# SIGHTLINE: WHAT A VISION–LANGUAGE MODEL READS FROM AN IMAGE IS NOT WHAT IT ANSWERS FROM

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Look inside a vision–language model and you can often find what it saw in the image. It is natural to conclude that the model answers from what we found there. We show that it does not: what we read out is the vision encoder's description of the image, and the language model answers from somewhere else. SightLine tests use by intervention rather than by reading: we install a different value in the model's state, verify with an independent probe that it is present, and check whether the answer follows. In a controlled legend-lookup task, the probe reads the label assigned to a symbol from the symbol's own visual tokens with perfect accuracy, and the vision encoder writes it there within two blocks. A different label installed at those tokens is verified present, yet the answer never follows it: the model overwrites the installed label by mid-stack and keeps its original answer on every pair, even when the label is re-imposed at every layer. The answer depends instead on the printed label in the legend, on eight attention heads at the answer position, and on a layout-specific row identity at the symbol's own tokens that tells the model which legend row it is looking at. That identity is a handle: injected on Qwen2.5-VL-7B with no training, it switches the answer to the predicted label on $68\%$ and $88\%$ of held-out pairs across the two legend layouts, and back. The separation and the attention-head pathway replicate causally in three model families. On real charts the answer still depends on the printed label, the row identity is readable at legend swatches, and errors arise after a correct identity. The same pattern appears on natural benchmarks: an image-quality feature is readable in four model families, survives changes of degradation and is already present at the vision–language connector, yet predicts nothing about which items the model gets wrong. Overwriting, not reading, is the test of use. More experiments and code are provided in the appendix and https://anonymous.4open.science/r/sightline-supplement/.

## 1 INTRODUCTION

Look inside a vision–language model (VLM) and you can often find what it saw in the image. Show it a legend that pairs symbols with labels, ask which label the symbol in the picture carries, and a probe pointed at that symbol's own visual tokens reads the right label with perfect accuracy. It is natural to conclude that the model answers from what the probe found. This paper tests that conclusion with one question, asked of every representation it touches: *does the answer run through what the probe reads?* The conclusion runs together three things that can be true of a representation:

- **Decodability.** A probe can read the property from the model's state.
- **Transferability.** The content can be moved: written into another context by direct intervention, or read out unchanged after a shift in the data.
- **Causal use.** Changing the representation predictably changes the model's behavior.

That a probe can read a property the model does not use is the settled lesson of a decade of probing critique in NLP (Hewitt & Liang, 2019; Belinkov, 2022; Elazar et al., 2021). The critique separates the two by *removal*: erase the property and watch the behavior. We separate them by *installation*: write a different value in, verify with an independent probe that it is there, and watch what the model



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.

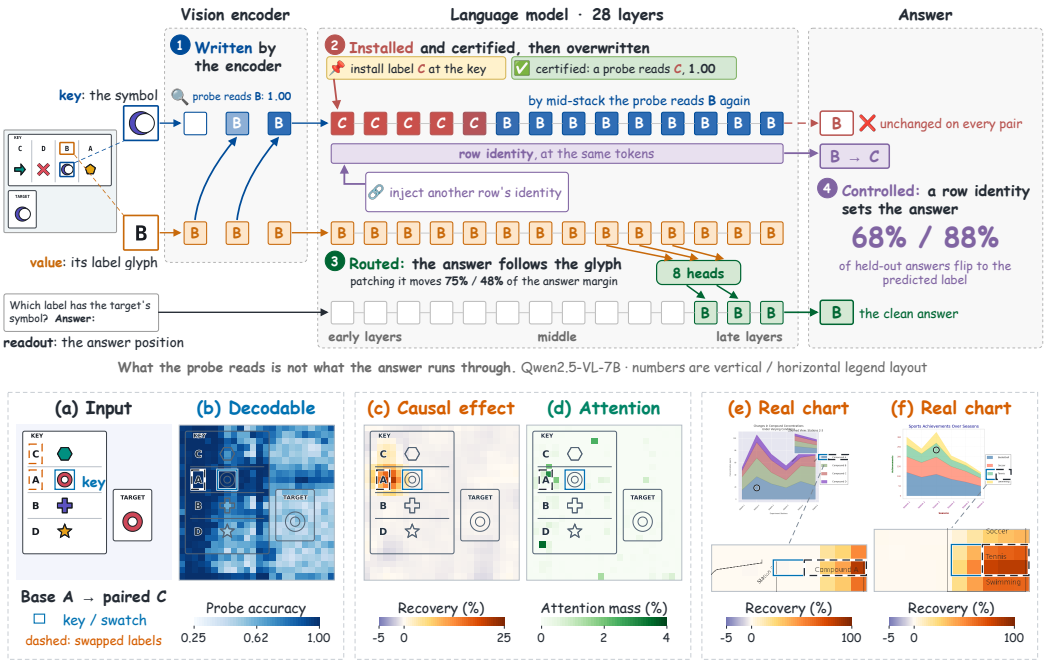


Figure 1: **What a VLM reads from a legend is not what it answers from.** Top: the pathway on Qwen2.5-VL-7B. The encoder writes the label at the key [blue]. A certified replacement is overwritten [red]. The answer follows the value glyph [orange] through eight committee heads [green], while a row identity controls which entry is read [purple]. Figure 2 reports the results. Bottom (Qwen3-VL-8B): **(a)** An example input. Blue boxes mark the key or swatch, dashed boxes the label glyphs (in (a), both swapped labels). **(b)–(d)** All $24\times24$ tokens in this image's coordinates: (b) probe accuracy over 144 pairs sharing its layout, (c) this image's relative recovery averaged over covering $3\times3$ patch windows, and (d) eight-head attention. Decodability spans the legend row and letter stripe, while causal effect and attention concentrate at the glyph and a few blank cells. **(e)–(f)** Real charts with enlarged legend patch maps on a shared scale. Color bars retain negative patch effects (Appendix C.7).

does with it. Installation comes with a check that removal cannot offer, since a removal that does nothing looks like one that failed. We call the test *certified installation* and the instrument SightLine: a probe has a clear line of sight to a representation, and whether the model shares it is the question. We run the test on the visual side of a VLM, where the vision encoder writes into every visual token before the language model exists. The write can be dated, and the question sharpens: is the probe reading the encoder's description of the image, or the language model's path to the answer?

**Natural benchmarks raise the question but cannot settle it.** The conclusion is drawn most often from pre-answer states on natural benchmarks. The pattern shows there too: an image-quality feature is readable in four model families, survives changes of degradation and beats a matched pixel baseline, yet is already present at the vision–language connector and predicts nothing about which items the model gets wrong (Appendix F). But no value of such a feature can be written in and read back, so use can only be bounded by prediction. A test needs content that can be rewritten exactly.

**One forward pass, four questions.** A legend task built from exact pixel counterfactuals supplies such content: two images differ only in which label the legend assigns to a symbol, so the visual tokens of the matched symbol (the *key*) cover identical pixels in both. We follow the label through one forward pass (Fig. 1) and ask: what does a probe at the key read, and what wrote it? If a different label is installed there, does the answer follow? What does the answer depend on instead? And can that be used to set the answer? Our findings, which are also our contributions:

- **The vision encoder writes the label into the symbol's tokens (Sec. 4).** A held-out probe reads it from the pixel-identical key with perfect accuracy, and the encoder writes it within two blocks, by a displacement so small that a cosine audit calls the states identical.
- **An installed label is verified present, then overwritten (Sec. 4).** Certified installation writes the other image's label at the key, and an independent probe verifies that it is there. The model still answers with the original label on every held-out pair: by mid-stack it has recomputed the label from the legend's own printed labels, and re-imposing it at every layer changes nothing. The null is not a coverage artifact (patching every pixel-identical token moves the answer, the symbol's own legend row moves it, and the key, inside that row, contributes nothing), and it is stated relative to its instrument, position-basis patching.
- **The answer depends on the printed label (Sec. 5).** On Qwen2.5-VL-7B, the checkpoint on which the causal chain runs end to end, the answer follows the label printed in the legend (the *value* glyph) and never the key at any layer, and is assembled at the *readout* position, where the answer is generated, by a committee of eight attention heads. On Qwen3-VL-8B, where attention was measured, the committee looks at the printed label, not the key.
- **A row identity at the key's own tokens sets the answer (Sec. 6).** A binding-ID estimator finds what position-basis patching cannot express: a *layout-specific row identity* at the key's tokens, a signal that tells the model which legend row it is looking at and carries causal evidence of a binding signal. Injected, with no training, it flips held-out answers to the letter it predicts ($68\%$ and $88\%$) and, reversed, flips them back, with slot-permuted and norm-matched controls flat.

Figure 1, bottom, shows this separation on the image, and Figure 2 quantifies each stage. The key/value separation, the dependence on the printed label and the committee replicate causally on three model families. On real charts the answer still depends on the printed label, the row identity is readable at legend swatches ($0.933$) and already correct on $85\%$ of the answers the model gets wrong, and amplifying it does not repair them: the errors are wrong bindings, not weak ones, downstream of a correct identity (Sec. 7).

The probe reads the vision encoder's description of the image, and the language model answers from something else. On the visual side of a VLM, a readable property is the default, not evidence of use: overwriting, not reading, is the test.

## 2 Related Work

**Probing and causal use.** That decodable need not mean used is established in NLP: control tasks separate what a probe extracts from what it learns (Hewitt & Liang, 2019), a survey documents it (Belinkov, 2022), amnesic probing tests behavior by *removing* a property (Elazar et al., 2021), and causal abstraction licenses an interpretation when interventions on the variable and on the network agree (Geiger et al., 2021). We take this as premise. Our delta is separation by installation, whose manipulation check removal cannot produce, on a visual pathway where the write can be dated.

**Visual information flow in VLMs.** Where a VLM's language model takes its visual information is known: extracted from the image tokens in early-to-middle layers and routed to the question and answer positions, after which the image tokens can be pruned (Chen et al., 2024a; Neo et al., 2025; Kaduri et al., 2025; Zhang et al., 2025; Basu et al., 2024), and video models route temporal information similarly (Kim et al., 2026). Those studies locate the *flow*. Logit-lens reads (Neo et al., 2025; Jiang et al., 2025) find object identity decodable at the object's own image tokens. Upstream, the encoder mixes the image globally and repurposes low-information tokens for that computation (Darcet et al., 2024), so a token can carry a label its pixels lack. We ask what such a read is reading: installing a certified value at a visual position, dating its write to the encoder and watching the answer bypass it turns "decodable at the image tokens" into a report on the encoder.

**Causal interventions.** Activation patching localizes causal representations by splicing counterfactual runs (Meng et al., 2022). Path and attribution patching extend localization from positions to edges of the computational graph (Goldowsky-Dill et al., 2023; Syed et al., 2024). Our positive results use position-basis patching and a binding-ID estimator. Our central negative (no transferable label at the key) is a negative for those instruments: a learned fixed-rank subspace (Geiger et al., 2024) leaves it intact (Sec. 4), and the edge-level search that could overturn it is named (Sec. 8).

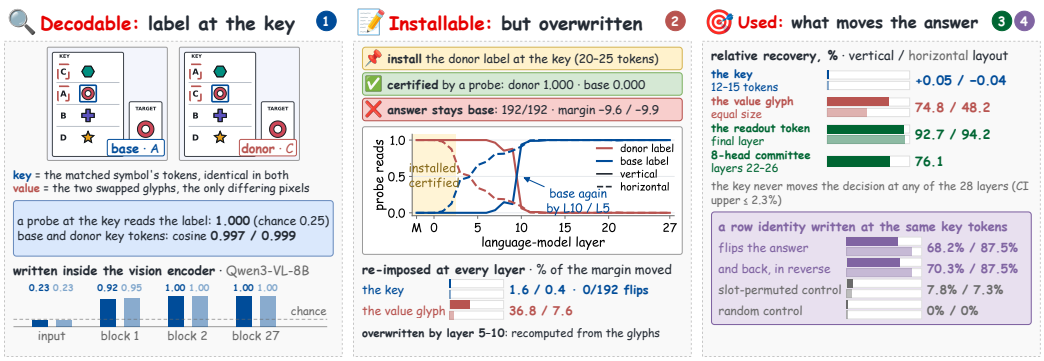


Figure 2: **Quantitative evidence for Fig. 1's pathway.** Qwen2.5-VL-7B. Recovery on 64 pairs per layout. Certification and answer control on 192 held-out pairs (encoder dating: Qwen3-VL-8B). Bars: vertical (solid), horizontal (pale). Circles: Fig. 1's stages. Left: the label is decodable at the pixel-identical key and written in the encoder. Middle: an installed label is certified, overwritten by mid-stack and never followed, even installed persistently. Right: the value glyph, readout and eight-head committee carry the answer. Row-identity injection controls it reversibly.

**Readout and binding.** In multiple-choice answering, sparse sets of mid-layer heads assemble the answer at the readout position (Lieberum et al., 2023; Wiegreffe et al., 2025), and binding IDs link entities to in-context attributes (Feng & Steinhardt, 2024). In VLMs, failures on multi-object tasks have been read as a binding problem (Campbell et al., 2024), binding IDs are shared by an object's image tokens and its textual mentions (Saravanan et al., 2025), and position IDs serve as content-independent spatial indices for binding (Assouel et al., 2026). Our account's destination side replicates the readout committee in a VLM. The binding-ID estimator we reimplement, certify on its published cells, and point at visual stimuli, where it recovers a signal position-basis patching cannot express. That row identity is content-independent in the same sense but is not the image position: it survives translating the legend, a position difference from blank cells of the same rows carries almost nothing (Sec. 6), and what it names is a layout-bound legend slot (Sec. 7).

**Internal states and self-knowledge.** Probes read calibration and truthfulness from language models (Kadavath et al., 2022; Burns et al., 2023; Azaria & Mitchell, 2023), and similar hopes attach to VLM states for hallucination detection (Li et al., 2023). Step-level verification (Lightman et al., 2023) motivates our matched verifier. We add the test it skips: privileged self-knowledge must beat *matched observable evidence* at equal capacity, and Qwen3-VL-8B's does not (Appendix F.1).

## 3 Certified Installation

### 3.1 Task, Notation, and Models

**Task and stimuli.** Each image shows four items and a legend mapping four symbols to letter labels. The paired image swaps the target symbol's legend row with a distractor's and leaves every other pixel untouched, so the answer must follow the mapping (Fig. 2, left). Stimuli come in two *visual families*, one per legend layout (vertical and horizontal), which differ in icons, fonts and backgrounds (Appendix A.1), and an unlabeled pair of numbers is vertical then horizontal. The prompt spells out the operation (find the identical symbol, read its label), and the claim is scoped to operation-specified retrieval, at ceiling on every reported family and checkpoint (Appendix B.4). We call the visual tokens covering the matched symbol the *key*, the label glyph in its legend row the *value*, and the answer position the *readout*. Within a pair, the image whose answer an intervention tries to change is the *base*, the other the *donor*. The key is nine tokens on Qwen3-VL-8B's 32-px grid. On Qwen2.5-VL-7B's unaligned 28-px grid it is 12–15 at equal size and 20–25 over the symbol's full region.

**Models.** Three checkpoints pass every behavioral criterion: Qwen2.5-VL-7B (Bai et al., 2025), Qwen3-VL-8B (Qwen Team, 2025) and gemma-3-12b-it (Gemma Team, 2025). Four others do not

and are excluded (Appendix E.1). Qwen2.5-VL-7B is the primary checkpoint: every stage from the key null to control of the emitted answer runs there, and so does the real-chart diagnosis of Sec. 7. Qwen3-VL-8B carries the instruments run only there (the encoder dating, the region ladder, the read maps, the attention audit and the commitment-depth control), and gemma-3-12b-it is the third lineage. Each section names its checkpoint, and no number is compared across checkpoints.

**Constructed counterfactuals.** Every inference in Sec. 4 rests on holding the key's pixels fixed while the binding changes. No collection of natural images supplies that, so the pair is constructed, and its ground truth lets a manipulation check certify what a patch installed. The ecological cost is measured: minimal pairs re-rendered from real chart corpora replicate the value locus under the same estimator, with the prior-attributable shortcut component bounded below $5\%$ per pair (Sec. 7).

### 3.2 Installation as a Test of Causal Use

Fix a position set $P$ and write $h$ for the model's state there on a base input, $z(h)$ for the property in question (here the label bound to the symbol at $P$), and $f(h)$ for the emitted answer. A probe is a map $g$ with $g(h) = z(h)$ on held-out inputs. The standard test of use is *removal*: apply an operator $\mathcal{R}$ meant to destroy $z$ and ask whether behavior changes (Elazar et al., 2021; Ravfogel et al., 2020). We use *installation*: apply $\mathcal{I}_v$, which writes the value $v \neq z(h)$ at $P$, and ask the same question. The two differ in what their null observations admit.

**Proposition 1** (installation separates what removal confounds)**.** *The removal observation $f(\mathcal{R}h) = f(h)$ is consistent both with $z$ being causally unused and with $\mathcal{R}$ having failed, i.e. $z(\mathcal{R}h) = z(h)$. The certified-installation observation $\big(g(\mathcal{I}_v h) = v,\ f(\mathcal{I}_v h) = f(h)\big)$, with $g$ read on the same forward pass, is inconsistent with the corresponding failure.*

The removal null cannot be told from a failed removal, whereas the first coordinate of the installation observation is a manipulation check on the intervened state itself (proof in Appendix B.7).

We implement $\mathcal{I}_v$ by activation patching: the donor run's states are spliced into the base run at $P$, and an out-of-sample probe $g$ reads the patched state on the same forward pass, at the patch site and at every later depth. What Prop. 1 leaves open, the paper measures. A single install leaves the model free to recompute $z$ downstream, which $g$ reads back at later depths, and re-imposing $\mathcal{I}_v$ at every layer removes even that freedom. The certificate speaks in $g$'s basis, and Sec. 6 finds a different write at the same $P$ that does move the answer.

### 3.3 Measuring Causal Effects

To find what the answer does depend on, we apply the same patch over nested image regions, layers, the readout position and attention heads, add a binding-ID injection (Feng & Steinhardt, 2024) for a code that base and donor share, and score every intervention with one estimand.

**Estimand.** For a base/donor pair, let $\mathrm{margin} = \mathrm{logit}(\text{donor answer}) - \mathrm{logit}(\text{base answer})$ on the answer token. A patch splices donor activations into the base run at chosen positions. *Relative recovery* is the percent of the base-to-donor margin it moves ($100\%$: answers like the donor, $0\%$: unchanged). *Reverse damage* is the same quantity with donor as base and the same positions patched. A localization claim requires its CI lower bound above zero, so no effect is an artifact of direction. Bootstraps are over pairs (10,000 resamples), never over conditions within a pair.

**Normalization and equivalence.** Recovery is normalized by each model's clean margin span over 128 pairs: 19.9, 53.9 and 28.6 on Qwen2.5-VL-7B, Qwen3-VL-8B and gemma-3-12b-it, so one logit unit reads $2.7\times$ larger on Qwen2.5-VL and $1.9\times$ on gemma-3 than on Qwen3-VL. Every table is read *within* a model. A null localization is claimed only when the recovery's bootstrap CI *upper* bound is below $15\%$ (a one-sided criterion fixed in advance), and only beside an equal-size positive control that moves the answer and, after Sec. 4, evidence that the patch installed something. Every key CI has half-width under $0.3\%$ (precision, multiplicity: Appendix B.1).

**Validity criteria.** Splits, primary estimands and criteria were fixed before any forward pass. We interpret an intervention only when its capability and manipulation checks pass. A failed check makes the result *invalid*, an instrument that could not run, not negative evidence (every instance in Appendix G, each claim's instrument and boundary in Table 3, measurement sites in Appendix B.2).

# 4 THE KEY: DECODABLE, INSTALLABLE, OVERWRITTEN

At the *key*, the label is decodable at ceiling, a different label can be installed and certified, and the answer does not follow it (Fig. 1, stages 1 and 2). The key's pixels are byte-identical across a minimal pair, so anything its tokens encode about the label was computed from context rather than read off the symbol. Numbers are Qwen3-VL-8B's, which carries the encoder dating and the region ladder. The probe, the installation certificate and its readback also run on Qwen2.5-VL-7B, the checkpoint whose answer Sec. 6 controls, and the key null holds on all three checkpoints.

**The label is decodable at the key.** A linear probe trained on held-out pairs reads the label bound to the matched symbol from the key's tokens at $1.000$ against a chance rate of $0.25$, in both visual families and against a max-statistic permutation null. Base and donor key tokens have cosine similarity $0.999$, so a cosine audit reports them as carrying the same content. The probe overrules the audit: the key carries its label at ceiling, and so does the rest of its legend row (Fig. 1, bottom b).

