SAME CAUSAL QUESTION, DIFFERENT LENSES: LOCALIZED VISUAL EVIDENCE AND ANSWER-SIDE COMPUTATION IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A vision–language model can answer a question correctly without using the image region that contains the answer. We introduce a region-anchored audit that connects visual evidence to answer-side computation and tests how much of that effect a proposed internal description reproduces. On compact-evidence questions, we compare answer-region masking with same-area control masking, restore clean hidden states during the masked run, and test candidate descriptions through paired clean-state restoration and masked-state writeback. Across Gemma3, Qwen2.5-VL, and LLaVA1.5, answer-region masks cause greater target-score loss than matched controls, and aligned hidden restoration selectively recovers that loss. At the tested interfaces, fitted sparse reconstructions, selected native channels, and example-shared low-rank subspaces do not reproduce their corresponding complete-reference effects. In contrast, a frozen LLaVA complete hidden-state interface restores 95.8% and removes 95.0% of the mean natural clean-to-mask score change on 60 examples with result-blind-reviewed masks, passing all four paired-bootstrap bounds. The audit thus identifies a shared visual answer dependence while distinguishing evidence-sensitive internal structure from a description that faithfully reproduces its measured answer effect.

1 INTRODUCTION

Vision–language models (VLMs) can answer questions about objects, visible text, and spatial relations [4, 19, 22]. Yet a correct answer does not show whether the model used the image region that contains it: a typical object color or the surrounding scene can suggest the same response. Language priors and the limits of saliency-based explanations make this uncertainty consequential [1, 2, 12, 15]. Recent VLM studies also locate consequential visual information and trace interpretable internal structures [11, 16, 27]. These observations occupy different evidentiary levels. A region may be associated with an answer; changing it may alter the answer; and a proposed internal object may or may not reproduce the score change caused by that alteration. We ask: *when compact answer-bearing evidence is removed, which internal descriptions account for the resulting answer-score change?*

The first challenge is to isolate visual evidence use from generic image damage. Masking any region can disrupt the input, so we compare a human-annotated answer region with an equally sized non-evidence region on the same image. We then restore the clean hidden states at corresponding model layers and token positions during the corrupted run to locate where the additional answer-score loss can be recovered. Across Gemma3, Qwen2.5-VL, and LLaVA1.5 [5, 10, 17], answer-region corruption causes greater loss than matched corruption, and this matched-position restoration selectively recovers it. The two interventions establish an evidence-to-answer response at the input and hidden-state levels and give the next challenge a concrete target: the score change produced by removing the annotated evidence.

The second challenge is to determine whether an internal description accounts for the effect it is proposed to explain. A complete hidden state, a fitted sparse reconstruction, selected model channels, and a low-rank subspace are different objects; fitted sparse features are analysis tools rather than components of the base VLM [3, 7, 9, 18]. We test each candidate at a specified interface in both directions: inserting its clean value into the masked run should recover answer support, and writing

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.

its masked value into the clean run should remove the corresponding support. Each candidate is compared with the appropriate complete intervention at that interface; the full-hidden test uses the natural clean-to-mask score loss itself. The tested compact descriptions do not reproduce their complete references under this criterion. A frozen full-hidden LLaVA interface does: on 60 separately reviewed examples, it restores and removes about 95% of the mean natural score change. The contrast reveals a *dependence–visibility gap*: visual evidence can affect answer-side computation even when a particular compact lens does not faithfully reproduce its reference effect.

We contribute a region-anchored, bidirectional test of how faithfully an internal description accounts for a visual answer effect. The same annotated region and matched control first establish the effect to explain; paired internal writes then compare each candidate with a complete reference at its own interface. Applied across three VLMs, this test reveals a shared evidence-specific answer response and measures how much of it different internal descriptions reproduce. The independently confirmed LLaVA complete hidden-state interface provides a constructive example of near-complete effect matching at a frozen answer-side location. Together, these results turn the question of whether a circuit is visible into a measurable question about how much of a specified visual effect it explains.

2 Related Work

Three lines of work lead to our question. Visual explanations ask which image content is associated with an answer. Causal tracing tests whether changing an internal state changes that answer. Sparse circuit methods make selected parts of the computation easier to inspect. Their observations are useful, but an associated region, a consequential state, and a faithful account of the score change caused by masking the evidence are different findings.

Visual relevance and regional evidence. VQA benchmarks report whether an answer is correct, but language priors can support the same answer without the relevant pixels [2, 4, 12]. Saliency and attention methods identify image content associated with a prediction, while human-supervised explanations and counterfactual images test alignment or output sensitivity [13, 15, 23, 24]. These are useful observations, yet a plausible map need not reflect the learned decision function [1], and masking any region can damage the image. For our question, the relevant input-side comparison is therefore not simply clean versus masked: it is the answer-score damage from masking annotated evidence versus masking comparable non-evidence content on the same image. Once this additional damage is identified, the next question is where the score lost under the evidence mask can be recovered inside the model.

Internal causal localization. Causal tracing and activation patching restore clean internal states during a corrupted computation and measure how much of the lost answer support returns [20, 25, 26]. Their conclusions depend on the chosen corruption and answer metric [28]. In VLMs, NOTICE uses paired semantic images and text replacements to identify consequential attention heads [11]; Fine-grained Cross-modal Causal Tracing tests visual and text positions and several component types across layers to study object representations [16]. These studies establish that changing selected multimodal states can alter an answer. We build on this intervention logic by fixing a human-localized visual effect against a same-area control and testing both clean-state restoration and corrupted-state replacement. This paired design also creates a measured effect for evaluating a proposed sparse or compact description.

Sparse descriptions and causal fidelity. Sparse autoencoders describe dense activations with a small number of learned features; transcoders instead use sparse features to approximate a model's feed-forward update [6, 7]. A per-layer transcoder approximates one layer's update, whereas a cross-layer transcoder can describe contributions to later layers [3, 7]. These are separately fitted measurement tools, not components that the base VLM uses to answer. VLM studies have traced visual concepts with attribution graphs and tested feature sets through steering or circuit patching [27, 29]. Circuit-faithfulness research already distinguishes structural overlap from functional reproduction of a circuit's effect [14]. We bring that criterion to a human-localized VQA effect: can selected features reproduce the answer-score change associated with removing the annotated visual evidence? Native circuits can use different components for analogous tasks across modalities [21]; low-dimensional subspace interventions test distributed variables but require care when interpreted as natural mechanisms [9, 18]. We compare these different candidate descriptions with a complete

reference at each tested interface rather than align feature identities across models. The next section introduces the base-model states and fitted interfaces; the Method defines how their candidate changes are tested.

3 PRELIMINARIES

A vision–language model receives an image I and question q and generates an answer token by token. We write $x = (I, q)$ for this input and $a = (a_1, \ldots, a_T)$ for a fixed target answer of T tokens. At layer ℓ and visual or text position i, the residual-stream state $\mathbf{h}_i^\ell$ is the model's complete internal vector at that location; the model uses the image, question, and preceding answer tokens to predict each next token. We call this full vector a *dense* state [8].

Sparse transcoders offer a second, separately fitted view of that computation. A per-layer transcoder (PLT) approximates one feed-forward update with a small set of active features [7]; a cross-layer transcoder (CLT) can describe contributions to later layers [3]. Neither is a component the base model uses to produce its answer. Our question is therefore twofold: does masking a human-annotated answer-bearing region lower the target-answer score more than masking comparable non-evidence content, and which specified internal object reproduces the score change caused by the evidence mask? A dense state, a fitted sparse reconstruction, selected native feed-forward channels, and a low-dimensional subspace are different candidate objects, not interchangeable circuits. Appendix A gives the full model notation and sparse-interface definitions used below.

4 METHOD

The audit first measures the additional score loss from masking annotated answer evidence rather than comparable non-evidence content. It then asks whether a specified internal object reproduces the score change caused by the evidence mask. Within each internal comparison, we hold the image–question example, answer target, corruption, and readout fixed. We compare effects at specified interfaces without assuming that sparse features or internal coordinates are shared across models.

4.1 ANSWER SCORES AND MATCHED VISUAL DAMAGE

For an image–question input x and fixed target answer $a = (a_1, \ldots, a_T)$, a condition c specifies one complete model run, including any image corruption or internal intervention. We supply the fixed prefix before each answer token so that all paired runs score the same answer. If p_t^c is the model probability of target token a_t under this condition, the primary complete-answer score is

$$S_{\text{seq}}(x, a; c) = \sum_{t=1}^{T} \log p_t^c. \tag{1}$$

First-token logit and rank scores and correct-versus-wrong margins provide complementary readouts; their exact definitions are in Appendix B. Let c_0 be the clean run, c_{evid} the annotated evidence-mask run, and c_{ctrl} a run masking a same-area random non-evidence region. For any fixed answer score S, define

$$D_S(x, a; c) = S(x, a; c_0) - S(x, a; c), \tag{2}$$

$$E_S(x, a; c_{\text{evid}}, c_{\text{ctrl}}) = D_S(x, a; c_{\text{evid}}) - D_S(x, a; c_{\text{ctrl}}). \tag{3}$$

The first quantity is the clean-to-masked answer-score loss. A positive second quantity means the annotated evidence mask causes more loss than its matched control. Figure 1 illustrates these matched inputs on one question before the aggregate tests.



(a) Original (b) Evidence masked (c) Same-area control

Figure 1: **The matched visual comparison on one compact-evidence question.** For "What language is the sign on the bear in?" the target answer is "Spanish." Gray corruption covers the annotated text in (b) and an equally sized non-evidence region in (c). The panels illustrate the paired input conditions, not a single-case mechanism result; the full human annotation is shown in Appendix G.

4.2 Testing internal objects in both directions

At a specified layer and position group, let o be the complete dense state or a restricted candidate derived from it. $\texttt{Restore}(o, c)$ writes the clean value of o into a corrupted run c. $\texttt{Delete}(o, c)$ is shorthand for writing the value recorded under c into the otherwise clean run; it does not zero the object. Their target-score effects are

$$B_S(x, a; o, c) = S(x, a; \texttt{Restore}(o, c)) - S(x, a; c), \tag{4}$$

$$A_S(x, a; o, c) = S(x, a; c_0) - S(x, a; \texttt{Delete}(o, c)). \tag{5}$$

Restoration measures recovered support; corrupted-state replacement measures support lost from the clean run. The two writes use the same paired state difference in opposite baseline runs, so they are complementary intervention tests rather than independent samples. A clean state copied into a damaged run can improve an answer without showing that the state was needed in the clean run; a masked state inserted into a clean run can lower the score without showing that the clean state would rescue a damaged input. We therefore compare both effects with the same natural clean-to-mask score loss. A large intervention effect alone is not enough: recovery far above the natural loss also fails to reproduce it faithfully.

Candidate classes include complete hidden states, a fitted sparse reconstruction and its remaining reconstruction error at the same feed-forward output, selected native feed-forward channels, and a low-rank hidden-state subspace shared across examples within one model. The sparse tool is a separate measurement interface; native channels belong to the base model. References differ: full feed-forward output for sparse reconstruction, all channels for native groups, and the natural clean-to-mask score change for shared subspaces and full hidden states. Each candidate must meet its own reference and control gates in both directions before we call it a complete path. Appendix B gives the feature selection, decoder reconstruction, residual, subspace projection, and intervention details; the next section fixes the evaluation sets and statistical criteria.

5 Experimental Setup

We evaluate three frozen vision–language models with model-specific sparse measurement tools: Gemma3 and Qwen2.5-VL use per-layer transcoders, and LLaVA1.5 uses a cross-layer transcoder [3, 5, 7, 10, 17]. The tools are fitted separately from the base models. The regional evaluations draw from compact-evidence questions in the OK-VQA validation split [19]; Appendix C specifies each evaluation set and its selection rule. Cross-model comparisons align the image–question examples, visual corruption, fixed answer target, intervention direction, and statistical interpretation, not raw feature identities or score magnitudes.

Table 1: **Evaluation roles and independent units.** Dash: no interface selection; 61+52: separate evaluations; 60/61: accepted/reviewed masks.

Comparison	Selection	Evaluation	Inference unit
Regional and hidden-position	–	61 + 52	image–question
Answer-position, sparse, native	60	240	48 clusters
Shared-subspace rank	40 fit + 20 select	40	image–question
LLaVA full-hidden confirmation	frozen interface	60 / 61 masks	image–question
Matched late-answer diagnostic	rule from LLaVA	same 60	image–question

Table 1 separates groups that answer different questions. The 61-example expanded set and 52-example set selected without model results test whether regional damage and hidden-position preferences repeat. The 240-image set reuses previously analyzed images to compare internal candidates under stronger intervention controls; it is not a fresh confirmation set. Forty, twenty, and forty disjoint images respectively fit, select, and replicate the shared-subspace candidates. Only the final LLaVA full-hidden test uses newly reviewed evidence masks after freezing its interface and decision rule. The same 60 masks are subsequently reused in a three-model comparison whose relative-layer rule was chosen after the LLaVA finding.

