POINTER–PAYLOAD–HANDOFF: TRACING HOW VISUAL EVIDENCE REACHES THE ANSWER IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A vision–language model can attend to the right region and still answer incorrectly. Spatial grounding identifies *where* a model looks, not *how* the selected fact reaches the computation that chooses the answer. In Qwen3.5, the strongest spatial pointer even appears *after* the image-to-question handoff, and transplanting the grounded attention contribution itself fails to repair answers. We trace one answer-determining visual fact by asking four questions: where it is first causally available, which computation moves it downstream, where the transferred content is carried, and through which state it reaches the output. **Pointer–Payload–Handoff (PPH)** answers these questions with three complementary interventions—restoration, blocking, and semantic interchange—at coordinates frozen before held-out confirmation. Independent analyses in Qwen3.5, Qwen3-VL, and Gemma 3 recover the same ordered role graph $S \to W \to C \to R$ at architecture-specific depths. Writers range from one dominant attention component to distributed modules, yet every family certifies the directed links $S \to W$, $S \to C$, and $C \to R$. The $W \to C$ adjacency holds in Qwen3.5-9B, where it is measurable. The frozen carrier is causally usable: replacing it with a state built from correct visual evidence repairs 19–35% of initially incorrect real-image answers across the three families, against near-zero matched controls. With the task-relevant option text present, the carrier mediates 58–89% of the source effect across the three families. Replacing that block with equal-length unrelated text reduces mediation by 54 points in each family; instructing the model to ignore the options leaves the point estimate close to the original. These controls isolate the effect of the option text, not the answer word within it. Interpreting a vision–language model therefore requires stating the prompt condition under which its mechanism was recovered.

1 INTRODUCTION

A vision–language model (VLM) can look at the right object and still give the wrong answer. Asked for the color of a red object, its attention may concentrate on the correct region while the model answers "blue." The spatial map does not reveal whether the model failed to extract *red*, extracted it but never passed it into the decision computation, or made the fact available internally but failed to use it at the answer position. These are different mechanisms behind the same wrong answer, and mechanistic interpretability needs to tell them apart.

Locating information is not the same as explaining its use. Spatial attention, probing, and state-level attribution establish where a fact *is*. A causal explanation must show how that fact changes the output. Prior work on visual information flow traces states across depth, but typically localizes states or paths rather than certifying which state plays which causal role on data that did not participate in the search.

We study the provenance of one answer-determining visual fact through four questions: where is the selected evidence first causally available, which computation moves part of its effect downstream, where is the transferred content carried, and through which answer-position state does the effect reach the output? We call these functional roles source S, writer W, carrier C, and readout R. They deliberately mix state and computation roles: S, C, and R describe where evidence is available,

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.

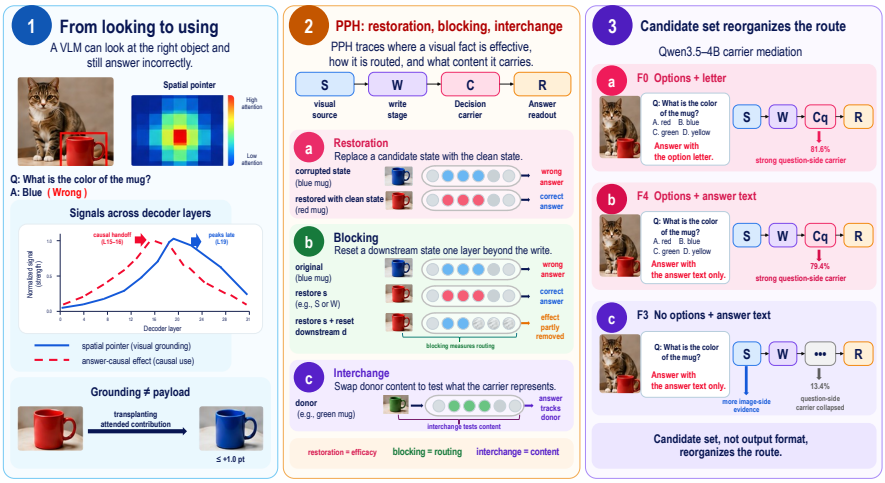

Figure 1: **From spatial grounding to causal use.** The spatial pointer peaks after the image-to-question handoff, while transplanting the attended contribution changes accuracy by at most $+1.0$ point. PPH assigns source, writer, carrier, and readout through restoration, blocking, and interchange; the option text in the prompt determines whether the route uses a question-side carrier.

whereas W identifies the computation that transfers an upstream effect. The carrier is not defined to occupy text tokens. Its token space is an empirical result.

The answer we find has two parts. When the answer string appears in the prompt text, the fact moves from an early image state through a mid-layer write into a question-side carrier that the readout consumes. When it does not, the causal leverage stays in image-side states.

Pointer–Payload–Handoff (PPH) assigns these roles with complementary interventions. The *pointer* is the localized visual support, the *payload* is its answer-relevant content, and the *handoff* records a change in where that content is causally available. Restoration establishes that a state is causally effective, directed blocking establishes that roles depend on one another, and semantic interchange establishes which content a state actually carries. Two design choices keep the graph falsifiable. First, coordinates are frozen on validation data before any role is evaluated on held-out examples. Second, every block is placed where the two roles occupy different positions or components, so the directed tests measure transfer rather than restating where a patch was written.

Qwen3.5 shows the separation directly. Enrichment over the correct region keeps rising after image-side causal effects begin to recede and peaks at a later layer, yet transplanting that localized attention contribution changes accuracy by at most $+1.0$ point. PPH instead recovers a directed route through the question-side carrier: the dominant writer intercepts 59.8% of the restored source effect, and 81.6% of that effect is routed through the carrier state. The carrier's content tracks the donor. Blocked shares measure routing between roles. Content specificity is established separately by GT-versus-control magnitudes, interchange, and real-image repair (Section 3).

The role order recurs after independent coordinate discovery in Qwen3-VL and Gemma 3. Qwen3.5 has a concentrated write, Qwen3-VL distributes it, and Gemma uses several adjacent modules, yet each family recovers the same ordered functional graph, and at the frozen carriers correct-evidence states repair held-out natural-image errors in all three families. What recurs is the causal role structure rather than a particular layer number or module.

The prompt text then selects which route dominates. Holding images and donor evidence fixed, task-relevant option text produces a strong question-side carrier; equal-length unrelated text does not. An instruction to ignore the options leaves the point estimate close to the original, with equivalence confirmed in two families. The paper makes three contributions:

- **A causal role protocol.** PPH combines state efficacy, directed dependence, and semantic specificity, evaluated at coordinates frozen before confirmation, to distinguish source, writer, carrier, and readout.
- **A cross-family causal role organization.** Independent discovery and held-out confirmation recover the ordered $S \to W$, $S \to C$, and $C \to R$ dependencies in Qwen3.5, Qwen3-VL, and Gemma 3 prefix-fusion VLMs, with family-specific writer implementations and carriers that repair real-image errors; the $W \to C$ adjacency is confirmed on Qwen3.5-9B, where the two roles are not adjacent.
- **A prompt-controlled routing result.** Task-relevant option text strongly increases the fraction of source effect routed through a question-side carrier. These tests do not isolate the answer word from the full option group; mechanistic interpretations are tied to the prompt condition tested.

2 Related Work

Grounding and explanation. VLM grounding work measures whether outputs or internal signals align with relevant regions (Peng et al., 2023; Schaumlöffel et al., 2026). Failure analyses reach the same conclusion from the other side: models that answer incorrectly can still encode the visual evidence about as strongly as those that answer correctly (Nooralahzadeh et al., 2026), and operation-level decompositions separate grounding failures from reasoning and language-prior ones (Gupta et al., 2026). Attention can reveal spatial selection, but spatial alignment alone does not establish that the selected signal causally determines the answer (Jain & Wallace, 2019). We study the downstream path from a grounded visual fact to the answer-causal state, and show that the two need not occupy the same layers or token spaces.

Causal interventions in multimodal models. Activation and path patching, attention-edge knockout, interchange interventions, and causal abstraction distinguish information presence from causal use (Meng et al., 2022; Geva et al., 2023; Wang et al., 2023; Geiger et al., 2025; Wu et al., 2024; Kramár et al., 2024). Prior studies typically apply them to localize states, trace paths, or align representations (Lieberum et al., 2023). PPH uses their conjunction to certify a typed role graph. Carrier assignment requires efficacy, mediation, and donor-semantic specificity, and every block sits beyond the layer a patch was written into, so the directed tests can return a negative result.

