4.0.

H CHECKPOINT ELIGIBILITY

Before the confirmation study, seven additional openly available model families with documented grounding capability were considered under a candidate list fixed in advance. Three were screened on the 120-image discovery cohort and did not meet the fixed interface criteria: GLM-4.1V-9B-Thinking because 49 of its 63 evaluable chains cited a single region for every step (R1 0.783, R5 0.222, R6 14), Ovis2.5-9B and MiniCPM-V-4.5 because their boxes mixed coordinate conventions (3.8% of boxes in the wrong frame; 8.1% degenerate under the declared frame). Four could not be run in the fixed software environment (DeepSeek-VL2-small, ERNIE-4.5-VL-28B-A3B, Kimi-VL-A3B-Thinking, CogVLM-grounding-generalist: their packages target `transformers` 4.x; we did not build a second software environment, so that new checkpoints and the frozen incumbent runs share one stack). For the robustness program a separate lock, written before download, considered three new candidates that had not been screened before: GLM-4.6V-Flash, Keye-VL-1.5-8B and a visually-grounded Gemma3-4B. GLM-4.6V-Flash (revision `411bb4d7`; full hash in Z.ai, 2025) was selected as the only released, runnable, natively box-grounding candidate from a distinct model family; Keye-VL-1.5-8B used a Qwen-derived language model and the Gemma3-4B weights were unreleased. It was screened on the same 120 images with the same gates (R1 parse rate, R2 valid-box rate, R3 multi-step rate, R4 region evaluability, R5 step-specific rate, R6 usable chains, R7a coordinate validity; Table 8). Its predecessor GLM-4.1V-9B-Thinking had failed R1, R5 and R6. GLM was scored only after the primary robustness results were frozen, using a scoring template frozen and checked on already-excluded images beforehand; its result cannot affect any primary decision. These outcomes concern interface and environment, not causal coupling, and support no claim about the corresponding architectures.

InternVL generation-budget sensitivity (post-hoc). All 19 InternVL3.5-4B parse failures on the robustness cohort, 11 of 11 on the discovery cohort and 13 of 14 on the confirmation cohort were generations that ran to the 384-token cap. The same checkpoint snapshot was regenerated on the robustness cohort with 2048 tokens and nothing else changed (protocol frozen before output): the 281

Table 8: **Checkpoint qualification gates for GLM-4.6V-Flash** on the 120-image discovery cohort: the same frozen gates and thresholds applied to every candidate checkpoint.

Gate	R1	R2	R3	R4	R5	R6	R7a
threshold	≥ 0.80	≥ 0.90	≥ 0.50	≥ 0.70	≥ 0.40	≥ 25	pass
GLM-4.6V-Flash	0.975	1.000	0.667	0.936	0.644	47	pass

previously parsed chains were byte-identical, 3 of the 19 truncated chains now parsed (the remaining 16 ran to the 2048 cap), eligibility rose from 163 to 166 chains, and cell A gave $\pi_{\text{own}} = 0.551$ $[0.505, 0.599]$ on the 163 matched chains (paired difference exactly zero) and 0.544 $[0.499, 0.589]$, median $+0.0025$ $[-0.0001, +0.0048]$, on the full 166-chain set. The three newly parsed chains claim much larger regions (median 16% of the image) than the retained ones (1.5%).