Every paired comparison uses the same fixed short-answer tokenization across clean, evidence-masked, control-masked, and internally intervened runs. The 61- and 52-example regional analyses resample image–question examples; their individual 95% intervals are unadjusted across panel cells and describe the direction and recurrence of the matched contrasts. The 240-image tests resample 48 entity–relation clusters and jointly correct their comparison families. Shared-subspace and complete hidden-state candidates must restore and remove 80–120% of the mean clean-minus-evidence-mask sequence-score change. For independent confirmation, we resample image–question examples while keeping each example's clean, masked, and intervened scores together. In each of the two write directions, the 95% lower bound must be nonnegative both for the candidate effect minus 80% of the natural loss and for 120% of that loss minus the candidate effect: four bounds in total. The later three-model result is diagnostic rather than jointly prospective. Appendix C gives all splits, target rules, reconstruction diagnostics, exact resampling counts, and model-specific layer settings.

6 Experiments and Results

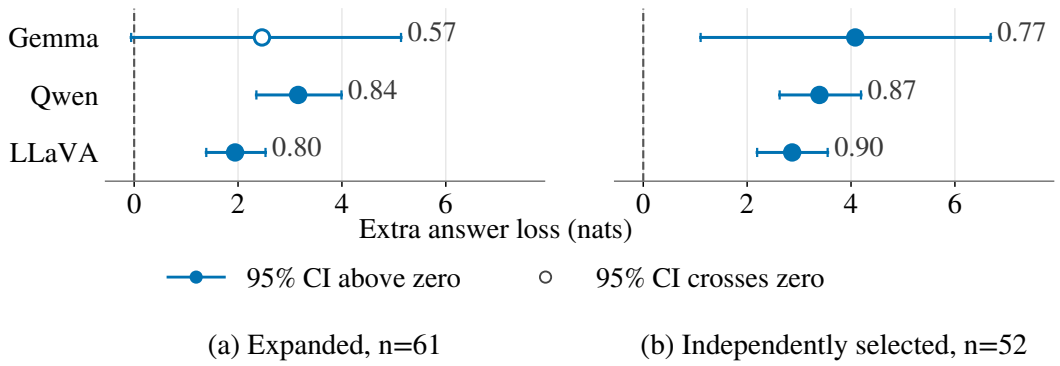
The experiments follow the three evidence levels introduced above. Matched masks establish whether the annotated region contributes more answer support than comparable image content; hidden restoration tests where that support is recoverable; bidirectional interventions compare proposed internal objects with their corresponding complete references. The final LLaVA confirmation directly matches a full-hidden intervention to the natural clean-to-mask score change. Appendix D contains the complete numerical results and controls.

6.1 Localized evidence changes answers beyond generic image damage

Figure 2 tests the first level using the complete-answer sequence score. In the expanded 61-example set, Qwen and LLaVA have intervals above zero and Gemma's interval crosses zero; all three intervals lie above zero in the independently selected 52-example set. The first-token rank intervals are above zero for all six model-by-set comparisons (Appendix D). Together, these scores show that the annotated region contributes answer support beyond generic equal-area image damage in this compact-evidence setting.

6.2 Clean hidden states recover the evidence-specific loss

Figure 3 establishes the second level using the raw target-logit score: aligned clean hidden states recover more under answer-region corruption than under same-area random corruption. Gemma and Qwen show positive intervals across pooled visual positions, while LLaVA shows a positive interval at predefined answer-adjacent positions rather than across all visual positions. The first-token rank contrasts support the same position pattern (Table 2). The independently selected 52-example grid retains this pattern.



(a) Expanded, n=61 (b) Independently selected, n=52

Figure 2: **Answer-region corruption reduces complete-answer support beyond matched occlusion.** Panels (a) and (b) show separate sets of 61 and 52 compact-evidence examples on a shared raw log-probability scale. Each mark is the mean additional complete-answer score loss from masking the answer region rather than a same-area random region; whiskers are unadjusted 95% sample-bootstrap intervals. Filled markers have intervals above zero; the hollow marker crosses zero. Numbers give positive-sample fractions. First-token rank results are in Appendix D.

A separate 240-image evaluation freezes answer-position layers on 60 development images, then compares correct restoration with equal-norm random directions, reversed differences, and wrong insertion positions. All 12 jointly corrected differences are positive for the aligned restoration relative to their baselines or controls, connecting the regional score loss to specific internal content, direction, and location (Appendix D).

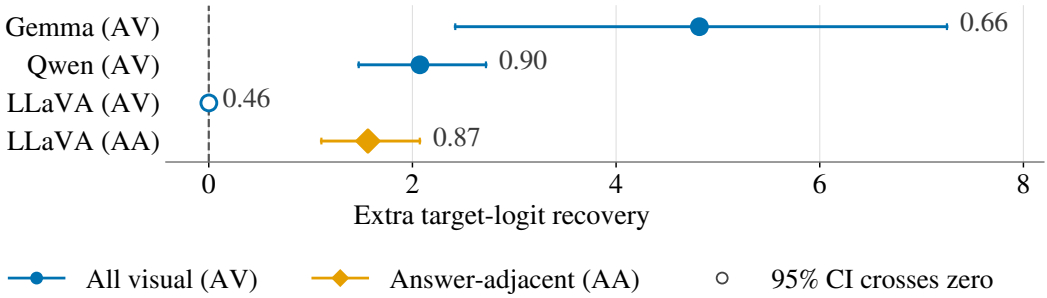


Figure 3: **Aligned hidden restoration recovers additional target-logit support under evidence masking.** Raw answer-region-minus-random restoration effects on 61 examples; whiskers are unadjusted 95% sample-bootstrap intervals. Filled marks lie above zero, the hollow mark crosses zero, and numbers give positive-sample fractions. AV pools visual positions; AA uses four answer-adjacent positions. Rank contrasts appear in Table 2.

Table 2: **Rank recovery follows the logit position pattern.** Δs_{rank} is the answer-mask restoration gain minus the same-area-random-mask gain on 61 examples. Brackets are unadjusted 95% sample-bootstrap intervals; AV pools visual positions, and AA uses four answer-adjacent positions.

Model–position	$\Delta s_{\text{rank}} \uparrow$	95% interval	Positive fraction
Gemma (AV)	14,444.7	[7,248.8, 23,404.2]	.738
Qwen (AV)	2,711.5	[549.4, 6,147.1]	.869
LLaVA (AV)	1.23	[-.25, 2.96]	.393
LLaVA (AA)	500.2	[124.9, 1,017.9]	.836

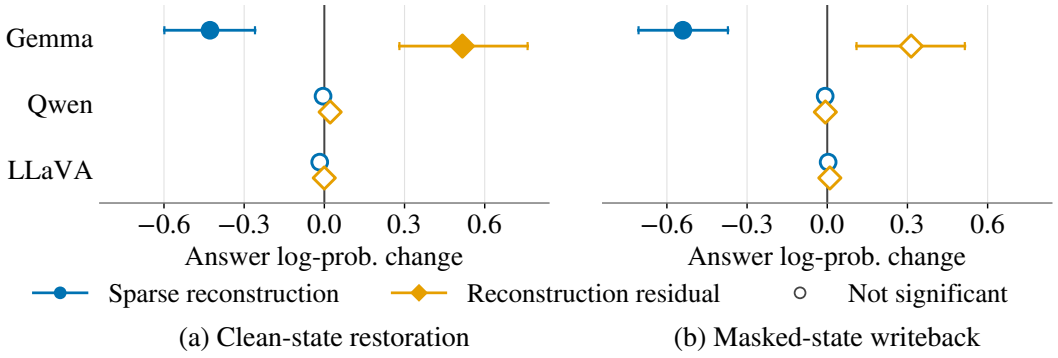
6.3 Effect matching distinguishes internal descriptions

The third level compares each candidate's two score effects with the complete intervention at its own interface. We first test compact fitted, native-channel, and shared-subspace descriptions, then use a frozen complete hidden state as a positive reference.

6.3.1 Readable structure does not guarantee effect matching

The sparse comparison begins with reconstruction fidelity. On a separate 60-image development set (59 valid images for LLaVA), at 128 selected features the fitted tools have mean relative squared reconstruction errors (residual squared norm divided by full-change squared norm) of 1.76, 1.87, and 1.62 for Gemma, Qwen, and LLaVA, with mean direction similarities of .23, .22, and .20. Each error exceeds one: the squared reconstruction residual is larger than the squared complete change. The fitted tools therefore poorly reconstruct the internal change they are asked to explain (definitions in Appendix C; per-model counts in Appendix D).

At matched feed-forward outputs on the same 240 evaluation images per model, Figure 4 shows no positive effect from the selected sparse reconstruction after correction over 72 tests. Yet the fitted tools' normalized feature summaries differ in 16 of 24 paired comparisons (Appendix Table 7). Readable structural variation and faithful reconstruction of a controlled answer effect are therefore distinct findings.



(a) Clean-state restoration (b) Masked-state writeback

Figure 4: **Selected sparse reconstructions do not yield a reliable positive answer effect.** Panels (a) and (b) show restoration and masked-state writeback at the same feed-forward output on 240 images per model. Whiskers are uncorrected 95% cluster-bootstrap intervals on a shared raw answer-score scale; filled marks survive the joint 72-test correction. Zero marks no score change.

Native feed-forward channels remove the fitted-tool reconstruction error and reveal local upstream–downstream links in all three models under positive-only selection. Their score effects, however, over-amplify the full-channel reference: Gemma's selected-group restoration changes the sequence score by 16.47, versus 3.79 for all channels (Appendix Table 8). The two frozen channel rules select separately for each image using its known answer, and neither meets the full set of effect-matching criteria.

An example-shared linear subspace offers a different compact description. Figure 5 shows the eight candidate ranks on the 20-image selection set: no rank from 1 to 128 enters the effect-matching band in both directions. Table 3 then reports diagnostic rank-128 effects on a separate 40-image set. Retaining hidden-change energy is not sufficient to match the answer effect: restoration can overshoot while masked-state writeback reverses direction. Appendix D reports the full controls and matched late-answer test.

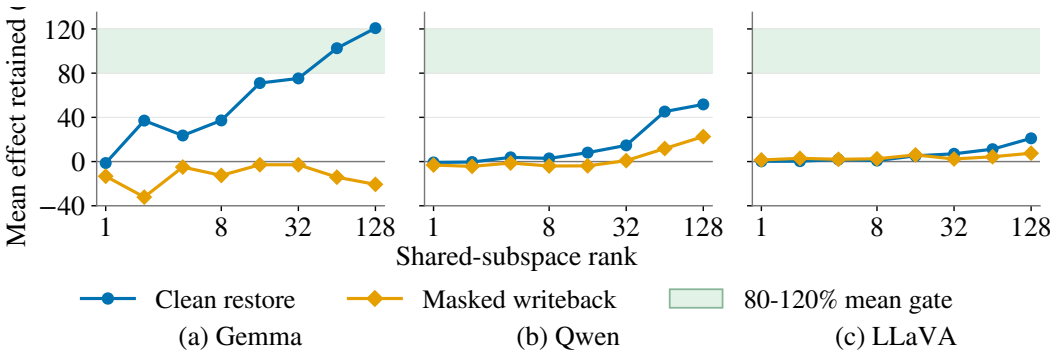


Figure 5: **No shared rank matches both directions.** Model-specific bases are fitted on 40 images; curves show restore/writeback effects for ranks 1–128 on a separate 20-image selection set, as percentages of mean natural clean-to-mask score change. Both must enter the shaded 80–120% band at the same rank; none does.

Table 3: **Hidden-change energy does not ensure answer-effect matching.** Diagnostic rank-128 values on 40 separate images. Energy is retained squared hidden-change norm; restore/writeback are percentages of the mean natural score change. No rank passed selection.

Model	Energy (%)	Restore (%)	Masked writeback (%)
Gemma	88.9	425.0	-357.7
Qwen	53.1	53.0	51.2
LLaVA	15.9	10.5	0.1

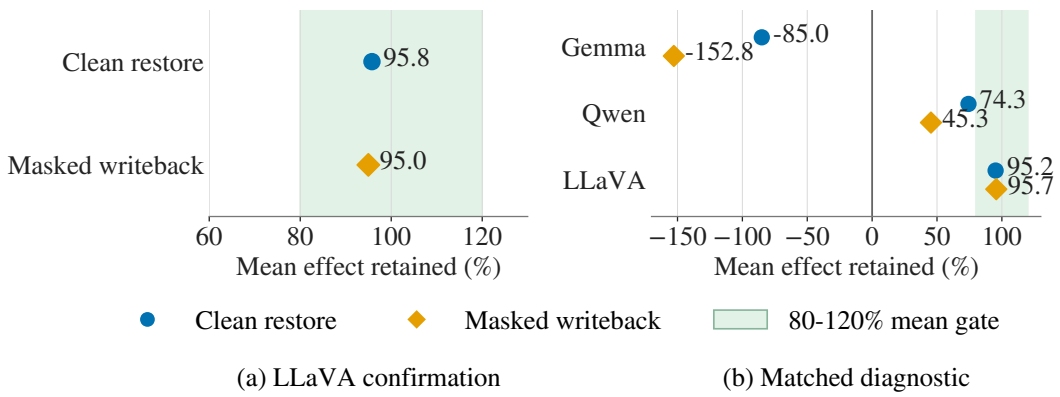
6.3.2 A COMPLETE LLAVA HIDDEN STATE MATCHES BOTH DIRECTIONS

The complete hidden state provides a positive reference for effect matching. At LLaVA layer 30 and four answer-adjacent positions, restoration and masked-state writeback retain 93.0% and 91.0% of the mean clean-to-mask score change on the 20-image selection set, then 94.9% and 92.4% on a separate 40-image replication set. We freeze that interface before testing 60 accepted evidence masks reviewed without model results. On this independent set, clean-state restoration recovers 95.8% and masked-state writeback removes 95.0% of the mean natural clean-to-mask sequence-score change. All four predeclared paired-bootstrap bounds pass (Appendix Table 10); a post-hoc 15–30% tolerance check gives the same decision (Appendix D.6). The same frozen internal location therefore nearly reproduces the natural score change in both directions.