Visual information flow. Prior VLM work traces visual and text states or components across depth (Palit et al., 2023; Li et al., 2026a). Neo et al. (2025) follows object information from visual tokens to the prediction position, and Zhang et al. (2024) reports transfer into question tokens followed by later propagation. Our experiments add held-out role confirmation, explicit writer mediation, donor-semantic tests, cross-family and cross-capacity replication, and a controlled interface manipulation that identifies which property of the prompt the routed graph depends on.

3 Pointer–Payload–Handoff

Attention maps and activation patches each expose one facet of a VLM's visual computation, but neither certifies the path this paper follows. A highly attended region need not provide a portable representation of the decisive fact, and a state that changes the answer need not carry the fact being traced. Following one answer-determining visual fact from the image to the output therefore requires a protocol rather than a single intervention. This section defines Pointer–Payload–Handoff (PPH).

3.1 Functional Roles

PPH starts from a localized answer-relevant image region supplied by the controlled annotation or an external grounding procedure; it does not discover that spatial support de novo. Given this region, consider a matched pair of inputs that differ in one answer-determining visual fact while keeping the layout and prompt fixed. PPH asks four questions about the downstream provenance of that fact: where it is causally available, which computation transfers it, where the transferred content is carried, and where it reaches the final answer. We denote the four roles by

$$S_{\text{visual}} \to W_{\text{write}} \to C_{\text{decision}} \to R_{\text{answer}}. \tag{1}$$

The source S is an image-position state whose restoration changes the answer. The writer W is one or more component updates whose reset removes part of the restored source effect. The carrier C is a downstream state that is causally effective, mediates upstream evidence, and tracks the donor's visual meaning. The readout R is the final answer-position state through which the carrier-induced effect reaches the output. S, C, and R are state roles; W is a computational role. A state receives the *causal carrier* role when it is causally effective and mediates an identified upstream effect, and the *semantically portable carrier* role when it additionally tracks donor content across examples. We report the two separately because portability is fact-dependent, whereas causal participation is not (Appendix A). Layer indices and recipient token groups are selected independently for each model and frozen before confirmation.

How to read the protocol. Restoration finds *where* a fact can change the answer. Blocking finds *how* its effect moves between states. Interchange finds *what* content a state actually carries. The blocked fractions in the result tables measure routing—where an effect flows—and do not by themselves establish that the carrier carries the specific visual fact. Content-specific claims require the separate evidence chain of GT-versus-control restoration magnitudes, semantic interchange, and natural-image repair (Appendix A).

3.2 Restoration Localizes Candidate States

Let A be the answer supported by the clean example and B the answer supported by its matched counterfactual. At the first answer position, the answer margin is

$$m = \ell_A - \ell_B, \tag{2}$$

where ℓ_y is the pre-softmax logit for answer token y.

For an internal state, we run the counterfactual example and replace that state with its clean counterpart. With margins m_{clean}, m_{cf}, and m_{patch}, normalized restoration is

$$\rho_{\text{restore}} = \frac{m_{\text{patch}} - m_{\text{cf}}}{m_{\text{clean}} - m_{\text{cf}} + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{3}$$

A value near zero recovers little of the clean–counterfactual margin gap; a value near one recovers approximately the full gap. Restoration localizes candidate states and, within a local transition window, candidate attention or MLP outputs. It does not establish a directed path by itself.

3.3 Blocking Tests Directed Edges

All edge tests are defined relative to the same unpatched counterfactual run. Restoring an upstream candidate s gives the raw effect

$$e_s = m_{\text{restore}(s)} - m_{\text{cf}}. \tag{4}$$

We repeat the restoration while resetting a downstream candidate d to its counterfactual value and measure the remaining effect

$$e_{s;\text{reset}(d)} = m_{\text{restore}(s),\text{reset}(d)} - m_{\text{cf}}. \tag{5}$$

The raw blocked effect is

$$B_{\text{raw}}(s \to d) = e_s - e_{s;\text{reset}(d)}, \tag{6}$$

and we report the normalized fraction

$$\text{Block}(s \to d) = \frac{B_{\text{raw}}(s \to d)}{|e_s| + \epsilon}. \tag{7}$$

A directed edge is confirmed from the paired raw blocked effect on held-out examples. PPH tests the links

$$S \to W, \qquad S \to C, \qquad C \to R, \tag{8}$$

together with $W \to C$ wherever the frozen carrier is not the writer's immediate next layer. In all three 4B-scale models the only residual entry between the two roles is the writer's own recipient layer, so a block there restates containment rather than testing transfer, and the adjacency edge is not separately measurable at those coordinates. We therefore report it on Qwen3.5-9B, whose frozen carrier sits five layers beyond the writer.

Each PPH link is tested where the two roles occupy different positions or components, so that a block measures transfer between roles rather than containment. For $S \to W$, we reset the writer's component output, leaving the residual stream and the sibling module as bypass routes. For $S \to C$, we reset the question-side state, which the source patch reaches only by transfer out of the image positions. For $C \to R$, we reset the answer-position state one layer before the readout.

The blocked fractions in the result tables measure *routing*—where an effect flows—rather than content. Any perturbation written at question positions propagates through the residual stream, so even a random patch produces a non-zero blocked fraction at the same site. Content-specific claims require the separate evidence chain of GT-versus-control magnitudes, semantic interchange, and natural-image repair, documented in Appendix A.

3.4 Semantic Specificity of the Carrier

Carrier assignment requires content as well as mediation. We first permute option contents while keeping labels fixed, and separately permute labels while keeping contents fixed. A content-bearing carrier should follow the moved content. We then perform four-value semantic interchange: a carrier state formed from one visual value is inserted into a recipient example, and the output shift toward the donor value is compared with norm-matched random updates. Together with the mediation requirement of Section 3.3, these tests operationalize the two carrier tiers defined in Section 3.1.

3.5 Discovery and Role Assignment

Discovery proposes coordinates; held-out tests assign the roles. In the controlled experiments, the visual candidate group is the set of image patches that overlaps the annotated answer-relevant region, with a matched zero-IoU region as a spatial control. Every decoder residual is swept; attention and MLP outputs are refined within the model's transition window, and the readout is anchored at the final answer-position residual. Model-specific grids nominate implementations, while the same held-out causal predicates assign roles. Exact grids, frozen specifications, and candidate counts appear in Appendix A.

For the image-to-question hypothesis we also sweep image and question token groups through depth, where a decline in image-side restoration with a rise in question-side restoration marks a candidate handoff region. These curves are diagnostic; role assignment uses the tests above, summarized in Appendix Table 6.

4 Experimental Setup

Models and controlled data. We study Qwen3.5-4B, Qwen3-VL-4B-Instruct, and Gemma-3-4B-IT, which use different attention organizations and visual encoders (Qwen Team, 2026; Bai et al., 2025; Google DeepMind, 2025). All three are prefix-fusion decoders (Liu et al., 2023), with visual tokens before text. A deterministic 200-pair diagnostic set changes one answer-determining visual fact while preserving the remaining layout and prompt. The primary semantic analysis uses 32 discovery, 32 validation, and 96 held-out confirmation pairs after excluding location-changing pairs.

Frozen inference protocol. Discovery performs broad residual sweeps and local component refinement. Validation freezes layers, token groups, components, controls, exclusions, and aggregation rules. Confirmation evaluates only those frozen targets, and each model is searched independently. Primary contrasts use paired bootstrap confidence intervals with 10,000 resamples. Appendix A gives the exact search hyperparameters and retained counts.

Counterfactual validity and scoring. The five controlled categories are color, number, spatial relation, object identity, and angle type. Prompts, answers, boxes, and images are generated from the same latent record, so each pair changes only the answer-determining fact; the appendix reports the regeneration and pixel-locality audit of this set. MCQ scoring uses the correct-minus-counterfactual option-letter margin at the first assistant position. The main results use teacher-forced scoring; a separately pre-registered free-generation check appears in Appendix A.11. Answer values are single-token, closed-vocabulary attributes (color, number, shape, angle), and the semantic arms are

scored by teacher-forced sequence log-probability over those same strings, so the evidence covers this answer form rather than open-ended generation.

Natural and interface tests. Real-image repair uses Vision-OPD-6K (Yuan et al., 2026). All donor variants share one pixel budget: the primary donor keeps the full image with an inset of the annotated answer-relevant region, and controls use a matched zero-IoU region, another example's GT region, or a norm-matched hidden direction. The interface experiment fixes the examples, images, donors, controls and pixel budget while changing only the answer interface. The original prompt is a multiple-choice question (F0); F4 keeps the same options with a semantic answer request; F3 removes the options and asks for the answer content. TextVQA/TextOCR supplies an independent direct-answer construction with a localized answer word (Singh et al., 2019; 2021). Capacity and post-training analyses appear in the appendix.