Qwen3-VL size comparison (post-hoc: specified after the 4B result, with its protocol committed before any output). Official dense Instruct checkpoints of the family: 2B (2.13 B parameters), 4B (4.44 B), 8B (8.77 B), 32B (33.4 B); MoE 30B-A3B and 235B-A22B. The 32B and MoE checkpoints exceed the 40 GB device's capacity for bfloat16 weights alone; running them would require quantization or offloading, which we did not use, so 2B and 8B were added to the 4B anchor. Same robustness cohort, prompt, 384-token budget, parser, region and DISTINCT rules, cell A, chain bootstrap; protocol, revisions and analysis plan committed before any output. Attrition (parse / fewer than two regions / no DISTINCT pair): 2B 30/80/99, 4B 34/17/67, 8B 21/150/25. Eligible chains 91 / 182 / 104; comparisons 368 / 916 / 424; π_{own} 0.541 $[0.486, 0.594]$, 0.607 $[0.568, 0.647]$, 0.696 $[0.644, 0.745]$; median Δ $+0.005$ $[-0.001, +0.015]$, $+0.022$ $[+0.013, +0.035]$, $+0.067$ $[+0.045, +0.101]$. The 2B interval includes 0.5. The condition that the protocol stated for describing the pattern as size-ordered (monotone point estimates, each larger size's interval above the smaller size's estimate) is met; no trend is fitted, and raw J is not compared across checkpoints.

Checkpoint descriptions. Table 9 lists the three primary checkpoints (the exploratory Qwen3-VL 2B and 8B are described above); parameter counts are the official safetensors totals; thinking status refers to the evaluated interface (frozen prompt, no system prompt).

Table 9: **Primary checkpoint characteristics.** Parameter counts are the official safetensors totals; thinking status refers to the evaluated interface (frozen prompt, no system prompt).

Checkpoint	Family / language model	Params	Thinking trace under the evaluated interface
Qwen3-VL-4B-Instruct	Qwen3-VL, dense; Qwen3 LM	4.44 B	none (Instruct edition)
InternVL3.5-4B	InternVL3.5; InternViT-300M vision, Qwen3-initialized 4.4 B LM	4.73 B	none (thinking requires a system prompt)
GLM-4.6V-Flash	GLM-V; GLM-4 LM	10.29 B	yes by default; stripped before parsing

I BEHAVIORAL VALIDATION OF THE FIXED-TEXT ASSAY

Timeline and freeze. The study is post-hoc relative to the fixed-text results above and *prospective* relative to its own cohort. The protocol, semantic endpoint, parser rules, invalid-output policy, analysis plan, sample-size rule and audit design were frozen before any behavioral generation; the sample size was fixed before the cohort manifest was created; the execution and tabulation code was frozen before formal generation; and the raw generations were frozen before the behavioral endpoint was computed. Every frozen file was re-verified by digest at the start of each execution stage. Exact commit identifiers are listed in Table 11 and, with file-level hashes, in the supplementary provenance manifest. Nothing in this appendix modifies any fixed-text result.

Protocol. Checkpoint: Qwen3-VL-4B only. Chains, claimed regions and DISTINCT pairs come from the frozen pipeline of Section 3; eligibility depends only on those pre-behavior rules. The

predictor $\Delta_{t,k}$ is the frozen cell-A margin (prefix context, blur) computed by the unmodified scorer. For each target step s_t and DISTINCT partner r_k, the current step is regenerated under the same prefix context, greedy decoding, batch size 1, bfloat16, a fixed budget of 64 new tokens and a stop at the closing quote of the step value, with only the image changed: x, $I_{r_t}(x)$ and $I_{r_k}(x)$ (the frozen blur). Baseline and own continuations depend on (t) only and are shared across partners. The budget is deliberately *not* the stored step's length, which would censor the divergence being measured; 64 is about twice the longest step seen in a technical pilot (median 12, 90th percentile 24, maximum 33 tokens).