**The encoder writes the label.** Probing the same positions at increasing encoder depth dates the write. At the encoder's input (patch embedding plus position, the only interface with solely the token's own pixels), the probe sits at chance ($0.23$), those pixels being identical within a pair. After one attention block it reads $0.92$ and $0.95$, and after two $1.00$. The displacement carrying the label is about $2\%$ in relative $L_2$, which is why the cosine audit missed it, and a geometric map places decodability along the matched legend row and letter stripe rather than in local ink. The probe at the key reads the encoder's description of the legend, finished before the language model exists.

**A different label can be installed and certified.** Running the key patch and reading the patched state with an out-of-sample probe in the same forward pass, the installed donor value decodes at $1.000$ at the merger and through the first layers, and the base value not at all. This is the manipulation check removal cannot produce: it separates "the intervention did nothing" from "the model ignored it." The same certificate holds on gemma-3 (Appendix E.2) and on Qwen2.5-VL-7B, with $1.000$ donor and $0.000$ base on $192$ pairs unseen by a probe trained on $384$ others, at the merger, where those tokens are the donor's own, and into the language model's first layers, where they are not.

**The answer does not follow the installed label.** In the same forward passes, the patched answer is the base letter in $192/192$ pairs in both families, with a large negative donor-minus-base logit margin. On Qwen2.5-VL-7B it is likewise $192/192$ and $192/192$, at margin $-9.6$ and $-9.9$. Nor is this a one-interface artifact: layer-wise residual patching at the key's positions never moves the decision at any layer (Fig. 3b), while an equal-size patch on the value glyph does.

**The installed label is overwritten by mid-stack.** The patched positions still carry the donor's label at the merger and through the early layers, then cross back: by layers $8$–$16$ the probe reads the *base* label again from the same positions. The intervention is not ignored: the model recomputes the value it was handed from the legend's own glyphs. This is the paper's one place where a behavioral null gets a mechanism rather than an absence, and it is not one model's quirk. Qwen2.5-VL-7B, which has no re-injection of encoder tokens into the language model at all, overwrites the same install at a comparable depth or sooner: the donor value holds to layer $6$ and crosses back at layer $10$ of $28$ in the vertical layout and by layer $5$ in the horizontal, with the base label read at ceiling from layer $13$ on (Fig. 2, middle). Whatever performs the recomputation, it is not that architecture's extra pathway.

**Re-imposing the label at every layer changes nothing.** A value installed once and recomputed by mid-stack leaves open that the model would read one that survived. It would not. Re-imposing the donor's value at the key at *every* layer, with the rest of the legend held clean, still leaves the answer where it was ($0.2\%$ and $0.1\%$ of the margin here, $1.6\%$ and $0.4\%$ on Qwen2.5-VL-7B, with no pair in $192$ flipping in either model), while the same operation one row over, at the value, moves it ($8.4\%$ and $6.3\%$ here, $36.8\%$ and $7.6\%$ on Qwen2.5-VL), and the row identity written into those rows decides the answer outright (Sec. 6). Those tokens are read, and read as content: exchanging what the two symbols *are* moves the answer $11\times$ and $6\times$ further here than importing a counterfactual in which only their labels differ (Table 5). What the key never supplies is the label bound to it.

**The null is not a coverage artifact.** An incomplete patch would explain an unmoved answer, so we patched everything the label could reach (Fig. 3a) and slid the key patch's footprint across the image (Fig. 1, bottom c). Replacing *every* pixel-identical visual token ($564$ of $576$, leaving only the two swapped glyphs) recovers $76.9\%$ and $61.0\%$. The matched symbol's own $66$-token legend row recovers $16.0\%$ and $18.4\%$, the same $66$ tokens from either entry the counterfactual never touched

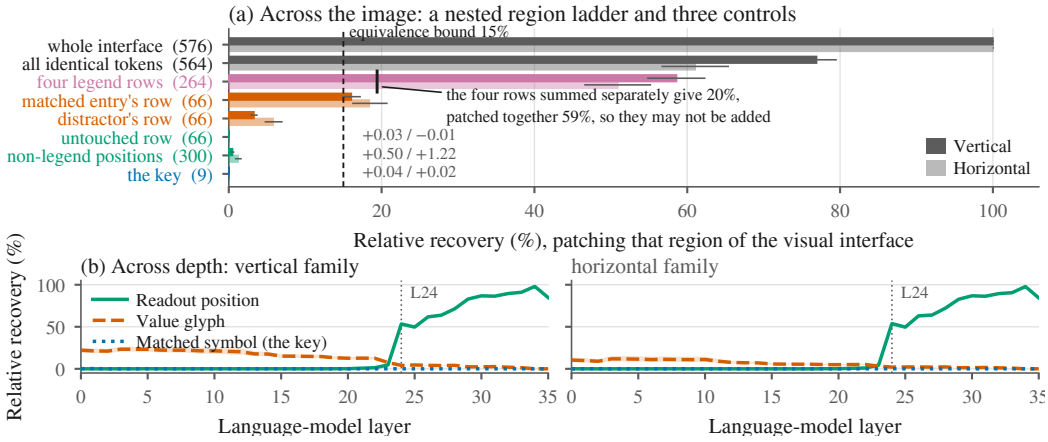


Figure 3: Causal localization on Qwen3-VL-8B. **(a)** Patching regions of the visual interface, largest first. The whole interface, the pixel-identical tokens, the four legend rows, the matched entry's row and the key are nested. The distractor's row, an untouched row and non-legend positions are controls. The two swapped glyphs carry 100% of the pixel difference, so the chain decomposes the recovery ceiling rather than sampling it. The interface is live at 66, 264, 564 and 576 tokens. The key's nine are the one region where it is not (Appendix C.4). **(b)** Layer-wise patching, pair-clustered 95% CIs: the key moves the decision at no layer (worst-case CI upper 0.53%), the value glyph contributes early and decays, the readout crosses 50% at layer 24, the model's generic commitment depth.

recover between $-0.01\%$ and $+0.05\%$, and all 300 non-legend positions (every chart item and every patch of canvas) recover $0.5\%$ and $1.2\%$. The carrier is the legend strip, specific to the entries that changed, and the key's nine tokens sit inside the most potent row in the image and contribute $0.04\%$. The rows are also where the mapping enters the language model: their merger tokens alone move the decision, with reverse damage, and the DeepStack taps do not (Appendix C.5).

**Scope.** The key null is relative to position-basis patching, which cannot see a binding code that base and donor *share*, the code Sec. 6 finds at these same tokens. Nor does it depend on the basis: on Qwen2.5-VL a learned subspace at the key (rank $\leq 1024$, three layers) flips at most $1\%$ of held-out answers, like a label-shuffled control, while the same search at the value flips $100\%$ and $99\%$ at rank 64 (Appendix C.8). Region effects are super-additive ($59\%$ for four rows that sum to $20\%$), so the carrier is localized *to* the matched row, not within it. Here the equal-size value control is the weakest of the three ($26.1\%$ and $11.9\%$ on the discovery split, the second below the $15\%$ bound), and the null rests on the within-run ratio (the key sits about $300\times$ below its value control), with the digit-task control and the stronger value controls on Qwen2.5-VL and gemma-3 (Table 12).

## 5 Localizing the Answer Pathway

Table 1 follows what the answer runs through (Fig. 1, stage 3) on Qwen2.5-VL-7B.

**The value glyph moves the answer, and the key never does.** Patching the key recovers $+0.05\%$ and $-0.04\%$, and stays inside the bound over the symbol's full region and at each of the 28 layers (CI upper $\leq 2.3\%$). The equal-size patch on the value glyph recovers $74.8\%$ and $48.2\%$, with reverse damage. With digit values and answer letters confined to a randomized option line, the digit glyph recovers $98.8\%$ and $82.1\%$: the answer follows the value where it lives, not the answer token.

**Eight heads assemble the answer at the readout.** Patching the readout token is sufficient from layer 24 of 28 and carries $93\%$ and $94\%$ of the margin by the last layer, after visual positions stop being sufficient. No single head suffices (the strongest reaches $12.4\%$). The top eight of a full sweep, ranked on one family and scored on the other, recover $76.1\%$. On Qwen3-VL-8B, where attention was measured, the committee looks at the glyph, not the key: $3.2\%$ and $1.9\%$ of its attention mass fall on the glyph's six tokens and $0.2\%$ and $0.1\%$ on the key's nine (Fig. 1, bottom d).

Table 1: Causal interventions on Qwen2.5-VL-7B. Entries are relative recovery (%, 64 pairs/layout). The last four rows give predicted flips (%, 192 held-out pairs/layout). Brackets: 95% CIs (layer sweep: worst CI upper bound). Key patches are equivalent to zero. Row writes use a $3 \times 3$ lattice.

| Intervention | Tokens | Vertical (%) | Horizontal (%) |
|---|---|---|---|
| **Key** / equal-size patch | 12–15 | 0.05 [−0.11, 0.21] | −0.04 [−0.18, 0.11] |
| Key / full region | 20–25 | 0.08 [−0.12, 0.28] | 0.18 [0.03, 0.34] |
| Key / worst of 28 layers | 12–15 | $\leq$ 2.28 | $\leq$ 0.48 |
| **Value** / equal-size patch | 12–15 | **74.84** [71.77, 77.79] | **48.16** [44.21, 52.07] |
| Readout, final layer | 1 | 92.7 | 94.2 |
| Top-8 heads, cross-family | – | 76.06 [75.11, 76.99] | |
| **Row identity** / forward | 9 | **68.2** | **87.5** |
| Slot-permuted control | 9 | 7.8 | 7.3 |
| Norm-matched random | 9 | 0.0 | 0.0 |
| Row identity / reverse | 9 | **70.3** | **87.5** |

Table 2: Cross-model replication. Cells: vertical / horizontal layout. Recovery normalized per model. Depth: first sufficient / total layers. Committees use a held-out layout. Flips use 192 held-out pairs/layout. Reverse injection restores the base answer. Dash: not tested. Qwen3-VL patch cells: the 192-pair confirmation split. The 64-pair discovery split gives 26.1/11.9 for the value.

| Measure | Qwen2.5-VL-7B | Qwen3-VL-8B | gemma-3-12b-it |
|---|---|---|---|
| Key recovery (%) | 0.05 / −0.04 | 0.04 / 0.02 | 0.04 / −0.15 |
| Value recovery (%) | **74.8 / 48.2** | **23.8 / 12.0** | **87.0 / 92.6** |
| Readout depth | 24/28 | 24/36 | 29/48 |
| Top-8 recovery (%) | 76.1 | 80.9 | 77.0 |
| Forward flips (%) | **68.2 / 87.5** | **0.0 / 0.0** | – |
| Reverse flips (%) | **70.3 / 87.5** | **0.0 / 0.0** | – |

## 6 Answer Control by a Row Identity at the Key

**A row identity at the key carries a binding signal.** Position-basis patching cannot see a binding code that base and donor share (Feng & Steinhardt, 2024). A binding-ID estimator, certified in-checkpoint on its published task, finds one: a row-identity difference injected at the key's own tokens recovers 34.1% and 54.0%, against 1.6% and 0.4% for a donor patch, 1.2% and 1.7% for a norm-matched random injection, and 12.4% and 14.4% for a slot-permuted twin. This is causal evidence of a binding signal, not a complete account of binding.

**The row identity controls the emitted answer.** With scale and layer window chosen on separate calibration pairs, the injection (Fig. 1, stage 4) flips 192 held-out answers per family to the predicted letter on 68.2% and 87.5% of pairs, against slot-permuted controls below 8% and random controls at zero, and its reverse flips donors back at 70.3% and 87.5%. No pixels or weights change (Fig. 2, right): the model reads a different legend entry because a representation the label probe never read, and position-basis patching could not see, tells it which entry it is looking at. The readable label is not the carrier, the value supplies content, and a distinct row identity selects the entry (Appendix D).

**The identity is a legend slot, not an image position.** With the legend panel moved by one to four layout cells, the same $\Delta$ still flips 81% and 90% of answers, while a position difference from blank cells of the same rows flips 12% and 0%. A six-layer window keeps 32% and 79% against permuted controls at 0% and 2%, and without the context freeze the injection is non-specific (Appendix D.7).

## 7 Generality Across Models, Codebook Sizes, and Real Charts

One invariant runs through every experiment: across model families, codebook sizes and real charts, the position that encodes the label never supplies the answer (Table 2).

**The account replicates across model families.** Qwen3-VL-8B adds DeepStack re-injection and a different token grid, and gemma-3-12b-it shares no lineage with either. Both replicate the key

null against a value control, readout assembly, and a readout committee found from scratch, since head indices do not port ($80.9\%$ on Qwen3-VL, $77.0\%$ on gemma-3). The ordering replicates, not the layer index. On gemma-3 the row delta beats its slot-permuted control but not a norm-matched random injection, so causal binding evidence is Qwen-only.

**Behavioral control is checkpoint-specific.** The grid of injection scales and layer windows that flips Qwen2.5-VL produces zero predicted Qwen3-VL flips in all $60$ calibration cells and both $192$-pair held-out sets, despite moving margins. Scale and DeepStack controls do not explain it.

**The separation holds at eight entries.** On Qwen3-VL-8B, doubling the codebook leaves behavior at ceiling, the key patch inside the bound with its manipulation check passed, and the value control moving the answer with reverse damage. The binding control is collinear at eight rows, so binding is claimed through four entries. At sixteen the model's accuracy falls, which bounds the scope.

**On real charts, errors arise after a correct row identity.** On minimal pairs re-rendered from real chart corpora (Yang et al., 2025), the matched swatch transfers nothing on Qwen3-VL-8B while the label control and a swatch from a different chart move the answer, with the shortcut component bounded below $5\%$. On Qwen2.5-VL-7B, which answers $0.739$ of $1{,}400$ legend questions there, the row identity estimated on synthetic legends is readable at the real swatches ($0.933$ of correctly answered pairs, chance $0.25$), and on $85\%$ of wrong answers the swatch already carries the correct identity. Amplifying it does not repair them ($+0.011$ accuracy, inside every control): the errors are wrong bindings, not weak ones, downstream of a correct identity (Appendix E).

## 8 Discussion and Limitations

Middle-depth assembly by a sparse committee replicates a text-only mechanism (Lieberum et al., 2023; Wiegreffe et al., 2025). The source side is new: a content-only visual key, a layout-bound row identity, and an encoder-written value certified at image tokens but bypassed by the answer.

**Natural benchmarks show the same pattern.** On pre-answer states, an image-quality axis is decodable in four model families and transfers between blur and downscale in $37$ of $40$ cells, beating a matched pixel baseline in each. It is already present at the vision–language connector, predicts none of the item-level accuracy loss, and on Qwen3-VL-8B a verifier reading it ties one reading observable evidence. Nothing can be installed there, so use is bounded by prediction (Appendix F).

The key's position is used, since a row identity written there sets the answer. The label a probe reads there is not, and the title contrasts the two. The null is instrument-relative: it survives a fixed-rank subspace search (Geiger et al., 2024), and edge-level analyses (Goldowsky-Dill et al., 2023; Syed et al., 2024) remain open. Binding evidence is Qwen-only, behavioral control checkpoint-specific, and the real-chart path incomplete (Appendix H).

## 9 Conclusion

We asked whether a VLM's answer runs through what a probe reads from its internal states, and tested it by certified installation: write a different value into the representation, verify it arrived and watch the answer. The probe read what the vision encoder wrote, and the answer came from elsewhere. In the legend task, the label at the matched symbol's tokens is decodable at ceiling, installed, certified and overwritten by mid-stack, while the answer follows the label's glyph, an eight-head readout committee and a layout-specific row identity whose injection sets the emitted answer on Qwen2.5-VL-7B. The core of this account replicates causally across three model families, and its value locus on real charts. The same pattern appears over natural benchmarks, where a decodable, transferable image-quality feature predicts none of the model's item-level errors. On a VLM's visual side, decodability is the default, not evidence of use, and a claim of use needs an intervention with a manipulation check. The row identity is a train-free handle on which entry the model reads and a diagnostic that locates real-chart errors in the value-side read a repair would target. Open directions are a binding instrument beyond four entries and the row identity's necessity.

### AI USE STATEMENT

We used generative AI tools only to polish the language of this manuscript. The authors reviewed every AI-suggested edit and take full responsibility for the content of this work.

### ETHICS STATEMENT

This work studies internal mechanisms of publicly released vision–language models using synthetic stimuli and charts re-rendered from public corpora. It involves no human subjects, personal data, or model releases. Insights about steering emitted answers via activation injection (Sec. 6) require white-box access to model internals and are reported with controls to support interpretability research.

### REPRODUCIBILITY STATEMENT

For every study, data splits, stimulus manifests, primary estimands and pass/fail criteria were written into a protocol document before any model forward pass, and failed capability or manipulation checks are recorded as invalid rather than negative. Appendix B gives the criteria and estimators used in the main text, including the estimator's model-dependent normalizer and the exact form of the equivalence criterion. Appendices C–E report the full per-arm numbers, including every control condition, and Appendix J lists every checkpoint and data source. All supplementary materials are available in an anonymized repository at `https://anonymous.4open.science/r/sightline-supplement/`: the stimulus generators, manifests with their SHA-256 hashes, patch runners, analyzers with their threshold regression tests, and the protocol documents with their revision history, so that the order of criteria and results can be checked.

## REFERENCES

AIDC-AI. Ovis1.6-Gemma2-9B. Hugging Face model card, 2024. URL `https://huggingface.co/AIDC-AI/Ovis1.6-Gemma2-9B`.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2506.15871`.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37, 2024. doi: 10.52202/079017-0237.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://arxiv.org/abs/2411.00238`.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 2024a. doi:

10.1007/978-3-031-73004-7_2. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10478_ECCV_2024_paper.php.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37, 2024b.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Matt Deitke, Dustin Schwenk, Jordi Salvador, et al. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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, 2021.

Sagi Eppel, Mor Bismut, and Alona Faktor-Strugatski. Shape and texture recognition in large vision-language models. *arXiv preprint arXiv:2503.23062*, 2025.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, et al. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

HuggingFaceM4. Idefics3-8B-Llama3. Hugging Face model card, 2024. URL https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=94kQgWXojH.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14558, 2025.

Minji Kim, Taekyung Kim, and Bohyung Han. Map the flow: Revealing hidden pathways of information in VideoLLMs. In *International Conference on Learning Representations*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Shiyin Lu, Yang Li, Yu Xia, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/900fb3439e4968df79a6f2bfedec49cd-Abstract-Conference.html.

Qwen Team. Qwen3-vl-8b-instruct. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

Darshana Saravanan, Makarand Tapaswi, and Vineet Gandhi. Investigating mechanisms for in-context vision language binding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. URL https://arxiv.org/abs/2505.22200.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

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

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, 2025.

Yuwei Yang, Zeyu Zhang, Yunzhong Hou, Zhuowan Li, Gaowen Liu, Ali Payani, Yuan-Sen Ting, and Liang Zheng. Effective training data synthesis for improving MLLM chart understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2653–2663, 2025.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19781–19791, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

## APPENDIX OVERVIEW

**A Stimulus and Task Details** . . . . . 14
Rendered pairs with legends, digit legends and text options, codebooks of 8 and 16 entries, real-chart pairs and the degradation ladder. Extends Sec. 3.

**B Protocols, Estimators, and Validity Criteria** . . . . . 17
The claim ledger, the causal estimator, measurement sites, validity criteria, cross-backend agreement, the proof of Prop. 1 and reproduction. Extends Sec. 3.

**C The Key: Additional Results** . . . . . 22
Decodability, the certificate and overwrite, an install that cannot be overwritten, the region ladder, the shape key, read maps and a learned subspace. Extends Sec. 4.

**D The Answer Pathway: Additional Results** . . . . . 30
Layer sweep, commitment depth, the readout committee, digit legends, the row identity, answer control on two checkpoints and position controls. Extends Secs. 5 and 6.

**E Generality: Additional Results** . . . . . 37
The behavioral screen, gemma-3-12b-it, codebooks of 8 and 16 entries, real charts and one table across settings. Extends Sec. 7.

**F Probing on Natural Benchmarks** . . . . . 40
Four model families, path selection, decodability and transfer, the matched verifier and a terminated direction. Extends Secs. 1 and 8.

**G Validity Checks That Rejected Something** . . . . . 44
Each validity check that rejected a model, task size, instrument, run or sentence. Extends Sec. 3.