5 From Spatial Grounding to Causal Handoff

5.1 Spatial selection and causal use peak at different states

At the answer-start query, GT enrichment within visual attention rises through the middle full-attention layers and peaks at L19. The peak exceeds equal-area random and shuffled-box controls by more than 10 enrichment units and then declines toward the final layer in 93% of examples. The random control remains approximately flat (Appendix A). This profile identifies a spatial pointer. The open question is whether the attended contribution is itself an answer-changing payload, and the direct test is to transplant it.

5.2 Grounding and answer-causal carriage separate

If spatial selection already produced a portable answer state, transplanting that state is where the repair should appear. We reconstruct the bbox-conditioned attention contribution and transplant it into later residual states over a pre-specified strength grid with five matched control directions. Across the grid, GT-specific accuracy change is capped at $+1.0$ percentage point, and grounding strength is uncorrelated with the response (Appendix A). The spatially selected contribution and the portable answer-changing state are different targets, so the trace must find where the fact becomes usable.

5.3 MCQ exposes the directed causal links

Does the frozen coordinate chain pass the directed link tests? On 96 held-out pairs, the frozen Qwen3.5 coordinates pass the node tests and the directed link tests of Eq. (8) (Table 1). The L1 GT-image source restores 77.8% of the clean–counterfactual margin gap, and the L16 question state restores 97.5% of it, while the L31 answer readout restores 103.1%, near identity by construction. The L15 attention component restores 42.9% of the margin gap under its own transplant. As a directed test, resetting it removes 59.8% of the restored *source* effect. 81.6% of the restored source effect is routed through the question-side state, and transplanting the L16 carrier state while resetting the L30 answer-position state removes 53.7% of the carrier's effect, locating the readout route.

Additional controls for the blocked shares are in Appendix A.

Table 1: Qwen3.5-4B chain confirmation at frozen coordinates. Values are normalized percentages; paired intervals and secondary controls appear in Appendix Table 7. The final answer-position state restores 103.1% by construction and serves only as a reachability check.

Test	Effect
L1 GT-image source restoration (S)	77.8%
L15 attention writer restoration (W)	42.9%
L16 all-question carrier restoration (C)	97.5%
$S \to W$: source effect blocked at L15 attention	59.8%
$S \to C$: source effect blocked at L16 question state	81.6%
$C \to R$: L16 carrier effect blocked at L30 answer state	53.7%

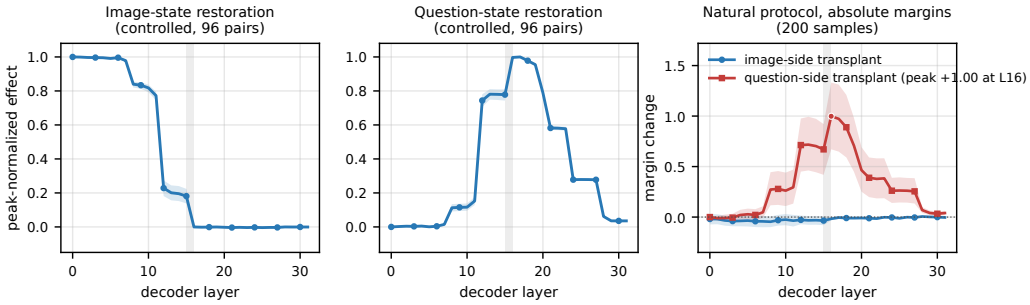

Figure 2: **Image-to-question transition.** *Left and centre:* controlled protocol, peak-normalized restoration at image-side and question-side positions (96 pairs, 95% intervals). *Right:* natural protocol in absolute margin units (200 samples). Transplanting the donor's image-side state does nothing at any depth, while its question-side state peaks at L16. The shaded band marks the independently localized L15–16 transition.

Table 2: Cross-family chain evidence within prefix-fusion VLMs. $S \to W$, $S \to C$, and $C \to R$ are normalized blocked fractions; the $C \to R$ block is applied at L_{R-1}, beyond the writer's recipient layer. Blocked fractions quantify where an effect is routed; content specificity is established by the GT-versus-control restoration magnitudes, the interchange tests, and the natural-image repair. Full intervals are in the appendix.

Model	Frozen $S/W/C/R$	Writer form	$S \to W$	$S \to C$	$C \to R$	Repairs
Qwen3.5-4B	1/15/16–17/31	concentrated	59.8%	81.6%	53.7%	42/120
Qwen3-VL-4B	2/19/20/35	distributed	35.8%	92.1%	96.4%	7/22
Gemma-3-4B	2/11/12/33	multi-module	15.9–23.1%	73.6%	88.4%	15/79

The same frozen carrier also repairs natural errors: replacing the L16 question state with a state from a GT-enhanced donor corrects $42/120$ held-out Qwen3.5 errors, $7/22$ for Qwen3-VL at L20, and $15/79$ for Gemma, with matched non-GT and shuffled donors near zero and no systematic effect on already-correct examples (Appendix A). The donor carries an inset of the ground-truth region, so this measures whether frozen carriers accept corrective evidence under a privileged construction, not the accuracy of a deployable correction method. Because the natural test reuses the carrier coordinate frozen by the controlled analysis, the agreement is out-of-sample transfer.

Full-depth sweeps show the same transition at a coarser level. Under the controlled protocol, image-state restoration falls as question-state restoration rises, with the crossover at the independently localized L15–16 boundary (Figure 2, left and centre). The natural protocol, measured in absolute margin units, reproduces the carrier independently: transplanting the matched donor's question-side state peaks at L16 with $+0.998$ $[0.671, 1.332]$, while transplanting its image-side state changes the margin by at most 0.046 in absolute value at any depth (right panel). Native spatial enrichment, by contrast, peaks later at L19, so the spatial pointer and the strongest answer-changing carrier occupy different depths and token spaces.

5.4 The role order recurs across prefix-fusion families

Does the same role organization appear with independent discovery? Independent analyses place $S/W/C/R$ at L2/L19/L20/L35 in Qwen3-VL and L2/L11/L12/L33 in Gemma. Their writer structures differ: Qwen3.5 has one dominant attention writer, Qwen3-VL distributes the write across more components, and Gemma spreads it across adjacent modules. The source-to-writer fraction falls from 59.8% to 35.8% and then to 15.9–23.1%, while the downstream carrier and readout links stay strong in every family (Table 2). In Qwen3-VL, L19 attention carries a positive source-blocking effect while the colocated MLP is near zero; Gemma instead has positive source interception spread across several adjacent attention/MLP updates. The three models share the role order and implement the write stage differently.

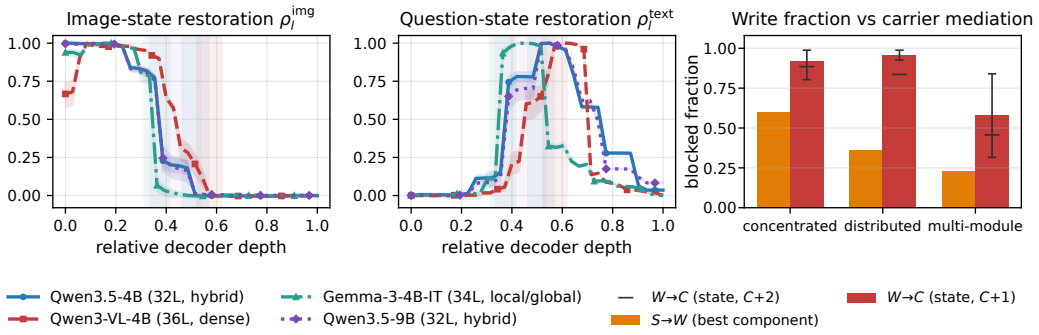

Figure 3: **Role-level replication across families and capacity.** Independently frozen sweeps recover the image-to-question transition under three attention organizations plus Qwen3.5-9B. The right panel summarizes the chain: the share of the source effect intercepted by the single best writer component falls as the write is distributed, while the question-side state one layer beyond the write still mediates most of it (bars, with 95% intervals; short marks show the state two layers beyond).

Table 3: **Interface modulation of the carrier.** Carrier mediation is the share of the restored source effect intercepted at the MCQ-frozen carrier state. F0 presents options and asks for a letter, F4 presents the same options and asks for the answer content, F3 removes the options. Scoring is within-prompt, so letter margins and semantic sequence log-probabilities are never compared on a shared scale; the last two columns are paired differences over the same 96 pairs.

