QUESTION TOKENS RELAY AN ANSWER-LEVEL SUMMARY IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation patching asks how vision-language models (VLMs) use visual information by copying hidden states from a donor input into a recipient input during inference. However, changing the donor image usually changes both the image content and the correct answer, confounding the two. To disentangle these factors, we introduce a controlled activation-patching setup that holds the question fixed while varying whether the donor shares the recipient's answer. We construct 538 quartets from VQAv2, each pairing one question with four images, two supporting each of two answers, e.g., two with "yes" and two with "no". Within each quartet, we patch the recipient with question-token states from either a same-answer or different-answer donor at a given layer, while matching per-token norms. In all seven checkpoints we study, different-answer donors pull the recipient toward the donor's answer; in six, same-answer donors change the full output distribution far less. Yet same-answer donors are not neutral: in four checkpoints, a recipient moves one-third to two-thirds of the way toward their donors' own answer margins. Later layers thus read an answer-level summary from the question tokens, namely which answer the image supports and how strongly, and a layer sweep in three models places this summary in a narrow band of middle layers. In each of the three VLMs, a learned rank-16 component preserves the donor effect when retained and suppresses it when removed, more strongly than magnitude-matched controls. The donor effect is weakest when the intervention layer lies beyond this window, so patching studies should control both the donor's answer and the layer.

1 INTRODUCTION

When a vision-language model (VLM) answers a question about an image, its visual evidence must be turned into a decision. Where this happens and what the model carries forward from the image in doing so matter beyond interpretability: they bear on whether answers are grounded in the image or in language priors, and on where errors arise (Goyal et al., 2017; Agrawal et al., 2018). Analyses of information flow indicate that visual information is routed into the question tokens at intermediate layers (Salazar et al., 2026; Zhang et al., 2025; Kaduri et al., 2024). This raises a sharper question: *when later layers read those question-token states, do they read the particulars of the image, or only what the answer requires?*

Activation patching is a standard tool for asking such questions. It replaces selected hidden states with those computed on another input and measures how the output changes. In VLMs, it has been used to locate object information, image–text communication, and modality-specific circuits (Neo et al., 2025; Serra et al., 2025; Nikankin et al., 2026). However, what it reveals depends on the donor input (Zhang & Nanda, 2024). Existing approaches commonly compare against a noise-corrupted input, as in causal tracing (Palit et al., 2023), or use a counterfactual donor that supports a different answer (Golovanevsky et al., 2025; Wang et al., 2025a). In either case, the intervention changes the image content, its fit to the question, and the answer at once (see Figure 1a), so a patching effect at a site cannot tell us which of these factors later layers read from that site. Isolating answer-level information requires a donor that changes the image but not the answer.

We obtain such donors by explicitly controlling whether the donor supports the same answer as the recipient. VQAv2 counters language priors by pairing each question with two images that call for different answers (Goyal et al., 2017). We combine two such pairs that ask exactly the same question

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

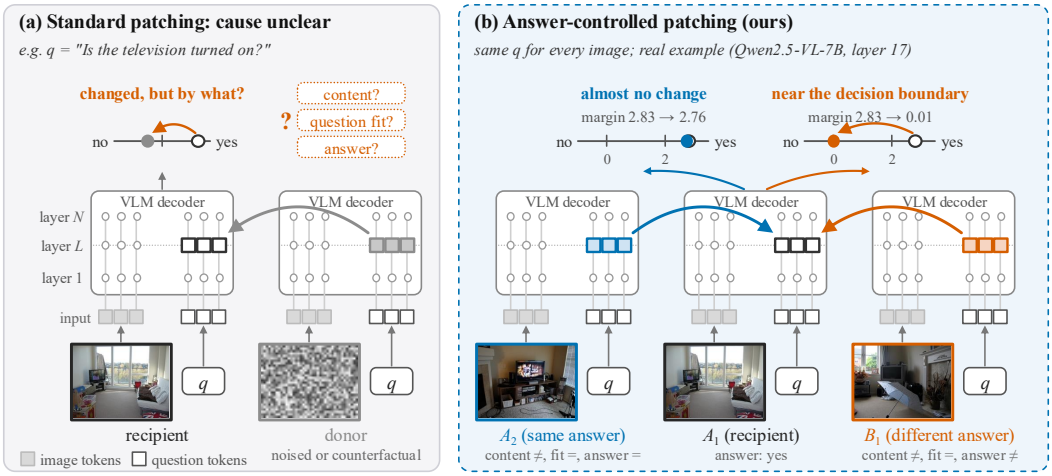


Figure 1: **Answer-controlled patching.** (a) A noised or counterfactual donor changes the image content, its fit to the question, and the answer at once (schematic). (b) With q fixed, the recipient receives the question-token states of a same-answer (A_2) or a different-answer (B_1) image. Hollow circles mark the recipient's margin without a swap and filled circles the margin after it: the same-answer donor barely changes it, whereas the different-answer donor pulls it to the decision boundary. Panel (b) is a real example from Qwen2.5-VL-7B at layer 17; more details on its selection and licences are provided in Appendix A.

into a *quartet*: one question and four images, two for each answer. In Figure 1b, the question "Is the television turned on?" comes with two images in which the television is on and two in which it is off. Each image in turn serves as the *recipient*, and another image from the same quartet serves as the *donor*, whose question-token states are copied into the recipient's run. A *same-answer donor* supports the same answer as the recipient and changes only the image content; a *different-answer donor* supports the other answer and changes both the image content and the answer. Because all four images share the same question text, their question-token states differ only in what they have read from the image. We patch these states at one layer, using one donor per run and matching per-token norms (see Figure 1b). Our central finding is that, within a narrow band of middle layers, later layers read an *answer-level summary* from the question tokens: which answer the image supports and how strongly, while other image content matters much less. Three results support this account:

The question tokens carry an answer-level summary (Section 4). In all seven checkpoints, different-answer donors shift the recipient's answer margin toward the donor's answer, far more than a random direction of equal norm; in six, same-answer donors change the full output distribution far less. Same-answer donors are nonetheless not neutral: in four checkpoints a recipient moves 33–67% of the way toward its donor's own margin, so the states transmit how strongly the image supports the answer, not only which answer it supports.

The summary resides in a middle-layer window (Section 4). A layer sweep in three models shows that answer disagreement moves the margin only within a few middle layers. In two of the three, the fixed patch layer lies past this window, which is consistent with their weaker fixed-layer results.

A low-rank component carries the summary and is used (Section 5). In each of three VLMs, a rank-16 projection learned at the patch layer retains and removes the donor effect more than norm-matched controls; it is not dormant on natural inputs, and its source carries the answer-aligned contrast. In Qwen2.5-VL-7B, a single direction within it flips 91.6% of yes/no decisions without any donor, and retaining it alone makes 32.5 percentage points more generated answers follow the donor than a norm-matched control.

For activation patching studies, this yields two concrete design implications: compare different-answer donors with same-answer donors at matched edit norms, while checking same-answer effects per recipient rather than only in the mean; and sweep layers before choosing a fixed patch site, since the donor contrast can shrink sharply outside the middle-layer window.

2 Related Work

Visual information flow in VLMs. Early work adapted causal tracing to BLIP for visual question answering (Palit et al., 2023). Subsequent analyses examined information storage and transfer in multimodal language models, including transfer through question tokens (Basu et al., 2024). Object localization and layer-wise analyses further show how visual evidence becomes available to language model computation (Neo et al., 2025), while other studies identify localized image–text communication (Serra et al., 2025), modality-specific circuits (Nikankin et al., 2026), and linear spatiotemporal mechanisms (Kang et al., 2026). Salazar et al. (2026) and Zhang et al. (2025) trace how visual information reaches the answer, including routes through the question tokens. V-SEAM uses visual semantic editing to trace object-, attribute-, and relation-level information within VLMs (Wang et al., 2025a). NOTICE uses semantic image pairs and symmetric text-token replacements to study multimodal processing without Gaussian corruption (Golovanevsky et al., 2025). These studies characterize the routes and semantic content of visual information, but leave open to what extent question-token states from different images supporting the same answer are interchangeable for answering the same question.

Patching methodology and learned interventions. Causal mediation analysis studies how internal model components transmit input effects to predictions (Vig et al., 2020). Interchange interventions test proposed causal roles by substituting representations across inputs (Geiger et al., 2021). Patching conclusions depend on how counterfactual inputs are constructed and how effects are measured (Zhang & Nanda, 2024). Distributed alignment search learns alignments between hypothesized variables and neural subspaces (Geiger et al., 2024), CDAS fits interventions to counterfactual output distributions (Bao et al., 2026), and learned subspaces can produce illusory effects (Makelov et al., 2024). Studies of in-context learning identify task vectors and function vectors whose activation-level transfer can elicit demonstrated task behavior (Hendel et al., 2023; Todd et al., 2024). Cross-modal task representations transfer across input modality and task format (Luo et al., 2025). The VLM patching studies discussed above commonly use answer-changing donors or corrupted images; to our knowledge, they do not systematically use same-answer image donors as controls for different-answer donors under a fixed question.

Probing and information removal. Probing studies distinguish information recoverable from representations from information relevant to model behavior (Belinkov, 2022; Ravichander et al., 2021). Iterative nullspace projection removes directions used by linear classifiers to decode a specified attribute (Ravfogel et al., 2020), while amnesic probing evaluates behavioral changes after information removal (Elazar et al., 2021). Both target an attribute specified in advance and ask whether it is encoded or used; a complementary question is which differences between natural images leave the downstream effect of their question-token states unchanged.

From internal diagnostics to task benefits. Analyses of visual attention sinks (Kang et al., 2025) and evidence localization (Liu et al., 2026) motivate interventions that are evaluated by task improvement. This criterion suits interventions intended to improve predictions, but diagnostic interventions can succeed while lowering accuracy, for instance by transferring an answer that conflicts with the recipient image.

3 Measuring Answer-Aligned Transfer

We ask how much a recipient's answer depends on whether the question-token states it receives come from an image with the same answer or with a different one. The procedure has three steps. We first fix a question and four images for it, two for each of its two answers. We then take one image as the recipient and, at one decoder layer, replace its question-token states with those computed on another image of the same set, the donor. Finally, we score how strongly the patched model still supports the recipient's own answer and compare same-answer with different-answer donors. We describe these steps in turn, followed by how the image sets are built from VQAv2 and which models we study.