**H Extended Limitations and Outlook** . . . . . 47
What is new and what replicates, the limitations in full, and the outlook. Extends Sec. 8.

**I Side Study: Base-Color Attraction in Geometry Judgments** . . . . . 48
A separate finding on geometry judgments, kept compact. The main text does not rely on it.

**J Models and Benchmarks** . . . . . 48
Every checkpoint and benchmark, and each model's visual-token grid. Extends Sec. 3.

## A STIMULUS AND TASK DETAILS

Every mechanism claim in this paper is measured on stimuli that were rendered, audited and hashed before any model produced an answer. This appendix catalogs the task families with rendered examples, because the strength of the causal design is easiest to judge by looking at what actually differs within a minimal pair.

### A.1 FOUR-ENTRY LEGEND FAMILIES

Each stimulus is a 768×768 canvas containing a legend panel (labeled KEY in the image, with four rows or columns mapping icons to letter labels) and a TARGET card showing one icon (Fig. 4). The prompt specifies the operation ("find the identical symbol and read its label"). Two pair types share this geometry:

**Value-locus pairs.** The pair fixes the target symbol and all icon pixels and transposes exactly two label glyphs, so the matched symbol's row carries a different letter. The only pixel difference between base and donor is the two swapped glyphs. The matched symbol is pixel-identical. These are the pairs behind the key null, the readout sweep, the committee, and the row-identity estimator.

**Query pairs.** The complement: the legend is pixel-identical and only the TARGET icon changes, so the model must re-run the match. The query is not a self-sufficient transplantable key (Appendix C.6).

Scale: 576 pairs per pair type per family, 2,304 questions per set (both mappings × two operation-specified prompts), 9,216 behavioral questions and 5,760 images in total, with image hashes unique, disjoint across families and from the earlier legend study of Appendix C.5. Balancing is exactly factorial: answer letters, target icons, base/donor column pairs, backgrounds, and the four target

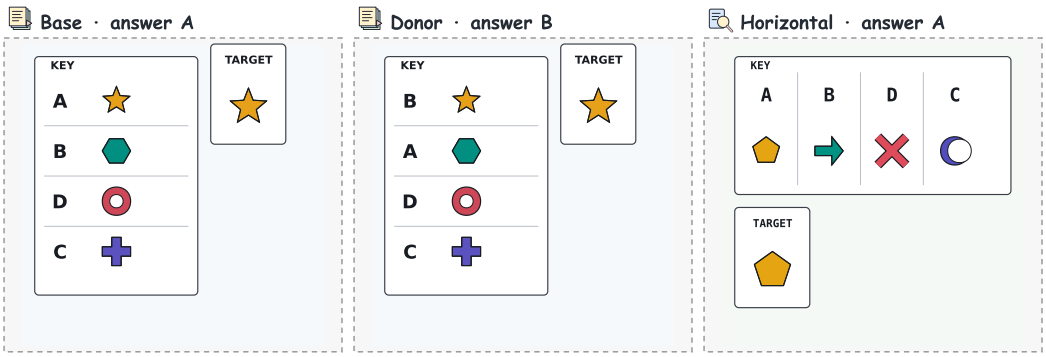


Figure 4: Vector redraws of the stimuli. Left, middle: a value-locus minimal pair from the vertical family (pair `car-v-v-000`, base and donor). The four icons, the target, the background, and every pixel outside two label glyphs are byte-identical. Only the labels of the matched row (star) and a distractor row (hexagon) are transposed (A↔B), so the correct answer must flip from A to B. Right: the base image of the corresponding horizontal family pair (`car-v-h-000`), with a different icon vocabulary (pentagon/arrow/x/crescent), fonts, layout axis, and backgrounds.

positions are each exactly balanced, and the target position is fully crossed with the answer letter (a first implementation left 4 of the 16 answer×position cells empty and was corrected before any model answer).

**Pre-answer render audits (per family, per pair type).** Every counterfactual pair's answers differ. Value-locus pairs differ *only* in the two swapped glyphs (all icon regions pixel-diff zero). Region masks are non-empty, grid-aligned, and equal-token-count where controls require it. Letter and symbol boxes share *zero* merged visual tokens (the 32-pixel grid alignment that a naive port of the predecessor layout would have violated: the letter and symbol boxes would have shared three merged tokens, and a payload verdict could have been produced by leakage alone).

### A.2 Digit Legends and Text Options

Because in the letter stimuli the retrieved value *is* the emitted answer token, a deflationary reading existed (the "value" results might be answer-token bookkeeping). Two derived families remove it:

**Digit legends.** The legend maps icons to digit glyphs from $\{2, 4, 7, 9\}$. The A–D letters leave the image entirely and appear only in a text option line `A)` $\sigma$`(A)` `B)` $\sigma$`(B)` `...` under a per-pair bijection $\sigma$, balanced so the emitted letter is exactly decoupled from the retrieved digit (each of the 24 bijections used exactly 24 times, and every answer-letter × answer-digit cell exactly $576/16 = 36$). Token geometry is untouched: digits render at the same size in the same cells.

**Text counterfactual.** Holding the digit-legend *image* byte-identical, the counterfactual transposes two digits in the text option line only, verified to tokenize to the same length and to differ at exactly the two option-digit positions. This is the pair type that located the value-to-answer-token mapping at the option-line digit token and measured its editing deadline (Appendix D.4).

**Non-legend MCQ controls.** Two further tasks keep the four-option format and remove retrieval entirely: a no-legend visual task (name the displayed shape, with options as shape names in text) and a text-only task (the target named in words). These are the controls that show the layer-24 landmark to be the model's generic commitment depth (Appendix D.2).

### A.3 Codebook Scaling to 8 and 16 Entries

The scaling generator extends the renderer lineage: same canvas, same per-entry cell sizes, same paired row-exchange construction, labels `A..H` / `A..P` audited as distinct single tokens per model, 576 pairs per size per family with hashes disjoint from all four-entry sets and from the other size. Pre-GPU audits additionally fix 64-pair mechanism manifests stratified by answer × target position,

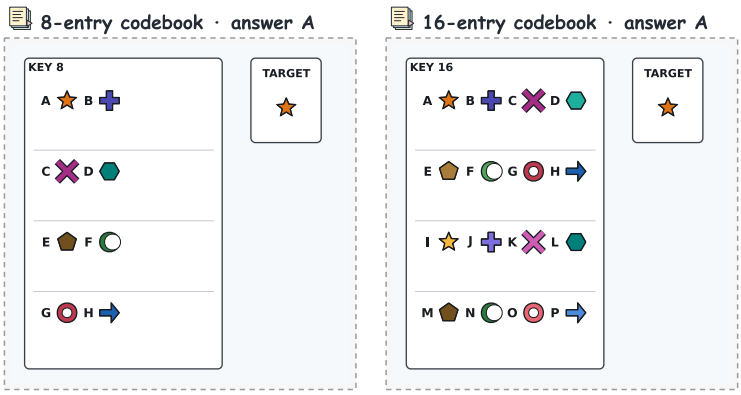


Figure 5: Vector redraws of the stimuli. Left: an 8-entry stimulus (`car-s8-v-000`, labels A–H, 4 rows × 2 columns). Right: a 16-entry stimulus (`car-s16-v-000`, labels A–P, 4×4). In the 16-entry codebook the icon vocabulary necessarily reuses shapes at different shades (compare rows 1/3 and 2/4: star/cross/x/hexagon and pentagon/crescent/ring/arrow in two color variants each). The size-16 capability failure concentrates exactly on these same-shape different-shade confusions (Appendix E.3).

and 64 content-balanced row-identity estimation samples per cell with all cyclic rotation contexts (512 images at size 8, 1,024 at size 16).

### A.4 Real-Chart Minimal Pairs (ECD)




Figure 6: Vector redraw of a real-chart minimal pair (`car-ecd-ecd10k-004647`), re-rendered from corpus-shipped matplotlib source under a fixed seed. The plot area, the ringed query point, and the option text are identical. The only difference is which series label sits beside which legend swatch (here "Software Adoption" and "Data Analysis Requests" are transposed), so the correct answer for the ringed teal bar changes. This pair is also one of the three on which the two inference interfaces disagree: it generates correctly under the deployment backend on both prompt variants, yet its donor form misses the exact argmax under the HF logit interface, reproducibly (Appendix G).

ECDBench and ECD-10k-Images ship the matplotlib source for every chart (1,224 and 10,535 scripts), which is what makes exact pixel counterfactuals possible on real charts: re-rendering under a fixed seed yields two forms that differ only in the legend's label–swatch pairing (Fig. 6). Content (data, series names, palettes, chart types) is untouched. Layout is normalized as a *disclosed* deviation: a fixed 768×576 canvas passes through the vision preprocessor unchanged, the legend is rebuilt outside the plot area (the corpora overwhelmingly use `loc='best'`, which occludes the data), and legend geometry is derived from the token grid so key and value never share a token. Twelve per-pair audits raise rather than warn. Among them, the pixel difference is confined to the

two swapped label boxes, swatches and texts are pairwise token-disjoint, the ring marker is identical across forms, and the option order is never the identity permutation.

Because ECD series names are semantically load-bearing (the data is generated to match them), a model can answer from priors without reading the legend. The behavioral gate therefore scores *pairs*, not charts: a pair survives only if all four cells ($\{$base, donor$\}$ $\times$ two prompts) are answered correctly, which is only possible by reading the legend. Of 350 audited script-disjoint pairs, 222 survive ($0.6343$ $[0.5829, 0.6829]$ against a $0.50$ floor fixed in advance). The survivor set is balanced on every design factor (all $p > 0.25$) and selected only on ring isolation: post-hoc analysis shows survival tracks the ringed point's clearance from other series ($p = 0.0245$), not swatch color separation ($p = 0.51$). The pair-level $2\times2$ table (both correct 227 / base-only 37 / donor-only 20 / neither 66 on the current gate outputs) bounds the prior-attributable component at $17/350 = 4.9\%$ $[0.6\%, 9.1\%]$ (historical first-pool snapshot: $11/150 = 7.3\%$, never mixed with the current denominator).

### A.5 Degradation Ladder for the Probing Experiments

The calibration stimuli behind Appendix F.1 hold the item, question, options, and token layout fixed and change only pixels: the clean image, five scale-invariant Gaussian blur levels (radius $f \cdot \min(H, W)$, $f \in \{0.0025, 0.005, 0.01, 0.02, 0.04\}$) and five downscale–upscale levels ($s \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$), plus a uniform blank image as the floor, applied *after* replicating the model's own image preprocessing so the degradation the model receives is exactly the one applied. Scale: five benchmarks (GQA, MMStar, POPE, HallusionBench, and MathVista) (Hudson & Manning, 2019; Chen et al., 2024b; Li et al., 2023; Guan et al., 2024; Lu et al., 2024), four model families, 250 base items per benchmark for the original two and 100 for the added three, 16 sampled paths per view. A third operator (five-level Gaussian noise, 1,250 views, 20,000 paths) exists for the transfer-breaking control.

### A.6 Environment and Backends

All experiments ran locally on 8$\times$ NVIDIA L20 (46 GB). Behavioral gates run on vLLM with greedy decoding and a strict single-letter parser. Mechanism runs use native HF Transformers forwards. Cross-backend agreement is measured before any scan on a new checkpoint and required to be $\geq 0.98$, after a case in which the two backends agreed on only $83.4\%$ of items for one candidate model (Appendix G).

## B Protocols, Estimators, and Validity Criteria

For every study, the stimulus manifests (with SHA-256 hashes), data splits, primary estimands, pass/fail criteria, control conditions and the capability checks that must pass before any effect is interpreted were written into a protocol document before any model forward pass. Those documents are part of the supplementary material. Table 3 summarizes, per claim, the instrument that decides it and the boundary of the conclusion it supports. The rest of this appendix gives the machinery behind that table.

Table 3: Claim ledger: for each claim, the instrument that decides it, the checkpoints and settings it holds on, and the boundary of what it supports. Recovery magnitudes are not comparable across models (Sec. 3). Read within a column.

| Claim | Decisive evidence (instrument) | Holds on | Boundary |
|---|---|---|---|
| *The key* (Sec. 4) | | | |
| The key's tokens carry the bound label | Held-out linear probe at 1.000, chance 0.25 (decoding) | Qwen3-VL, Qwen2.5-VL, gemma-3 | Decodability only, licensing nothing about use |

*continued on the next page*

*Table 3, continued from the previous page.*

| Claim | Decisive evidence (instrument) | Holds on | Boundary |
|---|---|---|---|
| The label can be installed there and verified | Independent probe reads the donor's label off the patched key at 1.000 (manipulation check) | Qwen3-VL, Qwen2.5-VL, gemma-3 | Certifies what the patch installed, not what the model does with it |
| No position-basis transfer at the key, at any layer | Layer-wise residual patching, CI upper below the 15% bound at every layer | Qwen3-VL, Qwen2.5-VL | Instrument-relative: survives a learned fixed-rank subspace at the key (next row), while the edge-level analysis stays open |
| No learned subspace at the key transports the label | Rank-$r$ interchange at the key, $r \leq 1024$, three layers: flips $\leq 1\%$, recovery CI upper $\leq 6.8\%$, label-shuffled control at the same level, while the same search at the value glyph flips $100/99\%$ at rank 64 | Qwen2.5-VL | Linear fixed-rank interchange applied once per layer at five layers, not covering non-linear readouts or pair-varying subspaces |
| The carrier is the legend strip, entry-specific, and skips the key | Nested region patch: all 564 pixel-identical tokens $76.9/61.0\%$, the matched row $16.0/18.4\%$, an untouched entry's row $\pm 0.05\%$, all 300 non-legend positions $0.5/1.2\%$ | Qwen3-VL | Region-completeness closed here only. Effects are super-additive, so regions may not be summed. Qwen2.5-VL closed at the symbol's region, never at row granularity |
| The installed value is overwritten from context | Same-pass probe reads the donor's label at the merger, the base label again by mid-stack (L8–16 of 36 on Qwen3-VL, L10 and L5 of 28 on Qwen2.5-VL) | Qwen3-VL, Qwen2.5-VL | Not measured on gemma-3. The checkpoint without encoder re-injection overwrites at a comparable depth or sooner, so the pathway is not that architecture's |
| The null survives an install that cannot be overwritten | Donor activations re-imposed at every layer with the legend frozen: $0.2/0.1\%$ and $1.6/0.4\%$, no pair flipping, against $8.4/6.3\%$ and $36.8/7.6\%$ one row over | Qwen3-VL, Qwen2.5-VL | Same instrument as the row-identity write. The two symbol swaps beside it are post-hoc diagnostics on Qwen3-VL only |
| *What the model uses* (Secs. 5, 6) | | | |
| The answer is assembled at the readout position by a small readout committee | Top-8 of a full head sweep, ranked on one visual family and scored on the other | Qwen3-VL, Qwen2.5-VL, gemma-3 | Head indices and the absolute layer do not port, but the ordering does |
| Attention identifies the route but does not rank the movers | Within-committee attention mass versus causal recovery, $\rho = -0.14$ | Qwen3-VL | One task, one committee. A caution about attention-based attribution, not a general law |
| A layout-specific row identity carries causal binding evidence | Binding-ID injection against slot-permuted and norm-matched controls, paired CIs excluding zero | Qwen3-VL, Qwen2.5-VL | Causal evidence of a binding signal, not a complete account of binding. The slot-specific part is layout-bound |

*continued on the next page*

*Table 3, continued from the previous page.*

| Claim | Decisive evidence (instrument) | Holds on | Boundary |
|---|---|---|---|
| The row identity is a legend-slot code, not an image position | The same $\Delta$ on legends moved by one to four layout cells ($81/90\%$ flips against $68/88\%$), a blank-cell position difference from the same rows ($12/0\%$), and a six-layer window ($32/79\%$) | Qwen2.5-VL | Within-layout translation only. Across layouts the slot-specific part stays layout-bound. Without the context freeze the injection is non-specific |
| *Scope* (Sec. 7) | | | |
| The row identity controls the emitted answer (Sec. 6) | Predicted-letter flip rate on zero-overlap held-out pairs | Qwen2.5-VL only | Zero flips on Qwen3-VL under the same grid. Neither cell extrapolates to the other |
| Neither magnitude nor the re-injection pathway explains that difference | The $\alpha$ grid, and the whole grid re-run with the DeepStack taps suppressed in context, estimation and injection alike | Qwen3-VL | Still zero predicted and zero any-flips in all 120 cells and both held-out sets, with capability inside the criterion and recovery *higher*. The difference is unexplained |
| The key/value separation survives ecological and load shifts | Same analyzer, unchanged criteria, on real chart minimal pairs and an eight-entry codebook | Real charts, size 8 | Real charts support the value locus only. The binding instrument and the real-chart readout check fail validity and are not read as negative |
| On real charts the row identity is readable, and the errors are downstream of it | Slot readout at the swatch ($0.933$ on correct pairs, $0.855$ on wrong pairs, chance $0.25$). Answer-agnostic amplification against norm-matched, subspace and reversed controls ($+0.011$, all inside) | Qwen2.5-VL, real charts | Decodability, not causal transport across layouts. The null characterizes the error (wrong, not weak) and is not a failed repair |
| *Probing on natural benchmarks* (Appendix F.1) | | | |
| A visual-accessibility axis is decodable and transfers across operators, yet knows nothing item-level and beats no matched observable evidence | Cross-operator transfer against a matched classical pixel probe, an item-level residual null, and a matched latent versus output-only verifier | 4 families, 5 benchmarks (transfer). Qwen3-VL-8B (verifier) | Low-pass operators only, and the axis already exists at the connector. No intervention on this axis was run: usability is bounded behaviorally, not representationally |

### B.1 Causal Estimator

For a base/donor minimal pair, all causal arms share one estimand. On the answer-token first-token logits,

$$\text{margin} = \text{logit}(\text{donor letter}) - \text{logit}(\text{base letter}), \qquad \text{recovery} = \frac{\text{margin}_{\text{patched}} - \text{margin}_{\text{base}}}{\text{margin}_{\text{donor}} - \text{margin}_{\text{base}}}.$$

Recovery $1$ means the patched run answers like the donor, and $0$ means unchanged. *Reverse damage* is the identical quantity computed with the roles exchanged (the donor run as base, the same positions patched from the base run), and every localization claim requires its CI lower bound above zero, so that a recovery cannot be an artifact of one direction. The bootstrap unit is always the *pair* (10,000 resamples). Strata, dual prompts, and conditions within a pair are never treated as independent samples.

**Normalization and cross-model comparison.** The normalizer $\text{margin}_{\text{donor}} - \text{margin}_{\text{base}}$ is a property of the checkpoint, not of the intervention. Measured clean spans are 53.9 on Qwen3-VL-8B, 19.9 on Qwen2.5-VL-7B and 28.6 on gemma-3-12b-it (128 pairs each, both families pooled, recomputed from the recorded patch shards), so an identical absolute logit movement yields a recovery $2.7\times$ larger on Qwen2.5-VL and $1.9\times$ larger on gemma-3. Two consequences are carried through the paper. Recoveries are compared *within* a checkpoint, never across (the $11$–$29\times$ cross-model row-identity gap in Appendix D.5 is reported with this factor of three already flagged). And the $< 15\%$ equivalence bound, being absolute, is a slightly different test on each model: it is easier to satisfy where the span is large. What makes the key nulls interpretable is therefore not the bound alone but the within-run ratio to the equal-size positive control, which runs from roughly $300\times$ to over $1000\times$ depending on the cell.

**Equivalence and multiplicity.** A null localization requires the bootstrap CI *upper* bound below the $15\%$ equivalence bound: a one-sided equivalence criterion fixed before execution, not a post-hoc reading of a non-significant point estimate. No power analysis or minimum detectable effect was fixed in advance. What the design supports is the precision it achieved, and in every key arm the CI half-width is below $0.3\%$. Recomputed from each arm's own pair-level dispersion, that precision is a minimum detectable effect at $80\%$ power (two-sided $0.05$) of $0.14\%$ and $0.12\%$ on Qwen3-VL-8B and $0.29\%$ and $0.22\%$ on Qwen2.5-VL-7B: between 50 and 125 times finer than the $15\%$ bound, and between 100 and 260 times finer than the equal-size value control measured on the same pairs, whose recoveries are $26.05\%$ / $11.89\%$ and $74.84\%$ / $48.16\%$. The observed key recoveries ($0.08\%$ / $0.04\%$ and $0.08\%$ / $0.18\%$) sit at or below that detection floor. These nulls bound a small effect. They are not a failure to see a large one. Each arm carries one primary estimand with its own criterion, fixed in advance. The specificity and control conditions reported beside it are uncorrected for multiplicity and are exploratory context for that primary, not a family of independent tests. Two hard instrument gates accompany every patch run: the full-interface patch must reproduce the donor margin exactly (per-sample deviation $< 10^{-3}$, in practice $0.0$ in all reported runs), and clean base/donor margins must be large and opposite in sign (a silent encoding failure produces a blind model with near-zero margins, a failure mode that was caught, Appendix G).

