REGIONTRACE: MEASURING HOW IMAGE REGIONS SHAPE MULTIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable multimodal assistance requires understanding whether answers are grounded in relevant visual evidence. We present *RegionTrace*, a framework for tracking how image regions support the same answer throughout generation. At each stage, it holds the text generated so far fixed and measures how restoring an image region changes the answer score. Keeping the answer and regions fixed reveals where attribute errors find support, how support evolves during explanation, and whether visual revisits restore it. Experiments across three models and four benchmarks reveal that models often borrow attributes from the wrong object: in color-misbinding errors, the color-bearing distractor outweighs the queried region in 68.0–97.9% of cases. Across 7,206 explanations, a complete-image probe matches the eventual answer in 86–92% of traces before explanation begins, while regional support weakens after the first sentence in all six model–task settings. Explicitly revisiting the image does not reliably restore this support or improve accuracy. RegionTrace connects spatial attribution with conditional answer-support trajectories throughout multimodal generation. Anonymous project homepage: `https://regiontrace-anon.github.io/regiontrace`.

1 INTRODUCTION

Multimodal assistants can produce answers inconsistent with image content (Li et al., 2023), and their explanations can rely more on text than on the image (Parcalabescu & Frank, 2025). Assessing their reliability therefore requires tracing the visual evidence behind an answer. Different image regions can support or suppress the same answer. In Figure 1, a model is asked whether a suitcase is in front of a couch. The suitcase and couch support the answer *Yes*, while the nearby chair pushes against it. As an explanation grows, the model's earlier words also guide what comes next; studies of multimodal hallucination show how this context can sustain errors and increase reliance on text (Zhong et al., 2024; Zheng et al., 2025). Understanding an answer therefore means tracing both where its visual support comes from and how that support develops during generation.

Existing methods provide several ways to examine this evidence. Image perturbation methods measure how predictions respond when visual content is hidden or altered (Fong & Vedaldi, 2017; Petsiuk et al., 2018; Covert et al., 2021), while multimodal attribution quantifies the contributions of images and text to predictions and explanations (Parcalabescu & Frank, 2023; 2025). For generated text, Token Activation Map links individual words to visual cues while accounting for earlier tokens (Li et al., 2025), and EAGLE identifies compact image regions that sustain selected outputs (Chen et al., 2026). These advances motivate a common question: how does support for the same answer change as generation unfolds? Answering it requires keeping the candidate fixed as the explanation changes and examining each region alongside the others, since the same content can strengthen an answer in one visual context and weaken it in another.

We introduce *RegionTrace*, a toolbox for measuring how image regions shape an answer throughout multimodal generation. Its central advance connects regional attribution to the changing support for the same candidate answer, including before that answer is written. At each stage, we keep the question and generated text fixed, hide selected regions, and measure the score change when a region or group is restored. We call this signed change *Region Effect*: positive values support the answer, and negative values suppress it. Across visual contexts, these effects reveal how regions work together; across stages, they trace how that influence evolves. Regions may be defined using dataset annotations, user-specified masks, or automatic segmentation, while the target score may correspond

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

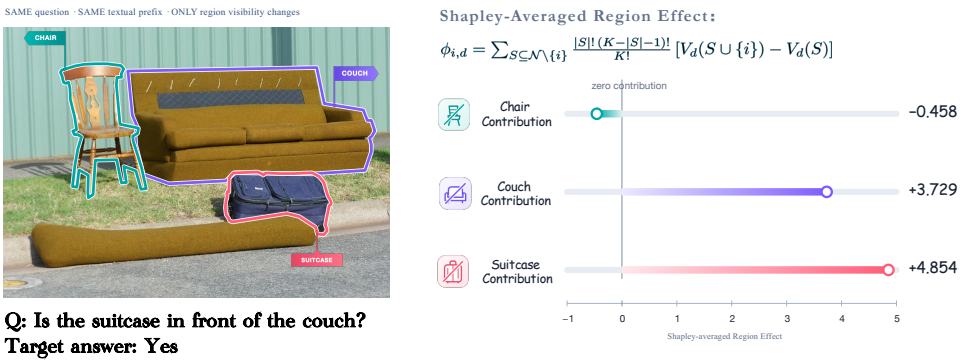


Figure 1: **RegionTrace in one example.** For a fixed question, preceding text, and target answer, RegionTrace changes only which selected image regions are visible and rescores the same target. Here, the general score V_τ is instantiated as the Yes–No log-score margin V_{YN} (that is, $\tau = \mathrm{YN}$), and exact Shapley averaging summarizes each region's signed contribution across visual contexts. Positive values support the fixed answer *Yes*, whereas negative values suppress it. The example uses InternVL3-8B on COCO 2017 image 81061 (Lin et al., 2014) with processor-mean masking.

to either a generated token or a fixed answer candidate. For small region sets, we enumerate all visibility subsets and optionally summarize their effects using exact Shapley averaging (Shapley, 1953).

We demonstrate these capabilities in three studies that follow visual evidence from an initial answer preference through explanation and a deliberate second look. Study I locates the visual source of attribute errors in 886 controlled scenes from PTR (Hong et al., 2021) and 500 photographs from PACO (Ramanathan et al., 2023) (Section 3). Wrong color answers often draw stronger support from the region carrying that color than from the region named in the question: the model finds a real property, but attaches it to the wrong place. Study II follows the eventual answer through 7,206 explanations on MM-GCoT (Wu et al., 2025) and VCR (Zellers et al., 2019) (Section 4). The image often favors that answer before reasoning begins; as the explanation grows, the written text increasingly sustains the same preference while the selected regions provide less additional support, even when the answer is wrong. Study III tests whether an explicit second look restores regional influence once the generated text already sustains the answer, using Look-Back (Yang et al., 2026) and equal-length matched continuations constructed from ordinary reasoning. (Section 5). Neither tested variant consistently restores this influence or yields a reliable accuracy gain across the two tasks. The three studies cover spatial attribute misbinding, answer-support trajectories during explanation, and visual re-examination.

2 REGIONTRACE

RegionTrace tracks how the same image regions support a fixed answer as generation proceeds. At each stage, we hold the existing text fixed and measure how restoring a region changes support for that answer in a given visual context. Comparing these effects across stages reveals how visual support changes as the text grows.

2.1 FOLLOWING THE SAME TARGET THROUGH GENERATION

RegionTrace records a response generated from the complete image. At selected stages, it saves the text generated so far, termed the *prefix*. For example, after the first explanation sentence, that sentence is held fixed while different image regions are restored. We also keep the question and scored output fixed across these image changes.

Formally, let $I \in \mathbb{R}^{H \times W \times 3}$ denote an image, q the question, and p_θ the evaluated multimodal model with fixed parameters. Given the complete image, the model generates a token sequence

$$\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_T), \qquad \hat{y}_t \sim p_\theta(\cdot \mid q, I, \hat{\mathbf{y}}_{<t}). \tag{1}$$

Here, $\hat{y}_t$ is the token at position t, and $\hat{\mathbf{y}}_{<t}$ is the text generated before it. When image regions are changed, we keep this preceding text and the target token fixed and recompute its score. We can also

score the same candidate answer at different stages of generation, including before the model writes that answer.

2.2 DEFINING THE REGIONS

Let $\mathcal{R} = \{r_1, \ldots, r_K\}$ be the regions to be tested, indexed by $\mathcal{N} = \{1, \ldots, K\}$. Each region r_i is represented by a binary mask $m_i \in \{0,1\}^{H \times W}$. The regions can come from dataset boxes, user drawings, external segmentations, or an automatic search from text. In the automatic procedure, a phrase ℓ_i is taken from the question or from a description of a reasoning step. Grounding DINO (Liu et al., 2024) uses this phrase to locate a bounding box. The Segment Anything Model (SAM) (Kirillov et al., 2023) then converts the box into a mask that follows the visible boundary. This combination of text-guided detection and box-prompted segmentation follows the Grounded SAM pipeline (Ren et al., 2024). When a dataset provides a search area, the box and final mask are kept inside it.

We process masks from smallest to largest, removing pixels already assigned to earlier regions. Let π denote this order and $\tilde{m}_i$ the raw masks:

$$m_{\pi_j}(u) = \tilde{m}_{\pi_j}(u) \prod_{h<j} \big(1 - \tilde{m}_{\pi_h}(u)\big), \tag{2}$$

which guarantees $\sum_i m_i(u) \le 1$ at every pixel u. We save the dataset search area, detected box, raw and non-overlapping masks, prompt, and mask area so that each region can be inspected.

Once the masks are fixed, a set $S \subseteq \mathcal{N}$ specifies which regions remain visible. The regions in S retain their observed content, the other selected regions are replaced by a baseline, and image area outside the selected regions remains unchanged. Thus $S = \varnothing$ replaces every selected region while retaining the rest of the image.

2.3 MODIFYING THE IMAGE

The next step constructs one modified image for every region set S. Let I^0 denote the fixed baseline image used in a given experimental configuration. In our experiments, I^0 is a constant RGB image, with the exact baseline fixed within each attribution run. Hidden pixels are replaced by the corresponding pixels of I^0, while all other pixels remain unchanged. For a visible region set S, the mask k_S indicates which pixels retain their original values:

$$k_S(u) = ind\left[u \notin \bigcup_i r_i\right] + \sum_{i \in S} m_i(u). \tag{3}$$

Since the region masks do not overlap, $k_S(u)$ is either 1 (keep the original pixel) or 0 (use the baseline). The modified image is

$$I_S = k_S \odot I + (1 - k_S) \odot I^0. \tag{4}$$

The modified image I_S then passes through the model's standard image preprocessing and visual encoder. The main analyses use the fixed constant baseline specified for each experimental configuration. We also repeated selected experiments using the average color of the surrounding pixels and, for PTR, a fill rebuilt inward from the mask boundary (Telea, 2004); Appendix D describes these alternatives and reports the complete results.

2.4 MEASURING THE EFFECT OF A REGION

For every modified image I_S, RegionTrace evaluates the same scalar score while keeping its textual context and output target fixed. We denote this score by $V_\tau(S)$, where τ specifies the target being evaluated and the scoring rule. Depending on the question, $V_\tau(S)$ can be the log-probability of a generated token, the candidate-normalized probability of a fixed answer, whether that answer is preferred, or a binary log-score margin. Study I uses a Yes-versus-No log-score margin, while Studies II and III use the candidate-normalized probability of the fixed eventual answer at saved points in an explanation. The token-level readout used only for the supplementary Look-Back audit is defined in Appendix C.4.

Region Effect compares two images that differ only in the content being tested. Let C be one region or a region group, and let S contain the other selected regions already visible, with $C \cap S = \varnothing$:

$$E_{C,\tau}(S) = V_\tau(S \cup C) - V_\tau(S). \tag{5}$$

For example, in Figure 1, let C contain the suitcase and S the couch and chair. Equation 5 compares the score for the fixed answer *Yes* before and after the suitcase is restored. A positive value means that the restored content supports the fixed target, whereas a negative value means that it suppresses that target. Changing S changes which other objects are visible; including both the suitcase and couch in C measures their joint effect.

For a small number of regions, we evaluate every subset and store $V_\tau(S)$ for all $S \subseteq \mathcal{N}$. Exact enumeration requires 2^K evaluations of the fixed score; when the target is a recorded trace, one teacher-forced model pass scores all of its tokens for a given region set.

2.5 Averaging a region across visual contexts

A region's effect can depend on what else is visible: its contribution may weaken or change sign as the visible context changes. Relying only on an empty or full visual context would therefore make the attribution depend on one extreme comparison (Appendix F.2). RegionTrace averages the Region Effect of region i (Equation 5) over all subsets S of the other regions, using the Shapley weights (Shapley, 1953):

$$\phi_{i,\tau} = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!(K - |S| - 1)!}{K!} E_{\{i\},\tau}(S). \tag{6}$$

The resulting Shapley-averaged Region Effect, $\phi_{i,\tau}$, summarizes the region's support for or suppression of the fixed target across contexts. The Shapley weights give equal total weight to each context size and equal weight to subsets of the same size. The contributions sum to the effect of restoring all selected regions: $\sum_i \phi_{i,\tau} = V_\tau(\mathcal{N}) - V_\tau(\varnothing)$. Our project website explains this averaging with detailed examples.

Keeping the answer, regions, and masking fixed lets us follow individual and joint support as the explanation grows. We first ask whether a model can recognize a visible property yet assign it to the wrong object or part.

3 Study I: Localizing Spatial Attribute Misbinding

Attribute binding errors assign a visible property to the wrong object or part, a known failure in vision–language models (Tang et al., 2023; Park et al., 2025; Xu et al., 2026). RegionTrace reveals how the visual evidence behind such errors is distributed across regions.

A white-rimmed, black-bodied mug gives the simplest version of the phenomenon (Figure 2). Asked whether white is the main color of the rim, InternVL3-8B answers Yes. Change only the color word to black, and it again answers Yes; change it to red, and it answers No. The black answer is wrong, yet black is plainly visible—on the body rather than the rim.

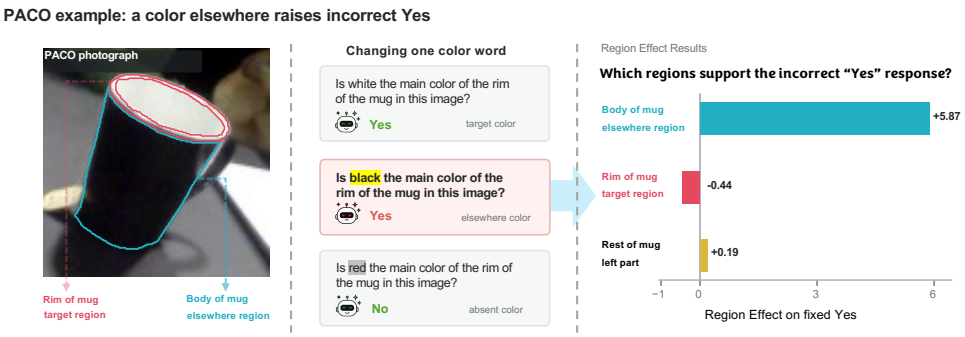


Figure 2: Motivating case: an answer can be visually supported yet spatially wrong. Only the color word changes across the three questions. The bars show each semantic region's Shapley-averaged Region Effect on the model's fixed false *Yes* relative to *No*; the black body, rather than the queried rim, provides the strongest support, and hiding it changes the constrained Yes–No decision to No. This illustrative development example is separate from the frozen PACO evaluation.

We call the rim the *target region* and white its *target color*. Black is the *elsewhere color*, carried by the body, its *elsewhere region*. Red is *absent* because it appears in no annotated region. The model's false *Yes* to black raises a sharper question: which part supports the error? RegionTrace answers by rescoring the model's original *Yes* relative to *No* as each region is hidden or restored. The body raises this score by $+5.87$, while the rim lowers it by -0.44; hiding the body changes the constrained Yes–No decision to No. The model has found a real color, but attached it to the wrong region. This example comes from PACO (Ramanathan et al., 2023).

3.1 Holding the image still and moving one word

We turn this example into a paired experiment in two complementary settings: 886 controlled PTR scenes with ground-truth part masks and colors (Hong et al., 2021), and 500 PACO photographs with annotated object-part masks and attributes (Ramanathan et al., 2023). Each image contributes three questions about the same target region, as illustrated in Figure 2. The first uses the target color and should be answered Yes. The second uses an elsewhere color and should be answered No. The third uses an absent color and should also be answered No. The image and target region remain unchanged, so the two No questions ask whether a color is more likely to elicit Yes when it is visibly present somewhere else. RegionTrace then measures how the target region, the elsewhere region, and the rest of the object affect support for Yes. We also hide the elsewhere region and rescore the constrained Yes–No decision. Appendix A gives the complete experimental setup.

Table 1: Yes rates (%) for the paired questions in PTR and PACO. Every entry uses all images in the corresponding dataset. Brackets show Wilson 95% confidence intervals; green values show elsewhere minus absent in percentage points.

Dataset	Model	Target color (correct Yes)	Absent color (false Yes)	Elsewhere color (false Yes)
PTR	Qwen2.5-VL-7B	85.4 [83.0, 87.6]	5.4 [4.1, 7.1]	30.9 [28.0, 34.0] (+25.5)
	InternVL3-8B	94.6 [92.9, 95.9]	14.6 [12.4, 17.0]	86.7 [84.3, 88.8] (+72.1)
	InternVL3-14B	90.7 [88.7, 92.5]	11.3 [9.4, 13.5]	53.6 [50.3, 56.9] (+42.3)
PACO	Qwen2.5-VL-7B	59.8 [55.4, 64.0]	7.6 [5.6, 10.3]	23.0 [19.5, 26.9] (+15.4)
	InternVL3-8B	77.0 [73.1, 80.5]	24.6 [21.0, 28.6]	57.4 [53.0, 61.7] (+32.8)
	InternVL3-14B	72.8 [68.7, 76.5]	18.8 [15.6, 22.5]	40.0 [35.8, 44.4] (+21.2)

3.2 The wrong answer follows where the color is

The relevant comparison in Table 1 is between the two questions whose correct answer is No. In PTR, making the named color visible on another region raises the false Yes rate over the absent-color question by 25.5, 72.1, and 42.3 percentage points across the three models. In PACO photographs, the corresponding increases are smaller but remain substantial: 15.4, 32.8, and 21.2 points. The error is therefore not explained by a general tendency to say Yes. In both controlled scenes and natural photographs, a wrong Yes is more likely when genuine evidence for the named color exists somewhere else in the image.