Model	F0	F4	F3	F0 − F4 [95% CI]	F4 − F3 [95% CI]
Qwen3.5-4B	81.6%	79.4%	13.4%	$+2.1\ [-8.9, +10.5]$	$+66.1\ [+63.3, +68.6]$
Qwen3-VL-4B	88.8%	79.4%	19.8%	$+9.5\ [+2.9, +13.8]$	$+59.6\ [+54.9, +63.9]$
Gemma-3-4B	72.5%	58.2%	6.3%	$+14.4\ [+4.0, +23.9]$	$+51.9\ [+43.2, +60.8]$

The replication criterion is the role graph, not numerical layer alignment: each family uses its own discovery, validation, and confirmation split, and capacity provides a within-family check. Qwen3.5-9B reruns the full pipeline without importing the 4B coordinates and freezes $S/W/C/R$ at L0/L11/L16/L31, so its carrier lies five layers beyond the writer, the setting where the $W \to C$ adjacency is measurable. Its four routed shares are all positive, and the $W \to C$ share is 0.858 $[0.814, 0.898]$ (Appendix A). The ordered organization therefore holds across all four links where the adjacency edge can be tested, and the 4B-scale models reproduce the same roles with a more concentrated write (Qwen3.5-2B adds a profile-level replication).

5.5 Task-relevant option text modulates the route

Which property of the prompt sets the carrier route? We now change only the answer interface while holding the images, donors, controls, and pixel budget fixed. The original MCQ prompt (F0) presents four options and asks for a letter; F4 presents the same options verbatim and asks for the answer content; F3 removes the options and asks for the answer content.

The carrier remains effective when the output format changes and weakens when the options are removed. Table 3 reports the share of the restored source effect that the MCQ-frozen carrier state intercepts under each prompt. F4 retains substantial mediation in every family, although it is below F0 by 9.5 and 14.4 points in Qwen3-VL and Gemma (both intervals exclude zero); the Qwen3.5 difference is 2.1 points with an interval spanning zero. Under F3, mediation at the MCQ-frozen carrier falls to 6–20%. The controlled images and Vision-OPD donors show the same ordering, and independent sweeps under F4 place the question-state peak at or near the MCQ carrier (L17, L22, L16), so the interface changes the strength of the carrier rather than its existence.

The natural side shows the same ordering: with the F4-frozen carrier and the same donors, GT-enhanced replacement repairs real-image errors in every family and exceeds non-GT, shuffled, and norm-matched random donors with every interval above zero (Appendix Table 9); the paired

F4−F3 gap is $+1.22$ $[0.67, 1.88]$, $+2.72$ $[1.50, 3.98]$, and $+7.83$ $[3.78, 12.44]$ margin units. A carrier selected under MCQ therefore keeps working when the answer's rendering changes, and weakens when the option text is removed.

To identify the relevant prompt property, we test four pre-registered prompt-control arms (Appendix A.8). An equal-length block of unrelated text drops carrier mediation from 79.4% to 24.6% (Qwen3.5-4B) and from 79.4% to 25.2% (Qwen3-VL-4B). Keeping the option text while instructing the model to ignore the list leaves it at 78.7% and 78.0%, intervals inside the equivalence margin; keeping it and asking for a description instead of an answer leaves it at 76.3% and 69.1%, intervals that do not confirm equivalence. The option text sets the route in the first two families even when the instruction says to ignore it; the arms do not separate the answer word from the option group that carries it.

6 Discussion

The cross-model result is an ordered causal role graph, not a match in layer numbers. Qwen3.5 concentrates the measured write in one attention component; Qwen3-VL and Gemma distribute it across updates. Independent discovery in each architecture recovers the same source, writer, carrier, and readout roles, even when no single component accounts for the write. The same role tests transfer across capacity within Qwen3.5 (4B and 9B, with a 2B transition profile).

Spatial enrichment and causal use also separate. In Qwen3.5 the attention localizer peaks at L19, after evidence has moved from image to question states at L15–17; its localized contribution does not reproduce downstream repair. Grounding identifies the visual evidence, while restoration and blocking trace where that evidence changes the answer and how its effect moves. Candidate transitions should therefore be selected from causal restoration across token groups, then interpreted against the grounded region.

The prompt determines how much causal work the question representation does. With task-relevant options, the question-side carrier mediates most of the restored source effect; removing the options shifts leverage to image states. Changing the output from option letter to answer content preserves substantial carrier mediation, although its fraction falls in Qwen3-VL and Gemma. Unrelated text does not restore the route. The carrier follows the option text rather than the answer alphabet or the instruction to compare options. The held-out test separating the answer word from its option group remains inconclusive (Appendix A.9).

Carrier claims require two distinctions. A state may mediate upstream evidence without transporting the donor's visual value across examples: color passes the causal tests but fails four-value interchange in all three families (Appendix Table 8). Semantic portability is established for the other four fact types, while the causal role holds across all five. Gemma's distributed writer yields smaller per-component source-to-writer fractions (15.9–23.1%); its frozen carrier ranks 6/34 among swept candidates, and the ignore-list control interval exceeds the equivalence margin. These bounds make the role graph more informative than a one-module description and delimit its current evidence.

The mechanism is thus conditional on the answer interface. Audits and edits of internal states should locate causal roles under the prompt of intended use and test transfer when that prompt changes.

7 Conclusion

PPH traces a localized visual fact from image to answer through a writer, question-side carrier, and readout. Frozen-coordinate dependencies recur in Qwen3.5, Qwen3-VL, and Gemma 3 despite distinct writer structures. Carriers accept corrective evidence from real images, and all four links replicate in Qwen3.5-9B. Removing task-relevant options shifts causal influence toward image states without changing weights. Thus functional roles under a specified prompt are the unit for comparing multimodal mechanisms across architectures and scales.

AI USE STATEMENT

Generative AI tools were used to assist with code writing, generate early visual references for figure design, organize relevant literature, and polish the manuscript. The early visual references were used only during design iteration and were not treated as scientific evidence. The authors reviewed and revised all AI-assisted content, verified the final code, figures, citations, and empirical claims, and take full responsibility for the final manuscript and its supporting artifacts.

ETHICS STATEMENT

This work uses synthetic diagnostic images and previously released research datasets; it introduces no new interaction with human participants and collects no new personally identifying information. Mechanistic interventions can support model auditing but may also inform model manipulation, so we report their scope and controls explicitly. Any released artifacts will exclude credentials, private infrastructure paths, and data that cannot be redistributed under the source licenses.

REPRODUCIBILITY STATEMENT

Appendix A documents data splits, preprocessing, prompts, token-group construction, frozen coordinates, controls, statistical procedures, retained counts, and criteria for inconclusive results. The supplementary package includes deterministic data-generation and audit code, experiment manifests or licensed-data metadata, frozen run configurations, intervention and aggregation scripts, and pair-level metrics. Discovery, validation, and confirmation artifacts are kept separate so that every reported confirmatory claim can be traced to a fixed specification.

REFERENCES

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

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

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

Google DeepMind. Gemma 3 model card, 2025. URL `https://huggingface.co/google/gemma-3-4b-it`. Hugging Face model card.

Navya Gupta, Bingjie Xu, Avinash Anand, Timothy Liu, and Zhengchen Zhang. How do VLMs fail? vision-operation misalignment in compositional VQA. *arXiv preprint arXiv:2607.16094*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? dissecting on-policy self-distillation. *arXiv preprint arXiv:2608.09228*, 2026.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *AAAI Conference on Artificial Intelligence*, 2026a.

Xirui Li, Ming Li, and Tianyi Zhou. What does RL improve for visual reasoning? a frankenstein-style analysis. *arXiv preprint arXiv:2602.12395*, 2026b.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision–language models. In *International Conference on Learning Representations*, 2025.

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

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. *arXiv preprint arXiv:2308.14179*, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Qwen Team. Qwen3.5-4B model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-4B`. Hugging Face model card.

Timothy Schaumlöffel, Martina G. Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. *arXiv preprint arXiv:2605.19792*, 2026.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation. *arXiv preprint arXiv:2608.18271*, 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026.

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 *International Conference on Learning Representations*, 2023.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah D. Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. *arXiv preprint arXiv:2403.07809*, 2024.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Gengwei Zhang, Jie Peng, Zhen Tan, Mufan Qiu, Hossein Nourkhiz Mahjoub, Vaishnav Tadiparthi, Kwonjoon Lee, Yanyong Zhang, and Tianlong Chen. Understanding the role of hallucination in reinforcement post-training of multimodal reasoning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. *arXiv preprint arXiv:2411.18620*, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Table 4: Role-search accounting reconstructed from the frozen run scripts and summary files. Counts are candidate cells, not examples. The question-state column counts residual layer–recipient-group cells; the component column counts local attention/MLP output cells. Matched non-GT and fixed readout controls are excluded from the candidate total.