A later same-sample comparison tests the penultimate language layer and four answer-adjacent positions in all three models. For LLaVA, this rule reuses the confirmed layer and positions under a common three-model run rather than selecting a new interface. At these matched relative locations, LLaVA again passes all four bounds (95.2% restoration; 95.7% writeback), Qwen reaches 74.3% and 45.3%, and Gemma's mean effects reverse direction. This common-location rule was chosen after the LLaVA confirmation, making the three-model comparison a diagnostic of the specified interfaces. Figure 6 places it beside the independently confirmed LLaVA result; Appendix D reports its gate margins and sample-level accounting.

7 DISCUSSION

The region-anchored audit provides a common target for comparing different internal descriptions. Matched corruption shows that changing the annotated region causes additional answer-score loss; aligned hidden restoration locates positions where part of that loss can be recovered. The paired internal writes then ask a stronger question: does a specified object reproduce the effect of the corresponding complete intervention in both directions? The evidence mask defines the visual change under study, while each internal candidate is judged against its own complete reference.



(a) LLaVA confirmation (b) Matched diagnostic

Figure 6: **One independently confirmed complete hidden-state interface; a separate matched diagnostic.** Each mark divides the within-model mean restoration or masked-state-writeback effect by the mean clean-minus-masked sequence-score change. The shaded 80–120% band is a mean gate, not a confidence interval. Panel (a) confirms the frozen LLaVA interface on 60 accepted masks with all four paired-bootstrap bounds passing. Panel (b) reuses those masks under a relative-layer rule chosen afterward; only the evaluated LLaVA interface passes all four bounds. Negative ratios reverse the mean effect; the panels use independent horizontal scales.

The results make this distinction consequential. The fitted sparse tools display different feature structures across models, yet poorly reconstruct the evidence-induced internal change and do not yield a reliable positive sparse-reconstruction score effect. Native channels reveal local interactions without fitted-tool error, but their selected groups overshoot the complete-channel reference. Shared low-rank projections test a distributed alternative and also miss the bidirectional gate. This **dependence–visibility gap** separates an evidence-specific answer response from incomplete reproduction by a specified internal description. Its size depends on the tested model–interface pair.

The independently confirmed LLaVA result gives a constructive answer-side example: at one frozen interface, complete hidden states nearly restore the natural score loss when copied into the masked run and remove nearly the same amount when their masked values are written into the clean run. This full-effect reference gives the compact-candidate comparisons a measured target. The resulting audit can compare internal descriptions across models without requiring their feature coordinates to align: measure the visual score change, test both directions of the candidate intervention, and compare with the appropriate complete reference. Appendix E details the evaluation scope, interface-specific boundaries, and annotation workflow.

8 Conclusion

We introduced a region-anchored audit that measures whether an internal description reproduces the answer-score effect of removing localized visual evidence. Across three vision–language models, annotated-region masks cause more score loss than same-area controls, and aligned hidden-state restoration selectively recovers it. The tested compact descriptions and the frozen LLaVA complete hidden-state interface span markedly different degrees of effect matching; the latter independently restores and removes about 95% of the natural score change. Bidirectional effect matching therefore provides a common, measurable standard for comparing what different representation lenses explain about visual evidence use.

AI USE STATEMENT

We used generative AI tools to assist with literature organization, experimental and figure-generation code, data analysis, and manuscript drafting and revision. The authors reviewed all AI-assisted text and code, verified reported measurements against saved result files, and compiled and visually inspected the final paper. We take responsibility for the complete content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The evaluated model and sparse-tool interfaces, human evidence masks, fixed answer targets, sample splits, intervention rules, and statistical decision criteria are specified in the Method, Experimental Setup, and Appendices A–K. The appendices preserve the full result tables, reconstruction and projection definitions, mask-review decisions, cluster-resampling procedures, and script and checkpoint identifiers needed to rebuild every reported figure and table. An anonymized snapshot of the underlying circuit-tracing code is available at https://anonymous.4open.science/r/tca-reasoning-EC3B/. External model and sparse-tool weights are identified by their public names and recorded hashes.

REFERENCES

[1] 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*, 2018. URL https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1712.00377.

[3] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[6] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

[7] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.11944.

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt,

Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html. Transformer Circuits Thread.

[9] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

[10] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and Morgane Rivière. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[11] Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025. URL https://aclanthology.org/2025.naacl-long.571/.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.00837.

[13] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1904.07451.

[14] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2403.17806.

[15] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/N19-1357/.

[16] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026. doi: 10.1609/aaai.v40i37.40431. URL https://doi.org/10.1609/aaai.v40i37.40431.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

[18] Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

[19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1906.00067.

[20] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

[21] Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms, 2025. URL https://arxiv.org/abs/2506.09047. arXiv:2506.09047.

[22] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2206.01718.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1610.02391.

[24] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1902.03751.

[25] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.12265.

[26] 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. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[27] Jingcheng Yang, Tianhu Xiong, Shengyi Qian, Klara Nahrstedt, and Mingyuan Wu. Circuit tracing in vision–language models: Understanding the internal mechanisms of multimodal thinking, 2026. URL https://arxiv.org/abs/2602.20330. arXiv:2602.20330.

[28] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

[29] Yunpeng Zhou. Sparse visual thought circuits in vision-language models, 2026. URL https://arxiv.org/abs/2603.25075. arXiv:2603.25075.

A Model computation and sparse-interface background

This section introduces the model-side objects needed to state our problem. We first describe how a vision–language model (VLM) produces an answer and what its dense internal state represents. We then introduce sparse transcoders as separate tools for viewing part of the dense internal state. With the base computation and the fitted measurement interface separated, we can state which internal objects may preserve the answer effect of localized visual evidence.

VLM computation. A VLM receives an image and a question and generates an answer one token at a time. Answer generation is autoregressive: after producing or receiving an answer prefix, the model uses that prefix together with the image and question to predict the next token. We write $x = (I, q)$ for the input image I and question q, and $a = (a_1, \ldots, a_T)$ for a fixed target answer containing T tokens. The index $t \in \{1, \ldots, T\}$ identifies one answer token, $a_{<t}$ is the prefix preceding a_t, and $\mathcal{V}$ is the set of tokens in the model vocabulary. The frozen parameters of the VLM are denoted by θ.

Before the model predicts a_t, it represents the image, question, and answer prefix as one joint sequence of visual and text positions. This sequence passes through L residual-stream layers indexed by $\ell \in \{0, \ldots, L-1\}$. At every position, the residual stream is the running internal vector that carries information from one layer to the next while the layer's attention and feed-forward calculations update it [8]. Attention combines information across positions, whereas the feed-forward calculation transforms the vector at each position. Let n_{pos} be the number of positions in the joint sequence and d the dimension of each residual-stream vector. We write

$$\mathbf{h}_i^\ell(x, a_{<t}; \theta) \in \mathbb{R}^d \tag{6}$$

for the residual-stream state at position $i \in \{1, \ldots, n_{\text{pos}}\}$ in layer ℓ. Because this vector retains all d coordinates at that position, we refer to it as a *dense* state.

After the final layer, the model produces a vector $\mathbf{z}_t(x, a_{<t}; \theta) \in \mathbb{R}^{|\mathcal{V}|}$. Each entry is the model's score for one possible next token before those scores are converted to probabilities; these entries are called logits. The entry associated with a_t therefore records the model's direct support for the next target token. Repeating this prediction step generates the complete answer.

Sparse representations. The dense state records the full internal vector at one layer and position, but its individual coordinates do not automatically correspond to separately understandable parts of the computation. A sparse transcoder provides another view. It is trained after the base VLM and kept frozen during analysis, so it is a measurement tool rather than a component that the VLM itself uses to generate the answer. The tool reads a dense state and describes a change that a layer adds to the dense state using a larger set of learned features, while allowing only a small number of those features to be active at once. This limited feature activity is what makes the representation sparse.

Two forms of transcoder are used in this paper. A per-layer transcoder (PLT) uses features to approximate the update produced by the feed-forward calculation within one layer [7]. A cross-layer transcoder (CLT) reads a feature in one layer and allows that feature to describe contributions written into later layers [3]. Applying these interfaces to VLM computation follows recent multimodal circuit-tracing work [27]. The distinction matters because a PLT feature has a within-layer target, whereas a CLT feature can have different target layers.

Let ϕ denote the frozen parameters of one model-specific sparse interface. In read layer ℓ, its encoder converts a dense state into a sparse activation vector,

$$\mathbf{f}_i^\ell(x, a_{<t}; \theta, \phi) = \texttt{Encode}^\ell\big(\mathbf{h}_i^\ell(x, a_{<t}; \theta); \phi\big)\,. \tag{7}$$

This vector lies in $\mathbb{R}^{n_{\text{feat}}^\ell}$, where n_{feat}^ℓ is the number of available sparse features in layer ℓ. The index $j \in \{1, \ldots, n_{\text{feat}}^\ell\}$ identifies one feature, and its scalar activation is $f_{i,j}^\ell = \big[\mathbf{f}_i^\ell\big]_j$. Each active feature is paired with a decoder direction that maps its contribution back into the VLM's dense vector space. The Method defines how features are selected, how PLT and CLT decoder directions are indexed by their target layers, and how selected features are reconstructed. A feature index belongs to one fitted interface and does not identify the same internal component in another model or interface.

Problem definition. A *human-localized answer-bearing region* is an image region that an annotator identifies as containing visual information needed for the target answer. Given an image–question

input, a target answer, a frozen VLM, and such a region, our first question is whether changing that region alters the model's score for the fixed answer more than changing comparable non-evidence content. Our second question is which internal object retains the answer-relevant internal change caused by the regional perturbation: the complete dense state, the part reconstructed by a fitted sparse interface, selected feed-forward channels that belong to the base model, or a low-dimensional subspace shared across examples.

These questions are related but not equivalent. Showing that changing the complete dense state changes the target-answer score does not show that a sparse reconstruction, a channel group, or a shared low-dimensional projection reproduces the same score change. Likewise, observing different sparse patterns across models does not by itself show that the underlying models use different complete mechanisms, because each restricted object may retain a different part of the dense computation. The next section defines the intervention protocol used to separate these questions.

B Detailed scoring and intervention construction

Our method compares how different vision–language models (VLMs) use localized visual evidence without assuming that their internal coordinates or sparse features are homologous. For each internal-object comparison, we hold the image–question input, evidence corruption, target answer, and answer readout fixed, then vary the object intervened on. This common input-side reference lets us evaluate dense states, fitted sparse reconstructions, native channel groups, and model-specific residual-state subspaces shared across examples without equating their coordinates or claiming that every object is tested on one identical dataset.

The method tests two hypotheses. The **regional-specificity hypothesis** predicts that corrupting a human-annotated answer-bearing region lowers the target-answer score more than matched non-evidence corruption, and that restoring aligned clean internal states recovers more score under evidence corruption than under matched control corruption. The **path-preservation hypothesis** asks whether a qualifying internal object can recover the score change caused by masking the evidence and remove a corresponding change when its corrupted value replaces the clean value. Compact sparse, native-channel, and shared-subspace candidates are tested before the complete dense state. Structural differences alone do not satisfy either requirement. This section first defines answer scores and matched causal contrasts, then the internal objects and intervention operations used to test the two hypotheses. Section 5 subsequently fixes the models, evaluation sets, selection roles, and statistical criteria.

B.1 Evaluation conditions and answer support

Our comparisons evaluate the same input x and target answer a under different complete forward-run configurations. Let $\mathcal{C}$ be the set of these evaluation conditions. Each $c \in \mathcal{C}$ specifies whether the image is unchanged or corrupted and whether selected internal states are unchanged or intervened on; it is an evaluation configuration rather than an additional text input. To score the same answer in every condition, we provide the fixed target prefix $a_{<t}$ before predicting each token rather than feeding back a condition-specific generated prefix. This teacher-forced forward pass produces condition-specific states $\mathbf{h}_i^\ell(x, c; \theta)$ and next-token logit vectors $\mathbf{z}_t(x, c; \theta)$. The target answer and its tokenization remain fixed across paired conditions, so the prefix is omitted from these condition-specific argument lists.