Endpoint. A closed step is normalized (lowercase, punctuation removed, stopwords removed with *yes*, *no*, *not* and spatial prepositions retained, a context-free plural rule) and mapped to typed units: numbers, spatial and relational terms, other content terms, and a final assertion when one is present. $d(A, B)$ is the Jaccard distance between unit sets; $D = d(\text{base}, \text{own}) - d(\text{base}, \text{other})$; $D = 0$ is a tie. Invalid continuations (unterminated, budget-exhausted, empty, JSON structure as content, refusal, or no units) are handled by a frozen rule: an invalid baseline excludes the comparison before either intervention is parsed; one invalid intervention receives $d = 1$; two give a tie. None of the 1,493 continuations of the formal cohort (315 baseline, 315 own-region, 548 other-region, and 315 repeated baselines as a determinism check) was invalid, so none of these rules fired and the invalid-as-missing sensitivity S2 coincides with the primary analysis. A secondary distance d_2 (normalized edit distance on the normalized string) was frozen for a direction check only. D is computed from a rule-based extractor and is a proxy for content change; it does not preserve relation roles, order or full meaning.

Sample size. A prospective simulation with the historical comparisons-per-chain distribution (182 robustness chains), the historical Δ distribution, a chain random effect, zero inflation $\pi_0 \in \{0.45, 0.60, 0.75\}$ and 400 replicates per cell gave, at the planning scenario $\pi_0 = 0.75$ and a prespecified moderate association $\rho = 0.20$, power 0.995 at the grid floor of 100 eligible chains (weak effect $\rho = 0.10$: 0.632). The frozen rule selected the minimum grid value meeting 80% power, $N = 100$ eligible chains, with a maximum source-image budget $M = \lceil 1.5 \times 100/(182/300) \rceil = 248$, where $182/300 = 0.607$ is the historical eligible-chain rate on the robustness cohort (coincidentally equal to that cohort's ordering probability, an unrelated quantity), and a NO-GO if M were exhausted first. Null calibration in the decisive cell gave 9/400 false detections (Clopper–Pearson $[0.010, 0.042]$). A structure-preserving sensitivity that resamples whole historical Δ vectors per chain gave power 0.990 at $N = 100$. The observed tie mass (0.473) and association (0.504) both differ from the planning values; the design was sized for a smaller effect than was observed.

Technical qualification. Before the formal cohort, the full pipeline ran on the second 80-image development set of Appendix B, already excluded from every cohort (44 eligible chains, 131 target steps, 587 continuations): baseline continuations were token-identical when regenerated before and after the interventions (131/131); chain generation was identical when repeated (80/80); semantic extraction failed on 0/131 parser-closed baselines; intervention format failure was 0/325; the rendered intervention, the processor pixel tensor at the generate call and the prefix token ids were verified for every continuation; a second run in a separate process, in reverse order with interventions before the baseline, reproduced every output byte for byte; no thinking segment appeared. One earlier launch aborted inside the manifest builder before any continuation was generated (a missing metadata field), and was classified as a pre-measurement runtime abort; its artifacts are preserved.

Cohort and attrition. Candidate pool: 8826 GQA validation images after excluding 1387 ids (every image of every earlier stage, the earlier behavioral attempts, the technical set, and A-OKVQA overlaps by COCO identity and perceptual hash); order fixed by seed 20260924; the first 248 were the budget.

Table 10: **Behavioral-validation primary, sensitivity and secondary results** (Qwen3-VL-4B, held-out cohort; 548 comparisons, 100 chains). Intervals are 95% percentile intervals from 2000 chain resamples; S1–S4 are the prespecified sensitivity analyses; the last row is exploratory.

	Quantity	n	Estimate	95% CI
primary	Spearman $\rho_s(\Delta, D)$, comparison-weighted	548	$+0.504$	$[+0.402, +0.600]$
S1	baseline-aligned subset (regenerated baseline reproduces the stored step's units)	369	$+0.602$	$[+0.463, +0.713]$
S2	intervention-invalid as missing (identical to primary: no invalid output)	548	$+0.504$	$[+0.402, +0.600]$
S3	ten most favorable chains removed	458	$+0.390$	$[+0.276, +0.497]$
S4	equal chain weighting (descriptive)	548	$+0.548$	$[+0.437, +0.650]$
secondary	regression slope of D on Δ	548	$+0.321$	$[+0.246, +0.422]$
secondary	$\rho_s(\Delta, D_2)$, edit-distance endpoint	548	$+0.521$	$[+0.425, +0.606]$
secondary	mean D for $\Delta > 0$	305	$+0.204$	$[+0.157, +0.252]$
secondary	mean D for $\Delta \leq 0$	243	-0.042	$[-0.097, +0.009]$
secondary	$\theta = P(D > 0) - P(D < 0)$	548	$+0.177$	$[+0.098, +0.264]$
exploratory	$\rho_s(\Delta_{\text{content}}, D)$, content-token margin	548	$+0.550$	$[+0.452, +0.639]$