### B.2 Measurement Sites in the Forward Pass

Figure 7 places each intervention of Secs. 4–6 at the stage of the forward pass it acts on, with the token regions of Fig. 3a drawn at their measured geometry.

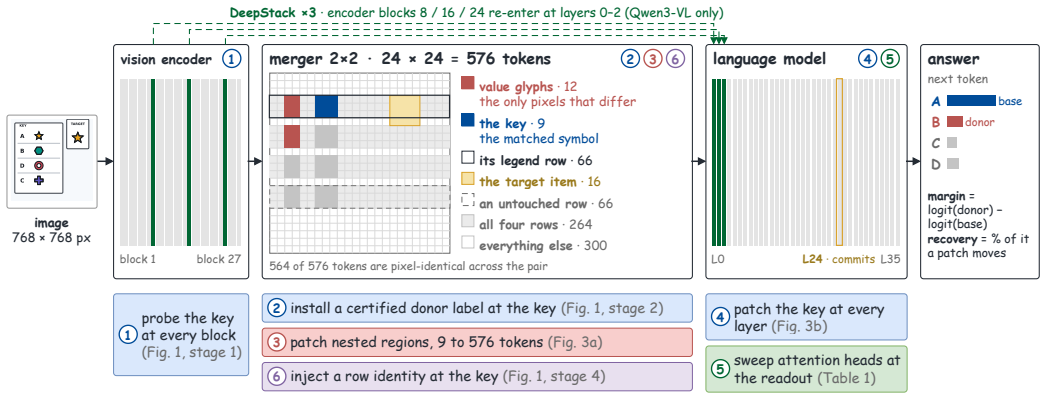


Figure 7: Measurement sites in the forward pass of Qwen3-VL-8B: a 27-block vision encoder, the $2\times2$ merger that leaves $24\times24 = 576$ visual tokens, and a 36-layer language model that commits by layer 24. The three DeepStack taps make "the visual interface" the merger plus three injection sites. The merger grid is drawn at the measured geometry, so the nested regions of Fig. 3a are exact (9, 66, 264, 564 and 576 tokens), and only the 12 tokens of the two swapped glyphs differ within a pair. Each of the six interventions sits under the stage it acts on. The answer is read as the margin, $\text{logit}(\text{donor}) - \text{logit}(\text{base})$, and every patch is reported as the percent of that margin it moves. Layer indices are Qwen3-VL-8B. The claim is the ordering, not the index.

### B.3 Invalid Instruments

A failed capability or manipulation check ends an arm as *invalid*: the effect is not interpreted in either direction, and the result is never reported as a negative mechanism finding. The rule is kept from becoming a shield in three ways. The checks reject something (the behavioral screen rejected four checkpoints and one task size, Appendix E.1). Every equivalence claim is valid only beside an equal-size positive control that does move the answer and, for the key, beside a manipulation check showing the patch installed decodable donor content. And each protocol records a prediction that is scored afterwards, misses included (Appendix G).

### B.4 Behavioral Capability Gates

The legend-task criterion (per pair type, per family): parse rate $= 1.00$ under a strict single-letter parser, overall accuracy $\geq 0.98$ with pair-clustered CI lower bound $> 0.95$, counterfactual pair consistency $\geq 0.95$ (CI lower $> 0.92$), answer flip rate $\geq 0.98$, dual-prompt consistency $\geq 0.95$, and every stratum (answers, icons, backgrounds, positions) $\geq 0.95$. Scaled codebooks use the same structure with per-size criteria fixed in advance. The real-chart criterion is the pair-survival rule of Appendix A.4, and the answer-control study adds zero-delta drift ($\leq 10^{-3}$), delta non-collinearity, and grid-completeness guards.

### B.5 Selection Protocol

Discovery and confirmation are separated everywhere. Committee heads are ranked on one visual family and scored out-of-family. Scale and layer windows for the answer-control injection are selected on a 64-pair calibration split under a tie-break rule fixed in advance and then confirmed on 192 zero-overlap held-out pairs. Row-identity deltas are estimated on stimuli that share no file with the evaluation manifests. Probe folds are grouped by pair and hashed before the first fit. Where a choice was nonetheless made while looking at evaluation data, it is disclosed as such (Appendix G).

### B.6 Cross-Backend Agreement

Behavioral gates run on vLLM, and mechanism forwards run on HF Transformers. After the Idefics3 case ($83.38\%$ item-level agreement between backends, Appendix G), agreement $\geq 0.98$ became a precondition for any scan on a new checkpoint: gemma-3 measured $1.0000$ vertical / $0.9983$ horizontal, and the two Qwen families reproduce vLLM behavior exactly under HF. The measured values matter more than the threshold, because a nominal $0.98$ would admit roughly four disagreeing pairs in a 192-pair held-out set, the same order as some of the effects reported here. What bounds the exposure is that the split is confined to the *gate*: every recovery, every layer sweep, and every flip is a comparison between two runs of the same HF harness on the same input, with the no-injection baseline recomputed in that harness rather than imported from the vLLM gate. A backend disagreement can therefore make a behavioral gate unrepresentative of the harness the mechanism runs in, but it cannot manufacture a recovery or a flip. The real-chart readout sweep (Appendix E.4) is this rule biting in the opposite direction: three pairs that generate correctly under vLLM miss an exact HF argmax criterion, and the sweep is not read rather than reinterpreted.

### B.7 Proof of Proposition 1

*Proof.* Under a failed removal nothing at $P$ changed, so $f(\mathcal{R}h) = f(h)$ identically. Under a genuine removal of an unused property the answer is likewise unchanged. The observation is the same in both, and no further reading of $\mathcal{R}h$ distinguishes them: a probe that now fails on $\mathcal{R}h$ shows only that $z$ is no longer decodable *in that basis*, which a successful removal and a displacement of $z$ both produce. For installation the first coordinate is a manipulation check evaluated on the intervened state itself: $g(\mathcal{I}_v h) = v \neq z(h)$ entails $\mathcal{I}_v$ took effect, so the second coordinate is a behavioral null under a demonstrably changed state. □

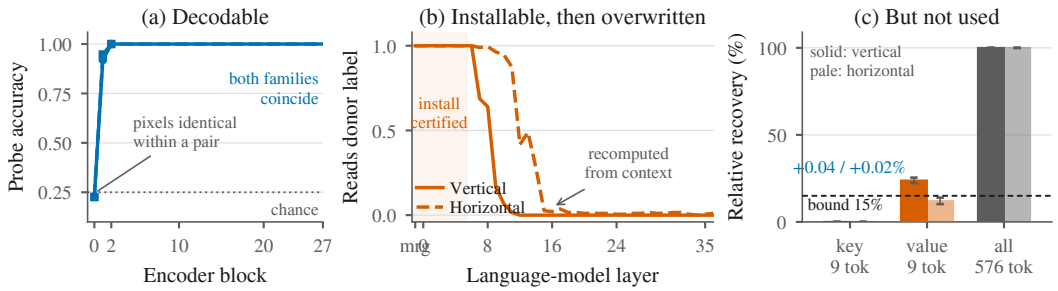


Figure 8: The three properties on Qwen3-VL-8B (Fig. 2 is the Qwen2.5-VL-7B version). **(a)** A held-out probe reads the bound label from the key at 1.000 against chance 0.25. At the encoder's input it is at chance, dating the write to encoder blocks 1–2. **(b)** With a donor's value patched in, an independent same-pass probe reads the donor's label at 1.000 through layer 5 and the *base* label again by mid-stack. **(c)** The answer does not follow: recovery at the key is $+0.04\%/+0.02\%$ against $23.8\%/12.0\%$ for an equal-size patch on the value glyph. Dashed line: the 15% equivalence bound. Confirmation split, $n$=192 per family.

### B.8 Reproduction

Each protocol document records its exact commands, environment, manifest hashes and output directories. Outputs are never overwritten (new replications write to new paths), superseded sets are retained rather than deleted, and analysis scripts pin their criteria to the recorded literals under regression tests. The anonymized supplementary package contains the stimulus generators, manifests, patch runners, analyzers and the protocol documents.

## C The Key: Additional Results

This appendix gives the full results behind Sec. 4: the decodability grid and its dating inside the vision encoder, the installation certificate and its layer-by-layer readback on both Qwen checkpoints, the install that cannot be overwritten, the nested region ladder, and the read maps drawn over the image. Unless a paragraph says otherwise the model is Qwen3-VL-8B and the stimuli are the value-locus pairs of Appendix A.1. Conventions are those of Appendix B.

### C.1 Decodability at the Key and the Encoder's Write

A linear probe trained on held-out pairs reads the bound letter from the key's 9 merger tokens at 1.000 in both families, against a max-statistic permutation null (p95 = 0.288). The earlier inference from the base/donor cosine of 0.999 (relative $L_2$ 0.045) that the key "does not encode" its label was wrong, and an offset-translated decodability map shows why the two families' shifted controls had disagreed: decodability is carried along the matched legend row or column and the letter stripe, not by local ink, so a shifted control lands inside that band in one layout and outside it in the other.

**Dating.** Probing the same positions at increasing encoder depth (Table 4, left): at the encoder's input (patch embedding plus position, the only interface containing solely the token's own pixels), the probe sits at chance (0.230 / 0.226, a structural zero, since the key's pixels are identical within a pair in all 576 pairs). After one attention block it reads 0.919 / 0.947, and after two, 1.000. The bound letter is written into the key by encoder blocks 1–2, before the language model exists, carried by a displacement of about 2% in relative $L_2$ that a cosine audit cannot see.

### C.2 Installation Certificate and Overwrite

**Manipulation check.** Running the key patch and reading the patched representation with an out-of-sample probe in the same forward pass, the installed donor value decodes at 1.000 (base at 0.000) at the merger and through layer 5 (Table 4, right): the patch demonstrably installs the value.

Table 4: Left: decodability of the bound letter at the key across encoder depth (the write happens in blocks 1–2, and depth 0 is a structural zero because the key's pixels are identical within a pair). Right: the manipulation check. After the key patch, an out-of-sample probe reads the installed donor value at ceiling while behavior never moves.

| Encoder depth | Vert. | Horiz. | rel. $L_2$ (vert.) |
|---|---|---|---|
| 0 (patch embed + pos.) | 0.230 | 0.226 | 0.0000000 |
| 1 | 0.919 | 0.947 | 0.001094 |
| 2 | 1.000 | 1.000 | 0.004439 |
| 3–27 | 1.000 | 1.000 | 0.006–0.023 |

| Interface | patched → donor | patched → base |
|---|---|---|
| merger | 1.000 | 0.000 |
| embed / L0–L5 | 1.000 | 0.000 |
| L8 | 0.167 (v) / 0.964 (h) | 0.833 / 0.036 |
| L12 | 0.000 (v) / 0.490 (h) | 1.000 / 0.510 |
| L16–L35 | 0.000–0.036 | 0.958–1.000 |

**Behavior and overwrite.** In the same forward passes the patched answer is the base letter in $192/192$ pairs in both families (mean donor-minus-base logit margin $-26.96$ / $-26.83$), and by layers 8–16 the probe reads the *base* letter again from the same positions (crossover at L8 vertical, L12 horizontal). The installed value is not merely ignored but recomputed from context by mid-stack. The supported sentence is: the key's tokens carry the bound value at ceiling, the value can be installed by intervention and verified as installed, and the model overwrites it rather than reading it.

**Replication on Qwen2.5-VL-7B.** The same probe recipe (standardize, PCA-256, standardize, $C = 0.1$) and analyzer were carried over unchanged to the checkpoint whose emitted answer Sec. 6 controls, and the certified cut is the one that produced that checkpoint's null (the full matched-symbol region). Two structural differences were declared before execution. First, there is no DeepStack, so the visual interface is the merger alone and the interface axis is 30 rather than 38. Second, a $28.44$-px cell does not align to the layout, so the region holds 20 or 25 tokens depending on the matched row's sub-cell phase (mean pooling is unaffected, and the per-token axis is undefined and was not written). The extraction anchor is this checkpoint's own base/donor merger cosine, $0.9929$ / $0.9960$, inside the $[0.9869, 0.9981]$ / $[0.9903, 0.9986]$ bands measured on its own records. On $192$ held-out pairs per family, with the probe trained on $384$ disjoint pairs, the bound letter decodes at $1.000$ at the merger (chance $0.25$). After the patch the installed donor value decodes at $1.000$ and the base value at $0.000$ there. The answer is the base letter in $192/192$ pairs in both families (margin $-9.57$ / $-9.91$), with clean base and donor accuracy both $1.000$, and the readback crosses back to the base letter at layer $10$ and layer $5$ of $28$, reading base at ceiling from layer $13$ on. The behavioral null on these $192$ pairs, measured in the same forward passes, is $-0.10\%$ / $+0.01\%$ for the full-region cut and $-0.07\%$ / $-0.07\%$ for the equal-size cut, beside the $+0.05\%$ / $-0.04\%$ reported on the $64$ calibration pairs (Table 1). The two pair sets are kept separate. The overwrite is therefore not a property of Qwen3-VL's re-injection pathway, which this checkpoint does not have.

### C.3 An Install That Cannot Be Overwritten

A value installed once and recomputed by mid-stack leaves open that the model would read one that survived. Table 5 measures this inside the row-identity instrument of Sec. 6: the moved activations are re-imposed at *every* layer and the rest of the legend is held to a clean forward at every layer, so the recomputation cannot erase them. The key's row still transfers nothing and flips nothing on either checkpoint, while the value's row and the row-identity write move the same answers in the same forwards. Two representational swaps on Qwen3-VL-8B, added as diagnostics, show those tokens are read as content: exchanging what the two symbols *are* moves the answer $11\times$ and $6\times$ further than importing a counterfactual in which only their labels differ.

### C.4 Region Completeness: The Carrier Is the Legend Strip

The decodability map shows the bound label readable along the whole matched legend row, so "the label spilled past the nine tokens we patch" is a hypothesis with geometry behind it. A nested

Table 5: The patch when it cannot be overwritten (192 held-out pairs per family per model). Every row is measured inside one instrument: the moved activations are re-imposed at every layer and the rest of the legend is held to a clean forward at every layer. Read within a model: the two checkpoints' margin spans differ by 2.7× (Appendix B.1). The two representational swaps were not run on Qwen2.5-VL-7B. Recovery rows are in percent of the margin, and "pairs flipped" rows are percent of pairs.

| | Qwen3-VL-8B | | Qwen2.5-VL-7B | |
|---|---|---|---|---|
| Condition | Vertical | Horizontal | Vertical | Horizontal |
| donor activations at the key's row | 0.20 | 0.09 | 1.60 | 0.42 |
| 95% CI | [.11, .28] | [.03, .16] | [1.25, 1.98] | [.27, .58] |
| pairs flipped to donor | 0.0 | 0.0 | 0.0 | 0.0 |
| donor activations at the value's row | 8.44 | 6.27 | 36.77 | 7.55 |
| pairs flipped to donor | 0.5 | 0.0 | 25.0 | 0.0 |
| norm-matched random at the key | 0.10 | 0.15 | 1.24 | 1.65 |
| row identity written at the key (Table 15) | 2.98 | 1.84 | 34.09 | 54.04 |
| exchange the two symbols | 2.16 | 0.51 | – | – |
| 95% CI | [1.76, 2.61] | [.40, .63] | | |
| exchange the two labels | 1.43 | 1.20 | – | – |

ladder of regions, each a strict superset of the last, tests it on the unchanged manifests. The pixel audit established that the two swapped glyphs carry 100.00% of the base-versus-donor pixel difference in every pair, so the image partitions exactly into a pixel-changed side (12 tokens) and a pixel-identical side (564), and since patching all 576 recovers exactly 100% per sample, the ladder decomposes that ceiling rather than sampling it. Each of the four legend entries occupies exactly three merger lines, so removing the two-line letter stripe leaves every entry's row at exactly 66 tokens with no equalization padding. On the discovery split the pipeline reproduces the published key null (+0.08% [−0.02, +0.18] and +0.04% [−0.05, +0.12]) and the equal-size letter control (26.05% and 11.89%) to the last printed digit. Table 6 is the confirmation split, plotted in Fig. 3a.

Table 6: Nested region ladder, confirmation split ($n$=192 per family, Qwen3-VL-8B). Relative recovery, in %, at the visual interface (merger plus all three DeepStack taps). Every region listed is pixel-identical between base and donor, asserted per pair, except the last two rows. The four legend rows are exactly equal in size by construction. Regions may not be summed: patching the four rows together gives far more than their sum.

| Region | tokens | vertical (%) | horizontal (%) |
|---|---|---|---|
| the key (matched symbol) | 9 | +0.04 [−0.02, +0.10] | +0.02 [−0.04, +0.09] |
| matched entry's row | 66 | +16.04 [+14.83, +17.29] | +18.39 [+16.15, +20.79] |
| same, minus the queried item | 54–66 | +16.04 | +18.37 |
| distractor entry's row | 66 | +3.30 [+2.82, +3.78] | +5.80 [+4.69, +7.05] |
| untouched entry's row (1) | 66 | +0.03 [−0.04, +0.09] | −0.01 [−0.09, +0.07] |
| untouched entry's row (2) | 66 | +0.05 [−0.02, +0.12] | +0.04 [−0.04, +0.12] |
| all four legend rows | 264 | +58.60 [+54.75, +62.39] | +50.93 [+46.52, +55.28] |
| all non-legend positions | 300 | +0.50 [+0.26, +0.73] | +1.22 [+0.80, +1.69] |
| all pixel-identical tokens | 564 | +76.92 [+74.12, +79.57] | +61.04 [+56.62, +65.47] |
| the two swapped glyphs | 12 | +22.14 [+19.60, +24.78] | +35.40 [+31.23, +39.71] |
| full interface | 576 | +100.00 (exact) | +100.00 (exact) |

Three readings follow. The key null is not a coverage artifact: the interface is live at 12, 66, 264, 564 and 576 tokens, and the key's own nine are the one region on which it is not, inside the row that recovers most. The carrier is entry-specific: the primary contrast, matched row minus distractor row, is +12.74% [+11.43, +14.07] and +12.58% [+9.95, +15.27], and the two rows belonging to entries the counterfactual never touched recover nothing at CI half-widths of 0.08%. The carrier is the legend strip: 264 of 576 tokens hold 76–83% of what the entire pixel-identical side holds, while every chart item and every patch of canvas together hold 0.5–1.2%. One fact bounds how the table may be read: patching the four rows together yields 59% where they sum to 20%, so the model resists partial installs of the mapping far more than a linear account predicts. No mechanism

is claimed for that. Reverse damage tracks every non-null rung in both directions ($+16.74\%$ and $+21.53\%$ at the matched row, $+3.65\%$ and $+7.65\%$ at the distractor's) and neither direction at the key ($-0.02\%$, $-0.00\%$). The full interface is exact to $0$ on every pair, and pixel-difference shares are exactly $0$ on every pixel-identical region.

### C.5 The Rows Enter at the Merger, and the Key Is a Shape Key

Patching only the $32$ merger tokens of the two exchanged legend rows recovers $57.1\%$ $[50.9, 63.4]$ / $49.3\%$ $[42.0, 56.9]$ in the two families with reverse damage and six flat equal-size controls, while the individual DeepStack taps are near zero: the earliest sufficient single interface is the final vision merger. This is also where the task's scope was set. The first confirmation family missed the behavioral criterion (accuracy $0.7135$), and the per-prompt decomposition showed why: on the same images an abstract prompt ("use the legend") scores $0.4271$ while the operation-specified prompt ("find the identical symbol and read that row's letter") scores $1.0000$. Token-length-matched generic/explicit prompt pairs both score $1.0000$ with a paired gain of exactly zero, so the gap is not instruction length, and the claim is scoped to operation-specified retrieval, revalidated on a third independent family ($480$ new pairs, horizontal layout, new icons, fonts and backgrounds: $1{,}920/1{,}920$ correct).