The probability assigned to target token a_t under condition c is

$$p_t^c = \frac{\exp([\mathbf{z}_t(x, c; \theta)]_{a_t})}{\sum_{v \in \mathcal{V}} \exp([\mathbf{z}_t(x, c; \theta)]_v)}, \tag{8}$$

where v indexes the tokens in $\mathcal{V}$. We use the answer-support set $\mathcal{S} = \{S_{\text{logit}}, S_{\text{seq}}, S_{\text{rank}}\}$. The first-token logit score is

$$S_{\text{logit}}(x, a; c) = [\mathbf{z}_1(x, c; \theta)]_{a_1}, \tag{9}$$

the complete-answer sequence score is

$$S_{\text{seq}}(x, a; c) = \sum_{t=1}^{T} \log p_t^c, \tag{10}$$

The first-token rank score is

$$S_{\text{rank}}(x, a; c) = -\texttt{Rank}(a_1, \mathbf{z}_1(x, c; \theta), \mathcal{V})\,. \tag{11}$$

The operation Rank returns the one-based descending rank of a_1 among the tokens in $\mathcal{V}$. The negative sign gives all three scores the same direction: larger values mean stronger support for the target answer. For a matched wrong answer $\bar{a}$ to the same input, the correct-versus-wrong margin is

$$M_S(x, a, \bar{a}; c) = S(x, a; c) - S(x, \bar{a}; c), \qquad S \in \mathcal{S}. \tag{12}$$

A positive margin means that the model assigns greater support to the correct target a than to the matched wrong answer $\bar{a}$.

B.2 Matched causal contrasts

Let $c_0 \in \mathcal{C}$ be the clean condition, $c_{\text{evid}} \in \mathcal{C}$ an evidence-region corruption, and $c_{\text{ctrl}} \in \mathcal{C}$ a matched control corruption. For any answer-support function $S \in \mathcal{S}$ defined above, the damage caused by condition c is

$$D_S(x, a; c) = S(x, a; c_0) - S(x, a; c). \tag{13}$$

Larger D_S means that the corruption removes more support for the fixed target answer a. The evidence-specific regional contrast is

$$E_S(x, a; c_{\text{evid}}, c_{\text{ctrl}}) = D_S(x, a; c_{\text{evid}}) - D_S(x, a; c_{\text{ctrl}}). \tag{14}$$

Thus, $E_S > 0$ means that corrupting the annotated evidence is more consequential than corrupting its matched control. The clean-to-evidence-mask damage $D_S(x, a; c_{\text{evid}})$ is a different quantity: E_S subtracts the damage from the control mask.

We next connect this input-side deficit to an internal model state. Let $\mathcal{O}$ be the set of intervention objects, where an object $o \in \mathcal{O}$ can be a complete hidden state, a projected hidden-state difference, a selected sparse reconstruction, a reconstruction residual, or a native-channel group. The condition $\texttt{Restore}(o, c)$ replaces object o in corrupted run c with its clean value. Its restoration benefit is

$$B_S(x, a; o, c) = S(x, a; \texttt{Restore}(o, c)) - S(x, a; c). \tag{15}$$

Restoration tests whether the selected object is sufficient to recover the target-answer score lost under corruption. The complementary condition $\texttt{Delete}(o, c)$ is shorthand for corrupted-state replacement: it writes the object's value from corrupted run c into the clean run c_0, rather than zeroing the object. The resulting deletion damage is

$$A_S(x, a; o, c) = S(x, a; c_0) - S(x, a; \texttt{Delete}(o, c))\,. \tag{16}$$

Deletion damage measures whether replacing the selected object's clean value with its corrupted value removes the clean-versus-corrupted target-score difference. This is a test of necessity under the specified corrupted-state replacement, not under every possible ablation. The comparisons specific to each result evaluate B_S and A_S against equal-norm random directions, reversed clean-state differences, wrong insertion positions, or full-channel references. Together, E_S and these internal controls separate evidence-specific input damage from score changes produced by restoring or replacing one internal object. Restoration alone does not test whether the clean run needs that object's clean value; corrupted-state replacement alone does not test whether restoring it recovers the masked deficit. An object passes the bidirectional interface criterion only if both directions meet their respective gates.

B.3 Dense and sparse representation changes

Applying Equation 7 to a condition-specific dense state defines $f^{\ell}_{i,j}(x, c; \theta, \phi)$. Let $\mathbf{u}^m_i(x, c; \theta) \in \mathbb{R}^d$ be the complete feed-forward output written at position i in target layer m. For a visually corrupted condition c, the prefix Δ denotes a clean-minus-corrupted difference at the same internal object and position:

$$\Delta\mathbf{h}^{\ell}_i(x, c; \theta) = \mathbf{h}^{\ell}_i(x, c_0; \theta) - \mathbf{h}^{\ell}_i(x, c; \theta), \tag{17}$$

$$\Delta\mathbf{u}^m_i(x, c; \theta) = \mathbf{u}^m_i(x, c_0; \theta) - \mathbf{u}^m_i(x, c; \theta), \tag{18}$$

The first quantity is the residual-state change used for hidden restoration. The second is the complete feed-forward-output change used for the same-object sparse comparison.

Let $\mathcal{J}^\ell = \{1, \ldots, n^\ell_{\text{feat}}\}$ be the complete feature-index set in read layer ℓ. For a fixed budget K, let $\mathcal{K}^\ell_i(x, c; K) \subseteq \mathcal{J}^\ell$ be the indices of the K largest nonnegative feature activations at position i under condition c. The clean and corrupted runs select their top-K indices separately. For feature j read in layer ℓ, let $\mathbf{w}^{\ell\to m}_j(\phi) \in \mathbb{R}^d$ be its decoder direction into target layer $m \in \{\ell, \ldots, L-1\}$. A PLT uses the within-layer case $m = \ell$; a CLT can write to a later target layer $m > \ell$. The selected-feature reconstruction of the clean–corrupted change at position i is

$$\begin{aligned}\widehat{\Delta\mathbf{u}}^{\ell\to m}_{K,i}(x, c; \theta, \phi) = &\sum_{j \in \mathcal{K}^\ell_i(x, c_0; K)} f^\ell_{i,j}(x, c_0; \theta, \phi)\, \mathbf{w}^{\ell\to m}_j(\phi) \\ &- \sum_{j \in \mathcal{K}^\ell_i(x, c; K)} f^\ell_{i,j}(x, c; \theta, \phi)\, \mathbf{w}^{\ell\to m}_j(\phi).\end{aligned} \tag{19}$$

The hat marks a reconstruction rather than a native model state. Any condition-independent affine decoder offset cancels in this difference. The corresponding reconstruction residual is

$$\mathbf{r}^m_{\text{err},i}(x, c; \theta, \phi) = \Delta\mathbf{u}^m_i(x, c; \theta) - \widehat{\Delta\mathbf{u}}^{\ell\to m}_{K,i}(x, c; \theta, \phi). \tag{20}$$

The subscript err identifies a reconstruction residual, not a second causal mechanism. Because the full change, selected reconstruction, and residual are inserted in separate nonlinear forward passes, their answer effects are not assumed to add.

B.4 Shared residual-state subspaces

Sparse reconstructions and native channels use particular fitted or axis-aligned coordinates. To test whether the clean-minus-masked hidden-state change is distributed across coordinates but lies in a low-dimensional space shared across examples within one model, we fit a separate orthonormal basis for each model. For a frozen layer ℓ and position group, let $\mathbf{B}^\ell_r \in \mathbb{R}^{d\times r}$ contain the first r orthonormal directions obtained by uncentered principal-component analysis of the clean-minus-masked differences from a designated fitting split. Section 5 fixes the fitting, selection, and replication roles. For fixed $(x, c; \theta)$, the projection of Equation 17 onto this basis is

$$\Delta\mathbf{h}^\ell_{r,i} = \mathbf{B}^\ell_r (\mathbf{B}^\ell_r)^\top \Delta\mathbf{h}^\ell_i. \tag{21}$$

The subscript r is the number of shared basis directions, and i ranges over the frozen position group. Restoration adds $\Delta\mathbf{h}^\ell_{r,i}$ to the masked run, whereas deletion subtracts it from the clean run. The full-hidden reference uses the unprojected $\Delta\mathbf{h}^\ell_i$ at the same layer and positions. The fitted basis is an analysis object constructed from intervention differences; it is not a native variable used by the VLM during generation.

B.5 Common intervention design

The primary visual treatment is corruption of the human-annotated *answer mask*, which covers the localized region supporting the short answer. When answer-bearing evidence and necessary adjacent context are separately annotated, their union provides a broader evidence condition. The primary control is a same-area random mask on the same image. Gray masking is the common corruption operator; shifted, shuffled, blurred, inpainted, patch-shuffled, dilated, and eroded variants are auxiliary checks reported in the appendix. Ambiguous or diffuse evidence is excluded rather than forced into a compact mask.

Internal interventions evaluate five distinct object classes. *Hidden-state interventions* restore or delete complete residual-stream states at predefined layers and token-position groups. *Sparse-component interventions* insert or remove a selected sparse reconstruction or its reconstruction residual at one matched feed-forward output. *Native-channel interventions* act on coordinates after the gated feed-forward calculation and before its output projection. *Native-channel path tests* jointly manipulate upstream and downstream channel groups. *Shared-subspace interventions* restore or delete the projected hidden-state difference in Equation 21. This ordering separates an internally recoverable

effect, the part visible to fitted sparse coordinates, compact groups of native coordinates, and a compact non-axis-aligned alternative before testing the complete hidden state.

Each result subsection specifies the locations and internal objects needed for its question. Hidden-state locations are compared through a predefined position grid in Section D.1. Sparse components are tested at a matched feed-forward output in Section D.2; Section D.3 then describes their normalized structures without treating them as faithful paths. Section D.4 tests native-channel candidates without fitted-tool reconstruction error. Section D.5 evaluates shared subspaces and the frozen LLaVA full-hidden candidate, ending the intervention sequence with the independent full-hidden confirmation and matched follow-up.

The next section fixes the model–interface pairs, evaluation sets, target construction, selection roles, and statistical criteria before any result is interpreted.

C Detailed evaluation setup and statistical criteria

The Method defines the common answer scores, intervention objects, and restoration–deletion operations. This section fixes the model–interface pairs, evaluation examples, selection and confirmation roles, answer targets, reconstruction diagnostics, and statistical criteria used to apply those operations. The order matters: we first specify what is compared and which data may select a configuration, then define the diagnostics and uncertainty calculations applied without changing that configuration.

C.1 Models and evaluation sets

The protocol is instantiated in three frozen model–lens pairs. **Gemma3-PLT** combines Gemma3 [10] with a per-layer transcoder (PLT), **Qwen2.5-VL-PLT** combines Qwen2.5-VL [5] with a PLT, and **LLaVA1.5-CLT** combines LLaVA1.5 [17] with a cross-layer transcoder (CLT). These names describe our measurement setups, not official variants of the base models. Because the sparse interfaces are model-specific, cross-model tests align the image–question example, corruption, answer readout, intervention direction, and statistical orientation; they do not compare raw feature identities or raw score magnitudes. Native-channel experiments use coordinates from the frozen base models and do not use the transcoders.

We use three groups of evidence-region examples for different experimental roles. The first contains filtered OK-VQA validation examples [19] whose short answer can be determined from a localizable image region. Every example contains an image, a question, a fixed target answer, and a human-annotated evidence mask. The expanded evaluation contains 61 image–question examples. A separate 52-example evaluation is selected and annotated without using model results and serves as the independently selected confirmation for the behavior and hidden-position comparisons. Appendix G reports the sample accounting and annotation protocol for these compact-evidence sets.

The answer-position, same-object sparse, native-channel, and descriptive sparse-structure experiments use a second 300-image evidence set. Sixty images select layers, feature budgets, or channel fractions; the remaining 240 images evaluate those frozen choices. The same 240 image identifiers are used for all three models. They belong to 48 predefined groups of related entity–relation questions, which are resampled as clusters rather than treating every image–question row as independent. Because these images appeared in earlier project analyses, we call this a reused confirmation set rather than previously unseen held-out data. Its primary readout is the total log probability of every token in the fixed correct answer.

The third group supports the shared-subspace and full-hidden tests through three disjoint roles. Forty images fit each model's basis, 20 different images select the smallest candidate rank that meets the restoration and corrupted-state-replacement mean gates, and a separate 40-image set is reserved for replication of a selected rank. The tested ranks are $\mathcal{R} = \{1, 2, 4, 8, 16, 32, 64, 128\}$, whose elements specify the number of retained basis directions. Earlier hidden-state screening fixed the layer and position group for each model before this subspace analysis: layer 1 at visual positions for Gemma, layer 14 at visual positions for Qwen, and layer 30 at the four answer-adjacent positions for LLaVA. The full-hidden reference is evaluated at the same interfaces. Any full-hidden interface whose mean effects meet the gate in both selection and replication is frozen for final confirmation on a separate

pool of 61 candidate masks. Human review is completed without access to model predictions or intervention results, and only accepted masks enter the final confirmation.