Setting. A *quartet* consists of one question and four images: A_1 and A_2 annotated with answer a, and B_1 and B_2 annotated with answer b. Each image in turn serves as the recipient. Its *same-answer*

donor is the other image with its answer ($A_1 \leftrightarrow A_2$, $B_1 \leftrightarrow B_2$), and its *different-answer donor* is the corresponding image with the other answer ($A_1 \leftrightarrow B_1$, $A_2 \leftrightarrow B_2$); each donor is used in a separate run. Because all four images are annotated for the same question, both donors differ from the recipient in image content but only the different-answer donor also differs in answer, and the question applies to every donor. Identical patch sites and matched patch norms (below) further rule out differences in where, or by how much, the state is edited.

Patching question-token states. Let $\boldsymbol{h}_t \in \mathbb{R}^d$ be the recipient's hidden state at the input of a fixed decoder block for question token t, and let $\boldsymbol{h}_{\mathrm{s},t}$ and $\boldsymbol{h}_{\mathrm{d},t}$ be the corresponding states computed on the same-answer (s) and different-answer (d) donor. We call this block the model's *interface*. Patching replaces $\boldsymbol{h}_t$ with a donor state for every original question token and runs the rest of the forward pass unchanged. Donor and recipient tokens are aligned by verified token identity and ordinal position within the question; image tokens, system text, the assistant prefix, and answer tokens are never edited, and the recipient keeps its own image, attention mask, visual grid, and positions. Because the two donors can move the state by different amounts, we rescale both displacements to the smaller of their two norms: for $u \in \{\mathrm{s}, \mathrm{d}\}$,

$$\begin{aligned} \boldsymbol{\delta}_{u,t} &= \boldsymbol{h}_{u,t} - \boldsymbol{h}_t, \qquad B_t = \min\big(\|\boldsymbol{\delta}_{\mathrm{s},t}\|_2, \|\boldsymbol{\delta}_{\mathrm{d},t}\|_2\big), \\ \boldsymbol{h}'_{u,t} &= \boldsymbol{h}_t + B_t \frac{\boldsymbol{\delta}_{u,t}}{\|\boldsymbol{\delta}_{u,t}\|_2}. \end{aligned} \tag{1}$$

The two patches thus edit the state by the same amount while keeping each donor's direction; neither is amplified. The interfaces were fixed in advance and not reselected after the layer sweep of Section 4; more details on each model's interface are provided in Appendix A.3.

Scoring the recipient's answer. After patching, we measure how strongly the model still supports the recipient's own answer a over the alternative b, which stay fixed across conditions. Each candidate answer is scored by teacher forcing: its tokens are fed after the prompt, and at each answer position we record how highly the candidate's token ranks against the whole vocabulary, as a standardized logit averaged over the candidate's tokens. Formally, for a candidate $c = (c_1, \ldots, c_{L_c})$ under condition k, let $\boldsymbol{\ell}_{k,j}$ be the full-vocabulary logit vector at answer position j, with entry $\ell_{k,j,v}$ for token v; then

$$S_{\mathrm{z}}(c \mid k) = \frac{1}{L_c} \sum_{j=1}^{L_c} \frac{\ell_{k,j,c_j} - \mu_{k,j}}{\sigma_{k,j}}, \tag{2}$$

where $\mu_{k,j}$ and $\sigma_{k,j}$ are the mean and standard deviation of $\boldsymbol{\ell}_{k,j}$ over the vocabulary $\mathbb{V}$; EOS is not scored. Standardizing each position makes the score invariant to shifting and positively rescaling the logits there. We also record the native token-mean log-probability $S_{\mathrm{lp}}(c \mid k)$, which Section 5 uses for training. The recipient's answer margin and the source contrast are

$$m_k = S_{\mathrm{z}}(a \mid k) - S_{\mathrm{z}}(b \mid k), \qquad D = m_{\mathrm{same}} - m_{\mathrm{different}}. \tag{3}$$

A positive D means that same-answer donors leave more support for the recipient's answer than different-answer donors do.

Building quartets from VQAv2. A same-answer donor requires a second image that answers the same question in the same way, which a single VQA example does not provide. VQAv2 pairs each question with two images that call for different answers, to counter language priors (Goyal et al., 2017); combining two such complementary pairs that ask exactly the same question yields one image pair per answer, as the setting requires. At least eight of ten normalized annotations agree for every image–question record. The 538 quartets of 2,152 distinct images comprise 371 yes/no, 74 number, and 93 Other questions, where Other is VQAv2's residual answer type; 466 use train and 72 use validation annotations. Selection relies only on annotations and input eligibility, never on model predictions or intervention outcomes. Idefics3 uses 520 quartets, because its native image tiling gives the four images of 18 quartets different token layouts; more details are provided in Appendix A.

Models. We study seven checkpoints: Qwen2.5-VL-7B and -3B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3.5-8B (Wang et al., 2025b), Phi-3.5-vision (Microsoft, 2024), Idefics3-8B (Laurençon et al., 2024), and Gemma3-4B (Team et al., 2025). They differ in vision encoder, connector, and language backbone, although several share Qwen-family language models.

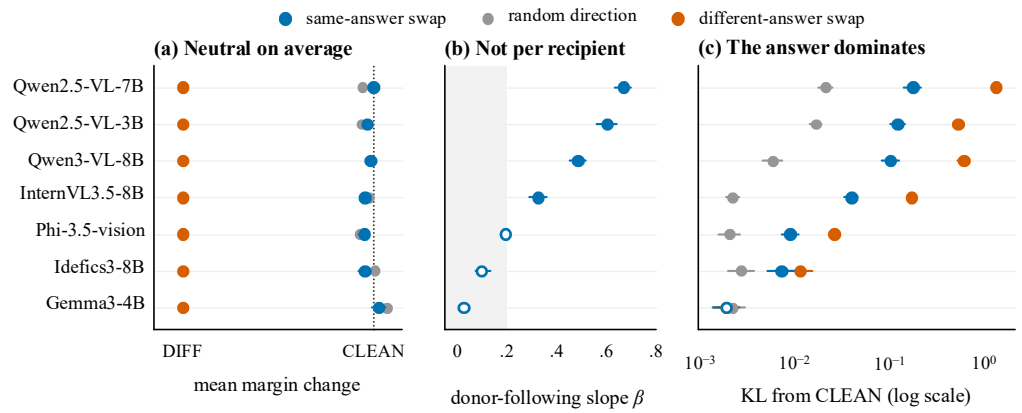


Figure 2: **Same-answer swaps are neutral on average but not per recipient.** (a) Mean margin change relative to CLEAN, in units of each model's CLEAN−DIFF gap. (b) Per-recipient donor-following slope β; hollow: following criterion not met; shaded: invariance region. (c) KL divergence from CLEAN at the first answer position. Corrected intervals; details in Appendix B.

Aggregation and inference. We average the four recipients within each quartet. For the two Qwen2.5-VL models, quartets are weighted equally; for the other five checkpoints, the three answer types are weighted equally so that frequent yes/no questions do not dominate, and Other questions are also analyzed separately. Primary intervals are bootstrap intervals over quartets or image pairs, Bonferroni-adjusted within each pre-specified family of endpoints at levels (95–99.875%) stated in each table and figure; secondary analyses report 95% descriptive intervals. Departures from the pre-specified plans, most notably that the CLEAN and random controls of the two Qwen2.5-VL models predate the protocol and are reanalyses, are flagged in the tables and listed in Appendix G.

4 The Question Tokens Carry an Answer-Level Summary

We ask whether changing the image in the question-token states affects the output mainly through a change of answer. Both donors supply states from a different image, but only the different-answer donor changes the answer; if later layers read an answer-level summary, same-answer swaps should affect the output less, while still allowing changes in how strongly the answer is supported. We examine this in four steps: the mean effect of each donor, the change per recipient, the change in the full output distribution, and the layers at which the effects arise.

Answer disagreement drives the contrast. The source contrast D is positive, with an interval excluding zero, in all seven checkpoints; it ranges from 0.038 in Gemma3 to 2.158 in Qwen2.5-VL-7B. More per-checkpoint details are provided in Appendix B. A positive D could mean that same-answer donors help or that different-answer donors hurt. Adding the unpatched input (CLEAN) and one norm-matched random direction per recipient separates the two, since $D = (m_{\text{same}} - m_{\text{CLEAN}}) + (m_{\text{CLEAN}} - m_{\text{different}})$. In every checkpoint the first term passes a pre-specified equivalence test within $\pm 0.1|D|$ and the second is positive (Figure 2a); in Qwen2.5-VL-7B they are -0.0025 and 2.161. The random direction moves the margin by at most 7% as much as a different-answer donor, so the disruption reflects what the donor supplies rather than the size of the edit.