Stage	Count
images processed (stopped at the 100th eligible chain)	159
chains excluded: parse or generation failure	19
chains excluded: fewer than 2 valid regions	6
chains excluded: no DISTINCT pair	34
eligible chains	100
target steps / DISTINCT pairs / ordered comparisons	315 / 274 / 548
comparisons lacking J; baseline-invalid; intervention-invalid	0 / 0 / 0
analysis population; ties in D	548; 259 (0.473)

Results. Table 10 reports the prespecified quantities; intervals are 95% percentile intervals from 2000 chain resamples (seed 20260922).

Directional summary: $P(D > 0) = 0.352$, $P(D = 0) = 0.473$, $P(D < 0) = 0.175$; $P(d_{\text{own}} > 0) = 0.544$, $P(d_{\text{other}} > 0) = 0.394$. No 0.5 reference is used for $P(D > 0)$ and no conditional $P(D > 0 \mid D \neq 0)$ is reported. By Δ quartile (cut points -0.048, $+0.008$, $+0.154$; 137 comparisons each): mean D -0.082 $[-0.175, -0.003]$, $+0.006$ $[-0.025, +0.044]$, $+0.087$ $[+0.028, +0.144]$, $+0.368$ $[+0.298, +0.437]$; $P(D > 0)$ 0.182, 0.124, 0.314, 0.788. Chain influence: the largest chain contributes 3.3% of comparisons; the leave-one-chain-out range of ρ_s is $[+0.488, +0.522]$. Baseline continuation versus the stored original step (diagnostic, not an inclusion criterion): EXACT 359 (65.5%), SEMANTIC (same units, different string) 10 (1.8%), APPROXIMATE ($0 < d \leq 0.5$) 112 (20.4%), DIFFERENT 67 (12.2%). Under the frozen claim-strength rule all four criteria are met (primary interval above zero; S2 positive and above zero; no chain above 10% and S3 above zero; d_2 direction positive), noting that S2 is met vacuously.

Ties and the fixed-text margin on this cohort (descriptive, added in revision). A tie in D means equal change, not necessarily no change. Of the 259 ties, 219 (40.0% of 548 comparisons) have both distances zero and 40 (7.3%) have equal positive distances; in one such case the baseline "Identify the computer to the right of the speaker in the middle of the picture" became "Identify the computer to the right of the speaker" under both blurs. On this cohort the fixed-text ordering probability, computed with the frozen chain bootstrap of the fixed-text analyses (2000 resamples, seed 20260826, chains ordered by identifier), is $305/548 = 0.557$ $[0.498, 0.616]$, median Δ $+0.008$ $[-0.001, +0.023]$. Neither computation is a prespecified quantity of this study, neither changes D or any reported statistic, and the fixed-text result here was not a prespecified confirmatory replication of the fixed-text endpoint. The 0.61 of Section 4.1 and the 39–42% of comparisons that favor the other region in the main text refer to the three earlier GQA cohorts and the robustness cells; on this cohort 243 of 548 comparisons (44.3%) do. The behavioral and confirmation cohorts happen to contain the same numbers of DISTINCT pairs (274) and comparisons (548) but share no image and no chain (none of the 159 processed behavioral images is among the 200 confirmation images), with 305 and 333 fixed-text wins respectively.