Table 2: Regional results among elsewhere-color questions answered Yes. Brackets show Wilson 95% confidence intervals. Elsewhere $>$ target asks whether the color-bearing elsewhere region has a larger Shapley-averaged Region Effect than the queried target.

Dataset	Model	False Yes cases (n)	Elsewhere $>$ target	Elsewhere Top-1	Hide elsewhere $\rightarrow$ No
PTR	Qwen2.5-VL-7B	274	97.1 [94.3, 98.5]	85.4 [80.7, 89.1]	83.6 [78.7, 87.5]
	InternVL3-8B	768	97.9 [96.6, 98.7]	90.8 [88.5, 92.6]	76.7 [73.6, 79.5]
	InternVL3-14B	475	92.0 [89.2, 94.1]	80.0 [76.2, 83.4]	76.0 [72.0, 79.6]
PACO	Qwen2.5-VL-7B	115	68.7 [59.7, 76.5]	55.7 [46.5, 64.4]	53.9 [44.8, 62.7]
	InternVL3-8B	287	77.0 [71.8, 81.5]	59.9 [54.2, 65.4]	37.3 [31.9, 43.0]
	InternVL3-14B	200	68.0 [61.2, 74.1]	53.0 [46.1, 59.8]	35.5 [29.2, 42.3]

Among false-Yes answers, the color-bearing elsewhere region has a larger effect than the queried target in 92.0%–97.9% of PTR cases and 68.0%–77.0% of PACO cases (Table 2). Hiding this region changes the constrained answer to No in 76.0%–83.6% of PTR errors and 35.5%–53.9% of PACO

errors. On correctly answered target-color questions, the target region instead exceeds the elsewhere region in 97.1%–98.9% of PTR cases. For PACO cases where no other annotated object region shares the target color, the corresponding range is 80.3%–81.8% (Appendix Table 4). These comparisons link support for the answer to the region carrying the named color.

Ranked deletion further tests whether the regions with the largest effects sustain the wrong answer. After removing the region with the largest effect, a positive Yes margin remains in 9.5%–21.2% of PTR errors and 23.5%–54.0% of PACO errors. Appendix Figure 8 shows the PTR curves and an independent PACO check; Appendix A reports the full 500-image PACO sequences. The regional contributions also depend on what else is visible. Across the six settings, 63.0%–67.2% of false-Yes cases contain a region whose effect switches sign when the other selected regions change from hidden to visible. The leading-region sets also differ between these contexts in 27.0%–30.5% of PACO errors (Appendix Table 20). Further checks examine image resolution and errors shared across models (Appendix A.2). Study II follows the same answer as the explanation grows, asking how its support changes with the accumulating text.

4 Study II: Tracking Answer Support Through Explanation Generation

Study I localized visual support for a fixed answer. Many methods use textual chain-of-thought (Text-CoT) to improve accuracy on tasks that require multi-step visual reasoning (Wei et al., 2022; Zhang et al., 2024; Chen et al., 2024). We now follow the same candidate through those reasoning steps, holding its identity fixed to distinguish changes in answer support from changes in the words being generated.

Figure 3 illustrates a visual-to-text handoff in one MM-GCoT trace. Before any explanation, the complete image assigns 98.0% probability to Yes, compared with 0.1% under I^{base}, the whole-image control that replaces the complete visual input with the mean-image baseline from Section 2.3; the selected regions contribute 93.2 probability points. After the first sentence is written, support under I^{base} rises to 89.3% while the regional contribution falls to 1.1 points; immediately before the answer, the corresponding values are nearly 100% and 0.0. The image initially favors Yes; the written explanation then sustains this preference with little additional support from the selected regions.

Experimental setup. To see whether this pattern holds beyond the example, we follow each explanation at the same three moments: before it begins, after the first sentence, and just before the model answers. At all three points, we ask how strongly the model favors the answer it will eventually give, whether that answer is correct or incorrect. We use the candidate-normalized probability defined in Section 2.4, and never include the answer itself in the text shown to the model. We then restore the chosen image regions while keeping the written text and the answer fixed. The resulting Region Effect measures how much those pixels change support for the same answer as the explanation unfolds. In MM-GCoT (Wu et al., 2025), the chosen regions are those referred to by the reasoning steps; in VCR (Zellers et al., 2019), they are the annotated people and objects named in the question. Across InternVL3-8B, InternVL3-14B (Zhu et al., 2025), and Qwen2.5-VL-7B (Bai et al., 2025), this gives

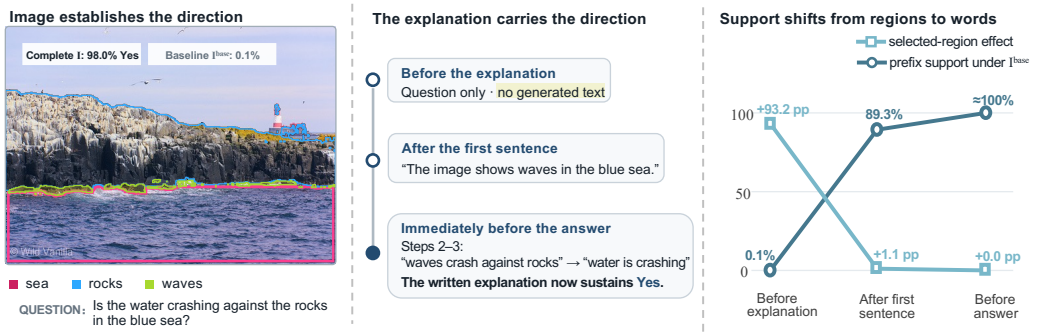


Figure 3: A visual-to-text handoff in one InternVL3-8B MM-GCoT trace (`judgement:673`). The left panel identifies the selected regions, the middle panel follows the generated explanation, and the right panel contrasts support retained under the whole-image I^{base} control with the probability-scale Region Effect. Both measurements track the same answer along the generated explanation.

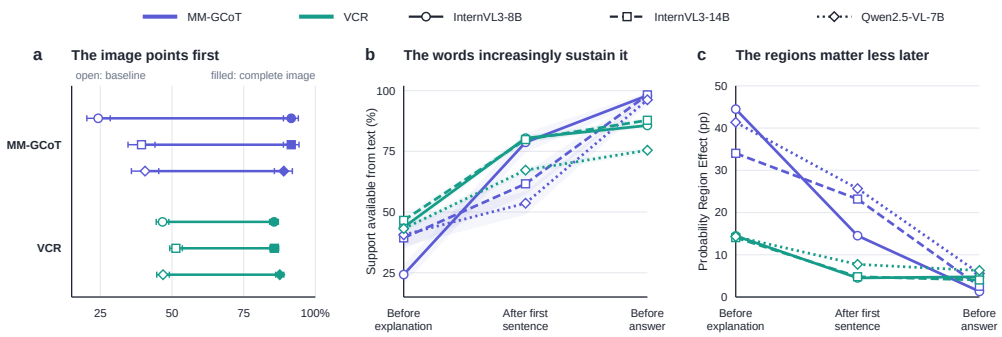


Figure 4: Population evidence for the visual-to-text handoff in Figure 3. **a**, Agreement between the answer favored before explanation and the final answer, with the complete image or mean-image baseline. **b**, Eventual-answer support under I^{base} as the explanation grows. **c**, Probability-scale Region Effect of the selected regions. Colors denote tasks and markers denote models; panel **b** is compared within task because MM-GCoT reports a favoring rate and VCR a mean probability. Error bars and bands in panels **a**–**b** show sample-level bootstrap 95% confidence intervals.

7,206 valid explanations. Appendix Section B.1 describes how the examples were chosen and how the scores were computed.

Results. Before any explanation, a simple answer probe uses the question and current image to score each possible answer. Figure 4a compares its top answer with the model's final answer. With the complete image, the two match in 86–92% of traces, but under I^{base} they match in only 24–51%. Thus the complete image often favors the eventual answer before the explanation begins.

Next, Figure 4b tracks support for that answer during explanation. As the explanation grows, support for the same eventual answer rises sharply even under I^{base}, and is highest just before the answer in every setting. At the same time, restoring the selected regions provides less additional support: their probability-scale Region Effect falls after the first sentence in all six settings and reaches only 1.3–6.3 probability points just before the answer (Figure 4c). Across all six settings, log-odds Region Effect also decreases from the first sentence to just before the answer, with every paired-change 95% confidence interval below zero (Appendix B.3). In the two MM-GCoT InternVL3 settings with complete regional decompositions, the mean sum of absolute probability-scale regional contributions also falls, showing that their decline is not solely due to signed cancellation (Appendix B.4). The same pattern also occurs in explanations ending in an incorrect answer. By the end of the explanation, more than 90% of MM-GCoT errors favor the wrong answer that the model is about to give, and support for the eventual wrong choice also rises on VCR (Appendix Figure 11).

We also test the link between regional support and the final answer by removing one region and asking the model to answer again. Across all 404 InternVL3-8B MM-GCoT traces, removing the region ranked highest by RegionTrace changes the final answer in 18.6% of cases, nearly twice the 9.9% rate when attention chooses the region and the 9.6% rate under uniform random selection. Appendix Table 9 gives the confidence intervals and paired comparisons. A changed answer is not necessarily a better one. The result shows that RegionTrace identifies regions with greater influence on what the model finally says.

Together, these results suggest that early visual evidence is carried forward in the generated text. As the explanation grows, the text can sustain the same answer, including an incorrect one, while the selected regions provide less additional support. Appendix B.2 repeats the main measurements with Qwen3.8-27B and Gemma 4 31B.

5 Study III: Auditing Visual Re-engagement During Look-Back

Once the generated text sustains an answer, can an explicit second look restore regional influence? Several methods pursue this idea by revisiting or zooming into the image during reasoning (Wu & Xie, 2024; Yu et al., 2025; Peng et al., 2026). We examine Look-Back (Yang et al., 2026), which further trains Qwen2.5-VL-7B-Instruct (Bai et al., 2025) to take a clearly marked second look at the

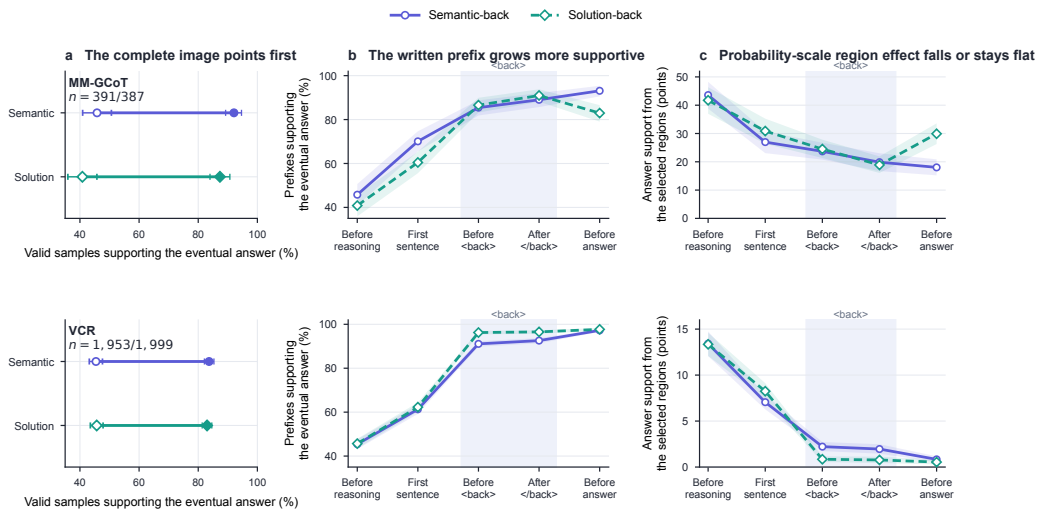


Figure 5: Study III results for Semantic-Back and Solution-Back on MM-GCoT (top) and VCR (bottom). **a**, Agreement between the answer favored before reasoning and the eventual answer; **b**, support supplied by the preceding text when useful image content is removed; **c**, probability-scale Region Effect. Shading marks the <back> span; panel **c** uses a separate scale for each task, and intervals use sample-level bootstrap.

image. Look-Back keeps the original image available and marks its visual check with a <back> span, during reasoning in Semantic-Back and after a preliminary solution in Solution-Back (see Appendix C.1 for details). We measure whether the selected regions supply more support for the fixed answer after this span and separately test changes in answers and accuracy.

Experimental setup. We keep the regions and I^{base} control from Study II, and follow the answer that each model eventually gives. We measure Region Effect at five moments: before reasoning, after the first sentence, immediately before and after <back>, and just before the answer. At each moment, we keep the words already written and the eventual answer fixed. Region Effect then measures what changes when the selected regions are restored. The main analysis contains 391 Semantic-Back and 387 Solution-Back responses on MM-GCoT, plus 1,953 and 1,999 responses on VCR, respectively. Appendix Section C.1 describes the included responses, the two models, and their training. We compare each genuine Look-Back span with an equal-length matched continuation constructed from ordinary reasoning generated from the same preceding text. Appendix C.3.1 also reports a direct <think>-marked control.

Results. Even before reasoning begins, the answer favored with the complete image matches the eventual answer in 83.0–92.1% of responses (Figure 5a). By the time <back> begins, the words written so far favor that same answer in 85.4–96.3% of responses, even without useful image content (Figure 5b). The selected regions do not consistently regain influence during the marked second look. Their Region Effect falls on MM-GCoT and remains near zero on VCR (Figure 5c). VisualSwap reports a related pattern: after the image changes during reasoning, models often continue from the earlier image even while saying that they are checking again (Shi et al., 2026).

Comparing only the start and end of <back> could miss a brief return to the image. In two MM-GCoT cases, attention to the selected regions rises by $1.93\times$ and $1.32\times$ during the check. Yet the regions' effect on the words written during that check rises by $2.52\times$ in one case and falls to $0.31\times$ in the other. The latter case retains the wrong answer (Appendix Figure 13). These examples illustrate that looking more at a region does not always make it more influential.

On MM-GCoT, the selected regions lose more influence during the genuine Look-Back span than under the primary equal-length matched control. The differences are -7.66 $[-9.15, -6.20]$ points for Semantic-Back and -8.21 $[-9.74, -6.77]$ for Solution-Back (Table 3). On VCR, the corresponding estimates are small and both intervals include zero. When the two tasks are combined, the accuracy differences are also small and uncertain: $+0.56$ $[-0.79, 1.80]$ and $+0.45$ $[-0.34, 1.35]$ points (Appendix C.3.1). In these settings, neither Look-Back variant consistently restores regional influence

Table 3: Equal-length comparison between genuine Look-Back and a matched ordinary-reasoning control. Values show the change in Region Effect during Look-Back minus that under the matched ordinary-reasoning control (percentage points; sample-level bootstrap 95% intervals).

Task	Variant	n	Genuine − matched (pp)
MM-GCoT	Semantic	391	-7.66 $[-9.15, -6.20]$
	Solution	387	-8.21 $[-9.74, -6.77]$
VCR	Semantic	500	$+0.23$ $[-0.37, +0.87]$
	Solution	500	$+0.13$ $[-0.23, +0.54]$

or improves accuracy. Additional analyses examine changes in answer correctness and results by VCR answer position (Appendix C.2).

6 Related Work

Reasoning traces and visual dependence. Chain-of-thought can improve performance without faithfully describing an answer's cause (Wei et al., 2022; Turpin et al., 2023; Balasubramanian et al., 2025). Truncating or altering reasoning traces tests how answers depend on the generated text (Lanham et al., 2023). Early errors can steer later text, visual use may decline, and Look-Back revisits the image (Zhong et al., 2024; Zheng et al., 2025; Jin et al., 2026; Yang et al., 2026). VisualSwap complements these studies by changing the image after a reflection cue and asking whether the model follows the new evidence (Shi et al., 2026). RegionTrace tests the same candidate at successive stages of generation.

Input and internal interventions. Interpretability intervenes on hidden states to locate mediators (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Palit et al., 2023), or on inputs to measure output effects. Recent work further decomposes attention heads into visual and textual routes and uses decision-aligned route interventions to identify text-dominant hallucination mechanisms (Cheng et al., 2026). Input-side work uses perturbations, semantic or multimodal Shapley values, and token maps (Fong & Vedaldi, 2017; Petsiuk et al., 2018; Cafagna et al., 2023; Parcalabescu & Frank, 2023; 2025; Goldshmidt, 2025; Samyoun et al., 2026; Li et al., 2025; Park et al., 2025). MultiSHAP uses Shapley interaction indices to quantify pairwise interactions between image patches and text tokens (Wang & Wang, 2025). EAGLE explains selected generated tokens using a sufficiency–necessity objective and analyzes their visual dependence (Chen et al., 2026). RegionTrace follows a fixed candidate through generation while retaining signed regional effects across visible contexts. On 37 PACO images with 2–6 named regions, RegionTrace averaged 4.11 s per image, including automatic box detection and mask generation with Grounding DINO and SAM. EAGLE's efficient search over 59–68 segments averaged 168.04 s on the same GPU (Appendix F.3). Appendices F–H give matched comparisons, formal attention conditions, and broader context.

7 Conclusion and Limitations

Conclusion. RegionTrace measures how image regions support or suppress the same answer across visual contexts and stages of multimodal generation. Our experiments show that incorrect attribute answers often draw support from real visual evidence outside the queried object or part. As explanations grow, the written text increasingly sustains the eventual answer while the selected regions provide less additional support. In the tested Look-Back settings, marked visual review does not consistently restore regional influence or improve accuracy relative to equal-length matched reasoning controls. An additional Study I experiment with GPT-5.2 shows that regional attribution can also be performed through a closed-model API that returns the required answer log-probabilities (Appendix A.4). RegionTrace connects where an answer finds visual support with how that support changes, providing a common framework for studying grounding errors and visual re-examination.