Model	GT-image residual	Question residual	Local Attn/MLP	Frozen confirmation
Qwen3.5-4B	32	42	98	10
Qwen3-VL-4B	36	36	4	18
Gemma-3-4B	34	34	28	22

A Experimental and Statistical Details

A.1 Splits, Inference Protocol, and Search Accounting

Splits. The controlled diagnosis set contains 200 pairs split into 40 discovery, 40 validation, and 120 confirmation examples. After excluding location-changing pairs, the semantic provenance experiments contain 32/32/96 pairs, 24 per non-spatial fact type. Primary natural repair uses a disjoint 40/40/120 split of 200 base-wrong Vision-OPD examples. The natural layer sweep uses another 200 examples balanced by base correctness, and the paired interface experiment uses those 200 examples unchanged.

Evidence process. Discovery sweeps decoder residuals to locate source and carrier transitions; the readout is anchored at the final answer-position residual. Component refinement tests attention and MLP outputs only in the local transition window, together with pre-defined recipient groups. Validation converts the result into a versioned specification containing four coordinates, recipient token groups, component choices, exclusion rules, and aggregation code; confirmation consumes that specification without updating it and writes pair-level metrics before aggregation. This split-sample selection is our control for the discovery search. Each architecture stores the search result, the frozen specification, and the confirmation record separately.

Frozen token groups. Image groups are GT-overlapping patches and a matched zero-IoU non-GT region. Text groups are the instruction, question stem, all option spans, each option span, all question tokens, and final readout. Each option group contains every tokenizer token in the complete option line, including its label; a unique contextual subsequence match is required, with a decoded-span fallback only for boundary merges. Alignment fails closed if a span is ambiguous or differs across the paired prompts. Answer letters must each tokenize to one token for first-position scoring, and fractional patch overlap is derived from processor metadata rather than a nominal image grid.

Layer convention. Ll denotes the residual stream at the entry to zero-based decoder block l. The donor tensor is `hidden_states[`l`]` and is inserted with a pre-hook on decoder block l, so hook, paper, and plot indices are identical. Component blocks are addressed at the same index as the residual entry they write into.

For Qwen3.5, the question refinement scans seven layers by six recipient groups, and the component refinement scans seven layers by seven groups by two component types. Qwen3-VL uses one all-question residual candidate per layer and four frozen local component alternatives around the transition. Gemma uses one all-question residual candidate per layer and a seven-layer by two-group by two-component local window. Candidate counts differ across models because token-group granularity and the width of each independently discovered transition window differ. Every model receives a full residual sweep followed by local component refinement, and validation then freezes one versioned specification before the same predicates are evaluated on held-out pairs.

A.2 Candidate Ranking Among Swept Layers

Role coordinates are frozen from a search, so we report where the frozen cell sits among the swept candidates rather than only its held-out value. Table 5 gives the peak-normalized restoration at the frozen coordinate, its rank among the swept residual candidates, and the best and second-best layers.

Table 5: Frozen coordinates among the swept candidates. Ranks are among all swept residual layers of that model and side, evaluated on the same held-out pairs as the confirmation statistics.

Model	Side	Frozen	Rank	Best / second-best
Qwen3.5-4B	image	L1 (0.815)	6 / 32	L6 (0.825) / L3 (0.824)
Qwen3.5-4B	question	L16 (0.990)	1 / 32	L16 (0.990) / L17 (0.987)
Qwen3-VL-4B	image	L2 (1.111)	2 / 36	L3 (1.134) / L2 (1.111)
Qwen3-VL-4B	question	L20 (0.717)	3 / 36	L22 (0.874) / L21 (0.760)
Gemma-3-4B	image	L2 (0.802)	4 / 34	L3 (0.837) / L4 (0.815)
Gemma-3-4B	question	L12 (0.743)	6 / 34	L15 (0.859) / L16 (0.857)

Table 6: PPH role assignment. Discovery proposes frozen coordinates; held-out confirmation determines whether each functional role is supported.

Role	Candidate evidence	Held-out confirmation
S	GT-image state restoration	Positive restoration and GT-specific effect over matched non-GT
W	Local component refinement near the image transition	Resetting the component removes a significant part of the restored S effect
C	Carrier state restoration	Positive restoration, upstream mediation, and donor-semantic specificity
R	Final answer-position state	Resetting the readout removes a significant part of the restored C effect

The ranks come from the same held-out sweep that supplies the confirmation statistics, so they locate the frozen cell among the candidates the search visited.

Three features of this table matter for reading the results. The image-side source sits on a broad plateau (ranks 6/32, 2/36, 4/34) with neighbouring layers within 0.02 of it, so the source claim is about an early-state region rather than one layer. The question-side carrier is the top-ranked cell in Qwen3.5 and within the top three in Qwen3-VL, while in Gemma it is sixth of 34 with the top five clustered within 0.12 of each other, which is why Gemma's writer fraction is reported as a range. By construction the held-out confirmation values are unbiased for the frozen cell: the confirmatory contrasts are pre-specified single cells, so no multiplicity correction is applied, and the reported intervals cover pair sampling.

A.3 Falsifiable Block Construction

PPH places each block beyond the writer's recipient layer. A block that reset the residual entry at the layer a patch was just written into would re-establish containment: the writer's update reaches layer $l{+}1$ only through that position's residual stream, so such a reset reproduces the target run for every patch. Placing the block beyond the write leaves the writer's bypass structure in place, and we verify on Qwen3.5/Qwen3-VL/Gemma that the resulting statistic is well behaved:

- **Position-overlap diagnostic.** Writing the donor into the option tokens and resetting the stem tokens blocks nothing (mediation 0.00 in all three families, blocked effect up to 38.5 margin units), whereas resetting the same option positions blocks the entire effect (blocked restoration exactly 0.0 on every pair).
- **Bypass routes stay open.** With the donor writer patch held fixed, resetting the L_C attention output, the L_C MLP output, and the L_{C+1} attention output yields mediation of $+0.44/-0.10$, $+0.24/+0.17$, and $-0.12/+0.12$ (Qwen3.5/Qwen3-VL; Gemma $-0.14/+0.07/+0.29$). The same write is therefore only partially removable when the block leaves the sibling module or the residual stream intact.
- **The delivered statistic varies with distance.** One and two layers past the writer's recipient layer, the blocked run differs from the target run on every pair (largest residual restoration

Table 7: Complete frozen confirmation cells in Qwen3.5-4B. Values are percentages of the clean–counterfactual margin gap or of the restored upstream effect.

Frozen cell	Effect	95% CI
L1 GT-image source restoration	77.8%	[71.4, 84.1]%
L15 attention writer restoration	42.9%	[37.4, 48.4]%
L15 MLP restoration	2.2%	[−1.7, 5.6]%
L16 all-question carrier restoration	97.5%	[95.7, 99.5]%
L17 all-question carrier restoration	98.3%	[95.8, 101.2]%
L31 answer readout restoration	103.1%	[102.3, 103.8]%
Source effect intercepted at L15 attention ($S \to W$)	59.8%	[49.8, 67.5]%
Source effect routed through L16 question state ($S \to C$)	81.6%	[70.2, 89.7]%
L16 carrier effect blocked at L30 answer state ($C \to R$)	53.7%	[51.2, 56.2]%

0.4–1.6 margin units), and the routed share decreases with distance ($0.92/0.95/0.58$ then $0.89/0.84/0.46$ for Qwen3.5/Qwen3-VL/Gemma).

- **Containment check.** Resetting the L_C residual entry at exactly the patched positions returns blocked restoration of exactly 0.0 on $96/96$ pairs in all three families; we report it only as a containment check.

What the blocked fractions measure. The blocked fractions locate where an effect flows. Any update written at question positions continues through the question-side residual stream, so a block placed there removes most perturbations regardless of content: at L_{C+1} the mediated share is $0.71/0.82/0.66$ for the random patch, $0.78/0.75/-0.08$ for the all-zero patch, and $0.89/0.83/0.51$ for the permuted patch (Qwen3.5/Qwen3-VL/Gemma), against $0.92/0.95/0.58$ for the donor patch. The blocked fractions are therefore reported as routing shares, and the content-specific claims rest on evidence that compares content directly: GT-versus-control restoration magnitudes at the same coordinates, four-value semantic interchange, and natural-image donor repair, where the GT donor beats matched non-GT, shuffled, and random donors in all three families. This division is the one the role table states: restoration and magnitude establish efficacy and specificity, blocking establishes routing, and interchange establishes carried content.

A.4 Controlled-Pair Audit

The 200 controlled pairs are regenerated from their latent records and compared byte for byte with the stored images: all 200 pairs regenerate identically, zero pixels differ outside the annotated fact region, and at least 2,025 pixels differ inside it (audit verdict `CONTROLLED_DATASET_AUDIT_PASS`; report in the artifact package). The shared latent record, from which prompt, answers, boxes, and images all derive, establishes that each pair changes one answer-determining fact and leaves the matched control regions pixel-identical, so a control-region transplant is a genuine null operation. Per-split and per-fact-type counts, answer-letter balances, and five representative cases are recorded in the same report.