A zero mean is expected; per recipient, same-answer donors transmit answer strength. The near-zero mean of SAME−CLEAN does not by itself establish invariance. Same-answer swaps are reciprocal within a quartet ($A_1 \leftrightarrow A_2$), so if each recipient partly adopted its donor's margin, the two changes would largely cancel, whatever the strength of adoption. We therefore test invariance per recipient by regressing each recipient's change in same-answer margin on the gap between the donor's and the recipient's clean margins. In Qwen2.5-VL-7B and -3B, Qwen3-VL, and InternVL3.5, the slope β is 0.67, 0.61, 0.49, and 0.33, with corrected lower bounds above 0.2, and absolute same-answer changes are 2.6–5.2 times those of a random direction (Figure 2b). A recipient thus inherits

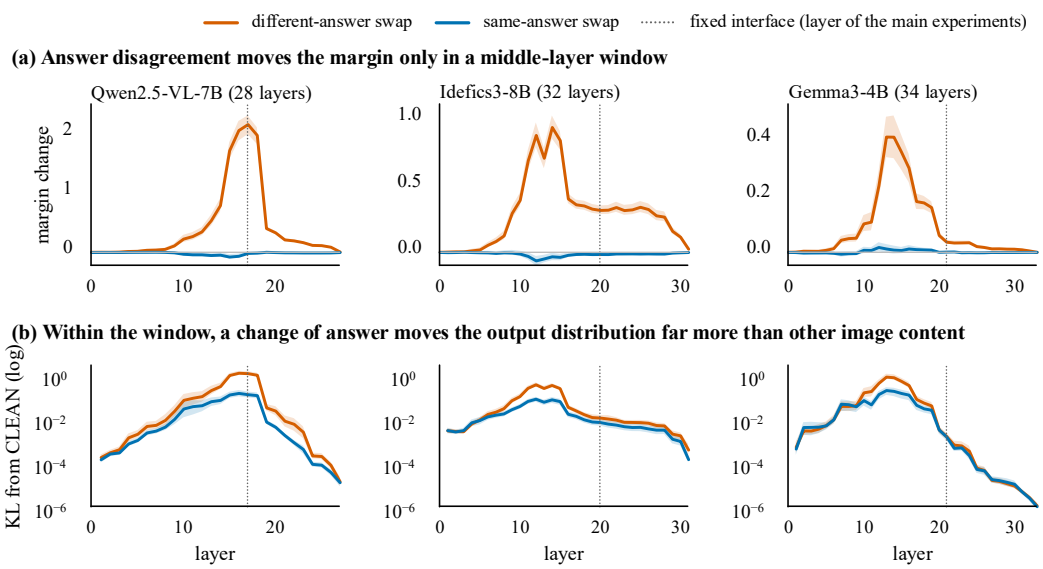


Figure 3: **Answer-level information passes through the question tokens in a middle-layer window.** (a) CLEAN−DIFF and SAME−CLEAN when the swap is applied at each layer (180 quartets per model; 95% bands). (b) KL divergence from CLEAN at the first answer position for the two swaps (log scale; descriptive). Where the curves separate, a change of answer moves the output far more than other image content; where they coincide, the two swaps have the same effect. The fixed interface is the layer patched in all other experiments.

one-third to two-thirds of its donor's answer strength. Gemma3 meets the pre-specified per-recipient invariance criterion ($\beta = 0.03$), and Idefics3 and Phi-3.5 are inconclusive. Where the states are not invariant, they carry how strongly the image supports the answer, not only which answer it supports.

Other image content matters much less than the answer. At the first answer position, the full-vocabulary KL divergence from CLEAN is smaller for same-answer than for different-answer swaps in six checkpoints, with mean ratios of 0.13–0.64 (Figure 2c); in Gemma3 the difference is not established. It is nonetheless larger than for a random direction in all six, so later layers are not blind to other image content, but in five of the six a change of answer moves the output at least three times as much. This concerns what reaches the output, not everything the states hold.

The summary resides in a middle-layer window. Sweeping the patch site over every layer of Qwen2.5-VL-7B, Idefics3, and Gemma3 shows that CLEAN−DIFF rises sharply in the middle of the network and falls again within a few layers, while SAME−CLEAN stays near zero throughout (Figure 3a). In Qwen2.5-VL-7B the peak coincides with the fixed interface (layer 17); in Idefics3 and Gemma3 it lies earlier (layers 14 and 13) than the fixed interfaces (20 and 21), where the effect is three to ten times smaller. The full-distribution comparison follows the same window (Figure 3b). In early layers, same-answer and different-answer swaps change the output about equally; inside the window, the same-answer KL falls to 0.11–0.22 of the different-answer KL at its lowest in all three models. This includes Gemma3, where the two nearly coincide at the fixed interface, so its weak fixed-layer results are consistent with where its interface falls rather than with an absence of the summary.

The effect extends to Other questions. Yes/no questions carry most of the magnitude (in Qwen2.5-VL-7B, 2.961 versus 0.647 for number and 0.156 for Other), but for Other questions alone the contrast is positive with an interval excluding zero in all seven checkpoints (post-hoc for the two Qwen2.5-VL models). More per-type details are provided in Appendix B.

Summary. Different-answer donors move the recipient toward the donor's answer, whereas same-answer donors leave the mean margin unchanged, transmit part of their answer strength per recipient,

and move the output far less; in the three swept models, this signal appears only within a middle-layer window. The question tokens thus carry an answer-level summary. Because answer agreement may co-vary with other image content, these results alone do not show that the summary is a distinct signal the model reuses across questions.

5 A Low-Rank Component Carries the Summary and Is Used

We next ask whether the answer-level summary is carried by a specific part of the question-token states that the model reuses across questions, rather than by whatever image content happens to co-vary with the answer. We address three questions: whether a single low-rank component, shared across questions and tokens, carries the donor effect more selectively than magnitude-matched edits (Section 5.2); whether it matters which donor supplies that component (Section 5.3); and whether the model uses it, both without any donor and in generated answers (Section 5.4). Section 5.1 describes the method. We study Idefics3, Qwen2.5-VL-7B (hereafter Qwen), and Gemma3.

5.1 Method: A Learned Projection and Magnitude-Matched Shams

Learning a response-matched projection. We fit one projection per model from its own states at its interface. Let $\boldsymbol{U} \in \mathbb{R}^{d\times 16}$ have orthonormal columns and $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top$. At each question token, the full displacement from the recipient to its different-answer donor is $\boldsymbol{\delta} = \boldsymbol{h}_{\mathrm{d}} - \boldsymbol{h}_{\mathrm{r}}$. FULL patches $\boldsymbol{\delta}$, and KEEP patches only $\boldsymbol{P}\boldsymbol{\delta}$. We optimize a matrix $\boldsymbol{A}$, take $\boldsymbol{U}$ as its reduced-QR factor, and keep the VLM frozen. The objective matches both candidates' log-probabilities under KEEP to their values under FULL:

$$\mathcal{L}(\boldsymbol{A}) = \mathbb{E}_{q,i}\left[\frac{1}{2}\sum_{c\in\{a,b\}}\big(S_{\mathrm{lp}}(c \mid \mathrm{KEEP}(\boldsymbol{P})) - S_{\mathrm{lp}}(c \mid \mathrm{FULL})\big)^2\right], \tag{4}$$

where q indexes fitting quartets and i their recipient images. Fitting both candidates, rather than only their margin, prevents the component from matching the margin while moving the two candidates differently. Training uses 96 quartets (32 per answer type) and one pass of 384 updates from a fixed initialization; we keep the final checkpoint. The approach is a response-matching variant of learned interchange interventions (Geiger et al., 2024; Bao et al., 2026). As a comparator, we build a fixed rank-16 *geometric* projection from the signed scatter between answer groups on the same fitting states (see details in Appendix C). Both projections are shared across examples and act independently at each question token.

Retention and removal against magnitude-matched shams. With $\boldsymbol{p} = \boldsymbol{P}\boldsymbol{\delta}$ and $\boldsymbol{c} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\delta}$, KEEP patches $\boldsymbol{p}$ and REMOVE patches $\boldsymbol{c}$. Each has a sham that follows the full direction $\boldsymbol{\delta}$ at the same norm: SHAM-KEEP (SK) at $\|\boldsymbol{p}\|_2$ and SHAM-REMOVE (SR) at $\|\boldsymbol{c}\|_2$. Two random directions at the KEEP norm serve as additional controls. A difference between KEEP and SK therefore reflects the selected direction rather than the size of the edit.

We evaluate on 320 Other image pairs (640 images, one question per pair) that were used neither for fitting nor for development, averaging the two patch directions within each pair; all three models use the same pairs. In 237 of the 320 pairs, neither answer string occurs in the fitting set. We measure the donor effect of condition k as the drop in the recipient margin, $E_k = m_{\mathrm{CLEAN}} - m_k$, so positive values mean movement toward the donor. The two primary contrasts are the *retention specificity* A_1 and the *removal specificity* A_2:

$$A_1 = \mathbb{E}[E_{\mathrm{KEEP}} - E_{\mathrm{SK}}], \qquad A_2 = \mathbb{E}[E_{\mathrm{SR}} - E_{\mathrm{REMOVE}}]. \tag{5}$$

Each contrast passes if its lower confidence bound exceeds $0.1\,s_{\mathrm{dev}}$, a tenth of the SD of FULL effects over 32 development quartets, and the FULL reference effect is positive.

5.2 Result 1: Selective Function Replicates, but Its Extent Varies

Both contrasts pass in all three models (Table 1); more details, including corrected intervals, are provided in Appendix C. How much of the full effect the component carries differs by model. In Qwen, KEEP matches FULL in mean (0.583 vs. 0.575; no equivalence test), and REMOVE is near

Table 1: **The learned rank-16 component is selective, source-sensitive, and shared across questions.** KEEP/FULL: share of the full donor effect retained. A_1, A_2: retention and removal specificity against norm-matched shams (320 held-out pairs). θ_{P}: projected-source effect for the learned (F) and geometric (G) projection (72 quartets). Variance: natural-input variance captured, relative to a random subspace. Flips: yes/no decisions flipped by a donor-free direction in the subspace or by a random direction in it (unit step). **Bold**: pre-specified criterion met.

	Carries the effect			θ_{P}			Flips (%)	
Model	KEEP/FULL	A_1	A_2	F	G	Variance	shared	random
Qwen2.5-VL-7B	1.02	**0.479**	**0.584**	**0.392**	0.207	4.19	**91.6**	4.4
Idefics3-8B	0.51	**0.276**	**0.359**	**0.077**	0.034	3.43	**11.1**	1.8
Gemma3-4B	0.29	**0.014**	**0.022**	**0.019**	0.0059	1.72	1.0	1.3

zero (-0.022 $[-0.062, 0.018]$), so the complement alone produces no detectable donor effect. In Idefics3, KEEP carries about half of FULL (0.370 vs. 0.726), and in Gemma3 less than half (0.0233 vs. 0.0794). What replicates is selectivity; the retained fraction is model-specific.

The component is not a dormant direction. Learned subspace patches can create effects by activating directions that are dormant on natural inputs (Makelov et al., 2024). The learned subspace captures 1.7–4.2 times as much natural-input variance of the question-token states as a random 16-dimensional subspace (Table 1), and in Qwen patching the complement alone produces no detectable donor effect, so the effect does not reside outside the subspace. A linear probe on the subspace coordinates separates yes from no answers with AUROC 0.91–0.99, although the geometric subspace does as well or better in Idefics3 and Gemma3.