A follow-up matched-interface analysis addresses the initial position asymmetry. On the same 60 accepted masks, all three models receive the same intervention at their zero-based penultimate language layer and four answer-adjacent positions: Gemma layer 32 of 34, Qwen layer 26 of 28, and LLaVA layer 30 of 32. The models use the same code, PyTorch runtime, answer-sequence scoring rule, and four paired gates. The shared-subspace test is also repeated at these matched relative locations on the same 40/20/40 fitting, selection, and replication image IDs. This relative-layer rule was chosen after the LLaVA result was known, so the follow-up is a matched diagnostic rather than a jointly prospective confirmation. Model architecture and tokenization remain different.

Answer targets are fixed before each paired comparison. Allowed short-answer aliases come from dataset answers and annotation metadata; once an alias is selected for an example, its tokenization is held constant across the paired corruptions and interventions. Empty targets, formatting failures, and cross-condition token mismatches are excluded from primary estimates and retained in the evaluation accounting. Sequence scores and correct-versus-wrong margins complement the first-token scores when an answer spans multiple tokens or has plausible alternatives.

C.2 Sparse reconstruction diagnostics

The same-object experiment first checks how accurately each sparse tool reconstructs the complete clean-minus-corrupted feed-forward-output change. We evaluate this reconstruction on the 60 development images at the four text positions immediately preceding the answer. Let $\mathcal{P}_{\mathrm{ans}}$ be this four-position set and $\mathcal{P}_{\mathrm{nz}} \subseteq \mathcal{P}_{\mathrm{ans}}$ the positions whose complete change is nonzero. The relative squared reconstruction error and mean direction similarity are

$$R_{\mathrm{recon}}(x, c) = \frac{\sum_{i \in \mathcal{P}_{\mathrm{ans}}} \|\mathbf{r}^m_{\mathrm{err},i}(x, c; \theta, \phi)\|_2^2}{\sum_{i \in \mathcal{P}_{\mathrm{ans}}} \|\Delta\mathbf{u}^m_i(x, c; \theta)\|_2^2}, \tag{22}$$

$$C_{\mathrm{dir}}(x, c) = \frac{1}{|\mathcal{P}_{\mathrm{nz}}|} \sum_{i \in \mathcal{P}_{\mathrm{nz}}} \frac{\left\langle \widehat{\Delta\mathbf{u}}^{\ell \to m}_{K,i}, \Delta\mathbf{u}^m_i \right\rangle}{\left\| \widehat{\Delta\mathbf{u}}^{\ell \to m}_{K,i} \right\|_2 \|\Delta\mathbf{u}^m_i\|_2}. \tag{23}$$

The cosine uses the same $(x, c; \theta, \phi)$ arguments as the reconstruction in Equation 19; these repeated arguments are omitted inside Equation 23 for readability. A zero-norm reconstruction contributes zero similarity, while an image with no complete change at any of the four positions is excluded from the image-level mean. We first compute each scalar within an image and then average across valid development images. A value $R_{\mathrm{recon}} > 1$ means that the squared reconstruction error exceeds the squared size of the complete change, while larger C_{dir} means closer directional agreement.

C.3 Statistical inference and decision criteria

The image–question example is the inferential unit. Multiple prompt, mask, feature, or intervention rows belonging to the same example are first averaged within that example; 95% bootstrap confidence intervals then resample image–question examples rather than raw rows. We report raw row counts only as computational coverage. The 61- and 52-example panel intervals are unadjusted across cells: an interval above zero describes a positive individual contrast, not a family-wise decision. A positive estimate whose interval crosses zero is reported as directional.

For the 240-image experiments, resampling and sign-flip tests operate on the 48 entity–relation clusters. Confidence intervals use 10,000 paired cluster-bootstrap draws and tests use 100,000 sign flips. The answer-position family corrects 12 comparisons together. The sparse-interface family jointly corrects 72 two-sided tests: 36 component effects (three models, three position groups, two components, two directions), 12 random- or wrong-position contrasts (three models, two controls, two directions), and 24 cross-model structural contrasts (three model pairs, two position groups, four summaries). The three component-effect groups are visual positions, answer-adjacent text positions, and their combined set; Figure 4 plots only the combined group. This single Holm correction spans effect, control, and structural questions; the 16-, 32-, and 64-feature budget curves are descriptive and are not part of the 72 tests. The native-channel family corrects 66 tests: three models times

nine contrasts under positive-only selection and 13 under positive-and-negative selection. A contrast passes only when it meets its predeclared direction or two-sided condition after the corresponding joint correction.

The shared-subspace test uses the complete-answer sequence score and image–question examples as the inferential unit. For a candidate rank, both the mean restoration and mean corrupted-state-replacement effect must fall between 80% and 120% of the mean clean-minus-masked sequence-score change on the 20-image selection set. This reference is the clean-to-evidence-mask damage in Equation 2, not the evidence-versus-control contrast in Equation 3. Only the smallest passing rank advances to the 40-image replication set, where four paired sample-bootstrap contrasts test the lower and upper bounds for the two directions. A rank that fails selection does not receive strict confirmation. The full-hidden reference uses the same mean gate to nominate one frozen interface. Its independent confirmation requires the 95% sample-bootstrap lower bounds of all four contrasts to be nonnegative; the mask review, interface, and decision rule are frozen before model evaluation. The matched follow-up applies these four bounds to each model's 60 rows without using its results to rescan layers or positions.

Cross-model conclusions compare score-change direction, positive-sample fraction, within-model retention, and gate status. Raw score changes are not ranked across models because architectures, vocabularies, layers, and intervention coordinates differ. Appendix H records lens checkpoints, layers, and feature counts; Appendix I reports the complete rank sweep and confirmation accounting; and Appendix B defines the intervention procedure.

D Detailed intervention results and numerical tables

This appendix retains the full numerical comparisons behind the three main-text findings. It reports the region-specific behavior and hidden-position grids, the same-object sparse effects and descriptive summaries, native-channel path tests, and the shared-subspace and complete-state comparisons. The main text states the scientific result of each contrast; the tables here preserve all means, intervals, controls, and selection boundaries.

D.1 Localized visual evidence has region-specific answer effects in all three models

The first question is whether removing the annotated answer region lowers target-answer scores more than removing a comparable non-evidence region in all three models. A second requirement connects that regional loss to internal computation: restoring the aligned clean hidden state must recover more score under answer-region corruption than under matched random corruption. The first comparison establishes regional specificity, the second connects the regional loss to an internal state, and a frozen control set tests whether the recovered content, direction, and location are specific.

D.1.1 The annotated region matters beyond generic occlusion

Figure 2 compares answer-region corruption with same-area random corruption on the same image–question examples. Positive values mean that the annotated region causes greater damage to the target rank or complete-answer score. In the expanded 61-example evaluation, the rank intervals lie above zero for all three models, and the complete-answer intervals lie above zero for Qwen and LLaVA; the remaining Gemma interval includes zero. Thus, five of the six model-by-metric estimates lie strictly above zero. In the independently selected 52-example evaluation, both outcomes have intervals above zero in all three models. Removing the annotated answer region therefore causes more damage than removing an equally sized arbitrary region.

D.1.2 Hidden restoration connects the regional loss to internal computation

Output damage alone does not show whether the target-score loss can be recovered from an internal model state. We therefore restore clean residual states during the answer-region-corrupted and same-area-random-corrupted runs. For each predefined position group, we compare the restoration benefit under answer-region corruption with that under matched same-area random corruption. The

main comparison reports answer-adjacent positions and all visual positions pooled together, so neither position group is selected by improvement in the target answer.

Table 4 reports the 61-example comparison. Gemma and Qwen are clearest when all visual positions are restored together, whereas LLaVA is clearest at the predefined answer-adjacent positions. Under those model-specific position groups, both the target-rank and target-logit intervals lie above zero. For LLaVA, the all-visual intervals instead cross zero. Thus, the same region-specific answer-score pattern need not be recovered most strongly at the same type of position in every model.

Table 4: **Hidden restoration recovers target-answer scores at different position groups.** Entries report the restoration benefit under answer-region corruption minus that under matched same-area random corruption on 61 examples. Δs is the mean score change, π the positive-sample fraction, and brackets the 95% sample-bootstrap interval. AV denotes all visual positions pooled together and AA denotes answer-adjacent positions. Rank changes are interpreted within model because vocabulary sizes differ.

Model–position	$\Delta s_{\text{rank}} \uparrow$ [95% CI]	π_{rank}	$\Delta s_{\text{logit}} \uparrow$ [95% CI]	π_{logit}
Gemma (AV)	14,444.7 [7,248.8, 23,404.2]	.738	4.818 [2.421, 7.249]	.656
Qwen (AV)	2,711.5 [549.4, 6,147.1]	.869	2.073 [1.473, 2.722]	.902
LLaVA (AV)	1.23 [−.25, 2.96]	.393	.001 [−.005, .007]	.459
LLaVA (AA)	500.2 [124.9, 1,017.9]	.836	1.563 [1.106, 2.072]	.869

The independently selected 52-example position grid reproduces the same position-group preference. The mean logit restoration benefit under answer-region corruption exceeds that under same-area-random corruption in all three models, with all-visual positions preferred for Gemma and Qwen and answer-adjacent positions for LLaVA. This replication establishes that the position difference is not specific to the expanded 61-example set.

D.1.3 Frozen answer-position controls confirm content, direction, and location

The position-grid result identifies where restoration is clearest, but it does not by itself show that the restored content and direction are specific. We therefore use a separate 60-image development set to choose, for each model, the earliest layer whose mean answer-position restoration reaches at least 90% of that model's development-set maximum. The selected layers are then frozen and evaluated on the same 240 confirmation images without changing the layer, target answer, four answer-adjacent positions, or control count. Table 5 reports the confirmation results for the restoration itself and the three specificity controls.

Table 5: **Frozen answer-position restoration is specific in all three models.** The Restore column reports the mean increase in complete-answer log probability over the corrupted baseline; the three control columns are paired contrasts between correct restoration and the named intervention, not additional restoration magnitudes. A control contrast can exceed Restore when that control lowers complete-answer log probability below the corrupted baseline. Brackets are 95% cluster-bootstrap intervals on the same 240 confirmation images. Layers were selected on a separate 60-image development set. All 12 comparisons remain positive after joint correction.

Model	Layer	Restore	vs. random	vs. reverse	vs. wrong pos.
Gemma	28	3.956 [2.150, 5.781]	3.743 [1.970, 5.587]	9.778 [6.960, 12.611]	3.135 [1.444, 4.924]
Qwen	23	1.831 [1.086, 2.628]	2.108 [1.327, 2.949]	5.179 [3.947, 6.432]	1.823 [1.082, 2.613]
LLaVA	20	1.060 [0.779, 1.359]	1.105 [0.825, 1.409]	2.367 [1.833, 2.946]	0.987 [0.720, 1.274]

Correct restoration is compared with the corrupted baseline, four equal-norm random directions, the reverse of the clean-state difference, and the same difference inserted at the wrong text positions. All four comparisons pass for each model after the 12 tests are corrected together. Randomizing the restored vector, reversing its direction, or inserting it at the wrong positions each reduces the

recovery; the consistent reduction ties the restoration gain to the correct internal content, direction, and location. Combined with the regional comparison, the four specificity controls establish the same operational pattern in all three models: answer-region corruption causes greater score loss than matched random corruption, and the aligned clean hidden state selectively recovers that additional loss. The next question is whether a fitted sparse reconstruction preserves the internal change caused by masking the annotated region and produces the corresponding target-score change at its tested feed-forward output.

D.2 The current sparse tools do not reproduce a positive target-score change

Restoring the complete hidden-state difference increases target-answer scores, but the increase does not show that a fitted sparse reconstruction produces the same score change. We compare complete and sparse interventions at one matched internal object: the feed-forward output at layer 17 for Gemma, layer 14 for Qwen, and layer 16 for LLaVA. For each model, the experiment measures the feed-forward output in the complete-image and answer-region-corrupted runs, reconstructs the output difference with the sparse tool, and writes each candidate difference back at the same position. The combined intervention uses 32 positions sampled uniformly across the visual tokens and the four text positions immediately preceding the answer. Keeping the examples, answer readout, and write locations fixed separates the complete change, sparse reconstruction, and reconstruction residual. We first qualify reconstruction accuracy and then evaluate how inserting or deleting each component changes the complete-answer log probability.