Post-hoc component / incremental-validity check (added in revision). On the 548 comparisons with `in_analysis=true`, with the frozen procedure of this study (comparison-weighted Spearman correlations; chains ordered by identifier; 2000 chain resamples; seed 20260922; 95% percentile intervals; the full statistic recomputed inside every resample): own-region sensitivity alone gives $\rho_s(J_{t,t}, D) = 0.282$ $[0.164, 0.397]$ and sibling-region sensitivity alone $\rho_s(J_{t,k}, D) = -0.232$ $[-0.329, -0.131]$. Partial Spearman correlations, defined by entering the Spearman rank correlations into the standard one-control partial-correlation formula, give $\rho_s(\Delta, D \mid J_{t,t}) = 0.444$ $[0.343, 0.537]$ and $\rho_s(J_{t,t}, D \mid \Delta) = -0.092$ $[-0.189, 0.006]$. All four quantities are reported; none was preregistered or prespecified, none changes a previously reported statistic, and the check is descriptive of this cohort only.

Blinded manual audit. 120 comparisons, 30 per Δ quartile, drawn with seed 20260923 from the analysis population after the generations existed; the worksheet contains only the baseline step and the two intervened steps as P/Q in an order randomized per item, with condition identity, Δ, D, d, boxes, step indices, chain and image identifiers withheld to a separate key; a whitelist assertion checks the worksheet before release. One author judged all 120 items without the key (responses P / Q / approximately equal / cannot judge; 0 cannot-judge). The reviewer wrote the protocol's gloss "approximately equal" where the scorer expects the token EQUAL; a copy with that single mapping applied was scored, the frozen scorer unmodified and the original file preserved. Agreement between the rule-based direction and the blinded judgment is 103/120 $= 0.858$ (Cohen's $\kappa = 0.771$ as a chance-corrected agreement, not an inter-rater reliability, since there was one reviewer); on the 54 items where both give a strict direction, 49 agree (0.907).

human \ rule	OWN	OTHER	TIE
OWN	32	2	1
OTHER	3	17	0
TIE (approximately equal)	6	5	54

Reproducibility artifacts. The frozen cohort manifest, comparison-level data, analysis outputs, power simulations, audit results and their full SHA-256 hashes are provided in the anonymous supplementary package; the complete file-to-hash mapping is listed in its provenance manifest.

J REPRODUCIBILITY RECORDS

Cohorts and inference. Discovery: 120 GQA images, seed 20260826. Confirmation: 200, seed 20260903. Robustness: 300, seed 20260916. A-OKVQA: 300, seed 20260917. The image-id list of each cohort was hashed in its cohort lock (supplementary provenance manifest). Bootstrap: 2000 resamples of chains, seed 20260826, with identical resample indices across the four robustness-cohort cells. The robustness and A-OKVQA analyses order chains by identifier; the discovery and confirmation analyses resample them in the order they appear in the frozen score file.

Freeze chronology. Discovery endpoints were fixed before any sensitivity value was computed. For confirmation, the cohort, the preregistered endpoints and the pre-sensitivity manifests were frozen before scoring. The robustness program additionally fixed the second intervention operator, the box-expansion levels, the checkpoint-selection rule, the cohort manifests and the analysis code before execution, and model generations were frozen before the primary analysis. The supplementary checkpoint and the later replication, size-comparison and budget experiments followed the same pattern: eligibility, protocols and analysis plans were committed before their outputs existed. The later audit of Proposition 1 and the confirmation-cohort temporal split postdated the corresponding confirmatory decisions and did not alter them. Exact commit identifiers are listed in Table 11; file-level SHA-256 hashes are in the supplementary provenance manifest.

Protocol deviations and later extensions. Table 12 separates the one preregistered analysis that was not carried out, a label-format normalization and a runtime abort from analyses added after the confirmatory results, a later theory audit and the checks added in this revision. Only the first three concern a written protocol; the others are labeled post hoc wherever they are reported.