5.3 Result 2: The Component's Source Carries Answer Selectivity

The projection was trained to reproduce the response to a different-answer donor, not to separate donors. We therefore choose the donors of its two parts independently: the crossed state $\boldsymbol{h} + \bar{\boldsymbol{p}}_u + \bar{\boldsymbol{c}}_v$ takes its projected part $\bar{\boldsymbol{p}}_u$ from donor u and its complement $\bar{\boldsymbol{c}}_v$ from donor v, with $u, v \in \{\mathrm{s}, \mathrm{d}\}$ and both parts rescaled to shared budgets, so every crossed patch has the same norm. The projected-source effect $\theta_{\mathrm{P}} = \frac{1}{2}\mathbb{E}[(m_{\mathrm{ss}} - m_{\mathrm{ds}}) + (m_{\mathrm{sd}} - m_{\mathrm{dd}})]$ is the margin gained by supplying the projected part from a same-answer rather than a different-answer donor. On 72 held-out quartets, θ_{P} is positive for the learned projection in all three models and exceeds that of the geometric projection, so both pre-specified tests pass (Table 1); it is positive under either complement source. By answer type, Qwen's projected-source effect is 0.812 for yes/no, 0.317 for number, and 0.0488 for Other questions, mirroring the ordering of the source contrast in Section 4; for Other questions, its advantage over the geometric projection is not resolved. In Qwen, supplying only the projected part from a same-answer donor raises the margin by 0.392, against 0.207 for the geometric projection. The complement also carries source information and the two parts interact, so the decomposition is partial rather than additive (see details in Appendix D).

5.4 Result 3: The Component Is Used

A shared direction in the component moves answers without a donor. Within the learned subspace, the mean yes-to-no displacement estimated on one half of the yes/no quartets defines a single direction. Added to the clean question-token states of the other half, with its sign set by the gold answer, it moves the margin toward the other answer by 5.41 more than a random direction in the same subspace in Qwen and by 1.15 more in Idefics3. In Qwen, it flips 63.8%, 91.6%, and 91.2% of yes/no decisions at half, one, and two steps, against 1.2%, 4.4%, and 23.5% for the random direction, so its effect saturates by one step (Table 1 reports the unit step). In Gemma3, it does no better than a random direction within the subspace. The component thus encodes a direction shared across questions in two of three models. More details on the direction are provided in Appendix C.

The component changes generated answers. Greedy decoding with a short-answer instruction on the 320 evaluation pairs shows that the component also changes what the model says (Table 2;

Table 2: **Retaining the component changes what the model says.** Top: eight Qwen2.5-VL-7B directions sampled at random by a rule fixed in advance, with the recipient's greedy answer under each condition (**bold**: the donor's answer). Bottom: share of all 640 answers that follow the donor, and the tested effects KEEP−SHAM-KEEP and SHAM-REMOVE−REMOVE (**bold**: pre-specified criterion met).

Question (recipient / donor answer)	CLEAN	SHAM-KEEP	KEEP	REMOVE	SHAM-REMOVE	FULL
Adult or baby? (adult / baby)	adult	adult	adult	adult	**baby**	**baby**
Are all the people male or female? (male / female)	male	male	male	male	**female**	**female**
Is it a tube TV or flat screen? (flat screen / tube)	flat screen	flat screen	**tube**	flat screen	**tube**	**tube**
Is it morning or evening? (morning / evening)	morning	morning	**evening**	morning	**evening**	**evening**
Is it morning or night? (morning / night)	morning	morning	**night**	morning	**night**	**night**
Is the batter a child or an adult? (child / adult)	child	child	child	child	child	child
Is the bus empty or full? (full / empty)	full	**empty**	**empty**	**empty**	**empty**	**empty**
Is the man holding the food in his left or right hand? (left / right)	**right**	**right**	left	left	left	left
Qwen2.5-VL-7B: answers following donor (%)	4.4	6.9	39.4	20.0	77.7	78.4
effect vs. sham, pp (98.75% CI)			retain **32.5 [28.0, 37.0]**		remove **57.7 [51.6, 63.8]**	
Idefics3-8B: answers following donor (%)	8.9	9.2	10.3	10.8	13.3	13.3
effect vs. sham, pp (98.75% CI)			retain 1.1 [0.2, 2.2]		remove **2.5 [1.1, 4.1]**	

the bottom rows give the tested effects). In Qwen, swapping the full question-token states at layer 17 makes 78% of answers follow the donor, against 4% without a swap, although the recipient keeps its own image tokens; later layers thus read the answer largely from these states. In Idefics3, whose interface lies past its window (Section 4), the same swap moves only 13%. In Qwen, KEEP makes 32.5 percentage points more answers follow the donor than SHAM-KEEP, and removing the component lowers donor following by 57.7 points relative to SHAM-REMOVE. In Idefics3, the removal effect (2.5 points) passes, whereas the retention effect (1.1 points) falls short of the 2-point criterion. In the random sample of Qwen directions in Table 2, the answer tends to change to the donor's when the component comes from the donor (KEEP, SHAM-REMOVE) and to stay when it does not (REMOVE, SHAM-KEEP). Following the donor makes answers less faithful to the recipient's image, so these results confirm use rather than improve accuracy. More generation details are provided in Appendix E.

6 Conclusion

We asked what later layers of a VLM read from the question-token states once visual information has reached them. To separate the answer from other image content, we held the question fixed and gave each image of a VQAv2 quartet the question-token states of an image with the same answer or with the other answer, at matched norms. Across seven checkpoints, different-answer donors pulled the recipient toward the donor's answer, whereas same-answer donors left the mean margin unchanged and moved the full output distribution far less; per recipient, however, same-answer donors transmitted one-third to two-thirds of their answer strength in four checkpoints. A layer sweep in three models placed this answer-level summary in a narrow band of middle layers, consistent with the weaker results of the two models whose fixed layer lies past it. In Idefics3, Qwen2.5-VL-7B, and Gemma3, a learned rank-16 component carried the donor effect more selectively than magnitude-matched controls and carried the answer-aligned contrast in its source. It was also used: a single direction within it moved yes/no answers without any donor in two of the three models, and in Qwen2.5-VL-7B retaining it alone changed generated answers. For patching studies, three practices follow: pair every different-answer donor with a same-answer donor at matched norm, test invariance per recipient rather than on average, and sweep layers before any fixed-layer analysis.

Limitations. The account is model-dependent, strongest in the Qwen family and weakest in Gemma3, and the fixed layers were not reselected after the sweep. Our readouts are taken at the answer, so they show what later layers read for this answer rather than everything the states encode. The data are English VQAv2 quartets dominated by yes/no questions, and the projections transfer to a new task in only one of four settings. More details on these points and on departures from the pre-specified plans are provided in Appendices F and G.

AI USE STATEMENT

We used generative AI tools to assist with polishing the manuscript, searching for and identifying relevant literature, and implementing and executing experimental code. The authors take full responsibility for the accuracy and integrity of the methods, experiments, results, code, and text presented in this paper.

ETHICS STATEMENT

This work aims to advance our understanding of internal representations and information processing in vision-language models. Our study uses pretrained models and existing visual question answering datasets to conduct controlled computational experiments. To the best of our knowledge, this work does not raise specific ethical concerns beyond those generally associated with research on such models and datasets.

REPRODUCIBILITY STATEMENT

Sections 3, 5, and 5.3 specify the intervention construction, answer scoring, projection objective, and primary evaluation contrasts. Appendix A documents cohort construction and data provenance, while Appendix A.3 describes model interfaces, token alignment, and numerical execution procedures. Further details of projection fitting and source crossing are provided in Appendices C and D. The cross-task, prediction, and generation evaluations are described in Appendices F and E, and Appendix G records the development history and protocol revisions. An anonymized code package is provided in the supplementary materials.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: answering complex counting questions. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33018076. URL https://doi.org/10.1609/aaai.v33i01.33018076.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, SUN Bing, Haiqin Weng, Liu Yan, and Jianwei Yin. Faithful bi-directional model steering via distribution matching and distributed interchange interventions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LoisXFZL3k.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 7400–7426. Curran Associates, Inc., 2024. doi: 10.52202/079017-0237. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0dfe31d6e703e138d46a7d2fced38b7c-Paper-Conference.pdf.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 03 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00359. URL https://doi.org/10.1162/tacl_a_00359.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

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 Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv.org/abs/2411.17491.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2zXRGiorSu.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. URL https://arxiv.org/abs/2408.12637.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755. Springer International Publishing, 2014.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JAI7afWA9e.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=77ziPGdQct.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Microsoft. Phi-3.5-vision-instruct, 2024. URL https://huggingface.co/microsoft/Phi-3.5-vision-instruct. Model card.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *Advances in Neural Information Processing Systems*, 38:66352–66385, 2026.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip, 2023. URL https://arxiv.org/abs/2308.14179.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of visual information flow in vision-language models, 2026. URL https://arxiv.org/abs/2607.03358.

Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in native multimodal models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 74684–74712. Curran Associates, Inc., 2025. doi: 10.52202/085713-2508. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6c420d6d2f19acbd44c219d244d892ee-Paper-Conference.pdf.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019. doi: 10.1109/CVPR.2019.00851.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Qidong Wang, Junjie Hu, and Ming Jiang. V-SEAM: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17396–17420, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.880. URL https://aclanthology.org/2025.emnlp-main.880/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

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

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025. doi: 10.1109/CVPR52734.2025.01842.

A Data, Models, and Notation

This appendix gives the setup behind Section 3.

A.1 Notation

Table 3 lists the symbols. Vectors are bold lower case, matrices bold upper case, and condition labels upright. Token indices start at one; layer indices are zero-based.

Table 3: Symbols used in the paper.