The key is a shape key. A difficulty ladder that holds token geometry byte-identical while varying icon discriminability finds that removing the color shortcut entirely costs nothing (accuracy $1.0000$), near-identical shapes at full size cost under one point, and only spatial scale breaks the model ($22$-px icons: accuracy $0.9275$). Those failures are upstream of the mechanism and are not analyzed further.

### C.6 The Query Symbol Contributes but Does Not Suffice

The complementary pair holds the legend pixel-identical and changes only the target icon, so the model must re-run the match. Transplanting the query symbol's merger tokens with the legend fixed is directional, specific and merger-localized (significantly above every control including equal-size cross and shifted patches, with reverse damage) but moves the decision only $46.2\%$ $[39.9, 52.6]$ / $34.9\%$ $[28.7, 41.4]$ of the way, below the $50\%$ sufficiency criterion, and symbol-specific tracking is $0.391$ / $0.297$ against a $0.90$ criterion. Supplementary controls are mixed (the non-member null passes vertically and fails horizontally at $0.078$). The query's tokens carry a real contribution, but they are not a self-sufficient transplantable key, and no sufficiency claim is made.

### C.7 Read Maps on the Image

The bottom row of Fig. 1 draws three measurements over one image, at every token of its $24\times24$ grid. Figure 9 is the same row with its scales, and Fig. 10 the population version. All are on Qwen3-VL-8B. The probe map is the single-token probe of Appendix C.1 restricted to the pairs whose matched symbol sits at the same grid cell as the shown image's ($144$ pairs of $576$, with the same folds by pair, PCA, penalty and bootstrap), so every token is the same offset for every pair and all $576$ tokens are scored with the same $n$. On the shown image the label reads at $1.00$ along its legend row and down the whole letter stripe, including blank canvas above and below the legend panel, at $0.84$ on the other symbols, and at $0.51$ on the target. The other two maps are the shown image's own, from re-runs of existing measurements per token on the $64$ pairs of the value-locus set. For attention, the forward pass of the audit (Appendix D.3) is repeated and the readout-position attention of the eight committee heads is kept per visual token instead of being summed by region. The region sums of the saved maps reproduce the audit's stored values exactly (maximum absolute difference $0$ over $64$ pairs, eight heads and four regions). For the patch map, the key patch of Appendix C.4 (merger tokens and the three DeepStack taps, donor into base) is applied to every $3\times3$ window of the grid at stride one ($484$ windows per pair), and each token is assigned the mean recovery of the windows that cover it. The key's own region recovers $0.08\%$ in the same passes, and the most effective window $32.2\%$ on average. The maps add no claim: they are the per-token form of the region ladder of Fig. 3a and of Table 11.

The population version (Fig. 10) averages the same three measurements over pairs in the frame whose origin is each pair's matched symbol, and reports only offsets that are in bounds for every

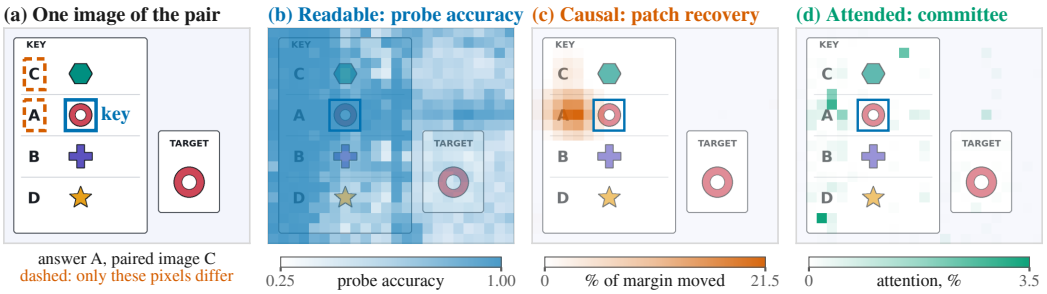


Figure 9: Read maps of one image with their scales (Qwen3-VL-8B, the maps of Fig. 1, bottom). **(a)** One image of a pair. **(b)** Accuracy of a linear probe for the key's label at each token, over the 144 pairs with this legend layout (chance 0.25). **(c)** Percent of the answer margin moved by a 3×3 patch from the paired image, this image. **(d)** Attention from the readout position, averaged over the eight committee heads, this image.

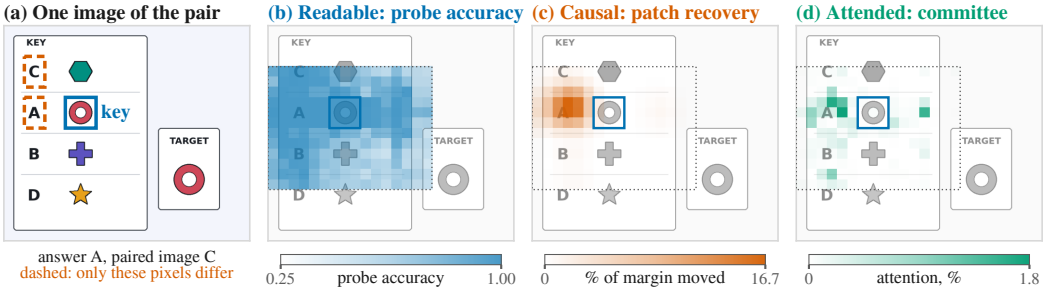


Figure 10: Read maps, population version (Qwen3-VL-8B, with means over pairs aligned on the matched symbol, drawn over one stimulus and shown where every pair is in bounds, dotted). **(a)** One image of a pair. **(b)** Accuracy of a linear probe for the key's label at each token offset (576 pairs, chance 0.25). **(c)** Percent of the answer margin moved by a 3×3 patch from the paired image (64 pairs). **(d)** Attention from the readout position, averaged over the eight committee heads (64 pairs).

pair, so that each cell has the same $n$: the offset map of Appendix C.1 (576 pairs) and the means of the 64 patch and attention maps.

Single pairs look like the mean for the patch and differ from it for attention (Fig. 11). The most effective window overlaps the target's own glyph in 64 of 64 pairs. The best window that touches neither swapped glyph averages 7.8%, inside the matched legend row of Fig. 3a. Attention exceeds an equal-size non-answer letter in every pair (Appendix D.3), but the single hottest token of the committee is on the glyph in only 6 of 64: the heads also place mass on a few blank cells at the foot of the legend's letter stripe and on one corner token that the other 568 heads share (on the shown image, the foot of the column takes 3.5% and the glyph's hottest token 3.0%). These sit at fixed image positions, so they wash out in the symbol-centered mean while the glyph, at a fixed offset from the symbol, does not.

**Horizontal layout.** The same maps in the horizontal layout (Fig. 12) put the glyph first again and add a second site: blank canvas five tokens below the key, under the legend panel, where the label is also decodable (0.89), a patch moves the answer, and the committee places its largest mean attention in the window (1.6%, against 1.1% at the glyph's hottest token). The key's own region recovers 0.04% and the most effective window 19.7% on average. That window overlaps the target's own glyph in 30 of 64 pairs, the other swapped glyph in 15 and neither in 19, and the committee's single hottest token is never on the glyph. The region contrasts of Table 11 are sums over these maps and are unchanged.

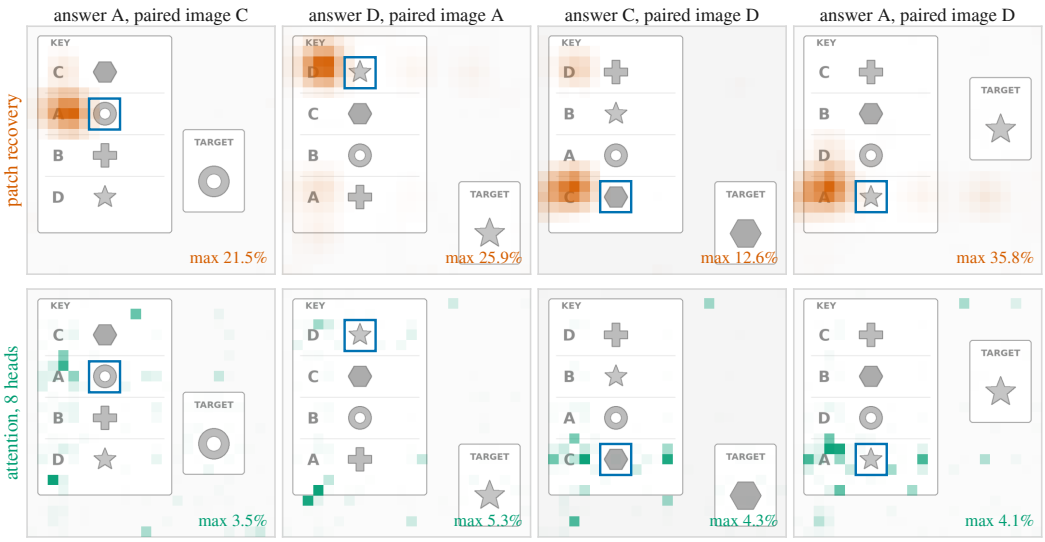


Figure 11: Single pairs behind Fig. 10c–d (Qwen3-VL-8B, the first four pairs of the evaluation set). Top: percent of the answer margin moved by each 3×3 patch. Bottom: attention of the eight committee heads from the readout position. The key is outlined in blue.

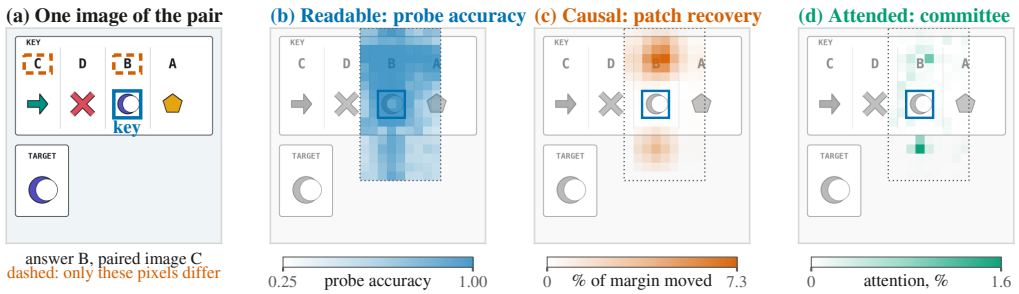


Figure 12: Read maps in the horizontal layout (Qwen3-VL-8B, with panels and conventions as in Fig. 10). The key's column position varies over pairs, so the offsets in bounds for every pair form a narrower window. The glyph above the key and blank canvas below it are both readable, causal and attended, while the key's own tokens are readable only.

**Real charts.** The same sliding patch was run on the first 24 pairs of the real-chart pilot pool (Appendix A.4, with 24×18 tokens and 352 windows per pair), all answered correctly in both images. The most effective window overlaps a swapped legend label in 24 of 24 pairs and the target's own label in 22. Patching the target's full label region recovers 77.8% of the margin on average, and its swatch 0.02% (Fig. 13). The full-region patch is larger than the equal-size control of Table 21, so the two recoveries are not the same quantity. Committee attention is not drawn for real charts: the committee was identified on the legend task, the real-chart prompt carries its options as text, and these heads place 7% of their mass on the image there.

### C.8 A Learned Subspace at the Key Transports Nothing

The null of Sec. 4 replaces every dimension of the key. A low-rank subspace could in principle carry the label while the remaining dimensions cancel it, so a null at full rank does not imply a null at lower rank. This section searches for such a subspace directly, on Qwen2.5-VL-7B and both layouts, with a distributed interchange intervention (Geiger et al., 2024). An orthonormal basis $B \in \mathbb{R}^{d\times r}$ ($d = 3584$) is trained so that $h'_p = h_p + BB^\top\left(h_p^{\text{donor}} - h_p\right)$, applied once at every token $p$ of the matched symbol's full region (20 or 25 tokens) in the residual stream entering language-




Figure 13: Patch maps on real charts (Qwen3-VL-8B, the first eight pairs of the real-chart pilot pool): percent of the answer margin moved by each 3×3 patch from the paired image, in which two legend labels are exchanged. The target's swatch is outlined in blue and the queried data point is circled.

model layer $\ell$, makes the base run answer with the donor's letter. The objective is the cross-entropy of the four-way answer softmax toward the donor letter (Adam, learning rate $10^{-3}$, batch 4, at most 20 epochs, early stopping on validation interchange accuracy and then validation loss, patience 4). The basis $B$ is fit on 320 pairs and selected on 64, and every number below is on the 192 held-out pairs of Table 1, which are never read during training. Ranks $r \in \{1, 4, 16, 64, 256, 1024\}$ (vertical) and $\{4, 64, 1024\}$ (horizontal) are searched at $\ell \in \{0, 6, 10\}$. The rank $r = d$, where $BB^\top = I$ and the intervention is the full patch, is evaluated untrained at $\ell \in \{0, 3, 6, 10, 14\}$ as an anchor. *Interchange accuracy* is the share of held-out pairs whose answer becomes the donor's letter. The clean base answer is the base letter on every pair, so it is a flip rate. The criteria were fixed in advance. A transferable subspace is found at a cell $(\ell, r)$ with $r < d$ if interchange accuracy reaches 50% and the recovery CI lower bound exceeds 15% while the label-shuffled control at that cell stays below 30%. The null is general within this class if no key cell reaches 30% interchange accuracy or a recovery CI upper bound above 15%, provided the same search at the value glyph reaches 50% at some $r \leq 64$.

**No key subspace moves the answer, while the same search at the value moves almost every one.** Table 7 gives every cell. At the key the largest interchange accuracy is 1.0% (two pairs of 192, at $\ell = 10$, $r = 64$, on both layouts) and the largest recovery CI upper bounds are 5.2% and 6.8%. The label-shuffled control, trained with the donor letters permuted across pairs so that the subspace can learn only a pair-independent bias, reaches the same levels (3.3% and 2.7% recovery at that cell, and on the horizontal layout a higher interchange accuracy, 2.1%, than the real arm), so the one to five percent the search buys carries no label. The same pipeline at the value glyph reaches interchange accuracy 100% and 99.0% at rank 64 with recovery 86.3% and 78.4%, above the full-rank patch (73.7% and 50.2%): the search finds a transferable label subspace where one exists, and finds a rank-64 one that beats replacing all 3,584 dimensions, because the full patch carries donor components that partly cancel. At the key it finds none.

**Decodability and use separate inside one basis.** The paper's probe recipe (standardization, PCA, standardization, multinomial logistic regression with $C = 0.1$), trained on the base and donor images of the 384 training pairs and tested on those of the held-out 192, reads the bound letter from the projection $B^\top h$ onto the learned key subspace at 1.000 (vertical) and 0.943 (horizontal) at rank 64, against chance 0.25 (Table 8). Installing the donor's component of that same subspace moves the answer on at most two pairs. The objection that the certificate holds only in the probe's basis is answered in the basis the search chose. The shuffled arm's basis reads the letter as well (1.000 and 0.930), so a high probe accuracy is generic to low-rank projections of this representation and is not evidence that the search recovered the probe's subspace.

The equal-size key subspace at $\ell = 0$, $r = 64$ reads 1.0000 (vertical), and the value subspace at the same cell 1.0000 and 0.9349.

Table 7: Learned-subspace interchange at the key on Qwen2.5-VL-7B, 192 held-out pairs per layout. Flips: interchange accuracy (%). Recovery (%) with pair-clustered 95% CIs. Dash: not run. Rows with $r = d = 3584$ are the untrained full patch, and rank 0 is the identity check.

| | | | Vertical | | Horizontal | |
|---|---|---|---|---|---|---|
| Site | $\ell$ | $r$ | Flips | Recovery [CI] | Flips | Recovery [CI] |
| key, identity | 0 | 0 | 0.0 | 0.00 [0.00, 0.00] | 0.0 | 0.00 [0.00, 0.00] |
| key, learned | 0 | 1 | 0.0 | 0.14 [0.05, 0.24] | – | – |
| key, learned | 0 | 4 | 0.0 | 0.42 [0.28, 0.56] | 0.0 | 0.27 [0.07, 0.50] |
| key, learned | 0 | 16 | 0.0 | 0.72 [0.52, 0.93] | – | – |
| key, learned | 0 | 64 | 0.0 | 1.07 [0.85, 1.31] | 0.0 | 0.80 [0.39, 1.30] |
| key, learned | 0 | 256 | 0.0 | 1.31 [1.05, 1.57] | – | – |
| key, learned | 0 | 1024 | 0.0 | 1.02 [0.76, 1.29] | 0.0 | 0.63 [0.45, 0.81] |
| key, full patch | 0 | $d$ | 0.0 | −0.10 [−0.22, 0.02] | 0.0 | 0.01 [−0.08, 0.09] |
| key, full patch | 3 | $d$ | 0.0 | −0.07 [−0.20, 0.05] | 0.0 | 0.02 [−0.07, 0.10] |
| key, learned | 6 | 4 | 0.0 | 0.52 [0.33, 0.72] | 0.0 | 1.61 [0.49, 2.98] |
| key, learned | 6 | 64 | 0.0 | 1.82 [1.42, 2.26] | 0.0 | 2.24 [1.14, 3.59] |
| key, learned | 6 | 1024 | 0.0 | 2.58 [2.13, 3.06] | 0.0 | 1.70 [1.23, 2.20] |
| key, full patch | 6 | $d$ | 0.0 | −0.04 [−0.16, 0.08] | 0.0 | 0.03 [−0.07, 0.12] |
| key, learned | 10 | 4 | 0.0 | 0.68 [0.51, 0.86] | 0.5 | 2.68 [1.23, 4.37] |
| key, learned | 10 | 64 | 1.0 | 3.59 [2.31, 5.16] | 1.0 | 4.66 [2.78, 6.79] |
| key, learned | 10 | 1024 | 0.0 | 3.38 [2.77, 4.04] | 0.0 | 2.00 [1.52, 2.53] |
| key, full patch | 10 | $d$ | 0.0 | −0.06 [−0.18, 0.05] | 0.0 | 0.01 [−0.07, 0.10] |
| key, full patch | 14 | $d$ | 0.0 | 0.79 [0.49, 1.11] | 0.0 | 0.52 [0.37, 0.67] |
| key, equal size, learned | 0 | 64 | 0.0 | 0.69 [0.50, 0.90] | – | – |
| key, equal size, full patch | 0 | $d$ | 0.0 | −0.07 [−0.17, 0.03] | 0.0 | −0.07 [−0.15, 0.02] |
| key, label-shuffled | 0 | 64 | 0.0 | 0.91 [0.62, 1.20] | 0.0 | 0.49 [0.24, 0.79] |
| key, label-shuffled | 6 | 1024 | 0.0 | 2.33 [1.85, 2.84] | 0.0 | 0.83 [0.59, 1.09] |
| key, label-shuffled | 10 | 64 | 1.0 | 3.32 [2.01, 4.87] | 2.1 | 2.74 [1.35, 4.51] |
| value, learned | 0 | 64 | 100.0 | 86.28 [84.56, 88.07] | 99.0 | 78.41 [76.77, 80.03] |
| value, full patch | 0 | $d$ | 94.8 | 73.69 [72.02, 75.34] | 50.5 | 50.24 [48.20, 52.26] |
| value, full patch | 3 | $d$ | 94.8 | 72.96 [71.29, 74.59] | 38.5 | 45.49 [43.41, 47.55] |
| value, full patch | 6 | $d$ | 79.2 | 64.28 [62.42, 66.11] | 7.8 | 30.81 [29.12, 32.57] |
| value, full patch | 10 | $d$ | 52.6 | 53.48 [51.24, 55.76] | 2.1 | 22.81 [21.38, 24.31] |
| value, full patch | 14 | $d$ | 38.0 | 45.36 [43.33, 47.41] | 0.5 | 19.65 [18.42, 21.00] |

Table 8: Accuracy of the paper's linear probe for the key's label on the projection $B^\top h$ of each learned key subspace (192 held-out pairs, both images, chance 0.25). On the full state $h$ the same probe reads 1.0000 in every cell. Dash: not run.

| | | Learned | | Label-shuffled | |
|---|---|---|---|---|---|
| $\ell$ | $r$ | Vertical | Horizontal | Vertical | Horizontal |
| 0 | 1 | 0.3203 | – | – | – |
| 0 | 4 | 0.6745 | 0.2865 | – | – |
| 0 | 16 | 0.9948 | – | – | – |
| 0 | 64 | 1.0000 | 0.9427 | 1.0000 | 0.9297 |
| 0 | 256 | 1.0000 | – | – | – |
| 0 | 1024 | 1.0000 | 1.0000 | – | – |
| 6 | 4 | 0.6667 | 0.3411 | – | – |
| 6 | 64 | 1.0000 | 0.9870 | – | – |
| 6 | 1024 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 10 | 4 | 0.5182 | 0.2917 | – | – |
| 10 | 64 | 1.0000 | 0.9479 | 1.0000 | 0.9427 |
| 10 | 1024 | 1.0000 | 1.0000 | – | – |

**Instrument.** The rank-0 intervention reproduces the clean logits bitwise. The untrained $r = d$ cell at $\ell = 0$ is the certified full patch of Sec. 4 on the same 192 pairs and reproduces its recovery to six decimals ($-0.001039$ and $+0.000087$, and for the equal-size key $-0.000735$ and $-0.000656$). The $r = d$ cells at the value glyph (73.7%, 50.2%) and at the key at $\ell = 14$ (0.79%, 0.52%) match the 64-pair calibration values (respectively 74.8% and 48.2%, and 0.91% and 0.39%) in sign and size on a separate denominator. Base and donor have bitwise-equal token ids and the same $27\times27$ grid, region masks are derived per pair, and the key's pixels are identical within every one of the 576 pairs while the value's differ. Two deviations were written before any result. Early stopping on validation interchange accuracy alone with patience 2 would have stopped the key arms, whose

validation accuracy is zero by expectation, at epoch 3 with an untrained $B$. The selection rule became interchange accuracy and then loss, with patience 4, which favors finding a subspace. The final evaluation runs at batch 1 so that the identity check is bitwise (at batch 4 the bf16 reduction order moved single logits by up to 0.5 while the margin moved 0.0003). Key arms train for 6–20 epochs with four-way training accuracy at zero throughout, while the value arm reaches 0.906 training accuracy in its first epoch.