Limitations. RegionTrace operates on a chosen set of image regions, with computational cost increasing as more regions are included. The spatial scope and precision of its attributions depend on the coverage and quality of the supplied masks. The reported analyses assume access to candidate-answer probabilities or token log-probabilities. The main longitudinal analysis conditions on the recorded prefix. Extending it to systematically regenerated reasoning trajectories after each image intervention remains future work.

AI USE STATEMENT

We used generative AI tools for literature and reference checking, feedback on methodology, experiments, and result interpretation, code debugging, figure preparation, and manuscript editing. AI was also used to identify potential inconsistencies and weaknesses during revision. All AI-assisted content was independently reviewed and verified against the underlying experiments, implementation, and source materials. The authors made all final scientific decisions and take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Section 2 specifies the score held fixed by RegionTrace, the region construction, the visual baselines, and the reported effects. The corresponding appendices document sample selection, prompts, model variants, comparison groups, denominators, and statistical checks. Our anonymous code repository is available at `https://github.com/regiontrace-anon/regiontrace/tree/main/code`. The released code exports masks, modified images, token identifiers, response tables, and source data for the reported figures, allowing each reported result to be traced to the corresponding model run.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9525–9536. Curran Associates, Inc., 2018.

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

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13406–13439. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.723.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJg1f6EFDB`.

Michele Cafagna, Lina M. Rojas-Barahona, Kees van Deemter, and Albert Gatt. Interpreting vision and language generative models with semantic visual priors. *Frontiers in Artificial Intelligence*, 6:1220476, 2023. doi: 10.3389/frai.2023.1220476.

Ruoyu Chen, Xiaoqing Guo, Kangwei Liu, Siyuan Liang, Shiming Liu, Qunli Zhang, Laiyuan Wang, Hua Zhang, and Xiaochun Cao. Where MLLMs attend and what they rely on: Explaining autoregressive token generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17057–17066, 2026.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 192–210. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.11.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. Spotlight.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL `https://www.jmlr.org/papers/v22/20-1316.html`.

Kawin Ethayarajh and Dan Jurafsky. Attention flows are shapley value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 49–54. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.8.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457. IEEE, 2017. doi: 10.1109/ICCV.2017.371.

Roni Goldshmidt. Understanding what vision-language models see in traffic: PixelSHAP for object-level attribution in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1794–1802, 2025.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Yining Hong, Li Yi, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17427–17440. Curran Associates, Inc., 2021.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357.

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8567–8608. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.387.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.574.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 48–58, 2025.

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

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 *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72970-6_3.

Gianluigi Lopardo, Frederic Precioso, and Damien Garreau. Attention meets post-hoc interpretability: A mathematical perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32781–32800. PMLR, 2024. URL https://proceedings.mlr.press/v235/lopardo24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022. URL https://www.jmlr.org/papers/v23/21-0439.html.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2848–2853, October 2023. doi: 10.1109/ICCVW60793.2023.00307.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4032–4059. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.223.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/37294f033582ac0064bf90fa557c2573-Abstract-Conference.html.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. HalLoc: Token-level localization of hallucinations for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29893–29903, 2025.

Lei Peng, Shuai Lv, and Wei Hu. ReGround: Restoring visual grounding in multi-step reasoning through self-diagnosis and visual re-examination. *arXiv preprint arXiv:2608.04385*, 2026. doi: 10.48550/arXiv.2608.04385. URL https://arxiv.org/abs/2608.04385.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. URL https://bmva-archive.org.uk/bmvc/2018/contents/papers/1064.pdf.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151, June 2023.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Sirat Samyoun, Yingtai Xiao, and Jian Du. Attribution-guided multi-object hallucination and bias detection in vision-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4529–4548. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.210.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, chapter 17, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

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. doi: 10.48550/arXiv.2605.15864. URL https://arxiv.org/abs/2605.15864.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14333–14348, 2023. doi: 10.18653/v1/2023.emnlp-main.886.

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34, 2004. doi: 10.1080/10867651.2004.10487596.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275.

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*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zhanliang Wang and Kai Wang. MultiSHAP: A shapley-based framework for explaining cross-modal interactions in multimodal AI models. *arXiv preprint arXiv:2508.00576*, 2025. doi: 10.48550/arXiv.2508.00576. URL https://arxiv.org/abs/2508.00576.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

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.

Yuanzhi Xu, Qian Gao, Jun Fan, Guohui Ding, Zhenyu Yang, Yuteng Xiao, and Sixue Lin. Diagnosing dense same-class attribute misbinding in large vision-language models. *arXiv preprint arXiv:2608.16805*, 2026. doi: 10.48550/arXiv.2608.16805. URL https://arxiv.org/abs/2608.16805.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in MLLM reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11694–11702, 2026. doi: 10.1609/aaai.v40i14.38154.

Xuan Yu, Dayan Guan, and Yanfeng Gu. Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*, 2025. doi: 10.48550/arXiv.2506.01663. URL https://arxiv.org/abs/2506.01663v2.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731, 2019.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. Why LVLMs are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4101–4113, 2025.

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

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

APPENDIX CONTENTS

Open Resources and Compute **15**

A Study I: Localizing Spatial Attribute Misbinding **15**

A.1 Datasets & experimental setup . . . 15

A.2 Supplementary checks . . . 16

A.3 Additional results . . . 17

A.4 Study I through a closed-model API . . . 17

B Study II: Tracking Answer Support Through Explanation Generation **18**

B.1 Datasets & experimental setup . . . 18

B.2 Additional validation with larger models . . . 20

B.3 Log-odds sensitivity analysis . . . 20

B.4 Absolute probability effects and cancellation . . . 21

B.5 Additional experiments . . . 23

C Study III: Auditing Visual Re-engagement During Look-Back **23**

C.1 Datasets & experimental setup . . . 23

C.2 Answer transitions after Look-Back . . . 25

C.3 Additional experiments . . . 25

C.4 Token-level audit and more examples . . . 26

D Changing How Hidden Regions Are Filled **28**

D.1 PTR attribution . . . 28

D.2 MM-GCoT answer-support trajectory . . . 29

D.3 VCR equal-length control . . . 29

E Approximating RegionTrace Beyond Exact Enumeration **30**

E.1 Sampling a small number of region orders . . . 30

E.2 Integrated gradients: a negative pilot . . . 32

F From Token Localization to Longitudinal Answer-Support Auditing **32**

F.1 Matched-region comparison with static attribution . . . 33

F.2 Endpoint attribution discards visual context . . . 34

F.3 Time and memory cost . . . 35

G Attention Routes Information; RegionTrace Measures Answer Support **36**

G.1 Attention and regional support . . . 36

G.2 Target scores and Shapley attribution . . . 37

G.3 Effects across textual and visual contexts . . . 38

H Extended Related Work **39**

Open Resources and Compute

Anonymous code and project website. The code associated with this work is available at `https://github.com/regiontrace-anon/regiontrace/tree/main/code`. The accompanying anonymous website, `https://regiontrace-anon.github.io/regiontrace/`, provides an interactive introduction to RegionTrace. It illustrates how regional effects are computed across visual contexts and presents precomputed examples from the three studies: attribute binding, answer-support trajectories during explanation generation, and visual re-engagement during Look-Back. The website replays saved experimental results and does not perform live model inference.

Compute. The experiments were conducted on two NVIDIA A40 GPUs and one NVIDIA RTX PRO 6000 GPU. The reported attribution tables and figures were computed from saved model outputs and intervention records; experiment-specific model, precision, masking, and sampling settings are documented in the corresponding appendix sections.

A Study I: Localizing Spatial Attribute Misbinding

A.1 Datasets & Experimental Setup

PTR. We retain 886 unique images containing one unambiguous instance of the queried object category. Within that object, the elsewhere color appears in exactly one annotated region in the complete scene, and the absent color appears in none. The matched elsewhere and absent color counts are brown 125, cyan 113, blue 112, purple 112, green 110, yellow 106, gray 105, and red 103. The object counts are 524 beds, 259 chairs, 76 tables, and 27 refrigerators. Every image and all three questions were retained regardless of model output. The first 500 cases were frozen before the final 386 eligible images were added; the same 886-image manifest is used for all three models. Figure 6 illustrates how a queried object is decomposed into a target part, an elsewhere part, and the remaining object pixels, and how the three matched color questions are formed.

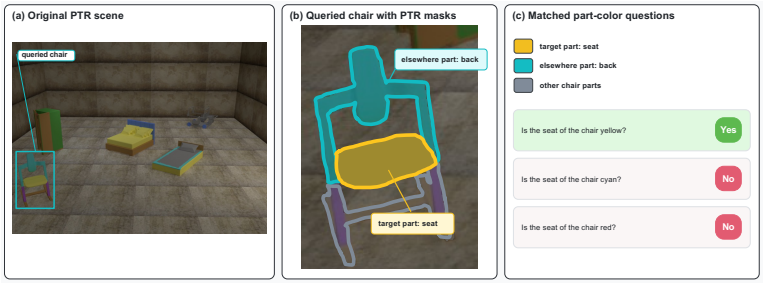


Figure 6: PTR construction example. **a**, the queried object in the complete scene. **b**, non-overlapping masks for the target seat, the elsewhere back, and the remaining chair pixels. **c**, matched questions using the target, elsewhere, and absent colors.

PACO. The main set contains 500 distinct PACO-LVIS training images, with one annotated object in each image. All cases were manually checked and frozen before model inference; one mosaic-like image was excluded during this image-only audit. The set covers 55 object categories, and no category contributes more than 30 images (6%). Each case has non-overlapping masks for the target region, the elsewhere region, and the remaining object pixels. The elsewhere color occurs in exactly one annotated region of the parent object, while the absent color occurs in no PACO-annotated region in the image. Elsewhere and absent color counts are exactly matched, and the elsewhere region is smaller than, similar in area to, or larger than the target in 167, 166, and 167 cases, respectively. The target color is also unique within the parent object in 157 cases; this stricter subset is used for the correct-question regional control. The earlier 30 validation and 70 test images do not overlap the 500-image set. The 30 images served as development cases, while the independently frozen 70-image test split is retained only as a held-out directional check. Figure 7 shows the analogous construction for a natural PACO photograph, where the target and elsewhere regions are semantic parts of the same annotated object.

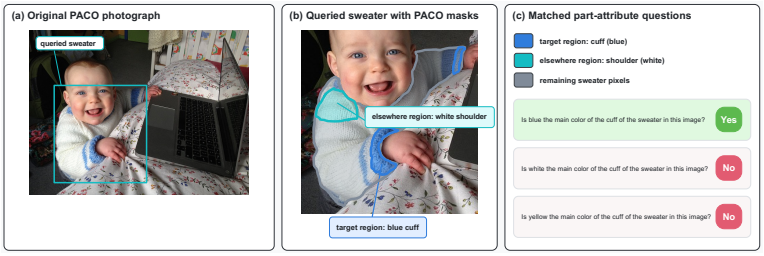


Figure 7: PACO construction example. **a**, the queried sweater in the original photograph. **b**, the target cuff, the elsewhere shoulder, and the remaining sweater pixels. **c**, matched part–attribute questions formed from the target, elsewhere, and absent colors.

Question construction. Each case contains a target region and a paired elsewhere region with different annotated colors. We ask three questions about the same target region while keeping the image fixed. The target-color question should be answered Yes, the elsewhere-color question should be answered No, and the absent-color question should also be answered No. The elsewhere color appears in the paired region; in PACO it is unique among the annotated regions of the parent object. The absent color appears in no annotated region in the image, which should not be read as a pixel-perfect claim about every background surface. Within each dataset, the elsewhere and absent questions have exactly matched color counts. Every question is wrapped as "Question: [question] Answer with exactly one token: yes or no. Answer:"

Region scoring. For every question, the object is divided into three non-overlapping pixel sets: the target region, the elsewhere region, and the remaining object pixels. For a visible-region set S, we score support for Yes relative to No as

$$V_{\mathrm{YN}}(S; q) = \log p_\theta(\mathrm{Yes} \mid q, I_S) - \log p_\theta(\mathrm{No} \mid q, I_S). \tag{7}$$

Yes denotes the largest next-token log probability over `yes`, `Yes`, and their leading-space forms; No is defined analogously. All 2^3 visible-region sets are evaluated. Equation 6 is then applied exactly to obtain one Shapley-averaged Region Effect for each pixel set. Top-1 means that a region has the largest of these three effects. For the behavioral check, we hide the elsewhere region and rescore the constrained Yes–No decision by next-token argmax. For the deletion curve, regions are hidden in descending order of their Shapley-averaged effects, with the ranking fixed before deletion begins.

Masking and statistical comparison. Each hidden region is filled with the fixed RGB baseline $[122, 116, 104]$, using the same replacement for every coalition score while leaving the image dimensions unchanged. Because the elsewhere-color and absent-color questions are paired within each image, we compare their answers with a two-sided exact McNemar test. For PTR, $p < 7 \times 10^{-45}$ for every model. For the 500 PACO images, $p = 1.97 \times 10^{-12}$ for Qwen2.5-VL-7B, $p = 1.14 \times 10^{-27}$ for InternVL3-8B, and $p = 5.13 \times 10^{-14}$ for InternVL3-14B. Wilson 95% intervals are used for the proportions in Tables 1 and 2 and Figure 8.

A.2 Supplementary checks

Two checks test whether the Study I pattern depends on the fixed-resolution scoring protocol or on separate, model-specific sets of errors.

Fixed image resolution. Exact coalition scoring requires every masked version of an image to preserve the same visual-token layout, so all reported Qwen2.5-VL-7B and InternVL3 region-scoring runs use a fixed 448-pixel input resolution. To check whether this choice itself creates the behavioral pattern, we reran a balanced 100-image answer-only subset with InternVL3-8B at its standard dynamic high resolution. The model produces 82 elsewhere-color false Yes answers under both the dynamic and fixed inputs, while only 11 absent-color questions are answered Yes under the standard input. Thus the fixed resolution leaves the central elsewhere-color error count unchanged in this check; the comparison tests answer behavior rather than equality of the numerical Region Effects.

Cross-model overlap. We also ask whether the aggregate results conceal disjoint failures by different models. This is a strict intersection: an image enters the comparison only when all three

models produce the same elsewhere-color false Yes. All three PTR models make the same elsewhere-color false Yes error on 169 images. In 152 of them (89.9%), the elsewhere region has a larger Shapley-averaged Region Effect than the target region for every model; in 120 (71.0%), it ranks first for every model. On PACO, all three models share 89 false Yes errors. The elsewhere region exceeds the target for all three models in 50 of them (56.2%), and for at least two models in 60 (67.4%). The shared errors therefore retain the same spatial signature across models, especially in the controlled PTR scenes; the lower agreement on PACO is consistent with evidence being more distributed in natural photographs. Together, the resolution and overlap checks show that the main finding is not created by one input format or by each model failing on a disjoint set of images.

A.3 Additional results

Table 4: Exact Study I results using the definitions from Section 3. The first two columns use all images. The next three condition on questions using an elsewhere color that were answered Yes. The last two condition on target-color questions answered Yes; for PACO, they additionally require that no other annotated region in the object shares the target color. All regional comparisons and Top-1 rankings use the exact Shapley-averaged Region Effect. Percentages are followed by numerator/denominator.

		All images		False Yes for an elsewhere color			Correct Yes	
Dataset	Model	Elsewhere color Yes	Absent color Yes	Elsewhere > target	Elsewhere Top-1	Hide elsewhere → No	Target > elsewhere	Target Top-1
PTR	Qwen2.5-VL-7B	30.9% (274/886)	5.4% (48/886)	97.1% (266/274)	85.4% (234/274)	83.6% (229/274)	98.9% (749/757)	98.2% (743/757)
PTR	InternVL3-8B	86.7% (768/886)	14.6% (129/886)	97.9% (752/768)	90.8% (697/768)	76.7% (589/768)	97.7% (819/838)	95.6% (801/838)
PTR	InternVL3-14B	53.6% (475/886)	11.3% (100/886)	92.0% (437/475)	80.0% (380/475)	76.0% (361/475)	97.1% (781/804)	95.5% (768/804)
PACO	Qwen2.5-VL-7B	23.0% (115/500)	7.6% (38/500)	68.7% (79/115)	55.7% (64/115)	53.9% (62/115)	80.3% (61/76)	72.4% (55/76)
PACO	InternVL3-8B	57.4% (287/500)	24.6% (123/500)	77.0% (221/287)	59.9% (172/287)	37.3% (107/287)	81.7% (76/93)	66.7% (62/93)
PACO	InternVL3-14B	40.0% (200/500)	18.8% (94/500)	68.0% (136/200)	53.0% (106/200)	35.5% (71/200)	81.8% (72/88)	68.2% (60/88)

For the main 500-image PACO set, the fraction of false Yes cases retaining a positive Yes margin after zero, one, two, and three ranked removals is $100.0\%, 23.5\%, 14.8\%, 19.1\%$ for Qwen2.5-VL-7B; $100.0\%, 54.0\%, 35.2\%, 34.8\%$ for InternVL3-8B; and $98.5\%, 45.0\%, 33.5\%, 35.5\%$ for InternVL3-14B. The 98.5% starting value reflects three exact Yes–No score ties among answers classified as Yes by the fixed decision rule. Figure 8 retains the independently frozen 70-image PACO test split as a separate directional check rather than mixing it into the 500-image main set.