Symbol	Meaning
a, b, c	Complete candidate answer sequences; c_j is an individual token
d, T, L_c	State dimension, question-token count, and candidate length
$\boldsymbol{h}_t, \boldsymbol{\delta}_t$	Question-state vector and source-to-recipient displacement
$\boldsymbol{p}, \boldsymbol{c}$	Projected and complementary displacement vectors
$\bar{\boldsymbol{p}}_u^m, \bar{\boldsymbol{c}}_u^m$	Their norm-matched forms in the source-crossing experiment
$\boldsymbol{H}^{(x)}$	Matrix whose columns are native decoder-input states for input x
$\boldsymbol{A}, \boldsymbol{U}$	Learned matrix and its orthonormal QR basis
$\boldsymbol{P}, \boldsymbol{Q}$	Projector $\boldsymbol{U}\boldsymbol{U}^\top$ and complement $\boldsymbol{I} - \boldsymbol{P}$
$\boldsymbol{I}$	Identity matrix of the implied dimension
$\boldsymbol{\ell}_{k,j}, \ell_{k,j,v}$	Full-vocabulary logit vector and its scalar entry for token v
$\mathbb{V}, \mathbb{R}$	Output vocabulary and the set of real numbers
$\mu_{k,j}, \sigma_{k,j}$	Full-vocabulary population mean and standard deviation
$S_{\mathrm{z}}, S_{\mathrm{lp}}$	Standardized and native log-probability sequence scores
m_k, E_k	Recipient-oriented margin and donor-oriented change $m_{\mathrm{CLEAN}} - m_k$
$B_t, b_{\mathrm{P}}, b_{\mathrm{Q}}, R$	Scalar displacement budgets
$D, A_1, A_2, \theta_{\mathrm{P}}, \theta_{\mathrm{Q}}, I$	Scalar source, functional, and factorial effects
$\mathcal{L}, \mathbb{E}, \nabla_{\boldsymbol{A}}$	Loss, expectation under the stated unit weights, and matrix gradient

A.2 Quartets and Cohorts

Each quartet fixes one VQAv2 question and contains two images annotated with answer a and two with answer b, each with at least eight of ten agreeing normalized annotations (Goyal et al., 2017). The 538 quartets (2,152 images) comprise 371 yes/no, 74 number, and 93 Other questions; 466

Table 4: Cohorts and inferential units. Repeated directions are averaged within their pair or quartet before inference, and model repetitions do not add independent images.

Comparison	Units	Type	Exposure / role
Source agreement	538 (520)	quartet	Seven checkpoints; Idefics retains 520 eligible quartets.
Component fitting / scale	96 / 32	quartet	Model-specific fitting; development groups are not new confirmation data.
Component evaluation	320	image pair	Other questions outside fitting/development; reused across models.
Crossed sources	72	quartet	24 per answer type; earlier source-cohort images, new interventions.

use training and 72 validation annotations. For Idefics3, a pre-specified, outcome-independent input check excluded 18 quartets (11 yes/no, 3 number, 4 Other) whose native image tiling (3×4 versus 4×4 tiles) gave the four images different visual-token layouts and hence different question-token positions; the check requires identical positions so that donor and recipient question tokens occupy the same positions. These quartets are excluded for Idefics3 only, without replacement; the other six models use all 538. Selection uses only annotations and input eligibility. The low-rank component is fitted on 96 quartets (32 per answer type), developed on 32 disjoint Other quartets, and evaluated on 320 further Other image pairs; the source-crossing study uses 72 quartets (24 per type) outside these sets. In 237 of the 320 evaluation pairs, neither answer string occurs in the fitting set. All models share one image pool, so adding models does not add independent images. Table 4 lists every cohort.

Figure 1 example. The example was selected by a fixed rule before candidate images were inspected. Among the 371 yes/no quartets, the first yes recipient of each quartet is taken, and quartets are ranked by $|x-\tilde{x}|/s_x + |y-\tilde{y}|/s_y$, where $x = m_{\mathrm{CLEAN}} - m_{\mathrm{different}}$ and $y = m_{\mathrm{same}} - m_{\mathrm{CLEAN}}$ in Qwen2.5-VL-7B at layer 17, and $\tilde{\cdot}$ and s are the population median and standard deviation. The top-ranked quartet asks "Is the television turned on?". The photos are unmodified COCO images (Lin et al., 2014) (train2014 ids 352821, 424342, 284756) distributed via Flickr under CC BY-ND 2.0, CC BY-NC 2.0, and CC BY-NC-ND 2.0.

A.3 Models and Patch Sites

We evaluate two Qwen2.5-VL checkpoints (Bai et al., 2025b) and one checkpoint each from Idefics3 (Laurençon et al., 2024), Gemma3 (Team et al., 2025), Qwen3-VL (Bai et al., 2025a), InternVL3.5 (Wang et al., 2025b), and Phi-3.5-Vision (Microsoft, 2024); Table 5 lists the patched decoder blocks. Donor and recipient question tokens are aligned by token identity and ordinal position, and the recipient keeps its own image tokens, attention mask, and positions. In Qwen2.5-VL, 125 of the 320 component pairs match exactly at the input; the other 195 differ only through native visual-token expansion, which alignment at the question tokens absorbs, and both subsets show positive sham-adjusted effects. All experiments run in FP32. If either displacement norm is at most 10^{-12}, both patches of a recipient are exact no-ops and the recipient stays in the data. For Gemma3, the numerical check was revised to separate unavoidable FP32 rounding from construction error, without changing the model, layer, examples, or effect criteria.

Table 5: Checkpoints and patch sites. Layer indices are zero-based decoder-block inputs; the mapping was fixed without reference to effects in the new models.

Checkpoint	Source layer	Functional layer	State role
Qwen2.5-VL-7B	17	17	Original question
Qwen2.5-VL-3B	22	—	Original question
Idefics3-8B-Llama3	20	20	Original question
Gemma3-4B	21	21	Original question
Qwen3-VL-8B-Instruct	22	—	Original question
InternVL3.5-8B	22	—	Original question
Phi-3.5-vision-instruct	20	—	Original question

B Answer-Level Results

This appendix gives the definitions and full results behind Section 4.

Source contrast and CLEAN decomposition. Table 6 summarizes each checkpoint and Table 8 gives Other questions separately. Relative to the unpatched input, the source contrast splits as

$$D = (m_{\text{same}} - m_{\text{CLEAN}}) + (m_{\text{CLEAN}} - m_{\text{different}}). \tag{6}$$

The first term is tested for equivalence to zero within $\pm 0.1|D|$ and the second for positivity; each forms its own seven-model family, Bonferroni-adjusted over models, and Table 7 gives the corrected intervals. The random direction is one fixed unit vector per recipient at the patch norm. For the two Qwen2.5-VL models, CLEAN and random controls existed before this analysis was specified, so their rows are reanalyses of known data. By type, the Qwen2.5-VL-3B contrast is 1.309 for yes/no, 0.478 for number, and 0.082 for Other questions.

Per-recipient donor following. For each recipient r we regress its same-answer margin change $m_{\text{same}}(r) - m_{\text{CLEAN}}(r)$ on $m_{\text{CLEAN}}(\text{donor}) - m_{\text{CLEAN}}(r)$ with an intercept; β is the slope, and intervals come from a quartet-level bootstrap. R is the mean absolute same-answer change divided by that of the random direction. A model shows donor following if the corrected lower bound of β exceeds 0.2, and per-recipient invariance if the interval for β lies within ± 0.2 and the upper bound of R is at most 1.5. Because same-answer swaps are reciprocal within a quartet, a near-zero mean SAME$-$CLEAN is expected under donor following, so the per-recipient test, not the mean, is the test of invariance.

Full-distribution change. We compute $\mathrm{KL}(p_{\text{CLEAN}} \,\|\, p_k)$ over the full vocabulary at the first answer position, in FP64. The tested contrast is $\mathrm{KL}_{\text{different}} > \mathrm{KL}_{\text{same}}$ (one-sided, seven-model family); the ratio of means and the comparison with the random direction are descriptive.

Layer sweep. Each of the three swept models uses 180 quartets (60 per answer type), and every decoder block input is patched in turn. Answer types are weighted by their population shares (371/538, 74/538, 93/538) for Qwen2.5-VL-7B and equally for Idefics3 and Gemma3, matching each model's fixed-layer estimand. Intervals are pointwise 95% bootstrap intervals and support no layer-wise significance claims. The first layer at which the lower bound of CLEAN$-$DIFF exceeds 20% of its maximum is 13 for Qwen2.5-VL-7B, 9 for Idefics3, and 12 for Gemma3.

Answer identity and answer strength. The per-recipient slopes show that the summary is graded: a recipient adopts part of its donor's margin, not only its answer. Combined with the reciprocity of same-answer swaps, this explains why their mean is near zero despite large per-recipient changes. In Qwen, a single direction within the component flips most yes/no decisions and saturates by one step, consistent with an axis whose sign gives the answer and whose position gives its strength. Our analyses do not establish whether answer identity contributes beyond such a graded variable; a joint donor-following fit to both donor types would separate the two.

Table 6: Answer-level transfer in seven checkpoints. $D = m_{\text{same}} - m_{\text{different}}$ (interval families: a, 97.5%; b, 95%; c, 99.1667%; d, 97.5%). D splits into SAME−CLEAN, which passes a pre-specified equivalence test within $\pm 0.1|D|$ in every row, and CLEAN−DIFF, which is positive in every row. β: slope of a recipient's same-answer margin change on the donor-minus-recipient clean margin; R: mean absolute same-answer change relative to a norm-matched random direction; *: donor answer-strength following (corrected lower bound of β above 0.2); †: per-recipient invariance (β within ± 0.2 and $R \leq 1.5$). KL ratio: mean KL from CLEAN at the first answer position, same- over different-answer swaps; ‡: a smaller same-answer KL is not established. Qwen2.5-VL rows reanalyze data known before the protocol was specified. Intervals are in Table 7.

Checkpoint	D	SAME−CLEAN	CLEAN−DIFF	β	R	KL ratio
Qwen2.5-VL-7B	2.158^a	−0.002	2.161	0.67^*	2.64	0.13
Qwen2.5-VL-3B	0.982^a	−0.035	1.017	0.61^*	2.78	0.23
Qwen3-VL-8B	1.591^c	−0.027	1.618	0.49^*	5.23	0.17
InternVL3.5-8B	0.976^c	−0.048	1.024	0.33^*	4.21	0.23
Phi-3.5-vision	0.405^d	−0.021	0.426	0.20	2.00	0.34
Idefics3-8B	0.301^b	−0.015	0.316	0.10	1.95	0.64
Gemma3-4B	0.038^c	0.001	0.037	$0.03^\dagger$	0.88	$1.00^\ddagger$

Table 7: Answer-level results with corrected intervals. $\Delta = 0.1|D|$ is the pre-specified equivalence bound for SAME−CLEAN. Each endpoint forms its own seven-model family (Bonferroni over models); the two decomposition endpoints are not a joint fourteen-endpoint guarantee. The KL ratio is descriptive; the tested contrast is $\text{KL}_{\text{DIFF}} > \text{KL}_{\text{SAME}}$, which holds in all rows except Gemma3.