D.2.1 The sparse tools reconstruct the relevant change poorly

The fitted sparse tools do not accurately reconstruct the change caused by evidence corruption in the compact-evidence questions studied here. At the primary $K = 128$ budget, Equations 22 and 23 give mean answer-position errors of 1.756 (Gemma, $n = 60$), 1.872 (Qwen, $n = 60$), and 1.617 (LLaVA, $n = 59$), with corresponding direction similarities of $.227$, $.221$, and $.199$. One LLaVA image is excluded from both image-level means because its complete change is zero at all four positions. We therefore interpret the following results as measurements of what each fitted tool reveals, not as direct measurements of a complete sparse mechanism in the base model.

Even under that restricted interpretation, the comparison itself is well controlled. Each model uses the same 240 confirmation images, the same complete-answer log-probability readout, and the same write location for the complete change, sparse reconstruction, and reconstruction residual. At most 128 nonnegative feature activations per token form the main sparse reconstruction; fixed budgets of 16, 32, 64, and 128 are all retained. Original-state writeback, vector decomposition, joint writeback, random-direction norm matching, repeated forward passes, and agreement with the native-channel answer baselines all pass their numerical checks.

D.2.2 Sparse and remaining components change target-answer scores differently

Table 6 reports the combined visual- and answer-position interventions at the main 128-feature budget. Positive restoration means that inserting a component from the complete-image run increases the

correct answer's complete-sequence log probability in the corrupted run. Positive deletion means that removing the corresponding complete–corrupted difference decreases the same sequence score.

Table 6: **The selected sparse reconstruction does not produce a reliable positive target-score change.** Results combine 32 uniformly sampled visual positions with the four text positions immediately preceding the answer and use at most 128 active features per token on the same 240 images per model. Each cell gives the mean change in complete-answer log probability and an uncorrected 95% cluster-bootstrap interval. Sparse is the selected reconstruction; residual is the reconstruction residual. $*$ marks results that remain different from zero after joint correction of 72 tests.

Model	Sparse restore	Residual restore	Sparse delete	Residual delete
Gemma	-0.428^{*} $[-0.599, -0.260]$	0.516^{*} $[0.281, 0.760]$	-0.540^{*} $[-0.706, -0.372]$	0.314 $[0.110, 0.515]$
Qwen	-0.005 $[-0.028, 0.017]$	0.022 $[0.002, 0.043]$	-0.008 $[-0.026, 0.010]$	-0.007 $[-0.028, 0.012]$
LLaVA	-0.017 $[-0.033, -0.002]$	0.000 $[-0.017, 0.018]$	0.003 $[-0.008, 0.015]$	0.010 $[-0.006, 0.025]$

No positive sparse-component score change survives the joint correction in any model. Gemma instead shows negative sparse restoration and deletion changes, while restoring the reconstruction residual increases the sequence score. Qwen and LLaVA show no component-level change after the same correction. The specificity controls reinforce this boundary: no sparse reconstruction outperforms equal-norm random directions or a wrong insertion position, and Gemma's combined-position sparse restoration is significantly weaker than both controls.

The four fixed feature budgets do not reveal a hidden positive pattern. Their mean sparse-restoration changes reverse direction across budgets and remain small for Qwen and LLaVA; Gemma's answer-position change remains negative from 16 through 128 features. These curves are descriptive because the primary joint tests use the frozen 128-feature setting. The conclusion is therefore about the current tools: the tested sparse reconstructions do not produce a reliable positive target-score change under restoration or deletion. This failure does not imply that the base models lack sparse computation, because reconstruction error may prevent the fitted interfaces from reproducing the complete internal change. Before removing the fitted tools from the next causal test, we examine what their sparse descriptions display across models and why those descriptions cannot by themselves establish a path.

D.3 The sparse tools display different normalized structures across models

The preceding same-object intervention shows that the fitted sparse reconstructions do not reliably recover target-answer scores. Their feature patterns may nevertheless describe how each fitted tool responds to the image corruption. We therefore ask how the clean–corrupted change is distributed within each sparse representation. Because feature identities are not shared across models, we compare normalized summaries rather than raw feature numbers, and we do not use these summaries to infer a faithful path.

The clean–masked support overlap is the fraction of active feature identities shared by the two image conditions. The feature fraction for 80% of change is the proportion of the tool's full feature set needed to account for 80% of the squared sparse change; a smaller value means that the change is more concentrated. The positive change-energy fraction measures how much of that squared change comes from features that increase rather than decrease. Finally, the log sparse/full norm ratio

compares the size of the sparse reconstruction with the complete internal change within the same model and position group. Table 7 reports these four summaries across models and position groups.

Table 7: **The structures seen by the sparse tools differ across models.** Values are model means on the same 240 images. "Pairs" counts how many of the three paired model comparisons remain different after the joint correction of 72 tests. These summaries describe each fitted tool and do not align feature identities across models.

Summary	Positions	Gemma	Qwen	LLaVA	Pairs
Clean–masked support overlap	Visual	0.351	0.406	0.407	2/3
Clean–masked support overlap	Answer	0.677	0.690	0.684	0/3
Feature fraction for 80% of change	Visual	0.0475	0.0660	0.0345	3/3
Feature fraction for 80% of change	Answer	0.0142	0.0099	0.0083	3/3
Positive change-energy fraction	Visual	0.512	0.525	0.510	1/3
Positive change-energy fraction	Answer	0.491	0.495	0.519	2/3
Log sparse/full norm ratio	Visual	-0.159	-0.247	-0.151	2/3
Log sparse/full norm ratio	Answer	0.079	0.125	-0.001	3/3

Sixteen of the 24 paired model comparisons remain different after all 72 tests are corrected together. The clearest pattern is concentration. At visual positions, the feature fractions needed for 80% of the change are .0660 for Qwen, .0475 for Gemma, and .0345 for LLaVA; at answer positions, they are .0142 for Gemma, .0099 for Qwen, and .0083 for LLaVA. All six pairwise comparisons for this summary remain different after correction. The other summaries are less uniform: clean–masked support overlap differs for two visual-position pairs but no answer-position pair, while positive change energy differs for one visual-position pair and two answer-position pairs.

These results establish model-conditioned structural differences in what the fitted tools display. They do not establish a shared feature identity, and they do not show that the base models themselves use the same or different complete sparse mechanisms. The earlier intervention found no reliable positive target-score change from these reconstructions despite their different normalized structures. We next remove the fitted tools from the intervention and test whether selected channels belonging to the base models account for the score change.

D.4 Selected native channels do not form a complete faithful path

The fitted sparse tools show different structures across models, but their small or inconsistent sequence-score changes may reflect the tools rather than the base models. Native channels provide a separate test without fitted reconstruction error. A native channel is one coordinate after the gated feed-forward calculation and before the layer projects that calculation back into the residual stream. It is therefore a pre-projection base-model coordinate, not a coordinate of the residual-space feed-forward output reconstructed in Section D.2. Native channels are not assumed to be individually meaningful features.

The upstream group covers all visual-token positions in the first half of each model's language layers. The downstream group covers the four answer-adjacent text positions in the second half. This split is fixed at half of the language network and is not chosen from answer effects.

On the 60-image development set, each image's known correct answer is used to score and rank its channels. The development results freeze the selection rule and retained fraction, rather than one shared list of channels. One rule retains only channels with the largest positive contributions to the answer. A second rule retains channels with the largest positive or negative contributions, preventing important suppressive contributions from being discarded. The first rule freezes 1% per layer in all three models; the second freezes 10% for Gemma, 5% for Qwen, and 10% for LLaVA. The same rule and fraction are then applied separately to each confirmation image using that image's known correct answer. The selected confirmation set is therefore answer-conditioned and image-specific. The experiment tests whether such compact per-image support sets satisfy additional path criteria; it does not identify one answer-independent channel set shared across images.

A compact path must satisfy more than a large selected-group sequence-score change. The selected upstream and downstream groups must jointly change the score more than equal-norm random groups; fixing the downstream group in its corrupted state must weaken the upstream restoration; and restoring the downstream group must recover part of an upstream deletion. A faithful path

must additionally stay within 80–120% of the corresponding full-channel score change, preserve the answer after channels outside the selection are removed, and show little recovery from those outside channels alone. Table 8 evaluates both selection rules against these criteria.

Table 8: **Strong native-channel effects do not establish a faithful compact path.** Restore and delete columns report two raw complete-answer log-probability effects separated by a slash: selected channels first, full-channel reference second. The slash does not denote a ratio. Positive-only selection retains 1% of channels per layer; the second selection retains the largest positive and negative contributions. "Local link" requires the upstream blocking and downstream bypass tests to beat matched random groups after the 66-comparison correction. A faithful path additionally requires effect matching and the tests of channels outside the selection.

Model	Selection	Kept	Restore selected/full	Delete selected/full	Local link	Faithful path
Gemma	Positive only	1%	16.472/3.791	50.084/3.294	Yes	No
Gemma	Positive + negative	10%	3.940/3.791	3.237/3.294	No	No
Qwen	Positive only	1%	4.506/1.859	12.709/1.706	Yes	No
Qwen	Positive + negative	5%	1.869/1.859	1.414/1.706	Yes	No
LLaVA	Positive only	1%	2.972/-0.288	7.082/1.935	Yes	No
LLaVA	Positive + negative	10%	0.474/-0.288	0.815/1.935	No	No

Positive-only selection produces a local upstream–downstream link in all three models, but its effects are not faithful to the full-channel reference. Gemma restoration is 16.472 compared with a full-channel effect of 3.791, and its deletion effect is 50.084 compared with 3.294. Qwen and LLaVA show the same over-amplification pattern. Selecting both positive and negative contributions brings the Gemma and Qwen restoration means close to their full-channel references, but the full set of criteria still fails.

The positive-only local links survive the 66-comparison correction in all three models, but their strong over-amplification fails full-channel effect matching. Under the positive-and-negative rule, only Qwen shows an answer-conditioned local upstream–downstream link: its blocking and bypass advantages over matched random groups have corrected p-values of $.0138$ and $.0438$. Qwen still does not pass the complete effect-matching and outside-channel tests. Gemma does not show a specific downstream handoff under this rule, and LLaVA's full-channel restoration reference is itself negative and uncertain. No model therefore provides a complete, faithful native-channel path under either frozen rule.

The native-channel test removes the fitted sparse tool from the intervention, yet no selected channel group forms a complete faithful path. This rules out the tested axis-aligned groups, not a distributed combination of channels. The remaining compact alternative is therefore a non-axis-aligned subspace shared across examples. The next section tests that alternative before returning to the complete hidden-state change.

D.5 Shared low-rank candidates fail; one LLaVA full-hidden interface passes

The sparse and native-channel tests leave a specific alternative unresolved. The clean-minus-masked hidden-state change may be distributed across individual coordinates while remaining confined to a low-dimensional residual-state subspace shared across examples within each model. We test this possibility before interpreting the complete hidden-state intervention. The experiment fits a separate basis for each model at frozen interfaces: layer 1 at visual positions for Gemma, layer 14 at visual positions for Qwen, and layer 30 at the four answer-adjacent positions for LLaVA.

D.5.1 No example-shared rank from 1 to 128 passes the bidirectional gate

For each model, 40 images fit the uncentered basis in Equation 21; 20 different images select among ranks $1, 2, 4, 8, 16, 32, 64$, and 128; and a separate 40-image set is reserved for replication. A candidate must restore and remove between 80% and 120% of the mean clean-minus-masked sequence-score change on the selection images. No tested rank passes both mean gates in any model,

so no rank advances to strict confirmation. Table 9 reports the largest tested rank on the replication images together with the full-hidden reference at the same interface.

Table 9: **An example-shared 128-dimensional subspace fails bidirectional effect matching at the initial interfaces.** Gemma/Qwen use visual positions and LLaVA uses answer-adjacent positions, each at its initially selected layer. Values are percentages of the mean clean-minus-masked sequence-score change on the separate 40-example replication set. Energy is the squared hidden-difference norm retained by the projection. Values above 100 indicate effect overshoot, and negative values indicate effect reversal. Rank 128 is the largest tested subspace; no rank from 1 to 128 passes both restoration and corrupted-state-replacement selection gates in any model. Full columns report the unprojected hidden-state reference at the same layer and positions.

Model	Rank-128 energy	Rank-128 restore	Rank-128 delete	Full restore	Full delete
Gemma	88.9	425.0	-357.7	192.3	83.2
Qwen	53.1	53.0	51.2	94.6	76.5
LLaVA	15.9	10.5	0.1	94.9	92.4

Retaining substantial hidden-difference energy is not sufficient for causal fidelity. Gemma's rank-128 projection retains 88.9% of the squared difference but over-restores the answer effect and reverses the deletion effect. Qwen retains 53.1% of the energy and approximately half of both causal effects. LLaVA retains 15.9% of the energy, restores 10.5% of the answer effect, and removes 0.1%. Failure at every selection rank therefore rejects an example-shared 1–128-dimensional complete subspace at these frozen interfaces. The tested rank range does not rule out higher-rank, nonlinear, or sample-specific representations.