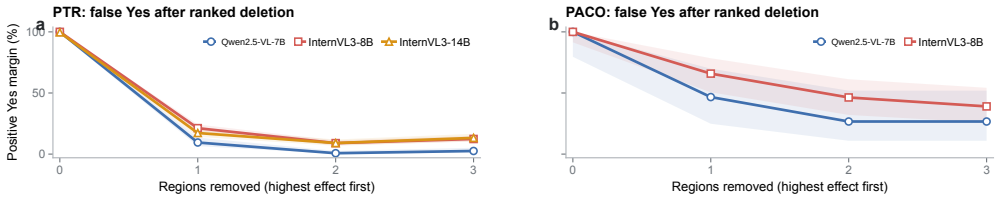


Figure 8: Removing regions in descending Shapley-averaged Region Effect reduces support for the false Yes. Curves show the fraction of false Yes cases for an elsewhere color whose fixed Yes-versus-No margin remains positive after zero to three regions are hidden. **a**, PTR errors are usually overturned by the first removal. **b**, the independent 70-image PACO test split shows the same direction but more diffuse support. Small late rebounds on PTR occur because a removed region can suppress Yes; the sequence is therefore not constrained to decrease at every step. Bands show Wilson 95% intervals; denominators are 274, 768, and 475 for PTR and 15 and 41 for the held-out PACO check.

A.4 Study I through a closed-model API

We additionally tested RegionTrace on GPT-5.2 through the OpenAI API, using only image inputs and returned answer log-probabilities. We used the first 50 images in the frozen 500-image PACO manifest, retaining the same three color questions and the target, elsewhere, and remainder masks. The first five images served as a pilot; we then fixed the next 45 images before querying them and reused the pilot responses. Selection followed manifest order, without filtering on the model's answers.

API setup. We used the snapshot `gpt-5.2-2025-12-11` through Chat Completions with reasoning disabled, temperature 1, and high image detail. For each question, we scored all eight region subsets and computed the same Yes–No log-score margin and exact Shapley summary as in Study I. Hidden pixels were replaced with fixed RGB $(122, 116, 104)$; this value is not assumed to match the API model's undisclosed processor mean. The prompt was unchanged from the open-model experiment: *Question: [question] Answer with exactly one token: yes or no. Answer:* The API returned five leading token log-probabilities. As in the original scoring rule, each answer branch used the highest score among its lowercase, capitalized, and leading-space forms. Both branches appeared among the returned candidates in all 1,200 evaluations, so their maxima were available without imputing missing scores.[1] All 150 questions therefore yielded complete regional attributions.

Table 5: Supplementary GPT-5.2 results on 50 PACO images. The first three rows use all 50 images; regional comparisons and deletion results condition on the 12 elsewhere-color false-Yes cases.

Measure	Count	Rate (%)
Target color: correct Yes	39/50	78.0
Absent color: false Yes	3/50	6.0
Elsewhere color: false Yes	12/50	24.0
Elsewhere $>$ target	11/12	91.7
Elsewhere ranks first	10/12	83.3
Hide elsewhere $\rightarrow$ No	6/12	50.0

Results. Elsewhere colors elicited false Yes answers more often than absent colors: 24.0% versus 6.0%, a difference of 18 percentage points (Table 5). Among the 12 elsewhere-color errors, the elsewhere region exceeded the target in 11 cases, ranked first in 10, and changed the answer to No when hidden in six. A higher signed effect need not be positive: in two of the ten first-ranked cases, all three regional effects were negative. The added 45 images alone contained 11 elsewhere-color errors; the elsewhere region exceeded the target in ten, ranked first in nine, and changed the answer to No when hidden in five. These observations agree with the direction of the main PACO findings and demonstrate regional attribution through a closed-model API. This small, ordered subset supports a supplementary check rather than a model ranking or a population estimate. The experiment verifies Study I answer scoring; it does not test API continuation from the fixed intermediate prefixes required by Study II.

Examples of API compatibility. Applicability depends on whether an image-capable endpoint returns the scores needed for the chosen answer comparison. OpenAI's GPT-5.2 Chat Completions API with reasoning disabled is the tested example above. Alibaba Cloud Model Studio also documents output log-probabilities and up to five candidate scores for Qwen3-VL-Plus and Qwen3-VL-Flash; these are potential Study I backends, subject to checking Yes–No candidate coverage, and were not tested here. Claude's standard Messages API does not currently expose token log-probabilities, so it cannot directly provide the Study I score used in this paper.[2]

B Study II: Tracking Answer Support Through Explanation Generation

B.1 Datasets & experimental setup

MM-GCoT. The MM-GCoT judgment test set contains 412 items. Study II examines multistep explanations, so the two one-step items are outside its scope, leaving 410 candidates. For the InternVL3-8B analysis, `judgement:917` produced an invalid answer format, and five items did not yield a detected region for every reasoning step under the fixed Grounding DINO and SAM procedure. The five region failures were `judgement:290`, `judgement:484`, `judgement:491`, `judgement:511`, and `judgement:640@2406690`. This left 404 valid InternVL3-8B explanations.

[1] The API returned a slightly positive log-probability in 88 evaluations. We retained these values under the pilot's preset 10^{-4} numerical tolerance and computed the original score differences without clipping.

[2] API documentation checked on 13 September 2026: OpenAI GPT-5.2, Alibaba Cloud Model Studio, and Claude Messages. These examples describe the named endpoints, not all models or services from each provider.

Qwen2.5-VL-7B was evaluated on the same 404 questions. Five outputs contained five reasoning sentences, outside the fixed two-to-four-sentence format used to define the three comparison points, leaving 399 valid explanations. The five items were `judgement:82`, `judgement:545`, `judgement:591`, `judgement:593`, and `judgement:674`. InternVL3-14B produced valid explanations for all 404 questions and uses the same frozen regions as InternVL3-8B. The main figure uses every valid explanation from each model. Restricting all three models to the 399 shared valid questions yields the same trajectories. The InternVL3-8B attention and answer-regeneration analyses use all 404 of its valid explanations. The dataset supplies 73 two-step, 328 three-step, and 3 four-step items among the 404 questions.

The analysis was first developed on 128 frozen InternVL3-8B items and then applied without changing the measurements to the remaining 276 valid items. The main text reports these 404 items together; the 128/276 division records how the analysis was developed. All retained items have non-empty, non-overlapping pixel masks. The complete-image token sequence matches the frozen explanation in every valid run, and the comparison immediately before the answer uses each model's own final sentence rather than imposing the same numbered sentence on explanations of different lengths. All four-step items are evaluated with all 16 region combinations. Figure 9 illustrates the correspondence between an MM-GCoT question, its step-level regions, and the grounded reasoning steps used to define the selected visual evidence.

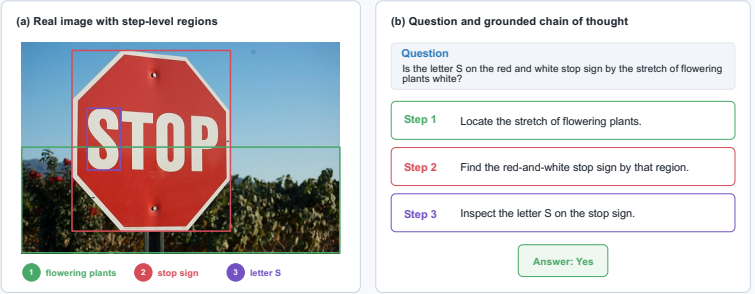


Figure 9: MM-GCoT example. **a**, a real image with one annotated region for each reasoning step. **b**, the question, grounded sequence of visual references, and final Yes/No answer.

VCR. VCR provides 26,534 validation questions over 9,929 image rows. Using a fixed random seed of 20260816, we shuffled the image rows and selected the first eligible question from each row until reaching 2,000 questions from 2,000 distinct images. An eligible question had to refer explicitly to at least one person or object in the official VCR annotations, allowing the visual content named by the question to be defined without consulting the answer choices or the model output.

Before model inference, a complete check of the locked source files found one image row with a Parquet CRC mismatch (row 4470). That row was excluded for this technical reason and replaced by the next eligible row in the same frozen shuffled order; no model result was available when this replacement was made. The final answer positions were balanced across the four choices (495, 484, 516, and 505 for A–D).

InternVL3-8B produced valid outputs for all 2,000 questions. InternVL3-14B also produced valid outputs for all 2,000 questions. Qwen2.5-VL-7B produced 1,999 valid outputs and one format failure, `vcr:val-16602`, whose generation contained fewer explanation lines than required by the fixed protocol. The failed output was recorded and was not regenerated or replaced. The main text reports every valid output from each model; restricting all three models to their 1,999 shared valid questions yields the same conclusions. The findings therefore apply to the fixed subset of VCR questions that explicitly identify annotated visual entities, rather than to every question in the full VCR validation set.

For each question, the selected visual content is the union of the official bounding boxes for the people and objects named in that question. This definition is fixed before inspecting the answer choices or model output. Figure 10 shows how these referenced entities define the visual content for a VCR question while the answer choices and rationale remain outside the region-selection procedure.

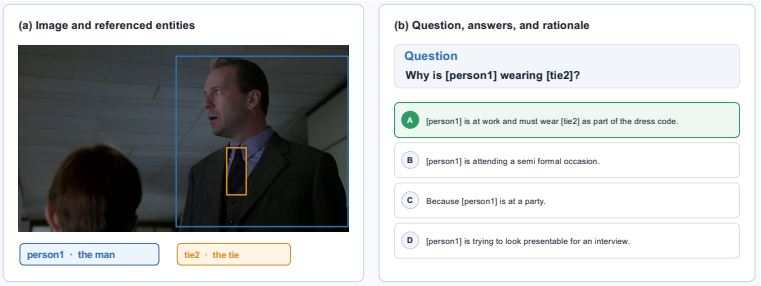


Figure 10: VCR example. **a**, the image and official boxes for the entities named in the question. **b**, the multiple-choice question, correct answer, and accompanying rationale.

Shared scoring. For both datasets, the answer eventually produced by the model is scored before the explanation, after its first sentence, and immediately before the answer. The answer itself is excluded from every saved prefix, and the answer target remains fixed whether it is correct or wrong. At each prefix, we compare the complete image, the whole-image I^{base} control, and the image with the selected regions removed. The main probability-scale analysis uses 10,000 sample-level bootstrap resamples for uncertainty intervals; the free-answer validation below uses 10,000 resamples with seed 20260815.

B.2 Additional validation with larger models

Setup. We repeated the three main Study II measurements with Qwen3.8-27B and Gemma 4 31B.[3] We used the same image regions and followed each model's own eventual answer at three moments: before the explanation, after its first sentence, and just before the answer. We evaluated all 404 MM-GCoT questions and the first 500 questions in the frozen VCR order; the VCR subset was fixed before inspecting either new model's outputs. Gemma produced valid explanations for every question; Qwen produced 403 valid MM-GCoT explanations and 500 valid VCR explanations. We retained the processor-mean replacement rule.[4]

Results. Table 6 reports the same three measurements as Figure 4. In all four settings, the complete image favored the eventual answer more often than the whole-image baseline before explanation began. As the explanation grew, support under the whole-image baseline increased. The mean probability-scale Region Effect of the selected regions fell after the first sentence and fell further just before the answer. The final Region Effect ranged from 0.47 to 1.58 probability points. These results extend the main pattern to both larger models: the written explanation increasingly sustained the same answer, with less additional support from the selected visual regions.

B.3 Log-odds sensitivity analysis

A probability difference can shrink as candidate probabilities approach one even when their log-odds contrast is unchanged. We therefore also examine selected-region support on the log-odds scale. For trace i at checkpoint t, let p_{it}^{full} and q_{it} denote the candidate-normalized probabilities assigned to the fixed eventual answer under the complete image and the selected-regions-removed image $I_{\varnothing}$, respectively. The corresponding Region Effect is

$$L_{it} = \operatorname{logit}\big(p_{it}^{\text{full}}\big) - \operatorname{logit}(q_{it}), \qquad \operatorname{logit}(p) = \log \frac{p}{1-p}. \tag{8}$$

For MM-GCoT, the odds compare the eventual Yes/No answer with the other candidate; for VCR, they compare the eventual answer with the remaining three candidates combined. The target answer remains fixed across the two post-initial checkpoints reported below. The checkpoint summaries are means of trace-level log-odds effects, rather than logit differences computed from mean probabilities. We also compute the change within each trace by subtracting its first-sentence effect from its final

[3] We used the official `Qwen/Qwen3.8-27B` and `google/gemma-4-31B-it` models in BF16, with additional thinking disabled.

[4] The replacement RGB values were $(127, 127, 127)$ for Qwen on MM-GCoT, $(128, 128, 128)$ for Qwen on VCR, and $(0, 0, 0)$ for Gemma on both tasks, following each processor's defaults and the existing task-specific rounding rule.

Table 6: Additional Study II validation with two larger models. Panels follow the three measurements in Figure 4. Brackets give 95% bootstrap confidence intervals. All scores concern each model's own eventual answer.

a. Early answer preference and sample counts

Model	Task	n	Acc. (%)	Complete image (%)	Whole-image baseline (%)
Qwen3.8-27B	MM-GCoT	403	86.10	93.30 [90.80, 95.57]	34.99 [30.38, 39.60]
Gemma 4 31B	MM-GCoT	404	81.19	91.58 [88.86, 94.15]	43.07 [38.29, 47.93]
Qwen3.8-27B	VCR	500	74.00	86.60 [83.40, 89.40]	47.00 [42.60, 51.40]
Gemma 4 31B	VCR	500	75.80	89.60 [86.80, 92.20]	49.40 [45.00, 53.80]

b. Answer support under the whole-image baseline (%)

Model	Task	Before explanation	After first sentence	Before answer
Qwen3.8-27B	MM-GCoT	34.99 [30.38, 39.60]	55.83 [51.01, 60.71]	100.00 [100.00, 100.00]
Gemma 4 31B	MM-GCoT	43.07 [38.29, 47.93]	65.35 [60.71, 69.88]	97.52 [95.91, 98.99]
Qwen3.8-27B	VCR	42.83 [39.98, 45.70]	88.46 [86.29, 90.51]	96.55 [95.42, 97.60]
Gemma 4 31B	VCR	49.11 [45.15, 53.05]	89.03 [86.40, 91.55]	97.91 [96.64, 98.99]

c. Selected-region effect (probability points)

Model	Task	Before explanation	After first sentence	Before answer
Qwen3.8-27B	MM-GCoT	32.54 [29.22, 35.98]	20.21 [17.50, 23.05]	0.47 [0.26, 0.74]
Gemma 4 31B	MM-GCoT	44.23 [39.36, 49.00]	23.60 [19.32, 27.90]	1.51 [0.53, 2.67]
Qwen3.8-27B	VCR	15.76 [13.30, 18.34]	2.02 [0.98, 3.08]	0.79 [0.26, 1.37]
Gemma 4 31B	VCR	20.77 [17.02, 24.48]	2.91 [0.78, 5.08]	1.58 [0.60, 2.73]

n counts valid explanations; accuracy uses the same denominator. Panel a reports how often the eventual answer ranked first before explanation. Panel b reports a favoring rate for MM-GCoT and a mean candidate-normalized probability for VCR, so it is compared within task. Intervals use 10,000 bootstrap resamples: whole images for MM-GCoT (349 for Qwen; 350 for Gemma), and questions for VCR (500 distinct images).

pre-answer effect. For each bootstrap resample, we average the trace-level effects and their paired changes using the same estimators as for the point estimates. Table 7 reports 95% confidence intervals for both the checkpoint means and the mean paired change.

Table 7: Mean log-odds Region Effect and its paired change between two explanation checkpoints. Paired change is before answer minus after the first sentence, computed within each trace. Means are followed by 95% bootstrap confidence intervals; all values are in log-odds units.

Task	Model	After first sentence	Before answer	Paired change
MM-GCoT	InternVL3-8B	1.35 [1.18, 1.52]	0.78 [0.65, 0.91]	−0.57 [−0.72, −0.42]
	Qwen2.5-VL-7B	1.28 [1.09, 1.47]	0.87 [0.72, 1.02]	−0.41 [−0.58, −0.24]
	InternVL3-14B	1.47 [1.30, 1.64]	0.88 [0.74, 1.02]	−0.59 [−0.74, −0.44]
VCR	InternVL3-8B	0.58 [0.52, 0.64]	0.35 [0.30, 0.40]	−0.23 [−0.29, −0.17]
	Qwen2.5-VL-7B	0.66 [0.59, 0.73]	0.41 [0.35, 0.47]	−0.25 [−0.32, −0.18]
	InternVL3-14B	0.57 [0.51, 0.63]	0.34 [0.29, 0.39]	−0.23 [−0.29, −0.17]

Mean log-odds Region Effect decreases from the first sentence to the final pre-answer prefix in all six model–task settings (Table 7). The paired changes range from -0.59 to -0.41 on MM-GCoT and from -0.25 to -0.23 on VCR, with all six 95% confidence intervals below zero. These paired reductions show that the regional log-odds contrast itself weakens during the later part of the explanation, beyond the probability compression that can occur when an unchanged contrast is mapped through the logistic function. All final means remain positive, indicating that selected regions retain support for the eventual answer even as their contribution diminishes.

B.4 Absolute probability effects and cancellation

A group Region Effect is a signed net quantity: contributions with opposite directions can cancel across regions, and signed averaging can also conceal effects with opposite directions across traces.

We examine these two possibilities using the saved probability scores underlying the Study II trajectories. The eventual answer, generated prefixes, selected regions, and replacement baseline remain fixed within each trace.