Checkpoint	SAME−CLEAN	Δ	CLEAN−DIFF
Qwen2.5-VL-7B	−0.002 [−0.014, 0.009]	0.216	2.161 [2.004, 2.317]
Qwen2.5-VL-3B	−0.035 [−0.043, −0.026]	0.098	1.017 [0.943, 1.090]
Qwen3-VL-8B	−0.027 [−0.047, −0.008]	0.159	1.618 [1.527, 1.711]
InternVL3.5-8B	−0.048 [−0.066, −0.032]	0.098	1.024 [0.954, 1.100]
Phi-3.5-vision	−0.021 [−0.034, −0.010]	0.041	0.426 [0.395, 0.461]
Idefics3-8B	−0.015 [−0.025, −0.007]	0.030	0.316 [0.286, 0.349]
Gemma3-4B	0.001 [−0.000, 0.002]	0.004	0.037 [0.031, 0.044]

Checkpoint	β	R	KL ratio
Qwen2.5-VL-7B	0.671 [0.637, 0.702]	2.64 [2.41, 2.89]	0.134
Qwen2.5-VL-3B	0.605 [0.564, 0.645]	2.78 [2.55, 3.01]	0.229
Qwen3-VL-8B	0.487 [0.453, 0.518]	5.23 [4.69, 5.82]	0.168
InternVL3.5-8B	0.327 [0.292, 0.362]	4.21 [3.83, 4.61]	0.232
Phi-3.5-vision	0.195 [0.176, 0.215]	2.00 [1.80, 2.21]	0.342
Idefics3-8B	0.099 [0.074, 0.134]	1.95 [1.73, 2.23]	0.637
Gemma3-4B	0.027 [0.022, 0.032]	0.88 [0.76, 1.01]	1.004

Table 8: Other analyses preserve their actual inferential identities. The original Qwen subgroup analyses were not new prospective confirmations.

Checkpoint	Other mean	Original reported interval
Qwen2.5-VL-7B	.1560	[.0966,.2239], 95%, post-hoc
Qwen2.5-VL-3B	.0821	[.0554,.1120], 95%, post-hoc
Idefics3-8B	.4451	[.3805,.5140], 95%, hierarchical
Gemma3-4B	.0232	[.0107,.0381], 99.1667%
Qwen3-VL-8B	.3761	[.2316,.5434], 99.1667%
InternVL3.5-8B	.2554	[.1332,.4045], 99.1667%
Phi-3.5-vision	.2506	[.1851,.3266], 97.5%

C THE LOW-RANK COMPONENT

This appendix gives the fitting, controls, and full results behind Sections 5 and 5.4.

Fitting. Only $\boldsymbol{A} \in \mathbb{R}^{d\times 16}$ is optimized; $\boldsymbol{U}$ is the orthonormal factor of its reduced QR decomposition and $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top$, with the model frozen. We use Adam (learning rate 10^{-3}, gradient clipping at norm 1) for one pass over the 96 fitting quartets, four recipient updates per quartet, and keep the final checkpoint. With responses $d_c^k = S_{\mathrm{lp}}(c \mid k) - S_{\mathrm{lp}}(c \mid \mathrm{CLEAN})$, the implemented loss is

$$\mathcal{L}(\boldsymbol{A}) = \mathbb{E}_{q,i}\Big[\tfrac{1}{2}\textstyle\sum_{c\in\{a,b\}}\big(d_c^{\mathrm{KEEP}(\boldsymbol{P})} - d_c^{\mathrm{FULL}}\big)^2\Big], \tag{7}$$

which equals equation 4 because the two responses share their CLEAN score.

Shams and thresholds. For $\boldsymbol{p} = \boldsymbol{P}\boldsymbol{\delta}$ and $\boldsymbol{c} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\delta}$, the shams follow the full displacement at the norm of each part:

$$\boldsymbol{d}_{\mathrm{SK}} = \frac{\|\boldsymbol{p}\|_2}{\|\boldsymbol{\delta}\|_2}\boldsymbol{\delta}, \qquad \boldsymbol{d}_{\mathrm{SR}} = \frac{\|\boldsymbol{c}\|_2}{\|\boldsymbol{\delta}\|_2}\boldsymbol{\delta}. \tag{8}$$

A contrast passes if its lower bound exceeds $0.1\, s_{\mathrm{dev}}$: 0.045 for Idefics3, 0.018 for Qwen, and 0.0043 for Gemma3. For Gemma3, the detailed statistical specification was finalized after the model runs but before effects were aggregated (Appendix G).

Geometric comparator. For quartet q and question token t, let $\boldsymbol{b}_{qt} = \frac{1}{2}(\boldsymbol{h}_{A_1,t} + \boldsymbol{h}_{A_2,t} - \boldsymbol{h}_{B_1,t} - \boldsymbol{h}_{B_2,t})$ be the between-answer difference and $\boldsymbol{w}_{A,qt}$, $\boldsymbol{w}_{B,qt}$ the within-answer differences. We average $\big(\boldsymbol{b}_{qt}\boldsymbol{b}_{qt}^\top - \frac{1}{4}(\boldsymbol{w}_{A,qt}\boldsymbol{w}_{A,qt}^\top + \boldsymbol{w}_{B,qt}\boldsymbol{w}_{B,qt}^\top)\big)/e_{qt}$, with e_{qt} the matching energy, over tokens, then questions, then answer types, and keep the top 16 positive directions. Earlier geometric candidates of rank 4 and 16 failed a pre-specified retention target, which motivated the learned projection.

Results. Table 9 gives the primary contrasts with their corrected intervals and Table 10 the individual conditions. Retention specificity is 0.276, 0.479, and 0.0143 in Idefics3, Qwen, and Gemma3, and removal specificity is 0.359, 0.584, and 0.0223. The retained fraction differs by model: KEEP/FULL is about 1.0 in Qwen, 0.51 in Idefics3, and 0.29 in Gemma3. In Idefics3, KEEP minus half of FULL spans zero; in Gemma3 it is -0.0164 $[-0.0243, -0.0086]$. Native log-probability contrasts point the same way as the standardized ones. In Gemma3, KEEP flips six two-candidate decisions from the recipient's answer to the donor's and one back, against four and zero for SHAM-KEEP. In Qwen, KEEP flips 123 two-candidate decisions from the recipient's answer to the donor's and two back, against 17 and one for SHAM-KEEP. In Gemma3, where KEEP carries less, the complement retains most of the effect (REMOVE 0.0564). Reconstruction MSE against FULL is 0.893 for the learned versus 1.138 for the geometric projection in Idefics3, 1.032 versus 0.904 in Qwen, and 0.534 versus 0.434 in Gemma3, so matching the full response and carrying selective effects need not coincide.

Diagnostics. Variance is the share of natural-input question-token variance inside the subspace, normalized so that a random 16-dimensional subspace scores 1. AUROC is the pooled out-of-fold score of an L2 logistic readout of yes versus no from the subspace coordinates. The shared direction is estimated on a random half of the yes/no quartets as the within-quartet mean difference between no and yes states projected into the subspace, averaged and normalized; its step is the median patch norm of that half, and its sign is set by the gold answer. On the other half, it flips 63.8%, 91.6%, and 91.2% of Qwen's decisions at 0.5, 1, and 2 steps, against 1.2%, 4.4%, and 23.5% for a random direction in the subspace; Idefics3 reaches 5.8–12.0% against 0.9–3.7%, and Gemma3 0.7–2.8% against 0.6–2.2%. The pre-specified test compares margin gains at unit step with a random direction in the subspace and in the full space (one-sided, six-contrast family); Table 11 gives all diagnostics, and Table 12 six randomly drawn Qwen recipients.

Table 9: Primary component and source-crossing contrasts. Retention specificity is KEEP minus SHAM-KEEP; removal specificity is SHAM-REMOVE minus REMOVE. The first two rows use 320 image pairs; the last two use 72 source-crossing quartets with answer types weighted equally. Intervals are 99.1667% for the Idefics3 six-endpoint family and 98.75% for the Qwen and Gemma3 four-endpoint families. Gemma3's interval specification was finalized after execution but before effect aggregation, and its numerical checks follow the revised protocol (Appendices A.3 and G).

Contrast	Idefics3	Qwen2.5-VL-7B	Gemma3
Retention specificity	.276 [.237,.319]	.479 [.421,.540]	.0143 [.0080,.0216]
Removal specificity	.359 [.312,.409]	.584 [.513,.658]	.0223 [.0155,.0302]
Projected source effect	.0765 [.0520,.1075]	.3924 [.3139,.4762]	.0189 [.0139,.0243]
Functional − geometric	.0421 [.0257,.0620]	.1857 [.1441,.2312]	.0130 [.0092,.0169]

Table 10: Available component-evaluation scores. All are donor-axis changes relative to CLEAN or explicit paired contrasts; reconstruction MSE has a different target. Intervals retain each model's original family confidence level, but auxiliary rows are not new primary tests.