D.5.2 A RESULT-BLIND EVALUATION CONFIRMS THE LLAVA FULL-HIDDEN INTERFACE

At the initial frozen interfaces, the unprojected reference gives different results. Gemma's layer-1 visual-position restoration is unstable across the selection and replication sets, while Qwen's layer-14 visual-position replacement effect remains below the 80% gate. The evaluated LLaVA layer-30 answer-position interface keeps both mean effects within 80–120% on both sets: restoration and replacement are 93.0% and 91.0% on selection and 94.9% and 92.4% on replication. We therefore freeze only this LLaVA interface for a separate result-blind test; no layer, position, rank, or threshold is rescanned. Because the three initial interfaces differ in both layer and position type, their cross-model contrast alone cannot assign the difference to model architecture.

The confirmation pool contains 61 masks reviewed without model predictions or intervention results. Human review accepts 60 masks and rejects one before model evaluation. On the 60 accepted examples, the mean clean-minus-masked sequence-score change is .9378. Restoring the complete hidden-state difference yields .8981, or 95.8% of that mean change; replacing the clean hidden states with their masked values removes .8907, or 95.0% of the same reference. Percentages use the unrounded means. Table 10 shows that all four predeclared paired-bootstrap gates have positive lower bounds.

Table 10: **The frozen LLaVA full-hidden interface passes separate result-blind confirmation.** The interface is layer 30 at the four text positions immediately preceding the answer. Restore and Delete are percentages of the mean clean-minus-masked sequence-score change; Delete writes masked hidden states into the clean run rather than zeroing them. R-L and R-U are raw sequence-log-probability lower bounds for restoration minus 80% of that mean change and 120% of that mean change minus restoration; D-L and D-U are the corresponding replacement bounds. All four bounds must be nonnegative, and Gate 1 denotes a pass.

Interface	n	Restore	Delete	R-L	R-U	D-L	D-U	Gate
LLaVA L30, AA-4	60	95.8	95.0	.050	.097	.047	.100	1

Restoring the evaluated LLaVA full-hidden interface recovers most of the clean-minus-masked sequence-score difference; replacing its clean value with the masked value removes most of that difference. The 60-example result therefore confirms this interface under both specified interventions.

It neither establishes a minimum dimension nor identifies an analogous interface in Gemma or Qwen. The initial comparison's layer-and-position asymmetry motivates one further test before we interpret the cross-model pattern.

D.5.3 A SAME-SAMPLE LATE-ANSWER COMPARISON NARROWS THE INTERFACE ASYMMETRY

For this follow-up, each model is tested at its penultimate language layer and the four text positions before the answer, using the same 60 reviewed examples and the same complete-answer score. No result is used to rescan the rule within a model. Table 11 reports restoration and corrupted-state replacement on the same image–question IDs. The relative-layer rule was motivated by the already observed LLaVA result, so this comparison is diagnostic; the original LLaVA result above remains its independent confirmation.

Table 11: **Only the evaluated LLaVA late-answer interface passes the matched bidirectional test.** All three models use the same 60 reviewed image–question examples, the zero-based penultimate language layer, four answer-adjacent positions, the same complete-answer score, and the same restored/replaced-state protocol in one runtime. n_+ counts examples with a positive clean-minus-masked sequence-score change. V is the mean of that change; R/V and C/V are ratios of within-model means in percent for restoration and corrupted-state replacement. R-L and C-L are raw sequence-score 95% bootstrap lower bounds for the respective effects minus $0.8V$; the upper bounds must also be nonnegative for Gate=1. Negative ratios denote reversed mean effects. Different models' raw score magnitudes are not compared.

Model	Layer	n_+	V	R/V	C/V	R-L	C-L	Gate
Gemma	32	29/60	0.407	-85.0	-152.8	-1.837	-2.196	0
Qwen	26	36/60	1.463	74.3	45.3	-0.483	-1.017	0
LLaVA	30	41/60	0.930	95.2	95.7	0.046	0.050	1

Only the evaluated LLaVA interface passes all four paired bounds in this matched run. Gemma's two mean changes reverse direction; Qwen restores 74.3% but removes only 45.3% of its mean clean-minus-masked sequence-score change under corrupted-state replacement. The corresponding LLaVA ratios are 95.2% and 95.7%. All 60 examples are scored in every model, but the clean-minus-masked sequence-score change is positive in 29, 36, and 41 examples for Gemma, Qwen, and LLaVA, respectively. Thus, the matched result concerns these frozen interfaces on this LLaVA-origin sample set, not the absence of another successful interface in either Gemma or Qwen. The original LLaVA confirmation and this unified-runtime rerun give slightly different point estimates but the same four-gate decision.

The same late-answer rule is also applied to the 40/20/40 shared-subspace splits. At the matched interfaces, none of the eight ranks from 1 to 128 passes both selection means in any model (Appendix I). This removes the initial visual-versus-answer position difference as an explanation for the tested low-rank failure, while leaving higher-rank and sample-specific objects open. Across the evaluated interfaces, all three models show additional damage from masking evidence rather than matched control content; the tested compact candidates fail their own complete-effect criteria, while one full LLaVA hidden-state interface passes independent bidirectional confirmation against the clean-minus-masked sequence-score change. The Discussion considers what can and cannot be inferred from that separation.

D.6 SENSITIVITY TO THE EFFECT-MATCHING TOLERANCE

The primary complete-effect rule uses the frozen 80–120% mean interval. As a post-hoc robustness check, we recompute the same four paired sample-bootstrap margins on the same 60 independently reviewed LLaVA examples at symmetric tolerances of 10, 15, 20, 25, and 30%. The 20% bootstrap lower bounds reproduce the original confirmation exactly. The low-rank column below rechecks only the two selection-set means for the eight ranks in each model; the Qwen column rechecks only the two means of the later matched-interface diagnostic. These auxiliary columns do not constitute independent confirmations.

Table 12: **Bidirectional decisions across tolerance widths.** The LLaVA minimum is the smallest of four 95% paired-bootstrap lower bounds on the independent 60-example set, in sequence-log-probability units. The low-rank count is the number of 24 model–rank candidates passing both selection means; Qwen is the later same-sample, same-relative-layer mean decision. The latter two columns do not use the independent LLaVA bootstrap gate.

Tol.	LLaVA min. bound	LLaVA 4/4	Low-rank / 24	Qwen both means
10%	$-.023$	0	0	0
15%	$+.015$	1	0	0
20%	$+.047$	1	0	0
25%	$+.076$	1	0	0
30%	$+.102$	1	0	0

The independent LLaVA point estimates satisfy the 10% mean gate, but the tighter paired-bootstrap bounds cross zero. Its four-bound decision is positive from 15% through 30%. None of the 24 low-rank selection candidates passes both mean gates in this range, and the matched Qwen interface remains outside the mean gate because masked-state writeback retains only 45.3% of its natural score change. This analysis checks the sensitivity of the original decision; it does not replace the predeclared 20% criterion.

E Detailed scope and limitations

The conclusions are bounded by six aspects of the evaluation: the compact-evidence task regime, the fitted sparse interfaces, the candidate families used for path identification, the interface-specific scope of the full-hidden confirmation and matched follow-up, answer-conditioned native-channel selection, and the annotation workflow. These boundaries restrict how the results should be generalized without changing the central finding within the evaluated setting.

Evaluation regime. The study targets compact-evidence visual question answering examples in which an answer-bearing image region can be localized and corrupted. It does not estimate how often such evidence occurs across the full OK-VQA distribution, nor does it extend the same conclusions to diffuse scene reasoning. The 61-example expanded and 52-example independently selected sets support the regional and hidden-position comparisons. The stronger content, direction, location, sparse-tool, and native-channel tests use a separate 60-image development and 240-image reused confirmation split; this split is confirmatory with respect to frozen choices but is not previously unseen project data.

Sparse tools remain model-specific approximations. The per-layer and cross-layer transcoders differ in training, target layers, and reconstruction quality. Their evidence-induced change-reconstruction errors exceed one in the tested answer positions, so the absence of a positive sparse-component answer effect does not establish that the base models lack sparse computation. Conversely, the observed structural differences describe the fitted interfaces and cannot be attributed uniquely to the base architectures.

The tested compact paths are not exhaustive. Native-channel tests cover two frozen answer-conditioned selection rules, while the shared-subspace tests cover uncentered linear projections of ranks 1–128 at the initial model-selected and later matched late-answer interfaces. Each basis is fitted separately within a model and shared only across that model's examples. Failure of these candidates excludes the corresponding channel groups and example-shared linear subspaces; it does not exclude higher-rank, nonlinear, layer-spanning, or sample-specific objects. Principal components are fitted from clean-minus-masked differences and are analysis objects rather than variables explicitly represented by the base models.

The full-hidden finding belongs to the evaluated interfaces. On 60 reviewed examples, restoring the complete hidden state at LLaVA layer 30 and four answer-adjacent positions recovers most of the clean-minus-masked sequence-score difference, while replacing its clean value with the masked value removes most of that difference. The later matched follow-up tests Gemma layer 32, Qwen layer 26, and LLaVA layer 30 at the same relative network depth and position type. Only the tested LLaVA interface passes all four bounds, but the rule was chosen after the LLaVA result and the 60

examples were not jointly unseen at that point. Different architectures, tokenization, and possible alternative layers remain. The finding does not establish minimum dimension, necessity of every coordinate, or absence of another successful interface in Gemma or Qwen. The confirmation masks were produced by GPT-assisted pre-annotation and reviewed individually without model results; 60 were accepted without correction and one was rejected before evaluation.

Native-channel selection is answer-conditioned. The native-channel experiment freezes the ranking rule and retained fraction on development images, but applies that rule to each confirmation image using its known correct answer. Both the positive-only links and the Qwen positive-and-negative link are therefore diagnostics of answer-conditioned, image-specific support sets. Establishing an answer-independent circuit would require selecting channels without using the evaluated image's target answer or confirming a shared channel set on separate images.

Annotation and wording. Evidence masks come from a single-pass annotation workflow, so independent inter-annotator agreement is unavailable. Same-area random, shifted, shuffled, dilation, erosion, blur, inpainting, and patch-shuffle controls reduce explanations based on one exact polygon or corruption operator but do not eliminate all mask uncertainty. Semantically similar prompts can also change which sparse identities or hidden locations appear most salient. We therefore claim robustness of the localized causal question, not invariance of one internal coordinate across wording.

The complete intervention tables, candidate-selection rules, sparse-tool settings, shared-subspace sweep, and additional mask and wording controls are given in Appendices D, H, I, J, and K. Within the stated evaluation regime, the three models show an evidence-specific answer-score response and the frozen LLaVA full-hidden interface passes independent bidirectional confirmation.

F Detailed reproducibility materials

For reproducibility, we separate the reported protocol into evaluation definitions, executable analyses, and external model dependencies. The evaluation definitions include the 61-example expanded and 52-example independently selected compact-evidence sets; the separate 60-image development and 240-image reused confirmation split; the 40-image subspace-fitting, 20-image rank-selection, and 40-image replication split; and the 61-mask result-blind review that yields 60 accepted confirmation examples. The follow-up uses those same 60 accepted examples for three-model full-hidden comparison and the original 40/20/40 image IDs for matched late-answer subspaces; its relative-layer rule was specified after the LLaVA result. The record also fixes mask and control construction, answer targets, model and sparse-tool identifiers, intervention layers and positions, feature and channel budgets, candidate ranks, decision gates, random seeds, and the unified follow-up runtime.

The executable analyses are packaged as scripts, manifests, mask metadata, sample-level results, and aggregate tables. The 240-image analyses retain the 48 entity–relation clusters, 10,000 paired bootstrap draws, 100,000 sign flips, and the complete 12-, 72-, and 66-comparison correction families. The subspace and full-hidden analyses retain their disjoint split manifests, all candidate ranks, paired sample-bootstrap summaries, result-blind mask decisions, and frozen protocol hashes. Existing main-text tables are rebuilt from frozen summaries by `figure_scripts/build_final_experiment_tables.py`; the original subspace and full-hidden tables are rebuilt from their JSON results by `figure_scripts/build_subspace_confirmation_tables.py`, and the matched follow-up tables by `figure_scripts/build_matched_interface_tables.py`. Rows that yield zero, negative, mixed, or unsupported outcomes remain in their corresponding comparison families rather than being silently dropped. Model weights and fitted sparse-tool checkpoints are external dependencies identified by their reported metadata and hashes.

Upon publication, we will release the human evidence masks, evaluation manifests, intervention and analysis code, and figure-source tables. Base-model and sparse-lens checkpoints remain external dependencies distributed under their original licenses; the release records their public identifiers, hashes, and feature metadata needed to reproduce the reported configurations.

G Compact-evidence annotation and mask selection

The evaluation uses image–question examples whose short answer can be supported by a human-localizable image region. Each example records the image, question, fixed short-answer target, and an answer-bearing evidence mask. Annotators mark the smallest visible region that directly supports the answer; when visually separable context is necessary, a broader union mask is also retained. Examples with diffuse scene-level evidence, visually underspecified support, or an unalignable answer target do not enter the compact-evidence comparisons. The expanded regional evaluation contains 61 examples. A separate 52-example evaluation was selected and annotated without using previous model results.