A.5 Ordered-Link and Cross-Family Confirmation

Figure 4 shows the recovered organization schematically; the held-out statistics behind each role and link are in Table 7.

Cross-architecture freeze. Qwen3-VL confirmation was run only after independent 32-pair discovery and 32-pair validation agreed; we then froze source L2, write-stage L19 attention, carrier L20 all-question residual, and readout L35 for 96 held-out pairs. Gemma used the same separation and froze L2/L11/L12/L33 after its own discovery and validation. The identical source-restore/writer-reset protocol gives 0.598 $[0.498, 0.675]$ at L15 attention in Qwen3.5, 0.358 $[0.312, 0.403]$ at L19 attention in Qwen3-VL with the colocated MLP near zero (0.024, $[-0.014, 0.060]$), and, in Gemma, fractions spread across adjacent components: 0.159 $[0.048, 0.261]$ at L11 attention, 0.195 $[0.138, 0.254]$ at L12 attention, 0.210 $[0.162, 0.261]$ at L13 attention, and 0.231 $[0.174, 0.289]$ at L14 MLP, all intervals excluding zero. A subsequent fresh-seed confirmation holds each model's carrier and readout

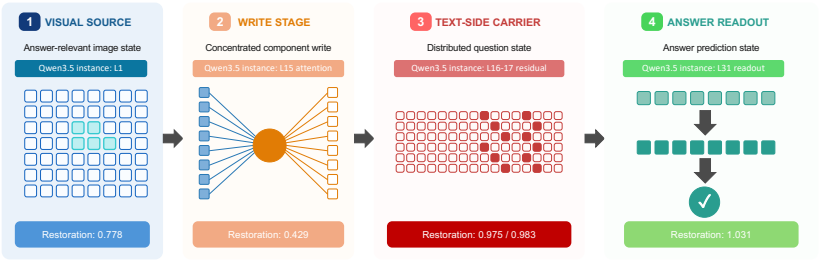

Figure 4: **Frozen PPH roles in Qwen3.5-4B.** Schematic of the recovered organization: the visual source, the attention write stage, the question-side carrier, and the answer readout, drawn at their frozen coordinates. Table 1 carries the held-out confirmation statistics, and Figure 2 shows the depth profile behind the coordinates.

fixed and yields raw carrier-to-readout blocked effects of 9.45 $[8.95, 9.94]$, 33.80 $[31.18, 36.35]$, and 26.83 $[23.04, 30.76]$.

A.6 Natural-Image Donors and Repair

Privileged donor construction. Every natural branch is rendered under an identical $1{,}003{,}520$-pixel budget. The primary GT donor is the full image plus an inset of the GT evidence region. The non-GT donor uses an identically styled, matched-size zero-IoU region; shuffled uses another sample's GT evidence; random uses a norm-matched hidden direction. A standalone GT crop is reported only as an upper bound.

Qwen3-VL natural transfer. The natural test contains 40 held-out real-image questions, yielding 22 initially wrong and 18 initially correct examples. GT-enhanced replacement is positive at writer L19 (0.75), carrier L20 (2.20), option carrier L20 (1.62), and post-carrier L21 (2.27), with every confidence interval above zero. At L20, GT exceeds matched non-GT by 2.00 ($[0.57, 3.61]$) and random direction by 2.28 ($[0.78, 3.95]$). On the 22 wrong examples, the carrier effect is 4.22 ($[2.07, 6.59]$), positive on 86%, and repairs seven answers with no adverse flip. On the 18 correct examples, GT has no systematic effect, whereas the random direction is harmful (-0.71, $[-1.13, -0.33]$).

Gemma natural transfer. The expanded test contains 120 questions, 79 of which Gemma answers incorrectly. On these errors, GT replacement yields $+2.93$ at L11 attention, $+4.50$ at the L12 question carrier, $+3.56$ at L12 options, and $+4.65$ at L13, with all intervals above zero and 11–16 repairs without an adverse flip. GT exceeds both same-image non-GT and shuffled evidence in all eight node–control contrasts; at L12 the differences are $+3.96$ ($[1.77, 6.39]$) and $+3.05$ ($[0.84, 5.49]$). The random-direction contrast is state-dependent, raising the margin on the error subset ($+0.82$, $[-1.75, 3.46]$) while lowering it on the correct subset (-4.14, $[-6.06, -2.27]$), consistent with a scale effect rather than content transfer.

A.7 Interface Analyses

Prompt construction and scoring. The same controlled images are rendered under four prompt interfaces: original MCQ (F0), option-content permutation with fixed labels (F1), label permutation with fixed content positions (F2), and direct answer without options (F3); the natural paired experiment adds F4, which keeps the F0 stem and option lines verbatim and replaces only the instruction with "Answer with the answer text only." Option permutations are deterministic functions of pair ID. F0 is scored by the option-letter margin at the first assistant position; F3 and F4 are scored by the teacher-forced log probability of the complete answer sequence over the same option texts, so raw margins are never compared across letter and sequence scales. Only within-prompt restoration ratios, mediator fractions, and paired differences of those ratios are reported across prompts.

Option permutation and four-value interchange. For F1/F2, the semantic tracking score is restoration on the target's semantically correct option content minus restoration on the content at its





Figure 5: **Two natural-image interventions at frozen coordinates. (a)** Under MCQ, replacing the frozen L16 question-side carrier with a GT-enhanced donor flips a Qwen3.5 answer from blue to the correct yellow (margin $-5.0 \rightarrow +1.4$); matched non-GT and shuffled donors leave the error. **(b)** Under direct answering, an image-side restoration from an inset of the correct word *Mike* raises the answer margin by $+36.3$, while transplanting the same evidence into the text-side carrier changes it by only $+0.04$.

Table 8: Four-value semantic interchange by fact type. Entries are mean donor selectivity ($n = 48$ receiver sides per cell). † marks an interval entirely below zero.

Model	Color	Geometry	Number	Identity	Spatial
Qwen3.5-4B	$-0.08^{\dagger}$	0.30	0.32	0.20	5.28
Qwen3-VL-4B	$-0.07^{\dagger}$	0.38	8.23	0.63	14.47
Gemma-3-4B	$-0.12^{\dagger}$	0.40	0.81	0.79	6.78

original F0 slot. All groups are anchored on the target prompt, so donor and recipient spans contain identical text and token counts. F1/F2 carrier scores are $1.75/1.94$ for Qwen3.5, $2.22/3.76$ for Qwen3-VL, and $7.61/6.31$ for Gemma, with all paired 95% intervals excluding zero, and label-only effects near zero relative to content effects. Four-value interchange uses each receiver side, the opposite-side semantic donor, and all four template values; its score is the donor-value log-probability change minus the largest non-donor change, with per-position random updates matched to the norm of the semantic donor update. Semantic scores are 1.20 $[0.92, 1.51]$, 4.73 $[3.91, 5.58]$, and 1.73 $[1.13, 2.44]$; semantic-minus-random differences are $1.30/4.95/2.36$, all above zero, and remain positive after excluding spatial-relation pairs ($0.27/2.51/0.94$).

Semantic portability is fact-dependent: all three families pass the controlled MCQ node and edge tests for color, and color is the one value type whose carrier state does not transfer across donors. The carrier role in the main chain rests on causal efficacy and upstream mediation, with donor-semantic specificity as an additional property that holds for the remaining value types.

Natural paired interface analysis. The DA layers are L16/L22/L18 for Qwen3.5/Qwen3-VL/Gemma. On the pre-defined donor-valid subsets, DA GT-minus-non-GT is 0.228 ($[0.043, 0.411]$), 0.250 ($[-0.245, 0.752]$), and 1.973 ($[1.296, 2.748]$); Qwen3.5 and Gemma also exceed shuffled and random controls, and the corresponding Qwen3-VL contrast is not separated from zero. Because donor validity uses observed input-level effects on the same examples, these subsets measure conditional transmissibility rather than unconditional prevalence.

F4 natural repair. The F4 arm reuses the E3' protocol with the options kept and the semantic instruction, and transplants at the F4-discovered carrier layer ($17/22/16$). Donor-valid subsets retain $98/103/77$ of 200 samples. GT-enhanced replacement is positive in every subset, and GT exceeds non-GT, shuffled, and norm-matched random donors in all three families:

Table 9: F4 natural-image carrier repair. Deltas are in semantic sequence-log-probability units; intervals are paired 95% bootstrap.