Checks added in this revision. Without any model inference, the following were recomputed from the released frozen outputs: every entry of Table 1, the robustness cells and the later experiments, compared with the shipped values by a strict comparator that checks key sets, types, counts, array lengths and every interval endpoint (it was first shown to flag injected missing keys, altered endpoints, shortened arrays and changed counts); the behavioral analysis and the blinded-audit scoring; the counterexamples of Appendix C and the four simulation studies behind Appendix D, whose reruns reproduced the frozen outputs exactly; and the descriptive quantities marked as added in revision. The earlier package's comparator checked only keys present in both files and skipped lists; the strict comparison found no difference, so no previously reported value changes. The synthetic calibration of the joint scale-interaction rule as applied (Appendix D, part iii) is new: its plan, code and unit tests were committed before its first run, and it is labeled post hoc. As a software check of the model-facing code, the frozen scoring script, run from the released package on one A100 GPU, re-scored the first three robustness-cohort chains of Qwen3-VL-4B in all eight cells and of InternVL3.5-4B in its five (cell A, the full-frame anchor and three box-expansion levels): all 424 re-scored values, including the unperturbed baselines, equaled the released values exactly, and every re-rendered intervened image matched its frozen hash. This checks the scoring path on six chains; no chain was regenerated and no cohort was re-scored in full.

Decision summary. H1 (question-only $\times$ blur), H2 (prefix $\times$ permutation) and H3 (question-only $\times$ permutation) each passed; the joint intersection-union claim passed; Holm correction rejected each individual null; H4 (A-OKVQA) passed. Cell A's replication category also requires both ordering and median intervals to exclude their nulls. The paper-wide sensitivity check treats the GQA conjunction (H1–H3) and H4 as two directional claim families. At family-wise one-sided $\alpha = 0.025$, Bonferroni allocates 0.0125 to each: the reported central intervals are $[q_{0.0125}(\hat{\pi}^*_{\text{own}}), q_{0.9875}(\hat{\pi}^*_{\text{own}})]$, whose lower endpoints are nominal 98.75% one-sided bounds. All H1–H3 lower endpoints, and H4's, exceed 0.5; cell A also does descriptively. These are not simultaneous 97.5% intervals for every table entry. Holm over H1–H3 uses the recorded bootstrap tail proportions $(1 + \#\{\hat{\pi}^*_{\text{own}} \leq 0.5\})/2001$ at one-sided 0.025; all three are $1/2001$. These are approximate percentile-bootstrap procedures, not exact finite-sample guarantees. An independent re-implementation of the comparison set, the win count, the ordering probability, the median and the chain bootstrap, written without reusing the analysis code, reproduces the win count, ordering probability, median and both interval endpoints of the eleven model–cohort–cell entries it covers: the three Qwen GQA cohorts, the four robustness cells, H4, GLM cell A and InternVL in all three cohorts. Its scope is those entries; the footprint, scale-diagnostic, temporal and localization tables are not part of that check. The post-hoc analyses of Appendix E are reproduced instead by the scripts released with them, and the re-scored J matrices agree with the frozen ones to 9.2×10^{-7} (Qwen) and 4.0×10^{-7} (GLM) with no change of win indicator in any of the 916 and 398 comparisons.

Environment. `transformers` 5.3.0, PyTorch 2.8.0 (CUDA 12.8), bfloat16. Discovery and confirmation: NVIDIA A100-SXM4 40 GB; robustness and A-OKVQA cohorts, including GLM: one NVIDIA A100-PCIE 40 GB, so every analysis unit ran on a single device. Raw J magnitudes are compared only within a run.

Table 11: **Freeze and provenance chronology.** Short git commit identifiers of the project repository; the supplementary provenance manifest gives full identifiers and the SHA-256 of every frozen file.