Let $p_{it}(S)$ be the candidate-normalized probability of trace i's eventual answer at checkpoint t with region subset S visible. We define the signed group contrast D_{it} and two absolute summaries as

$$D_{it} = p_{it}(\mathcal{N}) - p_{it}(\varnothing), \qquad G_{it} = |D_{it}|, \qquad A_{it} = \sum_{r \in \mathcal{N}} |\phi_{irt}^{(p)}|, \tag{9}$$

where $\phi_{irt}^{(p)}$ is the exact Shapley contribution computed with probability as the coalition score in Equation 6. Averaging G_{it} removes cancellation between traces; averaging A_{it} also removes cancellation between their regional Shapley contributions. Absolute values are taken before averaging over traces. For the binary MM-GCoT score, we convert each coalition's answer margin to probability before computing its Shapley values.

Group contrasts are available for all six settings. Complete coalition scores permit the regional decomposition for all 404 MM-GCoT traces from each InternVL3 model. The Qwen MM-GCoT and VCR runs retain group contrasts but do not provide the complete regional decomposition required for A_{it}. We compare each trace's first-sentence and final pre-answer checkpoints using 20,000 paired trace-level bootstrap resamples and percentile intervals, with fixed seeds 20260906–20260911 across the six settings.

Table 8: Absolute probability effects at two explanation checkpoints. Panel A averages the absolute group contrast G_{it}; Panel B averages the sum of absolute regional Shapley contributions A_{it}. All values are in probability points. Checkpoint entries are means; brackets in the final column give paired 95% bootstrap confidence intervals for before answer minus after first sentence.

Task	Model	n	After first	Before answer	Paired change [95% CI]
A. Absolute group effect: $\mathrm{mean}_i\, G_{it}$					
MM-GCoT	InternVL3-8B	404	16.20	1.65	-14.56 $[-16.92, -12.27]$
	Qwen2.5-VL-7B	399	28.26	6.59	-21.67 $[-24.17, -19.16]$
	InternVL3-14B	404	27.26	3.11	-24.15 $[-27.07, -21.30]$
VCR	InternVL3-8B	2000	7.53	6.92	-0.62 $[-1.19, -0.04]$
	Qwen2.5-VL-7B	1999	12.09	9.90	-2.19 $[-2.61, -1.80]$
	InternVL3-14B	2000	8.18	6.32	-1.85 $[-2.39, -1.32]$
B. Absolute regional contributions: $\mathrm{mean}_i\, A_{it}$					
MM-GCoT	InternVL3-8B	404	22.38	2.55	-19.83 $[-22.60, -17.06]$
	InternVL3-14B	404	34.05	4.61	-29.44 $[-32.60, -26.38]$

Mean absolute group effect decreases in all six settings (Table 8A). For the two complete regional decompositions, mean absolute contribution also decreases substantially: from 22.38 to 2.55 probability points for InternVL3-8B and from 34.05 to 4.61 for InternVL3-14B (Table 8B). Thus the probability-scale reductions in these two settings cannot be explained solely by increased cancellation between regional contributions or between traces.

We treat the six group comparisons and two regional comparisons as a family of eight exploratory tests. Bonferroni-adjusted 99.375% bootstrap intervals remain below zero for seven comparisons, including both regional decompositions; the interval for the smaller VCR/InternVL3-8B group change is $[-1.41, 0.17]$. Because the 404 MM-GCoT traces share 350 images, we also resample whole image clusters. The resulting paired 95% intervals for the change in mean A_{it} are $[-22.69, -17.07]$ for InternVL3-8B and $[-32.64, -26.26]$ for InternVL3-14B.

These are population summaries: $7/404$ InternVL3-8B traces and $6/404$ InternVL3-14B traces have a decreasing signed group effect but an increasing sum of absolute regional contributions. The supported conclusion concerns reduced additional probability-scale influence on the same eventual answer, conditional on the generated prefix. Image information may already be preserved in that prefix; this analysis does not determine whether later reasoning is useful or whether visual reasoning has failed.

B.5 Additional experiments

Free-answer deletion validation. For each of the 404 valid InternVL3-8B MM-GCoT traces, we removed one exclusive semantic region by replacing its pixels with the processor-mean RGB value, then deterministically regenerated the complete explanation and answer from the original prompt. A full-image rerun first had to reproduce the frozen visible-token sequence exactly. This auxiliary run predates the probability-scale Study II trajectories and uses its frozen binary answer-contrast selector: RegionTrace chooses the largest signed Shapley-averaged score toward the original answer at the final saved prefix, while attention chooses the largest raw region mass averaged over all heads in layers 24–27 at that prefix. Three RegionTrace maxima were tied and resolved by the lowest region index. For the random comparator, we enumerate all single-region deletions within each item and average their outcomes, giving the exact expectation under uniform region sampling without Monte Carlo noise. A flip is counted only when the regenerated output contains a parseable final Yes/No answer different from the original; unparseable outputs remain in the denominator as non-flips. The table reports only this discrete endpoint.

Table 9: Free-answer deletion validation on InternVL3-8B MM-GCoT. Flip-rate intervals are marginal item-level bootstrap 95% confidence intervals; the last column gives item-paired bootstrap intervals for the RegionTrace-minus-row difference.

Selection rule	n	Flip rate (%) [95% CI]	RegionTrace minus row (pp) [95% CI]
RegionTrace	404	$18.56\ [14.85, 22.52]$	–
Attention	404	$9.90\ [7.18, 12.87]$	$+8.66\ [5.69, 11.88]$
Uniform random (exact)	404	$9.63\ [7.69, 11.65]$	$+8.93\ [6.58, 11.41]$

The intervention units are semantic regions, not area-matched masks. The mean deleted image fraction is 22.59% for RegionTrace, 34.14% for attention, and 19.18% under uniform selection. Deletion size therefore cannot explain the RegionTrace–attention gap, but may contribute to the comparison with uniform selection.

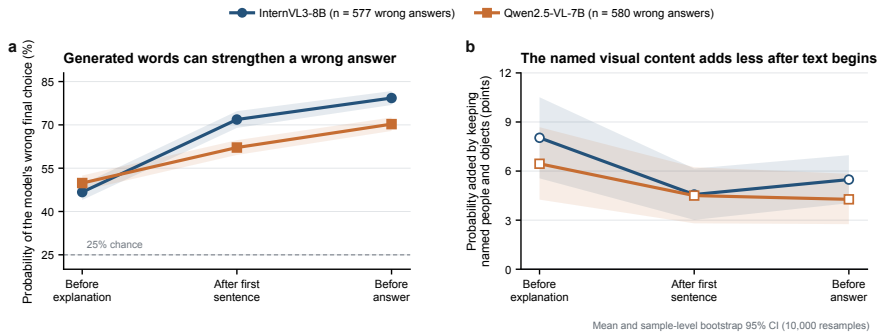


Figure 11: Generated explanations can also strengthen a wrong answer on VCR. **a**, Mean probability of the eventual wrong answer under the whole-image I^{base} control at the three saved moments. **b**, Group Region Effect of the people and objects named in the question, expressed as the difference between $I_{\mathcal{N}}$ and $I_{\varnothing}$. Results include every incorrect valid output ($n = 577$ for InternVL3-8B and $n = 580$ for Qwen2.5-VL-7B); bands and error bars show sample-level bootstrap 95% confidence intervals (10,000 resamples, seed 20260816).

C Study III: Auditing Visual Re-engagement During Look-Back

C.1 Datasets & experimental setup

Study III reuses the 404 MM-GCoT questions and the 2,000 VCR questions fixed for Study II. No failed trace was replaced. A trace enters the five-point analysis when it contains a complete `<back>` span, a parseable final answer, and valid candidate-probability readouts at all five prefixes. This

yields 391/404 and 387/404 MM-GCoT traces, and 1,953/2,000 and 1,999/2,000 VCR traces, for Semantic-Back and Solution-Back respectively.

Study III uses the released Semantic-Back-7B and Solution-Back-7B models from Look-Back (Yang et al., 2026). Both start from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and learn to generate a `<back>–</back>` span as an explicit boundary around a visual verification. Cold-start supervised fine-tuning teaches the output structure and the intended placement of this span; the authors report 4,000 marked examples for Semantic-Back and 10,000 for Solution-Back. GRPO training on 15,000 multimodal mathematics problems then combines answer accuracy with a format reward that encourages the model to invoke the learned review behavior. In both variants, the original image is supplied once and remains in the same multimodal context throughout generation: entering `<back>` changes the generated reasoning mode, but does not add a crop, a second image, or newly injected visual tokens.

Semantic-Back-7B. Semantic-Back places the visual check inside an unfinished chain of reasoning. An initial `<think>` span develops a partial plan or intermediate claim; `<back>` then returns to the image to verify the visual fact needed at that point; a second `<think>` span continues the argument from the checked evidence before producing the boxed answer. The review can therefore correct a local premise before it propagates through the rest of the solution.

Solution-Back-7B. Solution-Back delays the visual check until the model has formed a preliminary solution. Its `<back>` span re-examines the image against that provisional solution as a whole, after which a new `<think>` span rethinks the solution using the checked visual evidence and either confirms or revises the final answer. The distinction is thus one of timing and scope: Semantic-Back checks an intermediate visual step, whereas Solution-Back audits a nearly complete solution.

Figure 12 makes this difference concrete. In the airplane example, Semantic-Back verifies during reasoning that the image contains one airplane and immediately uses that fact to compute $1 - 1 = 0$. In the second example, Solution-Back first forms a tentative scene count, then checks that the image contains two bikes and two non-bike objects before applying the requested removal and addition to obtain six objects.

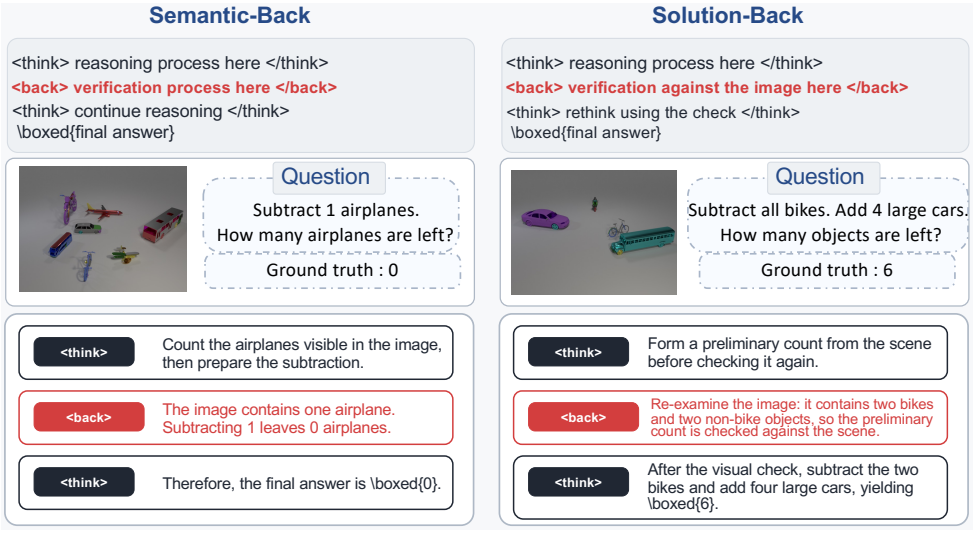


Figure 12: Output structure and illustrative behavior of the two Look-Back variants. **Left**, Semantic-Back inserts a visual verification while resolving an intermediate reasoning step. **Right**, Solution-Back checks the image after a preliminary solution and then rethinks the solution using that verification. Red spans denote the model-generated `<back>` verification; the original image remains available throughout both traces.

We evaluate both released models as provided, without further training, using their corresponding released prompts and the same question sets described above. We retain the regions from Study II—those named in the MM-GCoT reasoning steps and the annotated people and objects named in each

VCR question—and normalize answer probabilities over Yes/No for MM-GCoT and A/B/C/D for VCR at every saved prefix. The answer probe immediately before <back> appends the fixed answer cue only for scoring, rather than treating it as text already generated by the model. The main curves include every valid trace, differences are paired within each trace, and uncertainty is estimated with 10,000 sample-level bootstrap resamples.

C.2 Answer transitions after Look-Back

Figure 5 summarizes average answer support, but not whether the preferred answer actually changes after <back>. We therefore compare the answer favored immediately before the span with the final answer. Each trace either keeps its correctness state, is "Corrected" from wrong to correct, or is "Lost" from correct to wrong. Table 10 reports the changing outcomes; net is $(\text{Corrected} - \text{Lost})/n$ in percentage points, and p is the two-sided exact McNemar test on these paired transitions.

Table 10: Transitions from the answer favored immediately before <back> to the final answer written by the model. Net values are Corrected minus Lost as a percentage of all valid traces.

Task	Variant	Corrected	Lost	Net (pp)	p
MM-GCoT	Semantic	12	10	$+0.51$	.832
MM-GCoT	Solution	13	10	$+0.78$	.678
VCR	Semantic	48	36	$+0.62$	.230
VCR	Solution	12	16	-0.20	.572

On MM-GCoT, only 22 of 391 Semantic-Back traces (5.6%) and 23 of 387 Solution-Back traces (5.9%) switch between a correct and an incorrect answer state. Corrections are nearly balanced by losses: Semantic-Back records 12 versus 10, and Solution-Back 13 versus 10. The resulting net gains are only $+0.51$ and $+0.78$ percentage points, with $p = .832$ and $p = .678$. Thus neither variant shows evidence that the marked review systematically redirects an initially wrong preference toward the correct answer on MM-GCoT.

The same imbalance remains small on VCR. Semantic-Back changes correctness in 84 of 1,953 traces (4.3%), with 48 corrections and 36 losses, while Solution-Back changes only 28 of 1,999 traces (1.4%), with 12 corrections and 16 losses. Their net changes are $+0.62$ and -0.20 points, respectively, and neither paired comparison is significant ($p = .230$ and $p = .572$). These behavioral transitions agree with the main Region Effect result: a marked return to the image does not reliably translate into a corrected final answer.

Finally, we split VCR by the position of the correct option to check whether the aggregate trajectory is driven by a particular answer label. Semantic-Back shows a probability-scale Region Effect decline for each of A–D, whereas Solution-Back remains small and mixed around zero. The cross-option consistency indicates that the aggregate pattern is not an artifact of the answer-position distribution.

C.3 Additional experiments

C.3.1 Equal-length continuation control

The main trajectories compare two prefixes separated by a span of generated text. To distinguish the marked second look from added text alone, we branch from each trace immediately before <back> and construct two continuations with exactly the same number of tokens as its genuine <back> span. The model first generates ordinary reasoning text after a <think> marker while viewing the complete image. We then reuse those tokens inside either <back> or <think> markers; the two marker pairs have equal token lengths, and structural markers are excluded from the generated interior. The genuine span, the matched <back> span, and the matched <think> span therefore start from the same prefix and end at the same token position.

We score all three fixed prefixes with the complete image and with the question-referenced regions removed. The target is held fixed to the answer eventually written by the genuine trace, matching Figure 5c. The primary comparison subtracts the change under the matched <back> span from the change under the genuine span. It includes all 391 Semantic-Back and 387 Solution-Back MM-GCoT traces. For VCR, we select 500 traces per variant in proportion to pre-back correctness, answer identities, and genuine-span length; selection is deterministic with seed 20260822. All 1,778

matched examples completed successfully, and every pair has equal token length. Intervals use 10,000 sample-level bootstrap resamples with seed 20260822.

Table 11: Equal-length control for the change in eventual-answer Region Effect. Genuine and matched columns report the change from immediately before to immediately after the span. The final column is their paired difference. Values are conditional probability points with bootstrap 95% intervals.

Task	Variant	Genuine Δ	Matched Δ	Genuine $-$ matched
MM-GCoT	Semantic	$-3.82\,[-5.08, -2.58]$	$+3.84\,[+2.62, +5.06]$	$-7.66\,[-9.15, -6.20]$
MM-GCoT	Solution	$-5.61\,[-7.09, -4.21]$	$+2.61\,[+1.30, +3.94]$	$-8.21\,[-9.74, -6.77]$
VCR	Semantic	$-0.21\,[-0.59, +0.19]$	$-0.43\,[-1.06, +0.14]$	$+0.23\,[-0.37, +0.87]$
VCR	Solution	$-0.19\,[-0.63, +0.25]$	$-0.32\,[-0.75, +0.10]$	$+0.13\,[-0.23, +0.54]$

On MM-GCoT, ordinary matched text increases the eventual-answer Region Effect while the genuine span decreases it, so the decline is not explained by prefix length. On VCR, the estimated differences are small and their intervals include zero. Replacing the matched <back> markers with <think> markers gives the same task-level pattern: the differences between genuine and matched spans are -7.20 $[-8.80, -5.66]$ and -12.54 $[-14.55, -10.55]$ points on MM-GCoT, and -0.25 $[-1.06, 0.55]$ and -0.14 $[-0.58, 0.32]$ on VCR.

These differences in regional support do not produce a reliable behavioral advantage. Counting unparsed outputs as incorrect, genuine Look-Back exceeds its matched continuation in pooled accuracy by only 0.56 $[-0.79, 1.80]$ percentage points for Semantic-Back and 0.45 $[-0.34, 1.35]$ for Solution-Back, with both intervals including zero. The trajectory can nevertheless continue to change after the marked review: on MM-GCoT, the probability-scale Region Effect for Solution-Back rises from 18.89 points immediately after </back> to 29.91 before the answer. This is a delayed return during subsequent reasoning, not a recovery inside <back> itself. Together, the results distinguish renewed regional support from successful correction: one variant can recover visual influence later in the trace, yet Look-Back still does not reliably outperform an equal-length ordinary continuation.