Model	Outcome	Estimate [interval]
Idefics3	FULL	0.7263 [0.6213, 0.8350]
Idefics3	KEEP	0.3705 [0.3226, 0.4211]
Idefics3	Geometric KEEP	0.1893 [0.1582, 0.2219]
Idefics3	A1	0.2764 [0.2365, 0.3188]
Idefics3	A2	0.3590 [0.3119, 0.4086]
Idefics3	KEEP -0.5 FULL	0.00734 [-0.03859, 0.05407]
Idefics3	MSE (zero edit)	1.7763 [1.4085, 2.1913]
Idefics3	MSE (functional)	0.8935 [0.6809, 1.1442]
Idefics3	MSE (geometric)	1.1382 [0.8900, 1.4200]
Qwen-7B	FULL	0.5747 [0.5023, 0.6499]
Qwen-7B	KEEP	0.5834 [0.5131, 0.6563]
Qwen-7B	REMOVE	-0.02234 [-0.06241, 0.01786]
Qwen-7B	SHAM-KEEP	0.1043 [0.08725, 0.1222]
Qwen-7B	SHAM-REMOVE	0.5620 [0.4914, 0.6350]
Qwen-7B	Geometric KEEP	0.3868 [0.3367, 0.4394]
Qwen-7B	A1	0.4792 [0.4207, 0.5395]
Qwen-7B	A2	0.5843 [0.5133, 0.6579]
Qwen-7B	KEEP -0.5 FULL	0.2961 [0.2511, 0.3428]
Qwen-7B	MSE (zero edit)	2.0880 [1.6765, 2.5886]
Qwen-7B	MSE (functional)	1.0318 [0.8467, 1.2345]
Qwen-7B	MSE (geometric)	0.9036 [0.7339, 1.0957]
Gemma3	FULL	0.07938 [0.06276, 0.09653]
Gemma3	KEEP	0.02330 [0.01608, 0.03162]
Gemma3	REMOVE	0.05641 [0.04236, 0.07104]
Gemma3	SHAM-KEEP	0.00898 [0.00707, 0.01098]
Gemma3	SHAM-REMOVE	0.07873 [0.06225, 0.09574]
Gemma3	Geometric KEEP	0.02744 [0.02187, 0.03346]
Gemma3	A1	0.01431 [0.00796, 0.02162]
Gemma3	A2	0.02233 [0.01552, 0.03019]
Gemma3	KEEP -0.5 FULL	-0.01639 [-0.02431, -0.00861]
Gemma3	KEEP − geometric KEEP	-0.00414 [-0.01008, 0.00196]
Gemma3	MSE (zero edit)	0.5683 [0.4522, 0.6985]
Gemma3	MSE (functional)	0.5342 [0.4343, 0.6463]
Gemma3	MSE (geometric)	0.4343 [0.3439, 0.5407]

Table 11: Properties of the learned rank-16 component. A_1, A_2: retention and removal specificity on 320 held-out pairs (all pass). Variance: natural-input variance of the question-token states captured by the subspace, relative to a random 16-dimensional subspace (random = 1). AUROC: linear yes/no readout from the functional (geometric) subspace. Shared direction: margin change toward the other answer along a donor-free yes/no direction, minus that of a random direction in the same subspace, and its decision-flip rate. Generation: donor-following difference in percentage points under greedy short-answer decoding, KEEP−SHAM-KEEP and SHAM-REMOVE−REMOVE (98.75% intervals in Appendix E). *: pre-specified criterion met.

	Selectivity		Structure		Shared direction		Generation (pp)	
Model	A_1	A_2	Variance	AUROC	Δmargin	Flips	Retain	Remove
Idefics3-8B	0.276	0.359	3.43	0.984 (0.988)	1.15*	11.1%	1.1	2.5*
Qwen2.5-VL-7B	0.479	0.584	4.19	0.991 (0.991)	5.41*	91.6%	32.5*	57.7*
Gemma3-4B	0.0143	0.0223	1.72	0.910 (0.944)	−0.001	1.0%	–	–

Table 12: Donor-free direction on six Qwen2.5-VL-7B yes/no recipients drawn uniformly at random from the 680 held-out recipients. Entries are the gold-oriented margin and, in parentheses, the two-candidate decision; **bold** marks a flipped decision. The step is the fitted α (multiplier 1), with its sign set by the gold answer.

Question	Gold	CLEAN	Shared direction	Random in subspace
Is the dog sleeping?	no	2.82 (no)	−2.17 (**yes**)	1.60 (no)
Is the photo blurry?	no	2.02 (no)	−1.35 (**yes**)	1.19 (no)
Is there water in the picture?	yes	1.32 (yes)	−4.63 (**no**)	0.86 (yes)
Is this a museum?	no	5.50 (no)	−0.06 (**yes**)	5.07 (no)
Is this outdoors?	yes	2.56 (yes)	−4.63 (**no**)	2.87 (yes)
Is this person in the air?	no	4.76 (no)	−1.37 (**yes**)	4.13 (no)

D SOURCE CROSSING

This appendix gives the construction and full results behind Section 5.3.

Construction. For projection $m \in \{\mathrm{F}, \mathrm{G}\}$ with complement $\boldsymbol{Q}_m = \boldsymbol{I} - \boldsymbol{P}_m$ and donor displacement $\boldsymbol{\delta}_u$, $u \in \{\mathrm{s}, \mathrm{d}\}$, the shared budgets are

$$b_{\mathrm{P}} = \min_{m,u} \|\boldsymbol{P}_m \boldsymbol{\delta}_u\|_2, \qquad b_{\mathrm{Q}} = \min_{m,u} \|\boldsymbol{Q}_m \boldsymbol{\delta}_u\|_2, \tag{9}$$

and the crossed state takes its projected part from donor u and its complement from donor v:

$$\boldsymbol{h}^m_{uv} = \boldsymbol{h} + b_{\mathrm{P}} \frac{\boldsymbol{P}_m \boldsymbol{\delta}_u}{\|\boldsymbol{P}_m \boldsymbol{\delta}_u\|_2} + b_{\mathrm{Q}} \frac{\boldsymbol{Q}_m \boldsymbol{\delta}_v}{\|\boldsymbol{Q}_m \boldsymbol{\delta}_v\|_2}. \tag{10}$$

Since the parts are orthogonal, every crossed patch has norm $\sqrt{b_{\mathrm{P}}^2 + b_{\mathrm{Q}}^2}$; full-direction references REF-S and REF-D are matched to the same norm. With recipient margins M_{uv} averaged within quartets and equally over answer types,

$$\theta_{\mathrm{P}} = \tfrac{1}{2}\mathbb{E}\big[(M_{\mathrm{ss}} - M_{\mathrm{ds}}) + (M_{\mathrm{sd}} - M_{\mathrm{dd}})\big], \qquad \theta_{\mathrm{Q}} = \tfrac{1}{2}\mathbb{E}\big[(M_{\mathrm{ss}} - M_{\mathrm{sd}}) + (M_{\mathrm{ds}} - M_{\mathrm{dd}})\big]. \tag{11}$$

The primary tests require positive lower bounds for $\theta_{\mathrm{P}}^{\mathrm{F}}$ and for $\theta_{\mathrm{P}}^{\mathrm{F}} - \theta_{\mathrm{P}}^{\mathrm{G}}$, together with a positive REF-S−REF-D reference.

Results. Both tests pass in all three models (Table 9): $\theta_{\mathrm{P}}^{\mathrm{F}}$ is 0.0765, 0.3924, and 0.0189 in Idefics3, Qwen, and Gemma3, exceeding the geometric projection by 0.0421, 0.1857, and 0.0130. The learned projection's source effect is positive under either complement source (Table 13). The complementary-source effect $\theta_{\mathrm{Q}}^{\mathrm{F}}$ is 0.135 in Idefics3, 0.121 in Qwen, and 0.00598 in Gemma3 (positive at 95% but not at the corrected level in Gemma3), and the interaction is negative in Idefics3 and Qwen, so the decomposition is partial rather than additive. By answer type, Qwen's $\theta_{\mathrm{P}}^{\mathrm{F}}$ is 0.812, 0.317, and 0.0488 for yes/no, number, and Other; the learned-over-geometric advantage for Other questions is not resolved in Qwen or Gemma3.

Table 13: Projected-source background effects, complementary-source effects, and functional interaction. Every interval in this table is the reported 95% descriptive interval; the primary source contrasts retain their separate corrected intervals.

Model	Quantity	Estimate [95% interval]
Idefics3	$P^{\mathrm{F}}_{\mid Q=\mathrm{S}}$	0.07277 [0.05391, 0.09491]
	$P^{\mathrm{F}}_{\mid Q=\mathrm{D}}$	0.08030 [0.06050, 0.1028]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{S}}$	0.03108 [0.02322, 0.04003]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{D}}$	0.03782 [0.02859, 0.04832]
	$\theta^{\mathrm{F}}_{\mathrm{Q}}$	0.1352 [0.1029, 0.1735]
	I^{F}	-0.00753 [-0.01321, -0.00261]
Qwen-7B	$P^{\mathrm{F}}_{\mid Q=\mathrm{S}}$	0.3662 [0.3050, 0.4312]
	$P^{\mathrm{F}}_{\mid Q=\mathrm{D}}$	0.4185 [0.3506, 0.4903]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{S}}$	0.1641 [0.1366, 0.1938]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{D}}$	0.2492 [0.2063, 0.2943]
	$\theta^{\mathrm{F}}_{\mathrm{Q}}$	0.1210 [0.05096, 0.1997]
	I^{F}	-0.05231 [-0.09067, -0.01469]
Gemma3	$P^{\mathrm{F}}_{\mid Q=\mathrm{S}}$	0.01856 [0.01465, 0.02272]
	$P^{\mathrm{F}}_{\mid Q=\mathrm{D}}$	0.01931 [0.01529, 0.02351]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{S}}$	0.00568 [0.00443, 0.00711]
	$P^{\mathrm{G}}_{\mid Q=\mathrm{D}}$	0.00621 [0.00480, 0.00780]
	$\theta^{\mathrm{F}}_{\mathrm{Q}}$	0.00598 [0.00113, 0.01091]
	I^{F}	-0.00075 [-0.00157, 0.00008]

E GENERATION

This appendix gives the setup and further examples behind the generation results of Section 5.4.

Setup. On the 320 evaluation pairs, the instruction "Answer the question using a single word or phrase." follows the question, and only the original question tokens are patched, once at prefill. Decoding is greedy with at most 32 new tokens. An output counts as following the donor or the recipient if its full normalized string equals that answer; no output was truncated, and an independent token-level rescoring agrees with every label. CLEAN outputs match the recipient's answer in 90.5% of directions for Idefics3 and 95.0% for Qwen. Retention and removal form a four-endpoint family with the FULL−CLEAN reference; an endpoint passes if its 98.75% lower bound is positive, its estimate is at least 2 percentage points, and the reference qualifies. An earlier readout with the models' native prompts could not detect donor following because outputs rarely matched any answer string exactly (none did for Qwen); the short-answer instruction removes this floor. The generations reuse the evaluation images, so they are a follow-up rather than a replication on new images.