Stage	Frozen item	Timing relative to analysis	Commit
Discovery	scoring units and endpoints	before any sensitivity value	`46d23e9`
Confirmation	preregistration and cohort lock	before scoring	`7b738b9`
Confirmation	generations and pre-sensitivity manifests	before sensitivity computation	`6a27eb4`
Confirmation	result freeze	after frozen scoring	`a060ac9`
Robustness design	evidence-gap audit; synthetic scale-control audit	before any robustness experiment	`2d589ac`, `3d4eb9e`
Robustness design	second operator, box-expansion levels, checkpoint-selection lock	before robustness execution	`b41bd75`
Robustness preregistration	hypotheses, both cohort locks, analysis code	before execution	`0e2709b`
Robustness execution	execution code	before generation	`fd583ef`
Robustness execution	generations and pre-sensitivity manifests	before primary analysis	`ceca6bd`
Robustness result	primary results (all preregistered outcomes)	after frozen execution	`cadc9bc`
Supplementary checkpoint	GLM-4.6V-Flash eligibility and manifest	before its scoring	`fcb69cc`
Supplementary checkpoint	GLM-4.6V-Flash robustness-cohort result	after scoring	`083b3dc`
Later audit	Proposition 1 proof audit	postdates the confirmatory decision	`9352abd`
Later descriptive split	confirmation-cohort temporal split, pre-specified then run	postdates the confirmatory decision	`2bac7ff`, `812c72b`
Behavioral validation	protocol, endpoint, parser, invalid-output policy, analysis plan, audit design	before any behavioral generation	`baafb92`, `23fe620`
Behavioral validation	technical qualification (no formal data)	before the sample-size decision	`abd5cb1`
Behavioral validation	sample-size rule	before the cohort manifest	`f7d75c6`
Behavioral validation	cohort manifest	before formal execution	`02303a9`
Behavioral validation	execution and tabulation code	before any formal output	`cfe4940`, `d19a876`, `2ba07f2`
Behavioral validation	raw generations	before endpoint computation	`89cbbfb`
Behavioral validation	analysis; blinded audit	after the raw-data freeze	`be9cb7c`, `4f00bba`
Later experiments	protocols, run configurations and analysis plans (GLM replication; InternVL budget; Qwen size)	before any output	`363db7c`
Later experiments	result freezes (GLM replication; InternVL budget; Qwen size)	after frozen execution	`6878c6c`, `045d993`, `ac43f16`

Table 12: **Protocol deviations and later extensions.** None of these items changed a preregistered decision.

Type	What happened	Reported in
Preregistered analysis not implemented	Robustness-cohort localization strata were listed as descriptive, but the frozen analysis code did not implement them; no stratum result is reported for that cohort	App. E
Label normalization	The blinded reviewer wrote the protocol's gloss "approximately equal" where the scorer expects EQUAL; this single mapping was applied to a copy, the original file was kept and the scorer was not modified	App. I
Runtime abort	One early launch of the behavioral technical qualification stopped inside the manifest builder before any continuation was generated; its artifacts are preserved	App. I
Later extensions (new model runs)	GLM-4.6V-Flash on the robustness cohort and its confirmation-cohort replication; the GLM exploratory 2×2; Qwen3-VL 2B and 8B; InternVL at a 2048-token budget. Each protocol was committed before its own output, after the confirmatory results were known	Sec. 4.5, 4.6; App. E, H
Post-hoc analyses of frozen data	Content-only re-scoring; geometry and weighting strata; the confirmation-cohort temporal split; effect location, step length and answer correctness	App. E
Later theory audit	The audit of Proposition 1 restricted its scope and narrowed a corollary after the confirmatory decision	App. C
Checks added in this revision	Reruns of the Appendix C and D simulations; a post-hoc calibration of the joint scale-interaction rule as applied; strict re-comparison of reproduced statistics; a six-chain re-scoring check of the released scoring code; tie decomposition, the fixed-text ordering on the behavioral cohort and absolute own-region J (all descriptive)	App. D, E, I, J