**Boundary.** The search covers linear, fixed-rank, orthogonal projections of the base-to-donor difference applied once at the input of one layer, at five layers, on one checkpoint. Non-linear readouts and pair-varying subspaces are outside it. Not run: (i) Qwen3-VL-8B, (ii) ranks 1, 16 and 256 and the equal-size key on the horizontal layout, and (iii) low-rank cells at $\ell = 3$ and 14 (the full-rank sweep of Appendix D covers all 28 layers). One prediction failed: the horizontal layout was expected to be uniformly smaller in magnitude, and on the key side it is larger at $\ell = 6$, $r = 4$ and at every rank at $\ell = 10$.

# D The Answer Pathway: Additional Results

This appendix gives the full results behind Secs. 5 and 6: the layer sweep and the commitment-depth control, the readout committee and what it attends to, the digit legends that separate the value from the answer token, the row-identity estimator with its certification and controls, and the answer-control study on both Qwen checkpoints. Unless a paragraph says otherwise the model is Qwen3-VL-8B, on which the layer sweep, attention audit and digit control were run. Qwen2.5-VL-7B's numbers are given beside them wherever the arm was ported.

## D.1 Layer-Wise Patching

Residual-stream patching at every language-model layer over the fixed position sets decides between an addressable binding slot at the key and readout-side assembly. Table 9 gives the crossings plotted in Fig. 3b. The key's positions never acquire the label at any depth (worst case roughly $28\times$ inside the bound). The readout position carries nothing through layer 20 and crosses 50% at layer 24. Visual positions stop being sufficient at layer 19, exactly as the readout becomes decisive: the handoff window. On Qwen2.5-VL-7B the same sweep gives no binding slot at any of 28 layers (worst-case CI upper 2.28% / 0.48%) and a readout crossing at layer 24 of 28. The equal-size key patch is $+0.05\%$ / $-0.04\%$ and the full-region patch (20–25 tokens) is likewise null, against a value control of 74.84% / 48.16% (Table 1). Two build deviations forced by that checkpoint's unaligned 28.4-px grid were declared before any recovery was read, and both tightened the controls: minimum-count equalization after upward padding, which had turned an empty control positive, and an adaptive shifted-control displacement.

Table 9: Layer sweep on Qwen3-VL-8B (36 layers, 64 pairs per family, recoveries in %).

| Quantity | Vertical | Horizontal |
|---|---|---|
| key, peak recovery | 0.33 (L1) | 0.21 (L2) |
| key, worst-case CI upper over all layers | 0.53 | 0.35 |
| readout, first layer $\geq 50\%$ | 24 | 24 |
| readout, peak | 98.01 (L34) | 97.86 (L34) |
| all visual positions, first layer $< 50\%$ | 19 | 19 |
| value glyph, peak | 23.37 (L3) | 11.86 (L3) |
| equal-size controls, every layer | $\leq 3.2$ | $\leq 0.5$ |

## D.2 The Commitment Depth Is Generic

Four measurements converge on layer 24 of Qwen3-VL-8B: the readout crossing on letters and on digits, the two strongest committee heads, and the source-side editing deadline of the text counterfactual below. Two non-legend four-option controls, one visual without a legend and one text-only, cross at exactly the same layer ($\Delta = 0$, step-shaped, $< 12\%$ at L23 and $> 74\%$ at L24), so the landmark is the model's generic commitment depth rather than a signature of retrieval. Two other

third-checkpoint candidates failed capability (Appendix G) before gemma-3 broke the absolute-versus-relative tie: its crossing is layer 29 of 48, neither absolute 24 nor the Qwen relative depth, while the structure (early visual sufficiency, a five- to six-layer handoff, one-step readout sufficiency at its end) replicates on all three lineages. The claim is the ordering, never the index.

| Control run (Qwen3-VL-8B) | $L^*$ | $\Delta$ vs. 24 | mean @ L23 | mean @ L24 |
|---|---|---|---|---|
| visual no-legend, vertical | 24 | 0 | 6.92 | 79.42 [78.50, 80.32] |
| visual no-legend, horizontal | 24 | 0 | 11.71 | 77.27 [75.67, 78.87] |
| text-only, vertical | 24 | 0 | 7.48 | 74.63 [73.41, 75.82] |
| text-only, horizontal | 24 | 0 | 8.17 | 75.36 [74.14, 76.59] |

### D.3 The Readout Committee and What It Attends To

Per-head patching at the readout position (Qwen3-VL-8B: layers 12–29 $\times$ 32 heads $=$ 576 conditions, with the top-8 ranked on the vertical family only and scored on the horizontal) finds no single head above a 20% criterion (the strongest, `L24H29`, reaches 17.3%/17.5%), while the joint top-1/2/4/8 on the held-out family recover 17.6%/37.9%/60.8%/80.9% [79.4, 82.2]. The sum of all individually positive head recoveries (87.9%) is close to the top-8 joint value, so contributions are near-additive: 8 of 576 heads carry 81% and the median head carries 0.05%. Table 10 gives the same discovery on all three checkpoints, each from scratch, since head indices do not port.

Table 10: Head-committee discovery and confirmation on three checkpoints. Selection is always on one visual family, and the joint recovery (in percent of the margin) is scored on the family that played no part in selection. Slashed cells in the single-head row are discovery then confirmation, i.e. vertical then horizontal. No single head reaches the 20% criterion in both families on any checkpoint.

| | Qwen3-VL-8B | Qwen2.5-VL-7B | gemma-3-12b-it |
|---|---|---|---|
| LM layers $\times$ heads | 36 $\times$ 32 | 28 $\times$ 28 | 48 $\times$ 16 |
| Swept conditions | 576 | 448 | 288 |
| Strongest single head | 17.3/17.5 (L24H29) | 15.18/12.35 (L24H21) | 10.67/11.03 (L29H1) |
| Joint top-1 | 17.6 | 12.35 | 11.03 |
| Joint top-2 | 37.9 | 22.10 | 47.70 |
| Joint top-4 | 60.8 | 47.51 | 66.98 |
| **Joint top-8 (out of family)** | **80.9** [79.4, 82.2] | **76.06** [75.11, 76.99] | **77.02** [74.78, 79.24] |
| Committee layers | 23–29 | 22–26 | 26–32 |
| Readout crossing | L24/36 | L24/28 | L29/48 |

Head sets on Qwen3-VL: `L24H29`, `L24H31`, `L29H5`, `L28H0`, `L26H26`, `L29H11`, `L29H6`, `L23H4`. On Qwen2.5-VL: `L24H21`, `L26H24`, `L23H11`, `L26H26`, `L24H26`, `L24H27`, `L26H25`, `L22H1`. On gemma-3: `L29H1`, `L28H10`, `L29H11`, `L32H11`, `L26H11`, `L28H6`, `L32H7`, `L29H3`. Indices are not part of any claim, but the committee shape is.

**Attention identifies the route, and patching ranks the movers.** From the readout position, committee attention mass on the value glyph exceeds an equal-size, equally salient non-answer letter in the same legend (0.0316 vs. 0.0041 vertical, 0.0191 vs. 0.0020 horizontal), exceeds the 568-head non-committee population, and holds in 64/64 and 61/64 pairs, while attention on the key sits at population level: there is no key-lookup route to see (Table 11). Within the committee, attention mass and causal recovery are uncorrelated (Spearman $\rho = -0.143$, $p = 0.74$), and the head attending most to the value glyph is the weakest causal contributor. On this task attention is a worked counterexample to attention-based attribution: the right instrument for finding where to look, and no ranking of what moves the answer.

### D.4 Digit Legends and the Text Counterfactual

In the letter stimuli the retrieved value *is* the emitted answer token, so the value results could in principle be answer-token bookkeeping. With digit legends and randomized text option lines (Appendix A.2) all behavioral gates pass on both Qwen checkpoints and both decisive results replicate at unchanged criteria, with the digit-glyph positive control *stronger* than the letter control (Table 12).

Table 11: Committee-mean attention mass from the readout position (Qwen3-VL-8B). Letters task: the committee prefers the value glyph over an equal-size non-answer letter. Digits task: it prefers the matched digit in the image and the matching digit in the option line, and is statistically indifferent among option *letters*.

| Letters task | Vertical | Horizontal |
|---|---|---|
| value glyph | 0.0316 | 0.0191 |
| equal-size non-answer letter | 0.0041 | 0.0020 |
| the key | 0.0017 | 0.0012 |
| population (568 heads) on the value glyph | 0.0019 | 0.0012 |
| value glyph − other letter | $+0.0275\,[+0.0230, +0.0320]$ | $+0.0172\,[+0.0140, +0.0206]$ |
| **Digits task** | **Vertical** | **Horizontal** |
| image: matched digit glyph | 0.00123 | 0.00061 |
| image: other digit glyph | 0.00019 | 0.00014 |
| image: the key | 0.00024 | 0.00015 |
| text: option-line answer letter | 0.07586 | 0.07179 |
| text: option-line answer digit | 0.11213 | 0.11581 |
| text: other option letters (per token) | 0.07835 | 0.07884 |
| text: other option digits (per token) | 0.00665 | 0.00692 |
| answer letter − digit | $-0.0363\,[-0.0489, -0.0238]$ | $-0.0440\,[-0.0574, -0.0309]$ |

The letter-task committee, doubly out of sample (new value type, both families unseen), carries $86.5\%$ / $89.4\%$, and an independent re-sweep rediscovers 5–7 of its $8$ heads with the identical top three. Attention decomposes by content: the committee prefers the matched digit glyph in the image over its equal-size twin, prefers the correct digit among the option digits in text $16\times$, and is statistically indifferent among option *letters* (Table 11). One committee routes by the value wherever it lives, never by the key, and never by the answer token as such.

Table 12: Digit legends (answer letters only in text): both decisive results replicate on both Qwen checkpoints at unchanged criteria. The digit-glyph positive control is stronger than the letter-task control in three of the four cells (Qwen3-VL horizontal $18.16\%$ against $11.89\%$, and both Qwen2.5-VL cells by a wide margin) and marginally weaker in the fourth (Qwen3-VL vertical $24.47\%$ against $26.05\%$). Recovery cells are in percent of the margin, and dashes mark arms not run on that checkpoint.

| | Qwen3-VL-8B (V / H) | Qwen2.5-VL-7B (V / H) |
|---|---|---|
| key (equal size) | $+0.10$ / $-0.10$ | $-0.20$ / $+0.08$ |
| key (full region) | – | $-0.24$ / $-0.11$ |
| digit-glyph positive control | $+24.47$ / $+18.16$ | $+98.79$ / $+82.12$ |
| worst key CI upper over all layers | 0.59 / 0.16 | 0.85 / 1.56 |
| readout crossing | L24/36 both | L24/28 both |
| letter-task committee, joint top-8 | 86.5 / 89.4 | – |

**The source-side deadline.** A pure text counterfactual (image byte-identical, two option-line digits transposed) reaches the commitment depth from the source side: the changed digit tokens are fully sufficient through layer $\sim$15, decay across 16–23, and collapse at layer $24$. The non-legend text-only control shows the same collapse shape at the same layer. The option-line value token binds its own option letter from the earliest layers, which is why the committee never needed to look at letters.

| Option-digit sufficiency (%) | L15 | L16 | L17 | L18 | L19 | L20 | L21 | L22 | L23 | L24 |
|---|---|---|---|---|---|---|---|---|---|---|
| vertical | 99 | 95 | 95 | 94 | 79 | 69 | 64 | 57 | 37 | **1** |
| horizontal | 100 | 94 | 93 | 91 | 74 | 67 | 63 | 59 | 39 | **0** |
| text-only MCQ control | 97.7 (L12) → 58.9 (L16) → 21.0 (L19) → 10.0 (L23) → **0.7** (L24) | | | | | | | | | |

### D.5 The Row Identity: Estimator, Certification, and Controls

The key null has a principled blind spot: a binding code shared by base and donor moves with neither patch, so position-basis patching cannot see it (Feng & Steinhardt, 2024). A binding-ID estimator answers it.

**Estimator and certification.** Row-identity vectors $b(k)$ are estimated against cyclically shifted contrast contexts ($500$ samples $\times$ $4$ rotations, rendered by the same renderer as the stimuli and sharing no file with any evaluation manifest), so that content cancels exactly and $\Delta = b(j) - b(k)$ differs at the key's $3\times3$ lattice by a row identity and nothing else. It is injected at every layer with the legend context held to a clean forward, the published construction (Feng & Steinhardt, 2024). The reimplementation reproduces every cell of the published six-condition table on LLaMA-13b, Qwen3-8B and Qwen3-VL-8B's language model (control $1.0000$, attribute and entity interventions $0.0000$, both-restored $1.0000$, randoms $1.0000$), and the one remaining fidelity gap, RoPE position promotion, once implemented changes exactly the one cell it should ($0.1100 \rightarrow 0.0650$ on LLaMA-13b's attribute arm, all other cells unchanged to four decimals). A simplified estimator (one slot instead of two, one layer at a time, mean-based deltas and, decisively, no context freezing) fails its own positive control ($6.09\%$ against a $20\%$ criterion) and is not used. Under context freezing the reference operation itself drops from $26.05\%$ to $8.11\%$, so criteria for this instrument are set as fractions of the same-run reference operation.

**Confirmation.** On $192$ unseen vertical pairs and $192$ pairs of an untouched family (Table 13), the constructed $\Delta$ at the key's tokens recovers $2.98\%$ $[2.49, 3.51]$ / $1.84\%$ $[1.58, 2.13]$, against the donor patch at the same tokens ($0.20\%$ / $0.09\%$, the key null reproduced inside the harness) and norm-matched random ($0.10\%$ / $0.15\%$). The paired difference is $+2.78\%$ / $+1.75\%$ with CIs far from zero. Position-basis patching missed a causally real row identity, and on this checkpoint the effect is modest (roughly a third of the reference operation, and it never flips a decision).

Table 13: The row-identity confirmation (Qwen3-VL-8B, 192 held-out pairs per family, instrument fixed at its calibration configuration). The co-primary is the paired difference between the constructed row-identity injection and the donor patch at the same tokens. Entries are relative recovery in percent of the margin, and the last row is a ratio.

| Condition | Vertical (held out) | Horizontal (new family) |
|---|---|---|
| constructed $\Delta$ at the key | $2.98$ $[2.49, 3.51]$ | $1.84$ $[1.58, 2.13]$ |
| donor patch at the key | $0.20$ $[0.11, 0.28]$ | $0.09$ $[0.03, 0.16]$ |
| **paired difference** | $\mathbf{+2.78}$ $[2.30, 3.31]$ | $\mathbf{+1.75}$ $[1.49, 2.04]$ |
| norm-matched random at the key | $0.10$ | $0.15$ |
| constructed $\Delta$ at the value | $3.67$ | $10.68$ |
| donor patch at the value (reference op.) | $8.44$ | $6.27$ |
| swap symbols / swap letters | $2.16$ / $1.43$ | $0.51$ / $1.20$ |
| key $\Delta$ / reference | $0.353$ | $0.294$ |

**Characterization.** A slot-permutation control (own-family $\Delta$, wrong slots) separates identity from perturbation: correct slots beat permuted slots $2.2\times$ / $3.3\times$ ($2.98\%$ vs. $1.36\%$ vertical, $1.84\%$ vs. $0.56\%$ horizontal), with the permuted vector verified anti-correlated with the real one (mean cosine $-0.268$ / $-0.202$, never exceeding $+0.22$), so the control is not the same intervention in disguise. Cross-family transport withdraws the word "abstract": transported deltas beat random $4$–$5\times$ but no longer distinguish which slot they name (paired transport-minus-permuted $+0.22\%$ / $+0.02\%$, CIs touching or crossing zero), so the slot-specific component is layout-bound (Table 14). The representational decomposition agrees (cross-family matched-slot cosine $0.136$, of which $0.094$ is slot-generic) and localizes the slot-specific part to a depth peak at layer 23. A causal window sweep confirms the depth out of sample: the slot-specific effect lives at layers 15–20/23 and is exactly zero from layer $24$ on, while the letter reference operation peaks at 21–26: retrieval completes before commitment. (A window-restricted ratio to the reference operation is not a valid criterion: the window handicaps the reference $6\times$ and the primary $1.7\times$, so the ratio rises while the absolute effect falls from $1.84\%$ to $1.09\%$.)

Depth profile of the slot-specific component (paired correct $-$ permuted, 6-layer windows, discovery family): nothing through L11, $+0.28\%$ at 12–17, peak $+0.82\%$ $[+0.33, +1.32]$ at 15–20, $+0.65\%$ at 18–23, $+0.14\%$ at 21–26, and zero (CIs containing 0) for every window from $24$ on. The window

Table 14: The identity decomposition (Qwen3-VL-8B): slot-specific within a layout, slot-generic across layouts. Entries are relative recovery in percent of the margin.

| Condition | Vertical | Horizontal |
|---|---|---|
| own family, correct slots | 2.98 | 1.84 |
| own family, wrong slots (slot-permuted) | 1.36 | 0.56 |
| **paired correct − permuted** | $+\mathbf{1.61}$ $[1.13, 2.13]$ | $+\mathbf{1.28}$ $[1.03, 1.58]$ |
| other family, correct slots (transport) | 0.92 | 0.80 |
| other family, wrong slots | 0.70 | 0.78 |
| paired transport − permuted | $+0.22$ $[0.01, 0.44]$ | $+0.02$ $[-0.19, 0.23]$ |
| norm-matched random (transport norms) | 0.19 | 0.15 |

carried unchanged to the confirmation family gives $+0.73\%$ $[+0.58, +0.91]$. The letter reference operation peaks at 21–26 ($+3.89\%$): the two operations are separated in depth.

**Second checkpoint.** On Qwen2.5-VL-7B (instrument certified in-checkpoint: all six text cells exact) the same construction recovers $34.09\%$ / $54.04\%$ against the donor patch $1.60\%$ / $0.42\%$, random $1.24\%$ / $1.65\%$ and permuted $12.41\%$ / $14.38\%$, and already flips the emitted answer to the predicted row's letter on $24.0\%$ / $59.9\%$ of held-out pairs in this pilot form, while permuted flips land on the predicted letter almost never ($1.0\%$ of $6.2\%$ vertical, $0.5\%$ of $14.6\%$ horizontal) and random flips nothing (Table 15). The 11–29$\times$ cross-model magnitude gap is read with this checkpoint's $2.7\times$ smaller margin span in mind (Appendix B.1).