Model	Subset n	GT delta	GT − non-GT	GT − shuffled	GT − random
Qwen3.5-4B	200	0.76	0.70 $[0.49, 0.92]$	0.72	0.78
Qwen3.5-4B	98	1.71	1.66 $[1.34, 2.00]$	1.70	1.75
Qwen3-VL-4B	200	1.40	1.35 $[0.86, 1.85]$	1.33	1.43
Qwen3-VL-4B	103	3.49	3.44 $[2.80, 4.10]$	3.42	3.54
Gemma-3-4B	200	1.30	1.33 $[0.12, 2.61]$	1.54	1.12
Gemma-3-4B	77	7.56	7.49 $[5.48, 9.68]$	7.51	7.09

Table 10: Prompt controls. Carrier mediation is the fraction of the source effect that passes through the MCQ-frozen carrier; the columns give the three families on their own 96 held-out pairs. Entries are pre-registered means except Gemma-3 `x6c`, which is a descriptive median because one zero source anchor makes its pre-registered mean uninterpretable.

Arm	Prompt text between stem and instruction	Qwen3.5	Qwen3-VL	Gemma-3
`f4`	the option block, verbatim	79.4%	79.4%	58.2%
`x6a`	an equal-length, equal-format block of unrelated text	24.6%	25.2%	4.3%
`x6b`	the option block, plus "Ignore the list above."	78.7%	78.0%	58.1%
`x6c`	the option block, and a description request instead of an answer	76.3%	69.1%	66.6%

On the F4-wrong donor-valid strata the effects are larger still (1.82 $[1.41, 2.28]$, 4.25 $[3.38, 5.18]$, 9.07 $[6.43, 12.01]$), with $31/61$, $33/66$, and $19/53$ repairs.

A.8 Prompt Controls on Task-Relevant Option Text

The interface experiments change the answer interface, and with it several properties of the prompt text at once. Four prompt-control arms were therefore pre-registered — prompts, endpoints, thresholds, equivalence margin and decision rules frozen before any arm was run — to separate candidate comparison from the properties that co-vary with it. All four arms reuse the same images, donors, controls, MCQ-frozen coordinates and scoring; they differ only in the text between the stem and the instruction:

The equal-length unrelated block sharply reduces mediation, whereas retaining the option text with an instruction to ignore it leaves the point estimate nearly unchanged. Equivalence is confirmed for the latter contrast in Qwen3.5 and Qwen3-VL, but not in Gemma-3. The description-request arm retains a high point estimate where the ratio is stable; its contrasts do not establish equivalence with `f4`.

Section A.9 tests the answer word separately from the full option group.

The length control is built to be neutral and to match exactly: no filler word names a colour, a number, a shape, an angle or a position (checked against the manifest answer vocabulary), and every matched prompt equalled its `f4` prompt's token count exactly, with zero pairs outside the ± 2-token tolerance (768 of 768 pairs at a delta of zero). The length-control arm is indeterminate by the pre-registered strength bands in Qwen3.5 and Qwen3-VL, and weak in Gemma-3. Its paired contrast with `f4` excludes the equivalence margin in all three families; the conclusion rests on those contrasts.

The equal-length contrast is $+54.2$ points in Qwen3-VL and $+53.8$ $[+44.7, +62.6]$ in Gemma. The ignore-list arm differs from `f4` by $+1.4$ and $+0.1$ points, respectively; its equivalence interval lies inside the pre-registered margin in Qwen3-VL but extends beyond it in Gemma ($[-12.9, +8.6]$). The description-request arm retains mediation of 76.3%, 69.1%, and 66.6% across the three families,

Table 11: Answer-word controls, on new held-out samples (96 pairs per family, seed 43). Carrier mediation as in Table 10; each family is measured on its own pairs, and intervals and strength bands are given in the text.

Arm	Prompt text between stem and instruction	Qwen3.5	Qwen3-VL	Gemma-3
`f4`	the option block, verbatim	78.2%	76.0%	39.9%
`x7a`	a prose sentence naming the answer word, no list	21.0%	37.6%	36.7%
`x7b`	four labelled lines whose values are wrong on both sides	33.4%	30.1%	0.7%
`x7c`	the same prose frame with a neutral phrase	13.9%	16.0%	17.2%

without confirmed equivalence to `f4`. Gemma's value is a median: one near-zero source anchor produces a per-pair ratio of $-343{,}750$ and a pre-registered mean of $-3{,}580.1$. Because the description arm is scored over the answer word, it measures teacher-forced readability under that prompt, not generated behavior.

The location of the effect follows the prompt differently across families. Per-arm location sweeps on the discovery split put the image-side maximum at L6–L5 and the question-side maximum at L17–L16 in every Qwen3.5-4B arm, with profile correlations of 0.83–0.999, so there the prompt changes magnitude and not location. In Qwen3-VL-4B the same sweeps disagree (image side moving by up to sixteen layers, question side by up to thirteen, correlations as low as -0.69) and in Gemma-3 they disagree as well (up to eight layers on either side, correlations as low as -0.25). For those two families the routing claim is stated per arm in this table rather than as a shared location.

The pre-registration, the arm prompts and the analysis rules are part of the released material.

A.9 Separating the Answer Word from the Option Group

The controls above leave one alternative open: every arm that carries the answer also carries the option group, so the answer word on its own is never isolated. A second experiment was pre-registered before it ran, on fresh held-out samples drawn with a new seed (96 new pairs per family, disjoint from those of Section A.8), with prompts, endpoints, thresholds, equivalence margin and decision rules fixed in advance. It keeps `f4` as the reference and adds three arms that place the same fact in the prompt in three other forms: one prose sentence that names the answer word and contains no list (`x7a`); four labelled lines whose values are wrong on both sides, an option-shaped block that never contains the answer (`x7b`); and the same prose frame with a neutral phrase in place of the answer (`x7c`).

The three families return the same verdict at sharply different levels, and only the first two keep the verbatim block inside the strong band (78.2% and 76.0%); on Gemma-3 the reference itself reaches 39.9% $[22.5, 53.9]$, which is indeterminate, so nothing in that family is claimed at the strong level and its intervals are correspondingly wider. No alternative arm reproduces the block anywhere: the prose mention reaches 21.0% $[14.1, 27.4]$, 37.6% $[26.6, 46.7]$ and 36.7% $[19.6, 66.4]$; the wrong-candidate block 33.4% $[26.6, 41.5]$, 30.1% $[21.9, 38.7]$ and 0.7% $[-15.5, 10.6]$; the neutral frame 13.9% $[11.1, 16.8]$, 16.0% $[12.0, 20.2]$ and 17.2% $[-3.5, 49.0]$. On Qwen3.5-4B and Qwen3-VL-4B every paired contrast with `f4` excludes the equivalence margin ($+57.2$, $+44.8$, $+64.3$ and $+38.4$, $+46.0$, $+60.0$). On Gemma-3 only the wrong-candidate arm separates from the reference ($+39.3$ $[+17.5, +60.8]$); the prose mention ($+3.2$ $[-32.4, +27.1]$) and the neutral frame ($+22.7$ $[-16.5, +51.1]$) have intervals that span zero, so on those new samples that family separates neither from the block it is measured against. The prose mention sits above the neutral frame by $+7.2$ $[+1.4, +12.1]$ on Qwen3.5-4B and by $+21.6$ $[+11.7, +29.4]$ on Qwen3-VL-4B, while on Gemma-3 that contrast is $+19.5$ $[-18.3, +56.1]$; in the first two families the interval excludes zero but does not fall inside the ± 10-point margin.

The families also order the two alternatives differently, and the ordering is not stable. On Qwen3.5-4B the wrong-candidate block — which never contains the answer — sits numerically *above* the prose mention (33.4% against 21.0%); on Qwen3-VL-4B the order reverses (30.1% against 37.6%); and on Gemma-3 the prose mention is the higher of the two by a wide margin (36.7% against 0.7%). The pre-registered decision table returns "inconclusive" in all three families, so the routing claim is unchanged: the option text as a whole is what the evidence supports, and these arms license neither "the answer word routes it" nor "an option group routes it independently of the answer". What they add is that the effect does not survive translation of the same fact into prose in the two families where that contrast is resolvable, that the size of the residual prose effect is family-dependent, and that an option-shaped block of comparable size carrying only wrong values is not enough — the arm meant to stand in for "the option group without the answer" ranges from 33.4% to 0.7% across families. Per-arm location sweeps on the discovery split agree with the MCQ-frozen coordinates in Qwen3.5-4B (image side L5, question side L16, profile correlations 0.80–0.95) but not in Qwen3-VL-4B (argmax moves by three layers on the two arms that change the image-side maximum, correlations 0.92–0.98) nor in Gemma-3 (up to nine layers on the image side and four on the question side, correlations as low as -0.06), so as in Section A.8 the location statement is made per arm in those two families.