Relation to VisualSwap. VisualSwap changes the image after reasoning has begun and asks whether the model notices the new evidence. Study III leaves the image unchanged and asks whether the same answer-relevant regions regain influence during the marked <back> span. The two designs illuminate different parts of re-examination, but support the same caution: language that sounds like a second look does not by itself show that vision is redirecting the answer. VisualSwap also finds that a new user turn can restore grounding, so our result concerns autonomous Look-Back within a continuing response rather than an inability to inspect the image when explicitly redirected.

C.4 Token-level audit and more examples

This supplementary audit asks whether a marked revisit changes the generated verification words, rather than the probability of a candidate answer. For a fixed generated token $\hat{y}_t$ and its original preceding text, we use

$$F_t(S) = \log p_\theta(\hat{y}_t \mid q, \hat{\mathbf{y}}_{<t}, I_S)\,. \tag{10}$$

Evaluating all 2^K visible-region sets gives one fixed-token score per coalition and position. Substituting these scores into Equation 6 yields the exact contribution $\phi_{i,t}$ of region i to token t. The plotted curves retain its sign, whereas the interval summary uses $|\phi_{i,t}|$ to measure influence without cancellation. Regional attention $R_{i,t}$ follows the Look-Back ratio: causal attention is averaged over heads within each layer, the image-token numerator is weighted by the fractional mask coverage $w_{i,j} \in [0, 1]$, and the layerwise ratios are then averaged.

The blue interval $\mathcal{P}$ in Figure 13 is selected separately for each trace from its pre-back regional-attention trajectory. We locate the low-variation tail reached after the ratios decline, then extend its left boundary slightly to retain the end of that decline. Tokenizer fragments belonging only to structural markers, including the preceding </think>, are excluded, so the length of $\mathcal{P}$ can vary across traces. The revisit interval $\mathcal{B}$ contains the semantic tokens strictly between the detected starts of <back> and </back>, again excluding marker-only fragments. For either $X_{i,t} = R_{i,t}$ or $X_{i,t} = |\phi_{i,t}|$, let $m = \lceil 0.25|\mathcal{B}| \rceil$ and let $\mathrm{Top}_m(X_i, \mathcal{B})$ return the m largest values for region i inside $\mathcal{B}$. We summarize

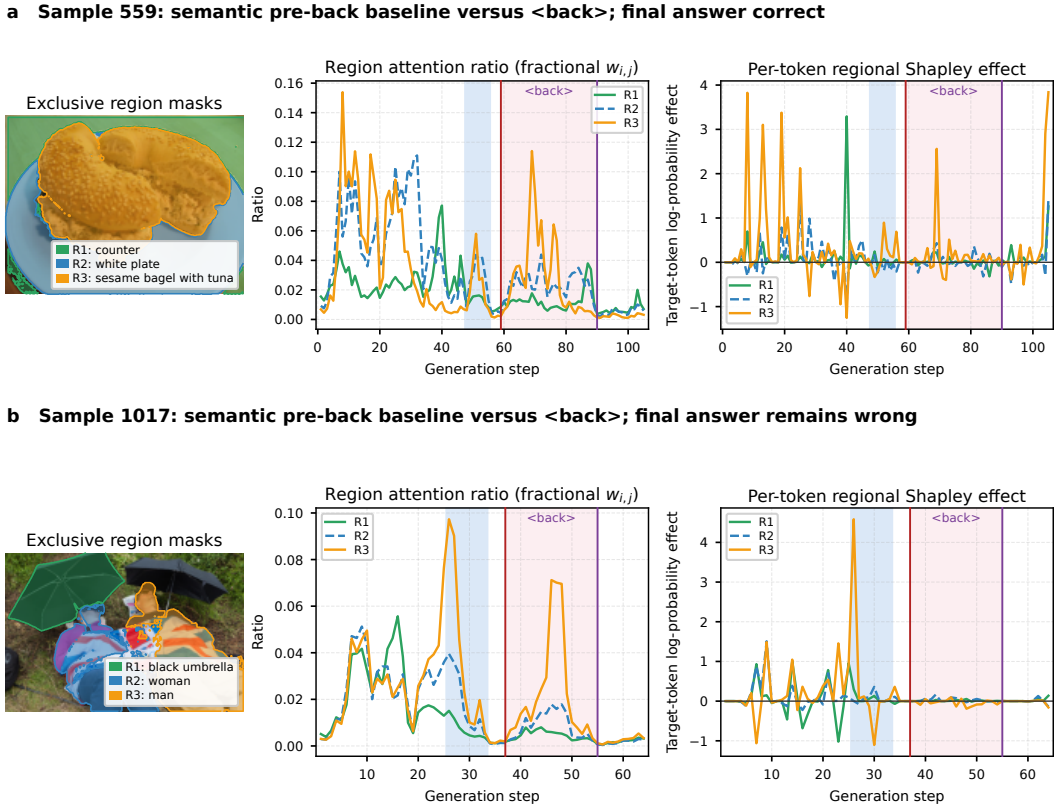


Figure 13: Token-level regional attention and effect in one correct and one incorrect Look-Back trace. The left column shows exclusive question-referenced masks; the middle and right columns show $R_{i,t}$ and signed $\phi_{i,t}$. Blue shading marks $\mathcal{P}$; red and purple lines delimit $\mathcal{B}$. **a**, Sample 559 ends with the correct Yes answer. **b**, Sample 1017 retains an incorrect No answer.

the two intervals by

$$\bar{X}_i^{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} X_{i,t}, \qquad \bar{X}_i^{\mathcal{B},\mathrm{top}} = \frac{1}{m} \sum_{x \in \mathrm{Top}_m(X_i, \mathcal{B})} x. \tag{11}$$

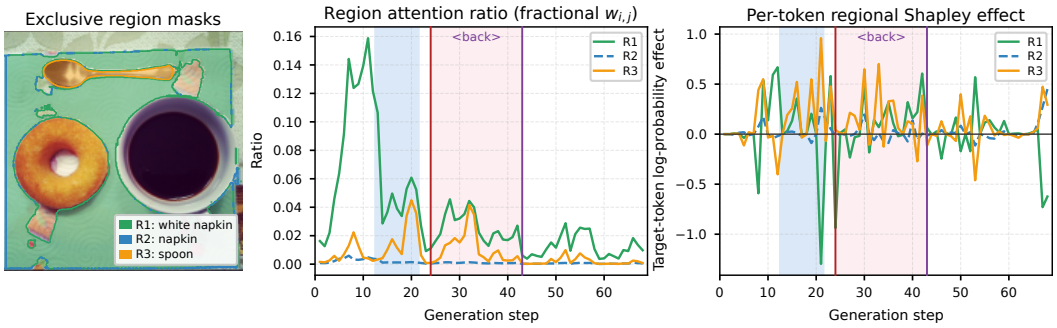
Table 12 averages these region-level values over i. Each arrow therefore compares the complete blue interval with the strongest quarter of the semantic tokens inside <back>.

Table 12: Blue-interval means and top-25% <back> means for the four traces in Figures 13 and 14. Attention and absolute token Region Effect are summarized per region and then averaged across regions.

Sample	Gold / final	Regional attention	Absolute token Region Effect
559	Yes / Yes	$0.022019 \rightarrow 0.042510$ (1.93×)	$0.1468 \rightarrow 0.3696$ (2.52×)
1017	Yes / No	$0.021215 \rightarrow 0.027968$ (1.32×)	$0.3012 \rightarrow 0.0922$ (0.31×)
193	Yes / Yes	$0.023766 \rightarrow 0.023760$ (1.00×)	$0.1846 \rightarrow 0.3591$ (1.95×)
562	Yes / No	$0.017204 \rightarrow 0.033543$ (1.95×)	$0.2197 \rightarrow 0.2411$ (1.10×)

Sample 559 revisits the bagel, plate, and counter: regional attention rises by 1.93×, absolute token effect rises by 2.52×, and the trace ends with the correct Yes answer. Sample 1017 behaves differently. Its peak regional attention rises by 1.32×, driven mainly by the region containing the man, yet peak absolute effect falls to 0.31× its pre-back level and the trace retains the incorrect No answer. The additional traces show that the two quantities can also separate in the opposite direction. In sample 193, average regional attention is nearly unchanged, while absolute token effect rises by 1.95× and the answer is correct. In sample 562, regional attention nearly doubles, but absolute token effect

c Sample 193: semantic pre-back baseline versus <back>; final answer correct



d Sample 562: semantic pre-back baseline versus <back>; final answer wrong

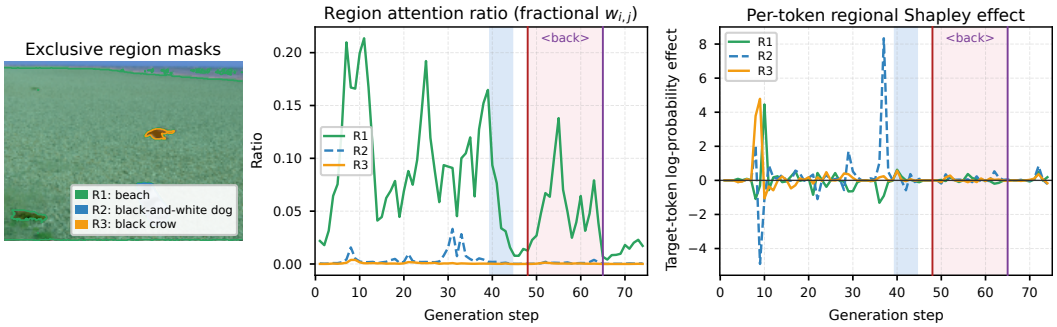


Figure 14: Two additional token-level traces using the same regions and interval definitions as Figure 13. **c**, Sample 193 ends with the correct Yes answer despite little change in average regional attention. **d**, Sample 562 retains an incorrect No answer even though regional attention rises inside <back>.

grows by only $1.10\times$ and the incorrect No answer remains. Together, the four cases sharpen the distinction suggested by the population results: attention may return without a comparable increase in the measured regional effect, while stronger regional effects can also appear without attention reactivation; neither change alone guarantees that the final answer is corrected.

D Changing How Hidden Regions Are Filled

A mask does not simply remove pixels; it also decides what appears in their place. The main experiments use the fixed constant baseline specified for each experimental configuration. We repeated selected analyses with fills taken from each image itself. The surrounding-mean fill uses the average color outside the hidden content. For PTR, we also used Telea inpainting, which reconstructs the covered object inward from its boundary. The questions, regions, saved text, and answers being scored stayed the same. Only the pixels shown in place of a hidden region changed.

D.1 PTR attribution

We selected 320 distinct PTR scenes from the full set of 886 using the scene layout, color, and relative region size, without reading model answers or Region Effects. The selection balances eight source–target part directions and smaller color–size groups. Because this gives rare groups more space than they have in the full dataset, each scene is weighted back to its original frequency. The sample and analysis rule were fixed before the alternative-fill results were inspected.

For each model, the main comparison uses the selected scenes in which the full image produced the source-color false Yes answer: 102 for Qwen2.5-VL-7B and 262 for InternVL3-8B. We ask how often the source (elsewhere-color) region's Yes-margin contribution exceeds the target region's

contribution by more than 10^{-6}. Intervals come from 10,000 paired resamples within the selection groups. Table 13 shows the result under each fill.

Table 13: PTR attribution under three ways of filling hidden pixels. Positive/n is the unweighted number of cases in which the source-region contribution exceeds the target-region contribution by more than 10^{-6}; ties count as non-positive. The last two columns use the sampling weights, with 95% intervals from paired, selection-group bootstrap resampling; the mean gap is measured on the Yes-versus-No log-score-margin scale.

Model	Hidden-region fill	n	Positive/n	Weighted source > target (% [95% CI])	Mean source − target (95% CI)
Qwen2.5-VL-7B	Main baseline	102	100/102	98.30 [95.42, 100.00]	5.856 [5.328, 6.351]
	Surrounding mean	102	101/102	99.62 [99.16, 100.00]	6.119 [5.534, 6.678]
	Telea inpainting	102	97/102	95.97 [92.44, 98.93]	3.586 [3.148, 4.039]
InternVL3-8B	Main baseline	262	253/262	96.88 [94.68, 98.80]	4.887 [4.625, 5.150]
	Surrounding mean	262	248/262	95.32 [92.74, 97.67]	4.963 [4.703, 5.226]
	Telea inpainting	262	241/262	92.15 [88.75, 95.14]	3.715 [3.435, 3.998]

The surrounding mean stays close to the original main-baseline result for both models. Telea produces smaller contributions, but the source region still outweighs the target in 95.97% of weighted Qwen cases and 92.15% of weighted InternVL cases. Table 14 compares each alternative directly with the main baseline on the same scenes.

Table 14: Paired PTR differences between each alternative fill and the main baseline. Rate and mean-gap differences use the sampling weights and are reported with paired bootstrap 95% intervals. The two p columns use the raw matched cases: the first is an exact two-sided McNemar test and the second adjusts the four tests together with Holm's method.

Model	Alternative fill	Rate difference (pp [95% CI])	Exact p	Holm p	Mean-gap difference (95% CI)
Qwen2.5-VL-7B	Surrounding mean	$+1.32$ [0.00, $+4.13$]	1.000	1.000	$+0.264$ [-0.050, $+0.569$]
Qwen2.5-VL-7B	Telea inpainting	-2.33 [-4.46, -0.37]	.250	.500	-2.270 [-2.761, -1.766]
InternVL3-8B	Surrounding mean	-1.56 [-3.69, $+0.45$]	.125	.375	$+0.077$ [-0.058, $+0.218$]
InternVL3-8B	Telea inpainting	-4.74 [-7.77, -1.95]	.00183	.00732	-1.172 [-1.422, -0.924]

The reading of the Qwen Telea difference is sensitive to how extremely small gaps are treated. If every gap above zero is counted as positive, rather than using the 10^{-6} buffer, the Telea difference narrows from -2.33 to -1.01 points and its interval reaches zero; the Holm-adjusted paired p value changes from .500 to 1.000. The InternVL Telea difference remains below zero under the same check (-3.95 points, Holm-adjusted $p = .0254$). The mean gap does not depend on this threshold and is clearly smaller under Telea for both models. For reference, the original strict-above-zero rates in the complete PTR records were 266/274 for Qwen and 752/768 for InternVL, close to the weighted processor-mean estimates in Table 13; the alternative-fill estimates were obtained from the frozen 320-scene sample and should not be extended beyond the original 886 scenes.

D.2 MM-GCoT ANSWER-SUPPORT TRAJECTORY

For all 404 InternVL3-8B MM-GCoT traces, we replaced the selected regions with the average color of the pixels left around them. The whole-image control likewise used the image's own average color. The recorded explanation and eventual answer were unchanged, so the comparison follows the same answer through the same three moments as the main analysis.

The trajectory is almost unchanged by the new fill. The two masks give the same effect direction in 96.5%–97.0% of traces at the three checkpoints, and every interval for the difference in change over time contains zero. Under the surrounding-mean fill, the probability-scale effect falls from 14.75 points after the first sentence to 1.58 points before the answer, closely matching the main trajectory.

D.3 VCR EQUAL-LENGTH CONTROL

The VCR check reuses the 500 Semantic-Back and 500 Solution-Back traces from the equal-length experiment. Their selections overlap on 389 traces and contain 611 different images, so we report the two variants separately rather than combining them. For each image, we measured the average color outside the question-referenced people and objects on the untagged image, then used that color only

Table 15: MM-GCoT masking check on all 404 traces. Panel a compares the mean Region Effect at each checkpoint under the main baseline and the image-specific surrounding mean. Δ is surrounding mean minus main baseline, with paired bootstrap 95% intervals. Panel b compares how much the trajectory changes between checkpoints.

Panel a. Checkpoint means

Checkpoint	Main baseline	Surrounding	Δ [95% CI]
Before explanation	44.483	44.823	$+0.340\,[-0.994, +1.662]$
After first sentence	14.520	14.748	$+0.228\,[-0.652, +1.119]$
Before answer	1.320	1.580	$+0.260\,[+0.027, +0.571]$

Panel b. Difference in the change over time (surrounding minus main baseline)

Transition	Probability points [95% CI]
First sentence − before explanation	$-0.112\,[-1.224, +1.054]$
Before answer − before explanation	$-0.081\,[-1.435, +1.275]$
Before answer − first sentence	$+0.031\,[-0.887, +0.935]$

inside the hidden area of the tagged image. The generated traces and the matched continuations were not changed.

Table 16 reports the genuine-minus-matched comparison used in the main text, both for all 500 traces and for the subset in which the hidden area covers at most half of the image. All effects are reported in probability percentage points.

Table 16: VCR masking check for the equal-length Look-Back comparison. Main baseline and surrounding columns show the genuine-minus-matched change in Region Effect, with bootstrap 95% intervals. Δ is surrounding mean minus main baseline. Agreement is the percentage of paired traces with the same non-zero direction, treating values within 10^{-12} of zero as ties.

Condition	n	Main baseline (95% CI)	Surrounding mean (95% CI)	Δ (95% CI)	Agreement (%)
Semantic-Back, all	500	$+0.229\,[-0.388, +0.857]$	$+0.158\,[-0.471, +0.802]$	$-0.070\,[-0.189, +0.043]$	88.00
Semantic-Back, mask ≤ 0.5	437	$+0.285\,[-0.323, +0.947]$	$+0.251\,[-0.384, +0.926]$	$-0.034\,[-0.150, +0.071]$	87.87
Solution-Back, all	500	$+0.129\,[-0.236, +0.544]$	$+0.100\,[-0.266, +0.516]$	$-0.029\,[-0.092, +0.030]$	89.20
Solution-Back, mask ≤ 0.5	434	$+0.125\,[-0.271, +0.571]$	$+0.114\,[-0.286, +0.567]$	$-0.011\,[-0.077, +0.050]$	89.17

On the probability scale, every interval includes zero under both fills, and every paired difference also includes zero. The probability-scale conclusion is unchanged in both coverage groups.