Table 15: The second checkpoint (Qwen2.5-VL-7B, 192 held-out pairs per family): the row identity replicates, with emitted-answer flips already visible in this pilot form (the answer-control study is Table 16). Magnitudes are not comparable with the Qwen3-VL tables above: this checkpoint's margin span is about a third as large. Recovery rows are in percent of the margin, and flip rows are percent of pairs.

| Condition | Vertical | Horizontal |
|---|---|---|
| constructed $\Delta$ at the key | 34.09 [30.67, 37.68] | 54.04 [50.80, 57.27] |
| slot-permuted | 12.41 | 14.38 |
| paired correct − permuted | $+21.68$ [18.22, 25.31] | $+39.66$ [36.46, 42.81] |
| donor patch at the key | 1.60 | 0.42 |
| norm-matched random | 1.24 | 1.65 |
| flip rate / lands on predicted letter | 24.0 / 24.0 | 60.4 / 59.9 |
| permuted: flip / predicted | 6.2 / 1.0 | 14.6 / 0.5 |
| random: flip / predicted | 0.0 / 0.0 | 0.0 / 0.0 |

### D.6 Answer Control on Two Checkpoints

The primary estimand is the percent of held-out pairs whose full-vocabulary greedy first token *changes* relative to the no-injection baseline *and* becomes the letter predicted by the injected row-identity difference (a baseline that already emits the donor letter does not count). Scale $\in \{0.5, 1, 1.5, 2, 3\}$ and layer windows (all layers plus all 6-layer windows, stride 3) were selected on the 64-pair calibration split only. Held-out evaluation uses 192 zero-overlap pairs per family, and every cell must also pass clean parse, accuracy, pair-consistency, zero-delta drift ($\leq 10^{-3}$) and delta non-collinearity checks (max positive cosine between the slot-permuted and correct deltas $< 0.50$). Table 16 is the result.

All four cells pass every check, so the Qwen3-VL cells are valid results below the criterion, not failed instruments. On Qwen3-VL all 60 calibration configurations already produce zero predicted flips, and the reported configuration is the tie-break fixed in advance (smallest scale, widest window).

**Scale does not explain the difference.** Across $\alpha \in \{0.5, 1, 1.5, 2, 3\}$ at all layers, Qwen3-VL's relative recovery is $1.1 \to 3.0 \to 4.6 \to 5.1 \to 5.4\%$ (vertical) and $0.9 \to 2.1 \to 3.2 \to 3.9 \to 3.8\%$ (horizontal): a saturating curve whose last doubling of $\alpha$ buys $0.3\%$ and $-0.1\%$, with no predicted flip anywhere in it. Qwen2.5-VL's clean margin span is roughly $0.37$ of Qwen3-VL's, so span-matching its selected $1.5$ would ask for $\alpha \approx 4$, past the end of a curve that has stopped rising,

Table 16: Predicted-letter flip rates, in percent of 192 held-out pairs per cell. Criteria fixed before execution: Qwen2.5-VL $\geq$20%/$\geq$50% (vertical/horizontal), Qwen3-VL $\geq$5%. Every cell also requires CI lower bound $> 0$, a $3\times$ margin over the larger control, and positive paired differences over each control.

| Model / family | Config | Predicted flip | Slot-perm. | Random | Reverse |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B vertical | $1.5\times$ all | 68.23 [61.46, 74.48] | 7.81 | 0.00 | 70.31 |
| Qwen2.5-VL-7B horizontal | $1.5\times$ all | 87.50 [82.81, 92.19] | 7.29 | 0.00 | 87.50 |
| Qwen3-VL-8B vertical | $0.5\times$ all | 0.00 [0.00, 0.00] | 0.00 | 0.00 | 0.00 |
| Qwen3-VL-8B horizontal | $0.5\times$ all | 0.00 [0.00, 0.00] | 0.00 | 0.00 | 0.00 |

and the instrument's own valid range ends first: at $\alpha = 3$ Qwen2.5-VL is still flipping answers (50.0% / 60.9%) but its slot-permuted control flips too (18.8% / 25.0%), a specificity failure rather than a stronger effect. There is no larger injection at which the two models could be compared.

**Neither does the re-injection pathway.** Qwen3-VL's DeepStack taps read vision-encoder blocks 8, 16 and 24 and are added to the visual positions after language-model layers 0, 1 and 2, so they cannot compete with the injection where it lands: the instrument restores the frozen legend positions to their clean values plus $\Delta$ at every layer's input, a tap firing after layer $\ell$ is overwritten at layer $\ell + 1$, and the perturbation entering each layer at the injected tokens is exactly $\Delta$ (the zero-$\Delta$ cell reproduces the clean logits bitwise on Qwen3-VL). What remains is the regime: the taps hold the rest of the image near the encoder's own description while the perturbation propagates. Suppressing them in the clean cache, the $\Delta$ estimation and the injection alike, and re-running the entire grid in that regime with the same stimuli, manifests, 500 estimation contexts, scales, windows, selection rule, seeds and criteria, costs the model almost nothing: on the 64 calibration pairs per family it still answers 1.000 (vertical) and 0.9844 (horizontal), with the one failing item (`car-v-h-175`, D read as A) the same item and the same error as with the taps in place. On the held-out sets, vertical stays at 1.000/1.000 and horizontal falls from 1.000/1.000 to 0.9896/0.9896, three pairs of 192 (`-510` on both sides, `-368` on base, `-168` on donor), inside the capability gate. The re-estimated $\Delta$ passes the same non-collinearity check (max positive cosine 0.2231/0.2032). All 120 calibration cells are capability-valid, and every one has a predicted flip rate of 0.0000 and an any-flip rate of 0.0000. The 192-pair held-out sets confirm, with relative recovery 1.21% [1.02, 1.39] and 0.93% [0.77, 1.10]. The injection is not weaker in this regime but stronger: recovery runs $1.37 \to 4.42 \to 6.50 \to 6.41 \to 5.58\%$ (vertical) and $0.88 \to 2.70 \to 5.00 \to 5.41 \to 4.73\%$ (horizontal) against $1.1 \to \cdots \to 5.4\%$ and $0.9 \to \cdots \to 3.8\%$ with the taps on. Removing the pathway lets the perturbation travel further and leaves the answer exactly where it was. Both candidate explanations of the model difference are excluded. The difference stands, unexplained.

### D.7 Position Controls and Lighter Injections

The row identity is estimated by cyclic shifts of content across legend slots, so it could in principle be nothing more than the absolute position of the slot's tokens. The injection that uses it also re-imposes $\Delta$ at every layer with the legend context held to a clean forward. Three controls test both, on Qwen2.5-VL-7B in both layouts, with the answer-control instrument of Appendix D.6, its 192 held-out pairs per layout, its configuration (scale 1.5, all layers) unless a row says otherwise, and its scoring (Table 17).

**A position difference from blank cells does not carry the identity.** The same estimator, run on a $3\times3$ lattice of blank cells in each slot's row, gives a $\Delta_{\text{blank}}$ whose injection at the key flips 11.5% and 0.0% of answers against the calibrated $\Delta$'s 68.2% and 87.5%. The lattice lies inside the legend panel in the vertical layout (columns 9–11 of the 32-px grid) and on pure canvas below the figure in the same columns in the horizontal layout (rows 21–23), the one blank placement each layout allows (a token-level ink audit over 2,384 stimuli finds no ink under any lattice token, and six other candidates were rejected by their ink counts). Matching its norm per layer to $\Delta$'s changes nothing (12.5% and 1.6%), so the gap is not one of magnitude, and the mirror agrees: the calibrated $\Delta$ written at the blank lattice flips 2.6% and 0.0%, at its own permuted control. The two vectors share a minority of their direction (per-layer cosine $0.372 \to 0.178$ with mean 0.336 in the vertical layout, and $0.344 \to 0.081$ with mean 0.233 in the horizontal), and what that share carries is the vertical

11.5%: blank cells inside the panel, in the same row, hold a small, real part of the identity, and pure canvas holds none.

**The identity survives translating the legend.** The same held-out pairs were re-rendered with the whole legend panel moved by one layout cell right and two down (vertical) or one cell diagonally (horizontal), and by four cells down and one cell the other way, respectively. The move is a pixel-exact panel copy audited on every image, and the model answers all four translated sets at 1.0000. The *untranslated* $\Delta$ flips 81.2% and 89.6% of answers on the moved legends, 60.4% and 93.2% at the second offsets, with reverse injection and controls as before. A $\Delta'$ re-estimated on translated contexts flips 81.8% and 91.7% and shares about 70% of its direction with $\Delta$ (global cosine 0.703 and 0.685, per-layer cosine 0.549–0.796 and 0.535–0.790, norm ratios 0.950 and 0.983), the remainder traceable to the sub-token phase that a 32-px shift necessarily changes on this checkpoint's 28.44-px grid. The identity is a code for the legend slot, not for the slot's place in the image. Its slot-specific part is still layout-bound across layouts (Appendix D.5).

Table 17: Row identity against position (Qwen2.5-VL-7B, 192 held-out pairs per layout). Predicted-letter flips in percent of pairs with pair-clustered 95% CIs. The norm-matched random control is 0.0 in every row. Offsets in pixels of the 768-px canvas (32 px is one layout cell) are $(+32, +64)$ vertical and $(+32, +32)$ horizontal for the first, and $(0, +128)$ and $(-32, -32)$ for the second.

| | Vertical | | | Horizontal | | |
|---|---|---|---|---|---|---|
| Injection at the key | flip | perm. | reverse | flip | perm. | reverse |
| $\Delta$ (Table 16) | 68.2 [61.5, 74.5] | 7.8 | 70.3 | 87.5 [82.8, 92.2] | 7.3 | 87.5 |
| $\Delta_{\text{blank}}$ from the same rows | 11.5 [7.3, 16.1] | 1.6 | 9.9 | 0.0 [0.0, 0.0] | 0.0 | 0.5 |
| norm-matched to $\Delta$ | 12.5 [7.8, 17.2] | 1.6 | 10.9 | 1.6 [0.0, 3.6] | 2.6 | 2.6 |
| $\Delta$ written at the blank lattice | 2.6 [0.5, 5.2] | 2.1 | 3.1 | 0.0 [0.0, 0.0] | 0.0 | 0.5 |
| $\Delta$, legend moved, first offset | 81.2 [75.5, 86.5] | 7.8 | 77.6 | 89.6 [84.9, 93.8] | 9.4 | 87.5 |
| $\Delta$, legend moved, second offset | 60.4 [53.6, 67.2] | 7.3 | 62.5 | 93.2 [89.6, 96.4] | 7.3 | 94.3 |
| $\Delta'$ re-estimated, first offset | 81.8 [76.6, 87.0] | 8.3 | 80.7 | 91.7 [87.5, 95.3] | 9.4 | 92.2 |

**A single window suffices, and an injection without the context freeze does not.** With the calibrated $\Delta$ and the best six-layer window of the calibration grid (layers 12–17, not re-selected), the injection still flips 32.3% and 78.6% of answers with slot-permuted controls at 0.0% and 2.1%, and every changed answer lands on the predicted letter (Table 18). Without the context freeze the injection accumulates layer by layer: over all layers it changes 46.9% and 69.8% of answers but lands on the predicted letter no more often than its permuted control (21.9% against 20.8% vertical, 20.3% against 15.6% horizontal), and restricted to the window it recovers a specific effect in the horizontal layout (29.2% against 3.1%) and a borderline one in the vertical (18.2% against 11.5%). The zero-$\Delta$ cell reproduces the clean logits exactly in every unfrozen configuration.

Table 18: Lighter injections of the calibrated $\Delta$ (Qwen2.5-VL-7B, 192 held-out pairs per layout, percent of pairs, pair-clustered 95% CIs). "Any" counts every changed first token, "flip" only those that land on the predicted letter.

| | | Vertical | | | Horizontal | | |
|---|---|---|---|---|---|---|---|
| Layers | Context frozen | flip | perm. | any | flip | perm. | any |
| all 28 | yes (Table 16) | 68.2 [61.5, 74.5] | 7.8 | 68.2 | 87.5 [82.8, 92.2] | 7.3 | 87.5 |
| 12–17 | yes | 32.3 [25.5, 38.5] | 0.0 | 32.3 | 78.6 [72.9, 84.4] | 2.1 | 78.6 |
| all 28 | no | 21.9 [16.1, 28.1] | 20.8 | 46.9 | 20.3 [15.1, 26.0] | 15.6 | 69.8 |
| 12–17 | no | 18.2 [13.0, 24.0] | 11.5 | 35.9 | 29.2 [22.9, 35.4] | 3.1 | 32.8 |

Not run: (i) $\Delta'$ at the second offset (the first already separates the mechanism from the direction), (ii) a rescaled legend, whose lattice would have to be defined on a rescaled pixel grid, (iii) a second blank placement per layout, which the target card and the letter and symbol rows leave no room for, and (iv) Qwen3-VL-8B, which has no emitted-answer flips to control for. The norm-matched blank arm was added after the blank arm had been scored, to rule out a magnitude explanation. No criterion or scoring rule changed.

# E GENERALITY: ADDITIONAL RESULTS

## E.1 BEHAVIORAL SCREEN ACROSS CHECKPOINTS

Table 19 gives the four-cell behavioral screen on every checkpoint the program touched. It passes three checkpoints at or near ceiling (Qwen2.5-VL-7B answers all 9,216 questions correctly, against a written prediction that it would fail) and rejects four: the two smaller Qwen3-VL scales fail monotonically in scale, Idefics3-8B-Llama3 parses perfectly (1.0000 on all 9,216 items) but misses the accuracy criterion on every cell, and Ovis1.6-Gemma2-9B emits chain-of-thought and never reaches a letter within eight tokens. Nothing about the rejected models' mechanisms is concluded.

Table 19: The four-cell behavioral screen across checkpoints (accuracy per cell, where a pass requires accuracy $\geq 0.98$ with CI lower $> 0.95$, pair consistency $\geq 0.95$, flip $\geq 0.98$, parse $= 1.00$, all strata $\geq 0.95$). Query cells change only the target icon, and value cells change only the legend's label pairing.

| Model | query, vert. | query, horiz. | value, vert. | value, horiz. |
|---|---|---|---|---|
| Qwen3-VL-8B | pass 1.0000 | pass 1.0000 | pass 1.0000 | pass 0.9935 |
| Qwen2.5-VL-7B | pass 1.0000 | pass 1.0000 | pass 1.0000 | pass 1.0000 |
| gemma-3-12b-it | pass 1.0000 | pass 0.9974 | pass 1.0000 | pass 0.9965 |
| Qwen3-VL-4B | pass 1.0000 | fail 0.9800 | pass 0.9970 | fail 0.9423 |
| Qwen3-VL-2B | fail 0.9371 | fail 0.8294 | fail 0.9102 | fail 0.6250 |
| Idefics3-8B-Llama3 | fail 0.9505 | fail 0.9401 | fail 0.9041 | fail 0.7474 |
| Ovis1.6-Gemma2-9B | fail: parse $\approx 0$ (chain-of-thought within 8 tokens) | | | |

## E.2 THIRD MODEL FAMILY: GEMMA-3-12B-IT

All arms reuse the 64-pair manifests per visual family. Recovery is relative recovery in percent of the base-to-donor margin, with pair-clustered bootstrap CIs (10,000 resamples). The manipulation check trains an independent donor-value probe on 512 non-evaluation pairs and tests on the 64 causal pairs: clean-base, clean-donor and patched-donor accuracies are all 1.000, so the key patch installs the donor value even though behavior does not move. Committee build-up on the held-out family: top-1 11.03%, top-2 47.70%, top-4 66.98%, top-8 77.02%. The readout crosses at layer 29 of 48 in both families. The row-identity arm is ambiguous: the correct delta beats its slot-permuted control but does not separate from a norm-matched random injection, so causal binding evidence is Qwen-only.

Table 20: gemma-3 causal arms. The committee is ranked on the vertical family only and scored out of family.

| Arm | Vertical (%) | Horizontal (%) |
|---|---|---|
| Key patch (bound $< 15\%$) | $+0.04$ $[-0.15, +0.23]$ | $-0.15$ $[-0.44, +0.13]$ |
| Value control | 87.00 $[84.16, 89.46]$ | 92.61 $[90.47, 94.72]$ |
| Reverse value damage | 87.98 $[85.74, 89.91]$ | 90.73 $[87.35, 93.88]$ |
| Strongest single head (criterion 20%) | 10.67 (L29H1) | 11.03 |
| Top-8 committee (ranked on vertical) | – | 77.02 $[74.78, 79.24]$ |
| Row identity: correct delta | 3.80 $[2.72, 4.93]$ | 8.53 $[6.42, 10.66]$ |
| vs. slot-permuted (paired) | $+1.89$ $[0.70, 3.13]$ | $+4.68$ $[1.82, 7.54]$ |
| norm-matched random | 4.47 $[3.38, 5.71]$ | 8.00 $[5.65, 10.88]$ |
| donor patch | 2.09 $[1.62, 2.60]$ | 6.25 $[4.72, 7.96]$ |

## E.3 CODEBOOK SCALING TO EIGHT AND SIXTEEN ENTRIES

At eight entries (labels A–H, 576 pairs per family, Qwen3-VL-8B) every behavioral criterion is met at 1.000. The key patch recovers $0.06\%$ $[-0.05, 0.18]$ / $-0.04\%$ $[-0.17, 0.08]$ with the manipulation check passed (worst probe cell $0.9844$ $[0.9531, 1.0000]$). The value control recovers $31.59\%$ $[28.40, 34.85]$ / $34.32\%$ $[30.87, 37.75]$ with reverse damage $32.81\%$/$34.88\%$, and the readout crosses at layer 26 of 36 in both layouts with a zero-deviation all-positions ceiling. The binding instrument

cannot run at eight rows: the slot-permuted control becomes collinear with the correct delta (max $|\cos|$ 0.8379/0.8195 against the $< 0.75$ criterion, concentrated on slot pair $2 \leftrightarrow 6$), reproduced exactly by an independent re-estimation, so no effect pairs were run and no binding conclusion exists at size 8 in either direction.

At sixteen entries the vertical family fails the capability criterion: accuracy $0.5677$ $[0.5356, 0.5998]$, paired consistency $0.4575$ $[0.4201, 0.4931]$, answer flip $0.6675$, worst label stratum $0.3750$. A 128-cell cross-backend recheck agrees at $0.96875$ with accuracies 0.6328/0.6172, so the failure is not a backend artifact, and confusions concentrate between same-shape, different-shade icon groups (Fig. 5). The horizontal family and all size-16 mechanism arms were not started: this is a task floor that bounds the testable scope, not a capacity ceiling and not a mechanism result.

### E.4 Real Charts: The Value Locus

On the 219 held-out real-chart pairs (Appendix A.4, Qwen3-VL-8B) the analyzer of the synthetic families runs unchanged. Table 21 gives every condition beside the synthetic families. The null is not a dead intervention: a *different chart's* swatch injected at the same positions moves the answer $+19.75\%$, so the region is causally live and what transfers nothing is precisely the own-donor swatch. That cross-chart condition is the one materially larger than in the synthetic family (a foreign color in a foreign context is far out of distribution), recorded for anyone porting a payload analysis. The positive control is conservative (equalized *down* to the swatch's 2–6 tokens on 210 of 219 pairs, against label regions of up to 24 tokens), and the swatch cosine has a minimum of $0.4530$ against its mean of $0.9931$: on at least one pair the representations differ substantially and still transfer nothing. The current 224-pair manifest gives the same picture (matched swatch $-0.14\%$ $[-0.34, +0.04]$, label control $+17.83\%$ $[+14.32, +21.54]$, cross-chart swatch $+19.56\%$), with the pair-level prior-attributable bound at $17/350 = 4.857\%$ $[0.571\%, 9.143\%]$ (historical snapshot $11/150 = 7.3\%$). The two denominators are never mixed.

Table 21: Real-chart value-locus arm versus the synthetic families, recoveries in % (last row: cosine). Historical 219-pair snapshot. The current 224-pair manifest gives $-0.14\%/ + 17.83\%/ + 19.56\%$ for the first, second and cross-chart rows.

| Condition | Real charts ($n$=219) | Synth. vert. | Synth. horiz. |
|---|---|---|---|
| matched swatch (the key) | $-0.10$ $[-0.27, +0.06]$ | $+0.08$ | $+0.04$ |
| matched label (the value) | $+17.33$ $[+13.91, +21.01]$ | $+26.05$ | $+11.89$ |
| both changed labels | $+9.96$ | $+10.48$ | $+13.10$ |
| another entry's swatch | $+0.10$ | $+0.04$ | $+0.09$ |
| background | $-0.01$ | $+0.23$ | $+0.14$ |
| shifted control | $+0.08$ | $+1.05$ | $+3.95$ |
| foreign swatch (another chart) | $+19.75$ | $+7.43$ | $+3.59$ |
| matched-swatch cosine | 0.9931 | 0.9990 | 0.9993 |

The layer-wise readout sweep on all 224 pairs produces a crossing at layer 24 of 36 with an all-positions ceiling deviation of 0, but the same HF forward classifies only $221/224 = 0.9866$ donors correctly in the A–D decision set against an exact criterion: three pairs that generate correctly under the deployment backend on both prompt variants miss the argmax under the logit interface, reproducibly (Appendix G). The crossing is therefore diagnostic only, the committee-transfer and row-identity arms on these charts were not run on Qwen3-VL-8B, and real charts support the value locus.