A.10 Per-Fact-Type Breakdown and the Gemma-3 Anchor Check

Two endpoints can be read side by side per fact type. Causal load-bearing is the mediation fraction of the carrier readout, taken from that run's per-pair rows; it varies by type on Qwen3.5-4B (colour 0.714, object identity 0.680, number 0.766, geometry 0.890) and sits at 0.99 for every type in the other two families (0.993–0.999 on Qwen3-VL-4B, 0.975–1.001 on Gemma-3). Semantic interchange — the receiver-side effect of writing the donor's value at the carrier rather than a random value — behaves differently, and colour is the type that does not interchange: its semantic and random values are both negative in the shipped aggregates (-0.079 against -0.050 on Qwen3.5-4B, -0.072 against -0.091 on Qwen3-VL-4B, -0.121 against -0.278 on Gemma-3), whereas number, geometry, object identity and spatial relation all show a positive semantic effect against a negative random control (number: $+0.322$ against -0.059, $+8.231$ against -0.193, $+0.809$ against -0.248). Colour is therefore routed but not exchangeable: the carrier passes colour evidence, and the four-value swap written through it does not move the receiver. These interchange numbers are the shipped aggregates of that run and are reported descriptively; the run ships no per-pair rows, so the semantic-minus-control contrast is not re-bootstrapped here.

The Gemma-3 prompt-control arm that asks for a description rather than an answer also explains its own uninterpretable mean. Its source anchors are two orders of magnitude smaller than every other arm's: median $+1.59$, upper quartile $+2.68$, thirty-six of ninety-six pairs below $|1|$ and one below 10^{-3}, where every other arm in every family has a median between $+7.2$ and $+41.1$ and at most four pairs below $|1|$. A ratio whose denominator is that small is not a rate, which is why that arm is reported by its median, with the pre-registered mean printed beside it rather than instead of it.

A.11 Free-Generation Check at the Frozen Carrier

Every carrier number above is a teacher-forced readout: the answer sequence is scored under the prompt rather than produced by the model. A pre-registered check asks whether that result survives when the model generates, with the smallest possible change — the same held-out pairs, the same prompt, the same frozen coordinate, the same token groups, and a greedy eight-token generation as the only new element. Three conditions are compared on each pair: no intervention; a patch of the carrier coordinate with the state captured from the donor view (*real donor*); and the same patch with the state captured from a donor image whose fact slot has been refilled by a same-size crop from a non-target region of that same image (*non-target donor*). The endpoint is the rate at which the generated answer parses to the donor's value; the rule, the margin and the parse rule were frozen before the runs.

Patching the carrier with the donor state raises the generated donor-value rate over baseline by 91.7, 59.4, and 46.9 points in Qwen3.5, Qwen3-VL, and Gemma, respectively; the corresponding advantages over the non-target donor are 47.9, 33.3, and 31.2 points (all paired intervals exclude zero). The pre-registered decision rule classifies this check as uninterpretable: the unparsed share exceeds its 20% ceiling in at least one condition for each family (Table 12). The main carrier results

Table 12: Free-generation check. Donor-value rate of the generated answer and, in parentheses, the share of generations that name neither value (unparsed), per family on its own 96 held-out pairs. Paired intervals, 10,000 resamples.

Family	`none`	`real_donor`	`non_target_donor`
Qwen3.5-4B	2.1% (5.2%)	93.8% (4.2%)	45.8% (30.2%)
Qwen3-VL-4B	2.1% (30.2%)	61.5% (31.2%)	28.1% (39.6%)
Gemma-3	5.2% (33.3%)	52.1% (33.3%)	20.8% (34.4%)

Table 13: Post-training carrier repair on the balanced 200-example set. Standardized repair divides mean L16 GT-minus-non-GT repair by the sample standard deviation of unpatched margins; intervals bootstrap both quantities.

Checkpoint	Raw repair	Margin SD	Standardized [95% CI]
Base	1.022	3.144	0.325 [0.216, 0.438]
SFT	0.522	2.054	0.254 [0.175, 0.347]
GRPO@150	2.524	13.940	0.181 [0.099, 0.267]
Vision-OPD	0.505	1.648	0.306 [0.213, 0.412]
OPSD	0.615	1.999	0.308 [0.217, 0.412]

therefore remain teacher-forced readouts. In a descriptive audit, the unparsed outputs include third valid values, digit forms of paired values, repeated digits, and bare option letters; the released per-pair records retain every generated string.

The pre-registration, the runner and the per-pair records, including every generated string, ship with the package.

A.12 Capacity, Post-Training, and Generalization

Capacity and depth profiles. Qwen3.5-2B has 24 decoder layers; on 96 MCQ pairs image restoration is high through L7, drops at L8, and is negligible from L9, while text restoration jumps at L8 and peaks at L12–13. Qwen3.5-9B repeats the complete pipeline without importing 4B coordinates and freezes $S/W/C/R$ at L0/L11/L16/L31; on 96 confirmation pairs raw node restoration is 14.56 $[13.51, 15.57]$, 2.58 $[2.21, 2.96]$, 17.35 $[16.33, 18.34]$, and 10.68 $[10.01, 11.33]$, and every paired interval is above zero. Because its carrier lies five layers beyond the writer, this is the setting in which the $W \to C$ adjacency is measurable; the normalized routed shares are 0.179 $[0.151, 0.209]$ for $S \to W$, 0.950 $[0.939, 0.960]$ for $S \to C$, 0.858 $[0.814, 0.898]$ for $W \to C$, and 0.571 $[0.555, 0.586]$ for $C \to R$, on the same scale as the 4B rows above.

Post-training profiles. All four training arms start from the same Qwen3.5-4B Base with the same data, processor, budget, evaluation code, and seed: supervised fine-tuning, GRPO at step 150, on-policy self-distillation (Zhao et al., 2026) from the broader self-distillation line (Hübotter et al., 2026), and its vision-conditioned form (Yuan et al., 2026). Independent 96-pair sweeps of SFT and GRPO@150 recover the Base transition region: the image cliff remains L11–12 and the text peak remains around L17, with early image-state magnitude approximately $5.6/9.9/19.1$ and peak text restoration approximately $7.4/12.3/24.0$ for SFT/Base/GRPO. Every checkpoint retains strong carrier recovery and source-to-carrier mediation at the frozen coordinates, while the leverage of individual nodes differs (Table 13). Asking the same frozen-coordinate readout after training complements concurrent analyses of what RL post-training changes in multimodal reasoning (Li et al., 2026b; Zhang et al., 2026).

Low-rank carrier compression. A PCA basis fitted on 1,748 token rows from 40 discovery donors in a 2,560-dimensional hidden space explains 43.9% of the variance at rank 32; rank-32 repair reaches 68.6% of full-state repair and significantly exceeds same-rank random, learned non-GT, and learned shuffled controls. The reported carrier results use the full state; the rank-32 basis is a compression check on the same content.

Setting and scope. PPH starts from an annotated or externally grounded region and traces what happens to it downstream, answering "how is this evidence used" rather than "where does the model look". All models are prefix-fusion decoders, where visual tokens precede text and causal attention already implies an image-to-text direction; the content of the result is therefore the role structure, the family-specific writer implementations, the fact-dependent portability of the carrier, and the interface dependence. The main scoring is teacher-forced over option letters or option texts; the separately pre-registered generation check in Section A.11 is uninterpretable under its pre-specified rule. The natural-image results use a privileged donor that carries an inset of the ground-truth region, so they measure whether frozen carriers accept corrective evidence under that construction rather than the accuracy of a general-purpose correction method. Privileged context is itself studied as a training signal in on-policy distillation (Tian et al., 2026; Shrestha & Tessier, 2026; Ichihara et al., 2026); here it serves as a diagnostic donor. The checkpoints compared in the post-training analysis are internal training runs of one base model. Quantitative reporting uses pre-specified single-cell contrasts with paired bootstrap intervals over pairs, without imputation: normalized restoration omits pairs whose clean–counterfactual gap is below ϵ, mediation ratios require an upstream raw restoration above 10^{-3}, and the effective n for each such statistic is the n printed with it.

Boundary conditions and reporting conventions. The direct-answer construction is also instantiated on TextVQA/TextOCR (Singh et al., 2019; 2021), where the answer is a localized OCR word. After answer-box corruption, source restoration remains strong and most of its effect lies outside the candidate text branch. This qualitative check is not used for the cross-family interface comparison. A GQA-200 construction with scene-graph boxes was not used as interface evidence because those boxes are frequently not answer-sufficient for global position, place, and activity questions. Every run writes a `RUN_CONFIG.json` with model and processor paths, manifest, frozen layers, sample cap, bootstrap draws, seed, and runtime, alongside pair-level metrics and aggregate statistics; discovery, validation, and confirmation artifacts are kept separate so that each confirmatory claim traces to a fixed specification. Because the specifications were frozen in dated internal logs rather than a public registry, we describe them as pre-specified.