Taken together, these checks leave the main patterns in place when the hidden area is filled from the image itself. The more different Telea fill weakens the PTR effect, but the source region remains stronger than the target in most cases under both models.

E Approximating RegionTrace Beyond Exact Enumeration

Exact Region Effect requires 2^K fixed-score evaluations for K regions. The main experiments use small region sets; including more regions increases the number of required model evaluations. We tested two possible shortcuts. Sampling a small number of region orders continues to estimate the discrete quantity in Equation 6, whereas integrated gradients follows a different, continuous path through partly restored images.

E.1 Sampling a small number of region orders

For an ordering π, let P_i^π be the regions that appear before region i. Averaging the change observed when i is restored gives

$$\widehat{\phi}_{i,\tau}^{(P)} = \frac{1}{P}\sum_{p=1}^{P}\left[V_\tau\left(P_i^{\pi_p}\cup\{i\}\right) - V_\tau\left(P_i^{\pi_p}\right)\right]. \tag{12}$$

Uniformly sampled orders therefore target the same Shapley average as exact enumeration. To reduce variation, we pair every sampled order with its reverse and cache any region subset that has already been scored (Mitchell et al., 2022). Here, P counts both directions, so $P = 4$ means two sampled orders and their two reverses. If the cache already contains all 2^K subsets, we return the exact

value without another model call. The regional estimates from every complete order also sum to $V_\tau(\mathcal{N}) - V_\tau(\varnothing)$; sampling can change how this total is divided among regions, but not the total itself.

Evaluation against exact effects. We applied this estimator to cached Qwen2.5-VL-7B scores from a separate frozen PACO localization benchmark (Ramanathan et al., 2023), not the color-question set in Study I. It contains 500 PACO-LVIS validation images with one positive annotated question per image: 250 ask whether an object is present, and 250 ask whether an annotated part is visible. Each object question compares the eligible annotated objects in the image. Each part question compares the eligible parts of the named object plus the rest of that object. The numbers of images with $K = 2, 3, 4, 5, 6$ regions were 161, 136, 119, 59, and 25, respectively. For every image and budget, we used 20 fixed random streams; smaller budgets were prefixes of the larger ones. Table 17 focuses on $K = 5$ and $K = 6$. Top-1 asks whether the region with the largest estimated effect matches the unique exact winner. Sign agreement is computed only for regions whose exact magnitude is at least 5% of the largest magnitude in that image. We also compare the numerical effects and the deletion curve obtained from the estimated ordering.

Table 17: Reverse-paired order sampling on the exact PACO localization scores. Brackets give image-level bootstrap 95% intervals for Top-1 after averaging the 20 fixed random streams. Rel. ℓ_1 is the median absolute effect error divided by the total exact effect magnitude. ΔAUC is the mean absolute difference between the exact and estimated deletion-curve areas among full-image Yes cases ($n = 44$ for $K = 5$ and $n = 19$ for $K = 6$). Calls count distinct cached subsets; the call ratio is exact calls divided by sampled calls, not wall-clock speed. Deletion uses cached exact subset scores only to evaluate the estimated order and is not included in the call ratio.

K	P	Top-1 (%)	Sign (%)	Rel. ℓ_1 (%)	ΔAUC	Calls	Call ratio
5	4	83.9 [77.4, 89.3]	92.3	22.9	.0181	15.68/32	2.04×
5	8	86.3 [80.4, 91.7]	95.7	16.1	.0131	22.79/32	1.40×
5	16	90.2 [85.4, 94.7]	97.9	9.2	.0085	28.93/32	1.11×
5	32	99.2 [98.4, 99.8]	99.9	0.6	.0009	31.72/32	1.01×
6	4	86.0 [75.2, 95.0]	95.3	17.9	.0178	19.96/64	3.21×
6	8	87.0 [76.6, 96.0]	97.2	12.4	.0137	31.71/64	2.02×
6	16	89.6 [81.0, 96.8]	98.3	9.6	.0126	45.73/64	1.40×
6	32	92.6 [85.4, 98.4]	99.4	6.7	.0069	57.57/64	1.11×

Across all 500 images, four orders gave 95.1% Top-1 agreement. The benchmark includes 297 images with only two or three regions, where caching frequently reconstructs the exact enumeration, so we additionally report the more informative $K = 5$ and $K = 6$ strata. With four orders, sampling retained the exact Top-1 region in 83.9% and 86.0% of cases while reducing model calls by $2.04\times$ and $3.21\times$, respectively. Increasing the budget steadily improved agreement: at $P = 16$, Top-1 reached 90.2% for $K = 5$ and 89.6% for $K = 6$; at $P = 32$, it reached 99.2% and 92.6%. For $K = 5$, the $P = 32$ cache covered every subset in 86.8% of images, explaining its near-exact result. These operating points allow the sampling budget to be selected according to use: small budgets prioritize faster identification of influential regions, whereas larger budgets more closely reproduce the exact ordering.

We separately checked whether fewer calls produced a real time saving on server GPU 0. For one frozen image at each of $K = 4, 5, 6$, with two timing repeats per method, four orders were $1.35\times$, $2.00\times$, and $2.91\times$ faster than exact enumeration. Eight orders were $1.16\times$, $1.46\times$, and $1.89\times$ faster. Model loading and free answer generation were excluded from both methods; image construction, preprocessing, and scoring were included. Peak allocated GPU memory was essentially unchanged at about 15.58 GiB. All three examples retained the exact Top-1 region, but three images are a timing and implementation check, not an accuracy estimate.

Taken together, exact enumeration remains convenient for small region sets and for analyses that require exact numerical effects. Paired order sampling extends RegionTrace to larger region sets with a tunable accuracy–cost balance: a small budget can identify influential regions at $K = 5$ or $K = 6$, and the budget can be increased when closer agreement is required. Natural $K \geq 7$ region sets provide the next setting in which to study this balance.

E.2 Integrated gradients: a negative pilot

We also replaced each binary region switch with a continuous gate and moved all gates together from the mean-filled image to the complete image. This is straight-line integrated gradients applied to the region gates (Sundararajan et al., 2017). It asks how the score changes along one path of partly restored images. Even with numerically exact integration, this need not divide interactions among regions in the same way as the discrete Shapley average (Sundararajan & Najmi, 2020).

The pilot was run separately on three frozen PACO part-color questions. Each image had three regions: the queried part, another part carrying the queried color, and the remainder of the object. We used Qwen2.5-VL-7B, the same first-answer Yes-minus-No score as Study I, and 16 integration intervals. It was a numerical check rather than a hallucination-rate experiment: the cases were not filtered by the model's answer, and only the knife item was a full-image Yes. Exact enumeration required eight forward evaluations. IG required 17 forward and backward evaluations. Only two images had a unique exact Top-1 region, and the 16-step estimate matched one of them. Across the nine region scores, the positive or negative direction matched in four cases (44.4%); the median relative ℓ_1 error was 682.7%. For an exact line integral of a score that is differentiable almost everywhere, completeness requires the three values to reproduce the score change between the two path endpoints. Our 16-step numerical estimate instead had absolute errors of 2.626, 4.555, and 3.091 score units. The corresponding median relative error was 88.3%.

One failure makes the problem concrete. For "Is purple the main color of the text of the jar?", exact enumeration assigned $+1.125$ to the queried text and -0.563 to the rest of the object. The 16-step IG estimate assigned -3.535 and $+6.499$, respectively, reversing both the signs and the largest region. In the knife example, the estimate did identify the same largest region, but its three values summed to 2.195 while the measured endpoint change was 6.750. In the guitar example, the exact target and object remainder tied at $+1.396$; the estimate instead assigned -9.391 to the target and $+16.208$ to the remainder. A 32-point Gauss–Legendre check on the jar image, in both bfloat16 and float16, still did not repair the endpoint-sum error.

The failure was not explained by using different prompts or endpoint images. Recomputed exact subset scores matched the saved scores within 2.4×10^{-7}, and the image-processor decomposition residual was below 0.07%. Inspection instead found narrow, opposing gradient spikes as the regions were partly restored, so practical low-precision integration missed large cancellations. More integration points may reduce this error, but they also weaken the intended speed advantage.

The cost result pointed in the same direction. Median model time was 2.95 seconds for exact enumeration and 16.17 seconds for 16-step IG, making IG $5.38\times$ slower in this three-region pilot. Median peak allocated memory rose from 15.66 to 19.84 GiB because IG also requires backward computation. These three images are too few to estimate general IG performance, and they do not show that integrated gradients fails in other models or along other paths. They do show that this implementation failed both the numerical check and the cost goal for RegionTrace, so we did not expand it. Random region orders remain the more direct approximation to test at larger K.

F From Token Localization to Longitudinal Answer-Support Auditing

An aligned decision-support field. RegionTrace follows visual support for the *same answer candidate* along the realized generation trace. Fixing the candidate separates changes in its support from changes in the output being explained; retaining the visible region set makes context dependence explicit. For a fixed scoring rule, let d denote the candidate and t the generation stage. The value function and regional marginal effect are

$$V_{t,d}(S) = \mathrm{score}\big(d \mid q, \mathrm{prefix}_t, I_S\big), \qquad \Delta_{i,t,d}(S) = V_{t,d}(S \cup \{i\}) - V_{t,d}(S). \tag{13}$$

For each checkpoint t, the field $\{\Delta_{i,t,d}(S)\}_{i,S}$ records whether a named region supports or suppresses candidate d under every visible context S. Because the candidate and semantic-region vocabulary stay fixed, these fields are directly comparable before, during, and after explanation generation, including checkpoints before the answer is emitted. RegionTrace therefore resolves three coordinates of answer support at once: *where* it lies, *when* it changes, and *under which visual context* it is sustained.

Table 18 situates this measurement object among representative visual-attribution outputs. Perturbation methods localize evidence for one score (Fong & Vedaldi, 2017; Petsiuk et al., 2018); semantic-output Shapley methods distribute changes in a regenerated response (Cafagna et al., 2023; Goldshmidt, 2025); multimodal Shapley profiles measure input contribution and explanation consistency (Parcalabescu & Frank, 2023; 2025); and token maps explain individual decoding states (Li et al., 2025). RegionTrace unifies spatial localization with candidate alignment, signed context dependence, and the generation axis.

Table 18: Primary measurement objects of representative visual-attribution approaches in their published formulations. RegionTrace follows the same decision through generation while retaining the signed effects of named regions in their visual contexts.

Approach	Primary measured object	Visual unit	Native generation alignment
Fixed-output perturbation / RISE	Target-score response to learned or random masks	Pixels / patches	One score at one input state
Semantic-output SHAP	Shapley attribution of change in a regenerated response	Semantic / object masks	One completed output
MM-SHAP / CC-SHAP	Input contribution or consistency profile for a prediction	Modalities / patches / input tokens	Prediction or each token's own prefix
TAM	Internal activation map for a generated token	Visual tokens / patches	Token's decoding state
EAGLE	Sufficiency–necessity ordering and token visual-dependence score	SLICO superpixels	Targets' original prefixes
RegionTrace	**Signed support field for the same answer candidate**	**Named semantic regions**	**Aligned prefixes before, during, and after generation**

F.1 Matched-region comparison with static attribution

We first evaluate the regional summaries at a fixed checkpoint to test their localization performance under matched inputs. The comparison uses the separate frozen PACO localization benchmark introduced in Appendix E, rather than the color-question set in Study I. It contains 500 positive object- or part-presence questions, $K = 2$–6 eligible semantic regions per image, and all $7{,}124$ coalition scores cached exactly for Qwen2.5-VL-7B. We evaluate the 366 cases in which the full-image answer is Yes: 222 object questions and 144 part questions. Every method receives the same fixed Yes probability, semantic partition, replacement operator, and coalition table, so the comparison isolates the regional summary rather than the proposal mechanism. No additional model inference is required.

The two RegionTrace summaries are exact Shapley averaging from Equation 6 and a uniform-context average,

$$\bar{\Delta}_i^{\mathrm{unif}} = \frac{1}{2^{K-1}} \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \left[V(S \cup \{i\}) - V(S) \right]. \tag{14}$$

We compare them with full-context leave-one-out occlusion and empty-context insertion,

$$O_i^{\mathrm{full}} = V(\mathcal{N}) - V(\mathcal{N} \setminus \{i\}), \qquad O_i^{\mathrm{empty}} = V(\{i\}) - V(\varnothing). \tag{15}$$

We also instantiate EAGLE's published sufficiency–necessity objective (Chen et al., 2026) on the same semantic-region lattice,

$$F(S) = V(S) + 1 - V(\mathcal{N} \setminus S), \tag{16}$$

and greedily add the region that maximizes F at each step. This equation-matched adaptation holds the semantic regions and score evaluation fixed, isolating the attribution objective from region proposal. Largest-area and random orders are non-attribution controls.

Gold Top-1 asks whether the annotated answer region is ranked first. Probability-drop AUC removes regions in the predicted order and integrates the drop from the full-image Yes probability against the fraction of candidate-region pixels removed; higher is better.

Table 19: Matched-region localization on the frozen PACO benchmark ($n = 366$ full-image positive cases). Every row uses the same semantic regions, fixed target, probability score, replacement operator, and exact coalition table.

Regional summary	Gold Top-1 (%) ↑	Drop AUC ↑
RegionTrace: exact Shapley	78.4	.387
RegionTrace: uniform contexts	78.1	.386
EAGLE-style sufficiency–necessity	76.5	.379
Full-context LOO occlusion	76.0	.364
Empty-context insertion	76.2	.370
Largest region	41.5	.204
Random expectation	34.5	.253

Exact Shapley achieves the highest point estimates in Table 19: 78.4% Gold Top-1 and .387 Drop AUC, compared with 76.5% and .379 for the matched sufficiency–necessity objective. The gains are 1.9 percentage points in Top-1 and .008 in AUC, approximately 2.5% and 2.1% relative improvements, respectively. Uniform context averaging yields 78.1% and .386, close to the Shapley estimates, suggesting that considering multiple contexts accounts for much of the observed localization benefit. The AUC gains over full-context LOO and empty-context insertion are .0227 [.0165, .0293] and .0172 [.0136, .0213], respectively, using $10{,}000$ paired image-level bootstrap resamples. Both intervals exclude zero. These results validate regional summaries at a fixed checkpoint; the next analysis examines the information retained before that summarization.

F.2 Endpoint attribution discards visual context

The added information is empirically consequential. Full-context LOO measures a region after all other selected evidence is already visible, whereas empty-context insertion measures it before any other selected region is restored. We computed both quantities from the exact Study I coalition tables for every false-Yes answer to the elsewhere-color question: $2{,}119$ model–question cases across two datasets and three models. For each case, we record whether the two endpoints produce different Top-1 sets and whether any region switches direction between the endpoints. To exclude numerical sign changes, a switch requires $O_i^{\text{empty}} > 0.05$ and $O_i^{\text{full}} < -0.05$, or the reverse, in Yes-minus-No log-score units.

Table 20: Endpoint context audit on Study I false-Yes cases. "Different Top-1 set" means full-context LOO and empty-context insertion produce different sets of maximally scored regions, with numerical ties retained. "Direction switch" requires both endpoint effects to exceed 0.05 in magnitude and to have opposite signs.

Dataset	Model	n	Different Top-1 (%)	Direction switch (%)
PTR	Qwen2.5-VL-7B	274	8.4	63.9
PTR	InternVL3-8B	768	10.7	67.2
PTR	InternVL3-14B	475	18.9	65.5
PACO	Qwen2.5-VL-7B	115	27.0	66.1
PACO	InternVL3-8B	287	30.3	66.9
PACO	InternVL3-14B	200	30.5	63.0

On natural PACO errors, changing only the occlusion endpoint changes the leading region set in 27.0–30.5% of cases—nearly one in three. More fundamentally, 63.0–67.2% of false-Yes cases in every dataset–model setting contain a region that supports the answer at one endpoint and suppresses it at the other. The result is not driven by near-zero effects: increasing the tolerance from 0.05 to 0.125 log-score units still yields direction switches in 50.0–60.7% of cases across the six settings. A single deletion score, insertion score, or compact ordering cannot encode this support-to-suppression transition. RegionTrace retains it directly in $\Delta_{i,t,d}(S)$ and then follows the same field over t.

Connecting localization, context, and generation. The two experiments evaluate complementary aspects of the measurement. The localization benchmark tests the regional summaries under matched inputs. The context audit shows that a single summary can conceal changes in both the leading region

and the direction of regional influence. Retaining the context-specific effects yields the trajectory

$$t \longmapsto \big\{\Delta_{i,t,d}(S) : i \in \mathcal{N},\ S \subseteq \mathcal{N} \setminus \{i\}\big\}, \tag{17}$$

which aligns the candidate and semantic regions across checkpoints. Study I uses regional effects to localize spatial attribute misbinding and examine visual-context dependence. Studies II and III use group contrasts to follow aggregate support for the same candidate during explanation and a marked visual revisit. These readouts connect the location of supporting evidence, its dependence on other regions, and its changing contribution as the textual prefix grows.