Examples. All examples were drawn by rules and seeds fixed before any example was selected. Table 2 shows a random sample for Qwen; Table 14 adds a random sample for Idefics3 and one example per outcome stratum of KEEP and SHAM-KEEP for both models. The stratified examples are selected by outcome and illustrate each stratum rather than its frequency.

Table 14: Further generation examples. Top: eight Idefics3-8B directions drawn uniformly at random by the same rule as Table 2. Middle and bottom: one direction drawn at random from each outcome stratum of KEEP and SHAM-KEEP, with the stratum's share of all 640 directions; these are selected by outcome and do not indicate prevalence. **Bold** marks the donor's answer.

Question (recipient / donor answer)	CLEAN	SHAM-KEEP	KEEP	REMOVE	SHAM-REMOVE	FULL
Idefics3-8B, random sample						
Are these adults or children? (adults / children)	Adults.	Adults.	Adults.	Adults.	Adults.	Adults.
Is it a man or woman's hand? (woman / man)	Woman.	Woman.	Woman.	Woman.	Woman.	Woman.
Is it dirty or clean? (dirty / clean)	Dirty.	Dirty.	Dirty.	Dirty.	Dirty.	Dirty.
Is the cat inside or on top of the suitcase? (inside / on top)	Inside.	Inside.	Inside.	Inside.	Inside.	Inside.
Is the dog small or large? (large / small)	Large.	Large.	Large.	Large.	Large.	Large.
Is the drink on the table hot or cold? (hot / cold)	Hot.	Hot.	Hot.	Hot.	Hot.	Hot.
Is the elephants trunk up or down? (down / up)	Down.	Down.	Down.	Down.	Down.	Down.
Is the hair human or animal? (human / animal)	Human.	Human.	Human.	Human.	Human.	Human.
Qwen2.5-VL-7B, one example per stratum						
Are the lights on or off? (off / on) stratum 1: KEEP donor, SHAM-KEEP recipient, 209/640	off	off	**on**	off	**on**	**on**
Color of men's shorts? (red / black) stratum 2: both recipient, 379/640	red	red	red	red	red	red
Is the area clean or dirty? (clean / dirty) stratum 3: both donor, 43/640	clean	**dirty**	**dirty**	**dirty**	**dirty**	**dirty**
Are these people competing or doing this for leisure? (competing / leisure) stratum 4: other, 9/640	compete	compete	compete	compete	**leisure**	**leisure**
Idefics3-8B, one example per stratum						
Is the person's planner open or closed? (open / closed) stratum 1: KEEP donor, SHAM-KEEP recipient, 7/640	Open.	Open.	**Closed.**	**Closed.**	**Closed.**	**Closed.**
Is it a man or a woman? (woman / man) stratum 2: both recipient, 570/640	Woman.	Woman.	Woman.	Woman.	Woman.	Woman.
Is the bus parked on the street or sidewalk? (sidewalk / street) stratum 3: both donor, 59/640	**Street.**	**Street.**	**Street.**	**Street.**	**Street.**	**Street.**
Are these kids or adult? (adults / kids) stratum 4: other, 4/640	Adult.	Adult.	Adult.	Adult.	Adult.	Adult.

F CROSS-TASK TRANSFER AND PREDICTION

This appendix reports the cross-task and prediction results summarized in the limitations of Section 6.

Cross-task transfer. The Idefics3 and Qwen projections are applied without refitting to 160 same-question pairs each from TextVQA and TallyQA (Singh et al., 2019; Acharya et al., 2019). The comparator is a fixed rank-16 projection built from local answer gradients on the fitting quartets: the clean answer-axis gradient at each question token, averaged as outer products within quartets, trace-normalized, and averaged within and then across answer types. Two further seeds repeat the original fit. The resulting 40 endpoints form one family with 99.875% intervals (Tables 15 and 16). The FULL references are 0.135 and 0.069 for Idefics3 on TextVQA and TallyQA, and 0.101 and 1.189 for Qwen.

Clean-state prediction. A ridge regressor ($\lambda = 1$) trained on the 96 fitting quartets predicts the standardized FULL effect from 24 native features (candidate scores, state geometry, and lengths), optionally adding ten projection summaries. On the Idefics3 evaluation pairs, the native, geometric, and functional feature sets reach MSE 0.533, 0.538, and 0.620; the paired native-minus-functional interval is $[-0.122, -0.054]$ and the geometric-minus-functional interval is $[-0.115, -0.052]$.

The component thus acts when patched without being predictive, a dissociation between decodability and use also noted in probing work (Belinkov, 2022; Ravichander et al., 2021).

Table 15: Frozen-component transfer on 160 pairs per task. Retain and Remove are the sham-adjusted specificities A_1, A_2; the right columns subtract the gradient comparator's corresponding effects. Asterisks mark endpoints that meet the pre-specified amplitude-and-reference criterion (thresholds 0.045001 for Idefics3 and 0.018009 for Qwen; 99.875% intervals), which is stricter than an interval excluding zero. All four teacher-forced FULL references are positive. Appendix F lists all 40 endpoints, the extra seeds, and the generation references.

		Selective effect		Advantage over gradient	
Model	Dataset	Retain	Remove	Retain	Remove
Idefics3	TextVQA	0.0034	0.0055	-0.0006	0.0029
Idefics3	TallyQA	0.0030	0.0077	-0.0277	-0.0198
Qwen-7B	TextVQA	0.0341	0.0450	0.0094	0.0093
Qwen-7B	TallyQA	0.4255*	0.7734*	0.1589*	0.2392*

Table 16: All 40 cross-task endpoints, with 99.875% intervals. FULL references qualify in all four model–dataset cells. A1, A2: sham-adjusted retention and removal specificity; B1, B2: their advantage over the gradient comparator; G1, G2: the corresponding generation effects.

Model/task	Analysis	Metric	Mean	Lower	Upper	Support
I/Text	transfer	A1	0.00339	-0.00631	0.01888	No
I/Text	transfer	A2	0.00548	-0.00043	0.02000	No
I/Text	gradient	B1	-0.00063	-0.00882	0.00754	No
I/Text	gradient	B2	0.00287	-0.00533	0.01111	No
I/Text	generation	G1	0.00000	0.00000	0.00000	No
I/Text	generation	G2	0.00000	0.00000	0.00000	No
I/Text	seed 102	A1	0.00254	-0.00573	0.01523	No
I/Text	seed 102	A2	0.00692	0.00143	0.01998	No
I/Text	seed 103	A1	0.00205	-0.00592	0.01492	No
I/Text	seed 103	A2	0.00585	0.00050	0.01876	No
I/Tally	transfer	A1	0.00298	0.00091	0.00503	No
I/Tally	transfer	A2	0.00768	0.00536	0.01024	No
I/Tally	gradient	B1	-0.02768	-0.03619	-0.01985	No
I/Tally	gradient	B2	-0.01983	-0.02894	-0.01156	No
I/Tally	generation	G1	0.00000	0.00000	0.00000	No
I/Tally	generation	G2	0.00000	0.00000	0.00000	No
I/Tally	seed 102	A1	0.00399	0.00162	0.00628	No
I/Tally	seed 102	A2	0.00865	0.00601	0.01159	No
I/Tally	seed 103	A1	0.00172	-0.00053	0.00372	No
I/Tally	seed 103	A2	0.00605	0.00368	0.00871	No
Q/Text	transfer	A1	0.03410	-0.00736	0.18404	No
Q/Text	transfer	A2	0.04498	-0.00135	0.21634	No
Q/Text	gradient	B1	0.00939	-0.01084	0.06732	No
Q/Text	gradient	B2	0.00930	-0.00673	0.04861	No
Q/Text	generation	G1	0.00000	0.00000	0.00000	No
Q/Text	generation	G2	0.00000	0.00000	0.00000	No
Q/Text	seed 102	A1	0.03341	-0.00723	0.18032	No
Q/Text	seed 102	A2	0.04699	-0.00032	0.22102	No
Q/Text	seed 103	A1	0.03367	-0.00691	0.18075	No
Q/Text	seed 103	A2	0.04604	-0.00050	0.21862	No
Q/Tally	transfer	A1	0.42550	0.32595	0.53605	Yes
Q/Tally	transfer	A2	0.77345	0.60364	0.96456	Yes
Q/Tally	gradient	B1	0.15892	0.09882	0.22684	Yes
Q/Tally	gradient	B2	0.23917	0.16714	0.32028	Yes
Q/Tally	generation	G1	0.00000	0.00000	0.00000	No
Q/Tally	generation	G2	0.00000	0.00000	0.00000	No
Q/Tally	seed 102	A1	0.45284	0.35349	0.56160	Yes
Q/Tally	seed 102	A2	0.79553	0.62665	0.98326	Yes
Q/Tally	seed 103	A1	0.43187	0.33353	0.53883	Yes
Q/Tally	seed 103	A2	0.77732	0.60853	0.96660	Yes

G Study Record and Reproducibility

This appendix records how the analyses were specified, where they departed from plan, and what is released.

Pre-specification. Each study fixed its data, interventions, endpoints, and interval family before it was run, and the studies form separate families. Four departures apply. For the two Qwen2.5-VL models, the CLEAN and random controls of Section 4 were computed before that protocol was written. The Qwen layer sweep was first pooled without population weights for the three answer types; after the pre-specified consistency check stopped, fixed weights were set before recomputation, with no change to samples, layers, or thresholds. For Gemma3, the component study's detailed statistical specification was finalized after the model runs but before effects were aggregated, and its numerical check was revised as described in Appendix A.3. The qualitative examples were drawn after all experiments were complete, under a sampling rule fixed beforehand. The interfaces were fixed in advance and were not reselected after the layer sweep.

Development history. The design grew out of exploratory patching and steering studies whose results motivated careful control of the donor but are not part of the reported evidence. Within the reported line, geometric projections and an early functional candidate failed their original retention targets, and the study then adopted the narrower partial-function question tested in Section 5; the clean-state predictor failed on independent images. These failures are reported rather than replaced.

Reproducibility. The release contains the prompts, patch hooks, candidate boundaries, eligibility rules, statistical scripts, and saved scores needed to rebuild every table and figure. Image redistribution follows upstream terms. The Qwen component study used about 3.2 GPU-hours and the native-prompt generation study about 9 GPU-hours.