Figure 7: **Example input, evidence annotation, and matched corruption.** For "What language is the sign on the bear in?" the target answer is "Spanish." The panels show the original image, human annotation, gray answer-region corruption, and a same-area random corruption. This image illustrates mask construction; aggregate answer-score and intervention results are reported separately.

The later full-hidden confirmation uses a different pool of 61 candidate masks. Following GPT-assisted pre-annotation, a human reviewed each mask without seeing model predictions or intervention results. One mask was rejected and the 60 accepted examples were evaluated with the interface and decision rule already frozen. This result-blind review is separate from the independently selected 52-example regional evaluation. The annotation workflow is single-pass, so no independent inter-annotator agreement estimate is available.

H Sparse-lens implementation and reconstruction details

The sparse analyses use frozen, model-specific PLT and CLT transcoders as post-hoc measurement interfaces; none is trained or updated on the compact-evidence evaluation samples. At read layer ℓ, a top-K reconstruction selects the K largest nonnegative activations independently at each position and under each image condition, then writes their decoded sum to target layer m. The superscript $\ell \to m$ identifies the read and target layers, while subscript i identifies the sequence position; $m = \ell$ for a PLT, whereas a CLT can use $m > \ell$. The resulting reconstruction residual is

$$\mathbf{r}^{m}_{\mathrm{err},i}(x, c; \theta, \phi) = \Delta\mathbf{u}^{m}_{i}(x, c; \theta) - \widehat{\Delta\mathbf{u}}^{\ell\to m}_{K,i}(x, c; \theta, \phi).$$

The decomposition experiments intervene separately with the full dense difference, its selected-feature reconstruction, and the reconstruction residual under the same answer target and mask/control setting.

The final same-object comparison uses $K \in \{16, 32, 64, 128\}$ active features per token for all three interfaces and treats $K = 128$ as the primary setting. These budgets are fixed before the 240-image confirmation results and are not selected separately for favorable answer effects. Because PLT and CLT are different fitted interfaces over different model streams, the main text compares normalized within-interface summaries rather than raw feature identities or raw cross-model effect magnitudes. Table 6 reports the primary answer effects, and Table 7 reports the normalized structural summaries.

The base-model identifiers are `google/gemma-3-4b-it`, `Qwen/Qwen2.5-VL-7B-Instruct`, and `llava-hf/llava-1.5-7b-hf`. The supplementary package records these identifiers together with the lens identifiers, checkpoint hashes, layer settings, and feature metadata paths. Table 13 reports the runnable settings directly. The dictionary column gives the total number of fitted sparse features; the fixed budgets select at most 16, 32, 64, or 128 active features per token. A zero in the evaluation-training column means that no evaluation row was used to fit the lens.

Table 13: **Sparse-tool settings for the final same-object comparison.** The layer column gives the zero-based middle language layer at which the complete feed-forward change, sparse reconstruction, and reconstruction residual are written back. No evaluation image is used to train the sparse tools.

Model	Tool	Layer	Dictionary	Budgets	Images	Evaluation training
Gemma	PLT	17	10,240	16, 32, 64, 128	240	0
Qwen	PLT	14	14,336	16, 32, 64, 128	240	0
LLaVA	CLT	16	16,384	16, 32, 64, 128	240	0

Earlier source-tracing exports provide graph and feature-node diagnostics under different model-specific interfaces. We retain those diagnostics in the supplementary record, but they do not determine the final cross-model path conclusion. The main text instead uses the same-object sparse comparison to measure tool visibility and the native-channel experiment to test path closure without sparse-tool reconstruction error.

I SHARED-SUBSPACE RANK SWEEP AND FROZEN HIDDEN CONFIRMATION

This appendix supports Section D.5 by reporting the complete selection sweep at the initially selected model interfaces. Every entry is the mean restoration/corrupted-state-replacement effect as a percentage of the mean clean-minus-masked sequence-score change on the 20-example selection set. A rank would be eligible for replication only if both percentages lie between 80 and 120. None does.

Table 14: **Complete example-shared subspace selection sweep.** Each model fits its own basis. Every cell is restoration/corrupted-state-replacement retention in percent, and Pass is 1 only when both values lie in the predeclared 80–120 interval.

Rank	Gemma R/D	Pass	Qwen R/D	Pass	LLaVA R/D	Pass
1	-1.3/-13.3	0	-0.8/-3.1	0	0.2/1.5	0
2	37.1/-32.1	0	-0.4/-4.4	0	0.4/2.9	0
4	23.6/-5.0	0	3.8/-1.3	0	1.8/2.0	0
8	37.2/-12.5	0	2.9/-4.0	0	0.9/2.5	0
16	71.0/-2.9	0	8.1/-3.9	0	5.2/6.0	0
32	75.2/-2.9	0	14.6/0.8	0	7.0/2.4	0
64	102.6/-14.1	0	45.2/11.8	0	11.2/4.4	0
128	120.7/-20.6	0	51.7/22.6	0	21.0/7.5	0

The fitting, selection, and replication images are disjoint within this experiment. The basis is fitted only on the 40 fitting images, and the candidate rank is chosen only from the 20 selection images. Because no rank passes selection, the 40-image replication summaries in Table 9 are diagnostic rather than a strict confirmation of any selected rank.

The full-hidden reference uses the same initial frozen model interfaces without projection. Only the evaluated LLaVA interface has restoration and corrupted-state-replacement means inside the target interval on both selection and replication. Its separate 60-example confirmation uses a frozen manifest produced after result-blind human mask review. The review accepts 60 of 61 candidates, requires no mask corrections, and excludes the rejected example before model evaluation. The interface, decision rule, and sample manifest are unchanged during that confirmation run.

The follow-up repeats the rank sweep at each model's penultimate language layer and four answer-adjacent positions, using the same 40 fitting, 20 selection, and 40 replication image IDs in all three models. Table 15 gives every rank's two selection-set means. No candidate reaches the 80–120%

interval in both directions. The relative-layer rule was added after the earlier LLaVA observation and is a matched diagnostic, not a second jointly prospective confirmation.

Table 15: **No shared low-rank projection passes at the matched late-answer interfaces.** Each model fits its own basis on the same 40 images and evaluates the same 20 selection images at its penultimate language layer and four answer-adjacent positions. Cells show restoration/corrupted-state-replacement as percentages of that model's mean clean-minus-masked sequence-score change. A rank advances only if both means lie in 80–120%; all 24 candidate-model combinations fail. The 40-image replication split is retained for diagnostic summaries, not for strict confirmation of a selected rank.

Rank	Gemma R/C	Qwen R/C	LLaVA R/C
1	-43.2/-17.4	-6.2/-4.9	1.5/1.6
2	-38.2/-15.3	-5.5/-5.6	1.2/1.7
4	-35.4/-17.6	-6.3/-6.1	3.0/2.4
8	-45.0/-27.3	-4.0/-5.4	2.0/4.2
16	-48.5/-23.1	14.7/0.0	6.7/5.0
32	-40.8/-25.6	19.0/-6.1	8.3/4.6
64	-28.0/-19.7	21.1/-3.5	11.0/6.0
128	-19.3/-12.9	25.3/-3.6	23.7/12.6

The unprojected full-hidden reference is reported separately from the selected low-rank candidates. In the follow-up's 20-image selection split, Gemma restores/replaces 86.6%/85.1% of its mean clean-minus-masked sequence-score change, Qwen 88.0%/68.3%, and LLaVA 93.4%/90.9%. Gemma does not retain that mean match on the 40-image replication split: its mean clean-minus-masked sequence-score change is only 0.0184, while restoration and corrupted-state replacement change the score by 0.3904 and 0.3708. Ratios to this near-zero reference are unstable and are not interpreted. On the same replication split, Qwen's replacement remains below the lower gate; LLaVA's unprojected interface passes all four paired bootstrap bounds. These reused splits complement, but do not replace, the separate 60-example LLaVA confirmation.

J Additional controls for localized visual evidence

The main input comparison contrasts answer-region corruption with same-area random corruption on the same image. Additional operators test whether the rank pattern depends on one gray-mask implementation. Table 16 reports target-rank damage relative to the clean image; larger values indicate a larger loss of target support. The expanded answer entry ranges over gray, blur, inpainting, and patch-shuffle corruptions, whereas the random entry uses a same-area region. The independently selected columns summarize the separate 52-example gray-mask comparison. These controls support the regional answer-score finding rather than the internal path result.

Table 16: **Answer-region versus matched-random rank damage.** Expanded answer entries give the range across four answer-region operators; other entries are mean damage for the named condition. Values are interpreted within model because target-rank scales differ.

Model	Expanded answer	Expanded random	Independent answer	Independent random
Gemma	9613–16342	951	5028	−120
Qwen	2795–3131	50	297	−39
LLaVA	73–98	−2.5	46	0

Mask morphology supplies a complementary localized check. Dilation and erosion preserve the direction of the earlier Gemma hidden-state diagnostic and Qwen sparse-feature diagnostic. These model-specific checks address sensitivity to one exact polygon boundary; the shared three-model, same-area comparison remains the primary regional control.

K PROMPT WORDING AND QUESTION-PARAPHRASE ROBUSTNESS

All prompt families append the same answer format, "`The answer is <short answer>`," to the original or paraphrased question. They differ only in whether they add no cue (`B_direct`), internal step-by-step reasoning (`C_step_only`), visual evidence (`D_visual_only`), or both reasoning and visual evidence (`A_step_visual`). Some main analyses use only the `B_direct` and `D_visual_only` subset, while this diagnostic uses all four families.

The wording and question-paraphrase grid tests whether the evidence-to-answer effect is merely driven by wording. These analyses predate the final same-object sparse comparison and are retained as auxiliary identity and hidden-state diagnostics. The unified Gemma3-PLT and Qwen2.5-VL-PLT analysis separates fixed-node strength from route-identity stability. Table 17 summarizes the diagnostic counts and model-appropriate stability statistics. For Gemma, the two counts are fixed-node rows and target-aligned rows; for Qwen, they are candidate-metric rows and top-K overlap rows; for LLaVA, they are evaluated condition keys and image–question samples. Node or feature overlap compares sparse identities, structural overlap compares Gemma graph edges or Qwen layer–position–feature identities, and the hidden positive fraction is the proportion of LLaVA condition keys with a positive hidden-restoration logit effect. NA denotes an unavailable quantity, not a zero effect. Route identity is only moderately stable, so these diagnostics do not establish fixed-node invariance or positive causal coverage by the fitted sparse tools.

Table 17: **Wording diagnostic matrix.** Counts use the row-specific units defined in the preceding text. Sparse-overlap columns apply to Gemma and Qwen; the hidden positive fraction applies to LLaVA.

Analysis	Full count	Analysis count	Node/feature overlap	Structural overlap	Hidden positive fraction
Gemma graph	524	414	0.280	0.183	NA
Qwen route rediscovery	4608	576	0.596	0.409	NA
LLaVA hidden grid	576	12	NA	NA	0.852

Changing the question wording or prompt family can therefore change *which sparse identities or hidden-effect pattern is observed*, even when the image evidence and answer remain the same. This observation establishes measurement sensitivity to wording; it does not by itself identify a different native computation. We therefore avoid claiming fixed-node invariance across prompts.

The LLaVA wording grid provides a hidden-level robustness check rather than a grouped-route identity test. Its hidden-logit positive fraction is 0.852 overall and ranges from 0.833 to 0.891 across the reported prompt-family and question-variant summaries. Wording changes therefore modulate effect size without removing the hidden-level intervention signal.

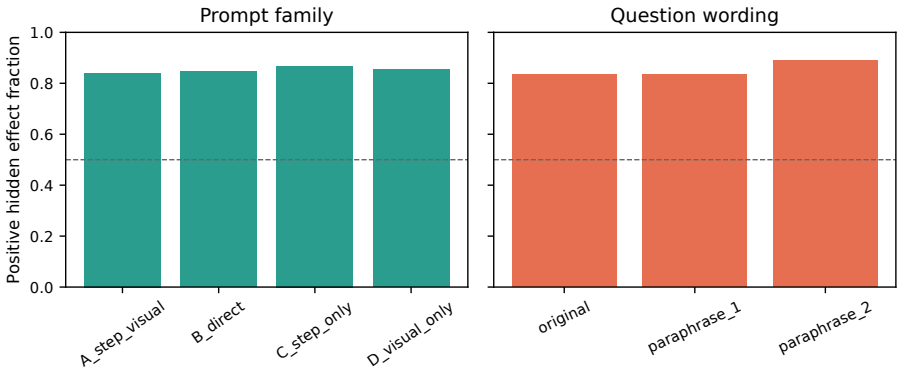


Figure 8: **LLaVA prompt and question-variant grid.** Cells summarize hidden-level effect direction across prompt families and paraphrases.

Together, these diagnostics support a careful secondary conclusion: prompt and wording changes modulate observable sparse identity and hidden-effect allocation, but do not explain away the hidden-state evidence-to-answer dependence measured by the final intervention tests.