F.3 Time and memory cost

Table 19 compares regional summaries using the same image regions. Here we compare time and memory when each method uses its own regions. All attribution runs use Qwen2.5-VL-7B on one NVIDIA A40, with BF16 and SDPA. RegionTrace uses 2–6 named regions, with either supplied masks or masks generated by Grounding DINO and SAM; EAGLE (Chen et al., 2026) divides each image into 59–68 SLICO segments. We test EAGLE's full search, full search with batch size 8, and efficient V2 search with batch size 8.[5]

Table 21: Time and memory per image on the same 37 PACO localization images for a fixed answer. Time covers the steps described below. Peak memory is the largest per-image peak in allocated GPU memory, including the models. The automatic-mask row includes region construction and attribution, with Grounding DINO, SAM, and Qwen2.5-VL all resident on the GPU.

Configuration	Mean time (s)	Median time (s)	Peak (GiB)
RegionTrace: supplied masks	3.21	1.70	**15.61**
RegionTrace: DINO + SAM + attribution	**4.11**	**2.56**	18.07
EAGLE: full, serial	921.29	844.34	16.44
EAGLE: full, batch 8	752.71	693.04	17.46
EAGLE: efficient V2, batch 8	168.04	159.65	18.05

RegionTrace directly scores subsets of semantic regions, avoiding EAGLE's iterative search over image segments. When boxes and masks are unavailable, automatically constructing the regions adds about 0.9 s per image. The complete pipeline averaged 4.11 s, about $41\times$ faster than the tested efficient EAGLE configuration. Thus the time advantage remains after including automatic region construction. With all three models resident, peak memory was comparable to efficient EAGLE (18.07 versus 18.05 GiB); starting from supplied masks required 15.61 GiB. RegionTrace scored 16.5 image versions per case on average, compared with 941.6 for efficient EAGLE and 4,147.7 for full search. These counts include images scored in batches. These results demonstrate an efficiency advantage over the tested EAGLE pipelines. Appendix E examines sampling when more regions are included.

Timing setup. Each image is timed once per configuration after warmup. Timing includes image preparation, model scoring, and the final region ranking. EAGLE's time also includes segmentation and heatmap construction. Model loading, disk input/output, and validation are excluded. RegionTrace scores Yes relative to the Yes–No alternatives; EAGLE scores one fixed Yes token over the full vocabulary. Hidden content is filled with the model's RGB mean for RegionTrace and black for EAGLE. RegionTrace runs first; the three EAGLE configurations rotate their order across images.

Automatic region construction. We additionally ran Grounding DINO-tiny and SAM ViT-B on the same 37 images and the same A40, using the fixed region names but no annotated boxes or masks. The detector searches the full image for each name, and SAM refines the predicted boxes into masks; overlapping pixels are assigned to the smaller region.[6] We then use these generated masks directly in exact regional attribution. The 4.11 s total is measured from detection through attribution in one run. A separate three-pass check of region construction averaged 0.90, 0.89, and 0.88 s per image. Both

[5]The batched full search uses the official chunked scoring routine. Efficient V2 uses search scope 8, 4 pending samples, and update step 10.

[6]We select the highest-scoring boxes, with non-maximum suppression for repeated object names. The 126 requested regions yielded 123 non-empty masks after overlap removal; all requested regions were retained for timing.

region models use their default full precision; timing includes their preprocessing and mask cleanup, with the same exclusions for model loading and disk access as above.

Retained sample. We initially sampled 50 of the 366 full-image Yes cases, grouping by task type and region count and sampling across image sizes. We stopped the run after finding other jobs sharing the GPU. We retain 37 images whose four runs finished before those competing processes started: 24 object and 13 part questions, covering $K = 2$–6. All table entries use these same 37 images. Known shared-use periods are excluded, but GPU use was not logged continuously; uninterrupted exclusive access therefore cannot be confirmed.

G Attention Routes Information; RegionTrace Measures Answer Support

Attention is a meaningful description of internal routing, but routing and answer support are different mathematical objects. The debate over attention as explanation has established that raw attention need not track output importance, while also cautioning against the blanket claim that attention is never explanatory (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Subsequent work separates attention weights from transformed values, identifies null spaces in which attention can change without changing a head output, and derives the relation between attention and post-hoc attribution in restricted architectures (Brunner et al., 2020; Kobayashi et al., 2020; Lopardo et al., 2024). Here we characterize the distinction for the object measured by RegionTrace: the signed change in a fixed candidate's score after an observed image region is restored. We first relate attention to regional support, then examine the role of the target score and changes across textual and visual contexts.

Fix a checkpoint t, candidate d, and visible context $S \subseteq \mathcal{N} \setminus \{i\}$, and abbreviate

$$\Delta_{i,t,d}(S) = V_{t,d}(S \cup \{i\}) - V_{t,d}(S). \tag{18}$$

For later use, write $w(S) = |S|!(K - |S| - 1)!/K!$ for the Shapley weight in Equation 6. We use the checkpoint-explicit notation

$$\phi_{i,t,d} = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} w(S)\Delta_{i,t,d}(S). \tag{19}$$

For a specified attention head and query, let j range over all key–value source positions, let $a_j \geq 0$ be the corresponding normalized attention weight, and let R_i be the subset of visual positions associated with semantic region i. The regional attention mass is $A_i = \sum_{j \in R_i} a_j$.

G.1 Attention and regional support

Proposition G.1 (Attention non-identification)**.** *There is no function of raw regional attention alone that recovers $\Delta_{i,t,d}(S)$ for every attention model and input. The same attention vector can accompany arbitrary regional-effect signs and rankings and, for an unconstrained scalar score, arbitrary magnitudes.*

Proof. Take any strictly positive attention vector $a = (a_1, \ldots, a_K)$ with $\sum_j a_j = 1$, and any desired signed effect vector $e = (e_1, \ldots, e_K)$. Consider one visual token per region and a single-head scalar readout whose attention weights are fixed across coalitions,

$$V(S) = \sum_{j=1}^{K} a_j v_j(S), \qquad v_j(S) = \begin{cases} e_j/a_j, & j \in S, \\ 0, & j \notin S. \end{cases} \tag{20}$$

Then $V(S) = \sum_{j \in S} e_j$, and hence $V(S \cup \{i\}) - V(S) = e_i$ for every S not containing i. The attention vector is unchanged while e is arbitrary, so attention alone cannot determine the regional effect. □

The construction already holds in the regime most favorable to attention: one head, fixed routing, and a linear target score. For example, attention masses $(0.9, 0.1)$ and target-projected value contrasts $(-1, 20)$ yield effects $(-0.9, 2.0)$: the high-attention region suppresses the target, while the low-attention region supports it more strongly. Attention remains informative about routing;

Proposition G.1 establishes that routing weights do not contain the target-directed content carried along the route.

The relationship can be made explicit without assuming a linear network. Let $I(\lambda)$ continuously restore region i into context S, with $\lambda = 0$ denoting I_S and $\lambda = 1$ denoting $I_{S\cup\{i\}}$. At one attention head, write

$$h(\lambda) = \sum_j a_j(\lambda) v_j(\lambda), \qquad V(\lambda) = G\big(h(\lambda), r(\lambda)\big), \tag{21}$$

where $r(\lambda)$ collects all other heads, residual streams, and paths that bypass this head.

Proposition G.2 (Attention-mediated effect decomposition)**.** *If every $a_j(\lambda)$, $v_j(\lambda)$, and $r(\lambda)$ is absolutely continuous, and G is continuously differentiable on a neighborhood of the path, then*

$$\begin{aligned}\Delta_i(S) = &\int_0^1 \nabla_h G^\top \sum_j a_j(\lambda)\dot{v}_j(\lambda)\, d\lambda \\ &+ \int_0^1 \nabla_h G^\top \sum_j \dot{a}_j(\lambda) v_j(\lambda)\, d\lambda + \int_0^1 \nabla_r G^\top \dot{r}(\lambda)\, d\lambda.\end{aligned} \tag{22}$$

The three terms are respectively value transmission, attention rerouting, and contribution through other paths.

Proof. Apply the fundamental theorem of calculus to $V(1) - V(0) = \int_0^1 \dot{V}(\lambda) d\lambda$, followed by the chain and product rules for $G(h(\lambda), r(\lambda))$ and $h(\lambda) = \sum_j a_j(\lambda) v_j(\lambda)$. □

Dots denote derivatives with respect to λ, and every gradient in Equation 22 is evaluated at $(h(\lambda), r(\lambda))$. The three components depend on the chosen restoration path, attention head, and computational cut; their sum, but not the individual decomposition, is fixed by the two endpoint scores. Equation 22 shows that attention is one routing factor inside the complete output effect, rather than a competing estimate of the same quantity. It also yields the exact special case in which the two become proportional.

Corollary G.3 (When attention equals Region Effect)**.** *Suppose that restoring region i leaves attention fixed, changes only values in R_i, has no effect through r, and the target score is affine in the head output, $V = \beta^\top h + b$. Then*

$$\Delta_i(S) = \sum_{j\in R_i} a_j(S)\, \beta^\top \big(v_j^{\text{visible}} - v_j^{\text{base}}\big). \tag{23}$$

If every target-projected value contrast in every region equals the same constant $c > 0$, then $\Delta_i(S) = cA_i(S)$.

Here v_j^{visible} and v_j^{base} are the endpoint values for the specified context S. Frozen routing, an affine readout, no alternative visual path, and uniform target-relevant value contrasts are sufficient to guarantee proportionality. If these conditions hold for every coalition with the same constant c, the Shapley summary is proportional to *coalition-averaged* attention,

$$\phi_{i,t,d} = c \sum_{S\subseteq\mathcal{N}\setminus\{i\}} w(S) A_i(S), \tag{24}$$

not necessarily to the full-image attention map. Equality with that single map additionally requires regional attention to remain invariant across visual contexts.

G.2 Target scores and Shapley attribution

Studies II and III use probability normalized over a fixed candidate set $\mathcal{D}$. This score imposes a conservation law that a nonnegative attention allocation does not express.

Proposition G.4 (Candidate-support conservation)**.** *Let $V_{t,d}(S) = p_{t,d}(S)$ with $\sum_{d\in\mathcal{D}} p_{t,d}(S) = 1$ for every coalition S. Then*

$$\sum_{d\in\mathcal{D}} \Delta_{i,t,d}(S) = 0, \qquad \sum_{d\in\mathcal{D}} \phi_{i,t,d} = 0. \tag{25}$$

For two candidates, the regional effects are exact opposites.

Proof. The first identity follows by subtracting two probability vectors whose components each sum to one. The second follows by averaging the first identity with the same Shapley weights for every candidate. □

A nonzero probability-scale regional effect therefore supports some candidates while suppressing others. Raw attention is a nonnegative route allocation; without target-specific values and an output readout, it does not encode how probability is redistributed among candidates. This proposition is specific to the candidate-probability scale and is not invoked for the binary log-score margin in Study I.

The choice of score also matters when comparing Shapley attributions. Attention rollout and attention flow aggregate routing across layers rather than using a single raw map (Abnar & Zuidema, 2020). Attention flow can itself be a Shapley value when the payoff is the internal flow through a layer (Ethayarajh & Jurafsky, 2021). RegionTrace instead uses the fixed answer score as its payoff. Being a Shapley value is therefore not sufficient to make the two explanations equivalent.

This distinction has a quantitative form. After mapping visual players to the same semantic-region set, let $F(S)$ be an internal-flow payoff and write its Shapley value as $\phi_i(F) = \sum_S w(S)[F(S \cup \{i\}) - F(S)]$. Suppose constants γ, b satisfy

$$\max_{S \subseteq \mathcal{N}} \left| V_\tau(S) - \big(\gamma F(S) + b\big) \right| \leq \epsilon. \tag{26}$$

Then Shapley linearity and the fact that its marginal weights sum to one give

$$|\phi_i(V_\tau) - \gamma \phi_i(F)| \leq 2\epsilon. \tag{27}$$

Uniform payoff closeness is therefore a sufficient condition guaranteeing that flow attribution approximates answer attribution. It is not necessary: two payoffs can differ substantially while their residual game has zero Shapley value. The fact that both methods use Shapley averaging does not itself establish payoff alignment.

G.3 Effects across textual and visual contexts

Attention maps at two decoding points are two internal snapshots. RegionTrace instead compares the same intervention and target across those points.

Proposition G.5 (Prefix–region mixed difference)**.** *For checkpoints $t < t'$, define*

$$\begin{aligned} M_{i,t\to t',d}(S) &= \Delta_{i,t',d}(S) - \Delta_{i,t,d}(S) \\ &= V_{t',d}(S \cup \{i\}) - V_{t,d}(S \cup \{i\}) \\ &\quad - V_{t',d}(S) + V_{t,d}(S). \end{aligned} \tag{28}$$

This is the discrete interaction between prefix growth and region restoration on the fixed score scale. Moreover,

$$\phi_{i,t',d} - \phi_{i,t,d} = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} w(S) M_{i,t\to t',d}(S). \tag{29}$$

Proof. The first identity follows by substituting the definition of Δ and rearranging. For the second, define $W(S) = V_{t',d}(S) - V_{t,d}(S)$ and apply Shapley linearity to W. □

Negative M means that the region's marginal effect is smaller at the later fixed prefix, while positive M means that it is larger. When both endpoint effects are nonnegative, these cases can additionally be read as conditional substitution and complementarity, respectively. Summing Equation 29 over regions also gives

$$\begin{aligned} \sum_i (\phi_{i,t',d} - \phi_{i,t,d}) &= [V_{t',d}(\mathcal{N}) - V_{t,d}(\mathcal{N})] \\ &\quad - [V_{t',d}(\varnothing) - V_{t,d}(\varnothing)], \end{aligned} \tag{30}$$

so the total attribution change equals the change in complete-versus-replaced image contrast. This is the longitudinal quantity used in Studies II and III. Because the later prefix was originally generated

by the model, M is a teacher-forced conditional interaction, not the causal effect of generating that prefix.

The context dependence in Study I admits a sharper interpretation. For fixed t, d, define the Harsanyi dividend of coalition U (Harsanyi, 1963) as

$$h(U) = \sum_{W \subseteq U} (-1)^{|U|-|W|} V(W). \tag{31}$$

The Möbius inversion identity gives

$$\Delta_i(S) = \sum_{T \subseteq S} h(T \cup \{i\}). \tag{32}$$

The Shapley summary equivalently distributes each dividend equally among the regions participating in it,

$$\phi_i = \sum_{U \ni i} \frac{h(U)}{|U|}. \tag{33}$$

Proposition G.6 (Interaction-dominance certificate)**.** *Let*

$$H_i = \sum_{\substack{U \ni i \\ |U| \geq 2}} h(U) \tag{34}$$

be the sum of all higher-order dividends involving region i. The two occlusion endpoints satisfy

$$O_i^{\text{empty}} = h(\{i\}), \qquad O_i^{\text{full}} = h(\{i\}) + H_i. \tag{35}$$

If the endpoints have opposite signs, then H_i has the opposite sign from $h(\{i\})$ and $|H_i| > |h(\{i\})|$.

Proof. Equation 35 follows from Equation 32 at $S = \varnothing$ and $S = \mathcal{N} \setminus \{i\}$. If $h(\{i\})$ and $h(\{i\}) + H_i$ have opposite signs, H_i must cross zero and therefore oppose and exceed the singleton term in magnitude. □

The direction switches in Table 20 are therefore not merely disagreements between two attribution heuristics. In 63.0–67.2% of Study I false-Yes cases, at least one region has a higher-order interaction with the other visible regions whose aggregate contribution is strong enough to reverse the region's singleton direction. This is precisely the context structure retained by the RegionTrace field and discarded by a single attention map or occlusion endpoint.

These propositions formalize the distinction between attention-based routing and intervention-based answer support. Attention describes the internal routes through which visual information may travel. RegionTrace asks what happens after routing, value transmission, downstream transformations, and bypass paths have taken effect: how much does restoring the observed content of a region change a fixed target? Its answer is conditional on the selected regions, replacement baseline, prefix, candidate, and score. It does not identify a unique internal mechanism or an intervention-independent notion of causality.

H Extended Related Work

Reasoning traces and grounded explanations. Chain-of-thought prompting and its multimodal variants expose intermediate text and can improve multistep performance (Wei et al., 2022; Zhang et al., 2024). Grounding methods test or improve the visual consistency of these steps (Chen et al., 2024; Wu et al., 2025), while controlled evaluations show that a rationale can still omit factors that changed the answer (Turpin et al., 2023; Balasubramanian et al., 2025). During long generation, early multimodal errors can snowball, linguistic context can increasingly govern later tokens, and measured visual use can decline (Zhong et al., 2024; Zheng et al., 2025; Jin et al., 2026). Look-Back explicitly revisits visual evidence (Yang et al., 2026). RegionTrace makes this evolution measurable by saving the realized prefix, fixing one answer candidate, and comparing its regional support before, during, and after the explanation.

Attribution and intervention. Attention and activation maps show large signals; input perturbations test output sensitivity to image content (Jain & Wallace, 2019; Adebayo et al., 2018; Fong & Vedaldi, 2017; Petsiuk et al., 2018). Multimodal and semantic Shapley methods attribute a prediction or regenerated response to modalities, patches, or regions (Cafagna et al., 2023; Parcalabescu & Frank, 2023; 2025; Goldshmidt, 2025; Samyoun et al., 2026), while token methods map generated or hallucinated spans (Li et al., 2025; Park et al., 2025). RegionTrace aligns a fixed candidate and semantic-region vocabulary across prefixes, preserving the visual context of each measured effect. Mechanistic methods locate internal mediators (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Palit et al., 2023); RegionTrace instead compares semantic regions, visual contexts, and checkpoints at the input–output boundary.