### E.5 Real Charts: Row-Identity Readout and Steering (Qwen2.5-VL-7B)

This study asks whether the row identity of Sec. 6 can be used, without access to the answer, to raise Qwen2.5-VL-7B's legend accuracy on the real-chart pool. Every split, readout definition, grid and selection rule was fixed before any forward pass on this checkpoint.

**Headroom.** Table 22 gives the behavioral picture on all 350 pairs (1,400 questions: two forms $\times$ two prompts). A *partner* error is a wrong answer equal to the other form's correct letter, i.e. the swapped legend entry. The headroom criteria (failing pairs $\geq 0.15$, partner errors $\geq 0.30$ of wrong cells) pass. The population for the readout and steering measurements is defined by the HF clean

forward over all $1{,}400$ cells (the same forward the intervention's clean condition runs) rather than by the deployment backend, after the two interfaces agreed on only $124/128$ cells of a subsample (below the $0.98$ criterion, with prompt construction shared code, pixel values byte-identical across backends, and the four disagreements near-ties on the bf16 logit grid). On the full population they agree on $1{,}382/1{,}400$ cells with identical accuracy to four decimals, and the headroom criteria re-pass there ($0.383$, $0.388$). On this checkpoint the $768 \times 576$ canvases resolve to a $21 \times 27$ grid of $567$ visual tokens. Of the $350$ items, $349$ clear the grid-dependent audits (one ring shares a token with a swatch), giving $134$ error pairs ($44$ calibration, $90$ held-out, split by hash order) and $215$ survivor pairs.

Table 22: Headroom on the $350$-pair real-chart pool. Qwen3-VL-8B is the existing record re-typed with the same analyzer. Qwen2.5-VL-7B is shown under the deployment backend and under the HF forward that defines the population for the readout and steering measurements.

| | Qwen3-VL-8B | Qwen2.5-VL-7B (dep.) | Qwen2.5-VL-7B (HF) |
|---|---|---|---|
| Accuracy (1,400 questions) | $0.7707$ $[0.7314, 0.8071]$ | $0.7386$ $[0.6964, 0.7786]$ | $0.7386$ |
| Joint survival (350 pairs) | $227/350 = 0.6486$ | $219/350 = 0.6257$ | $216/350 = 0.6171$ |
| Failing pairs | $0.3514$ | $0.3743$ | $0.3829$ |
| Wrong cells | $321$ | $366$ | $366$ |
| Partner errors (of wrong) | $119 = 0.3707$ | $143 = 0.3907$ | $142 = 0.3880$ |
| $2 \times 2$: both/base/donor/neither | $227/37/20/66$ | $219/22/27/82$ | $216/26/27/81$ |
| Prior-attributable bound | $0.0486$ | $0.0000$ | $0.0000$ |

**Readout.** The row-identity deltas are the answer-control cache for this checkpoint, estimated answer-agnostically by cyclic shift over $500$ synthetic contexts per family. For an image, the per-image mean over the four entries' swatch tokens is subtracted, the matched swatch's residual is projected onto the spatially averaged row-identity deltas, and the slot with the largest projection is read out. The layer window was swept on the calibration manifests only. Validity on the synthetic $192$-pair held-out sets: $1.000$ in both families (chance $0.25$), every window near ceiling, window selected as all layers. On real charts the vertical deltas were transported exactly as specified in advance and never tuned (spatial mean over the $3 \times 3$ lattice, broadcast to each chart's equalized swatch tokens). The averaged deltas are mutually anti-correlated as in the source study (mean cosine $-0.266$, max positive $0.278$). Readout accuracy for the matched entry is $0.9326$ $[0.8977, 0.9628]$ over $215$ survivor pairs and $0.8545$ $[0.7948, 0.9105]$ over $134$ error pairs. Recorded before the steering measurement: the prediction that amplification would fall below its criterion, because the identity is already correct on most wrong cells.

**Steering.** A hook adds to the residual stream at the matched swatch's tokens in a layer window. Arm A (primary): $\alpha$ times the projection of the model's own residual onto the delta span. Arm A2: the same operation at all four swatches, using no site information. Arm B: $\alpha \cdot (\Delta(k^*) - \bar{\Delta})$ with $k^*$ from the readout. Controls, all scored on emitted first-token accuracy: clean, a norm-matched random direction, a random subspace of matched norm (A) or slot-permuted reinforcement (B), and the reversed sign. Grid $\alpha \in \{0.5, 1, 1.5, 2, 3\} \times$ nine windows (all layers plus six-layer windows at stride three), selected on the $44$ calibration pairs (clean $0.3011$) by treatment minus the strongest control, then confirmed once on the $90$ held-out pairs ($360$ cells, clean $0.3250$, Table 23). At $\alpha = 1$ the added vector is $8$–$20\%$ of the residual norm, the same order as the successful flip injection on this checkpoint. Criteria: (i) gain $\geq 0.10$ absolute with CI lower $> 0$, (ii) treatment minus strongest control CI lower $> 0$, (iii) survivor accuracy $\geq 0.98$ under the selected configuration, and (iv) reversed sign below treatment. Pair-clustered bootstrap, $10{,}000$ resamples.

The primary arm is a valid null: every capability and instrument check passed, the zero-delta self-check was exact, the grid was complete ($45/45$ configurations on an identical cell set), and calibration and held-out pairs are disjoint. The calibration surface is itself informative: in every arm the maximum sits at the smallest $\alpha$ and accuracy falls close to monotonically as $\alpha$ grows (Arm A from $0.3466$ at $\alpha = 0.5$ to $0.2159$ at $\alpha = 3$, and Arm B to $0.1705$), against a clean $0.3011$: a perturbation, not a weak repair signal. The two secondary arms harm survivors ($19/860$ and $24/860$ cells) and are reported as such with no re-selection. Arm A2 answers the site-oracle objection in the only available direction: the identical operation with no privileged information produces the same nothing. An optional committee-boost stage was not run, because the anatomy below shows that the selection the committee transports is not what is wrong.

Table 23: Steering, held-out confirmation on 90 error pairs (360 cells), clean accuracy 0.3250, one configuration per arm, selected on calibration. Arm A amplifies the model's own identity at the matched swatch (primary), Arm A2 does so at all four swatches, and Arm B reinforces the slot the readout names. Brackets are pair-clustered 95% CIs.

| | Arm A | Arm A2 | Arm B |
|---|---|---|---|
| Configuration | $\alpha\,0.5$, L9–14 | $\alpha\,0.5$, L9–14 | $\alpha\,0.5$, L6–11 |
| Treatment | 0.3361 | 0.3444 | 0.3306 |
| Norm-matched random | 0.3278 | 0.3389 | 0.3361 |
| Subspace / slot-permuted | 0.3194 | – | 0.3083 |
| Reversed sign | 0.3222 | 0.3389 | 0.3000 |
| Gain vs. clean (criterion 0.10) | +0.0111 | +0.0194 | +0.0056 |
| 95% CI | [−0.0194, +0.0389] | [−0.0222, +0.0611] | [−0.0306, +0.0417] |
| vs. strongest control | +0.0083 | +0.0056 | −0.0056 |
| 95% CI | [−0.0194, +0.0333] | [−0.0361, +0.0472] | [−0.0444, +0.0306] |
| Survivors (criterion 0.98) | 0.9872 | 0.9779 | 0.9721 |

**Error anatomy (exploratory, with no criterion and no new forward).** Each wrong cell among the 134 error pairs was decomposed from the readout records and the clean scoring, asking whether the readout names the ringed series' entry and whether the emitted letter is the label currently at the entry the readout named. Of 366 wrong cells, 311 (0.850) carry the correct row identity at the swatch and answer another entry's label, 16 (0.044) read a wrong entry and answer it, and 39 (0.107) read a wrong entry and answer yet another. Licensed: the row identity is decodable and transportable to real charts, and the steering null characterizes the error as a wrong binding rather than a weak one. Not licensed: causal transport of the identity across layouts (Appendix D.5), anything about Qwen3-VL-8B, or steering interventions in general.

### E.6 One Table Across Settings

Table 24 is the numeric companion to the claim ledger (Table 3): one row per result, one column per setting.

Table 24: The retrieval mechanism across settings. Entries are relative recovery in percent of the margin, and slashes separate the two visual families. The Qwen3-VL patch cells are the 64-pair discovery split, which is why they differ from the 192-pair confirmation split of Table 2 (.04/.02 and 23.8/12.0). Both are reported and neither replaces the other. Dashes mark arms that were never run. **Columns are not comparable to one another**: recovery is normalized by each checkpoint's own clean margin span, which differs by 2.7× between the two Qwen models (Appendix B.1). The row-identity rows give the injection's recovery beside its norm-matched random control (slot-permuted controls and CIs are in Tables 13, 15 and 20, and on gemma-3 the injection does not separate from the random control). The flip row is the answer-control study (Table 16), and the Qwen2.5-VL pilot form gave 24/60%.

| Result | Qwen3-VL-8B | Qwen2.5-VL-7B | gemma-3-12b-it | Real charts |
|---|---|---|---|---|
| Key (symbol) patch | .08/.04 | .05/−.04 | .04/−.15 | −.10 |
| Value (label) patch | 26.1/11.9 | 74.8/48.2 | 87.0/92.6 | +17.3 |
| Label at key, any layer | < .53 | < 2.28 | – | – |
| Readout crossing | L24/36 | L24/28 | L29/48 | – |
| Committee top-8 (out of family) | 80.9 | 76.1 | 77.0 | – |
| Row identity Δ at the key | 2.98/1.84 | 34.1/54.0 | 3.80/8.53 | – |
| norm-matched random control | 0.10/0.15 | 1.24/1.65 | 4.47/8.00 | – |
| Predicted flips (%), calibrated | 0.0/0.0 | 68.2/87.5 | – | – |

## F Probing on Natural Benchmarks

The inference from decodable to used is made most often on pre-answer states over natural benchmarks, in the hope that a VLM's states know what it cannot see, so we ask the paper's question there as well. Decodability and transferability are established representationally. Use is bounded

*behaviorally* (by what the signal predicts about the model's own errors and by what a verifier can do with it), because no intervention along the axis exists in this setting. The four model families are Qwen3-VL-8B (Qwen Team, 2025), LLaVA-OneVision-7B (Li et al., 2024), Idefics3-8B-Llama3 (HuggingFaceM4, 2024), and InternVL3-8B (Zhu et al., 2025). Appendix F.1 gives the results. The subsections after it give the designs, the full tables, and the two behavioral directions the program ran beside them.

### F.1 Main Results

**An accessibility axis is decodable and transfers.** Varying only the pixels (five blur and five downscale levels plus a blank floor, with item, question, options and token layout fixed), a pre-answer Qwen3-VL-8B (Qwen Team, 2025) state decodes the ladder well, and the feature precedes answer generation, so answer-text leakage cannot explain it (Appendix F.3). The signal generalizes: across four model families and five benchmarks, a probe trained on one operator and tested on the other clears its bound in 37 of 40 cells. A third operator, Gaussian noise, breaks transfer cleanly, so the axis is a *low-pass* accessibility axis rather than corruption invariance. Decodability and transfer come apart here as well: Idefics3 posts the highest $R^2$ on all five benchmarks and accounts for all three non-transferring cells, so here decodability is a manipulation check rather than the result.

**The axis is the vision encoder's feature.** Blur and downscale share a low-pass signature, so the baseline is a probe over statistics designed to capture it (Laplacian variance, mean gradient magnitude, FFT high-frequency energy ratio and a $32\times32$ grayscale thumbnail) through the identical probe class, pipeline, folds and held-out items (Appendix F.3). The model feature beats it in every transferring cell with paired bootstrap CIs excluding zero (Fig. 16). A second comparison locates the advantage: a connector-token baseline (the state at the vision–language boundary, before any language-model computation) matches or exceeds the later prompt state. The axis is an image-quality feature computed by the vision encoder that survives into the residual stream, not knowledge the language model constructed about its own perception. The probe is reading the encoder.

**The axis predicts none of the item-level errors.** Predicting the item-level accuracy loss after removing per-condition means (five benchmarks, 250 base items, sixteen sampled paths per degraded view) is null on all five. The null is not a weak probe: a capacity sweep on the sibling residual-accuracy target finds signal, and it decomposes into a question-difficulty prior, which survives blanking the image, or into perceiving the answer itself: on POPE the ground-truth answer decodes from the sighted state at $\mathrm{AUROC}\ 0.979$ against $0.591$ blind. Predicting correctness by perceiving the answer is not self-knowledge.

**Hidden states never beat matched observable evidence.** A verifier reading hidden states should beat one reading only observable evidence if those states carry privileged evidence about the model's own errors. Our matched comparison holds training data, splits, capacity and multimodal access equal: a *policy-latent* verifier reads the answering model's internal states, an *output-only* verifier the image, step text and observable probability features. On a controlled factorial corpus the two are indistinguishable while both beat text- and probability-only baselines: parity, not a floor effect. On natural on-policy paths the policy latent is significantly *worse* at step-level grounding diagnosis and ties on reasoning diagnosis and selection, and the equal-budget repair comparison does not separate them (Appendix F.4). On Qwen3-VL-8B the policy latent never beats matched observable evidence.

**Relation to the legend task.** Decodable, transferable, written by the encoder, carrying nothing item-level about the model's own errors: that is the pattern of Secs. 4–5, measured where the inference is made in practice. This setting cannot install a different value of the axis and certify that it arrived, so the usability judgment rests on prediction rather than intervention. Certification needs a representation whose content can be rewritten exactly and read back, which the legend task provides.

### F.2 Path Selection and Vision Ablation

Sampling 16 reasoning paths per question and probing correctness from generated-token states, a text-state probe reaches best-of-$N$ accuracy $0.7455$ on HallusionBench, above majority voting ($0.7266$). Adding visual and cross-modal summaries does not help ($0.7371$). Removing the image before generation lowers best-of-$N$ by $0.339$ ($[0.325, 0.354]$), so vision's causal role is large, but

a proposed perception-versus-reasoning differential *changes sign* under answer-format and within-benchmark controls: the cross-dataset axis was confounded with dataset identity and answer format, the class of confound the within-item degradation design below removes by construction. Four increasingly strict tests reject the fallback that the model is uniquely blind to *confident* visual errors: (i) sighted-versus-blind answer agreement does not detect wrong answers on hallucination tasks, (ii) sighted-minus-blind probe confidence adds nothing beyond sighted confidence, (iii) conditioned on the model's own answer, perception errors separate no worse than reasoning errors, and (iv) on questions where at least 13 of 16 paths agree, confident-wrong cases become *easier* to separate (AUROC $\approx$ 0.97–0.98 on MMBench).

### F.3 Accessibility Axis: Decodability and Transfer

Table 25 gives the full decodability grid for the degradation ladder of Appendix A.5: four model families $\times$ five benchmarks, pre-answer states, held-out probes with GroupKFold by base item. Figure 14 shows the behavioral ladder the grid is decoded from, and Fig. 15 one item of it.

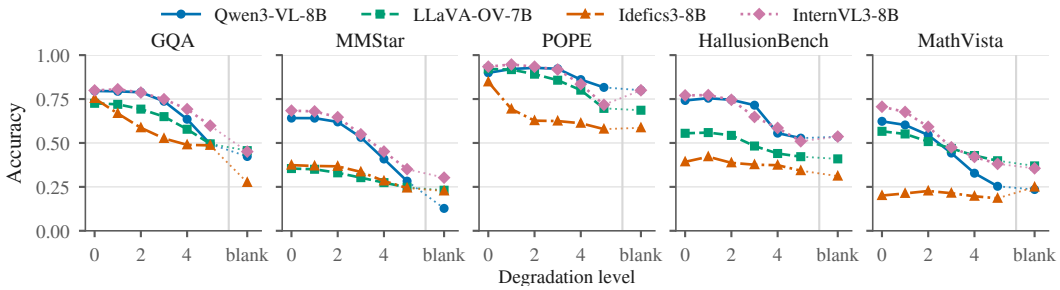


Figure 14: The within-item degradation ladder, averaged over blur and downscale. Only pixels change across a row, and every model–benchmark pair falls toward the blind floor (right of the rule, uniform gray image). On $8/20$ pairs accuracy first rises over the gentlest one or two rungs, by at most 0.03, before it falls. On $5/20$ pairs the floor sits slightly *above* the last rung (most visibly InternVL3 on POPE, $+0.083$), so a heavily degraded image can actively mislead.

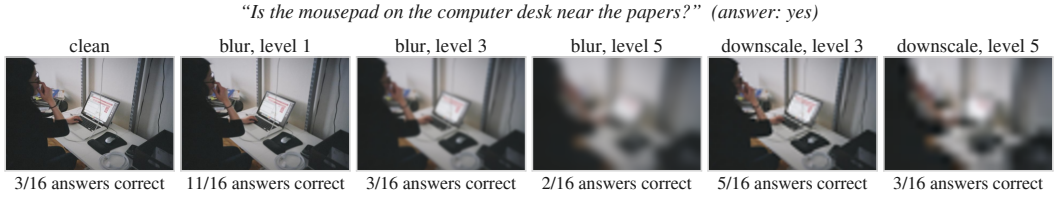


Figure 15: One real item of the ladder (the first GQA item of the evaluation file, on Qwen3-VL-8B with 16 sampled answers per level). The degradation is plain in every panel, and the model's answers do not track it: this is the item-level gap the bound of Appendix F.1 measures over all items.

**Pixel baseline.** The matched control is a probe over classical low-pass statistics chosen to be the strongest cheap explanation of this manipulation: Laplacian variance, mean Sobel gradient magnitude, FFT high-frequency energy ratio (one minus the fraction of shifted spectrum mass inside the central quarter-radius box), and a $32\times32$ grayscale thumbnail, concatenated into a 1,027-dimensional feature and run through the identical probe class, ridge penalty, GroupKFold folds, and held-out items as the model feature. Cells where its $R^2$ collapses below $-1$ are extrapolation failures on held-out base items, not a missing baseline. The transferring ones among them are drawn clipped in Fig. 16.

Three structural facts qualify the grid. Transfer, not decodability, is the claim: a probe trained on one operator (blur) and tested on the other (downscale) clears its bound in 37 of 40 cells, and beats the matched pixel-only probe in all 37 transferring cells (paired bootstrap CIs excluding zero).

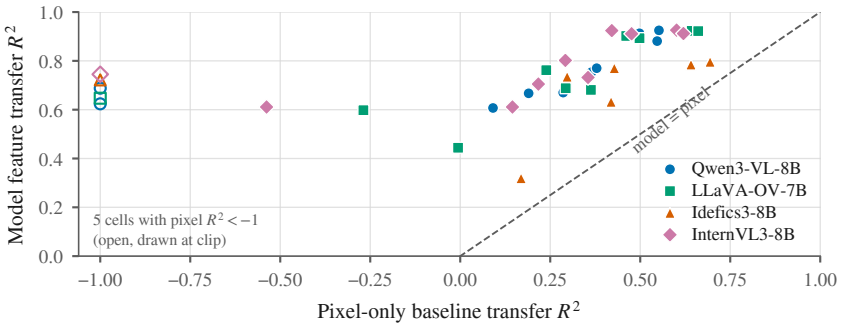


Figure 16: Cross-operator transfer: a probe trained on one degradation operator, tested on the other, on held-out base items. The 37 transferring cells of the 40 are drawn, and every one lies above the identity line: the internal state beats the matched classical pixel probe run through the same pipeline, folds, and held-out rows. Cells whose pixel baseline collapses below $R^2 = -1$ are drawn open at the clip boundary.

Table 25: Pre-answer decodability of the degradation ladder: the last prompt token's hidden state at median depth (layers 21, 16, 19 and 16 in column order), from the full-resolution extraction Fig. 16 reads. $R^2$ predicts the imposed level, and AUROC separates high- from low-access views. Every cell is a held-out probe with GroupKFold by base item.

| | Qwen3-VL-8B | | LLaVA-OV-7B | | Idefics3-8B | | InternVL3-8B | |
|---|---|---|---|---|---|---|---|---|
| Benchmark | $R^2$ | AUC | $R^2$ | AUC | $R^2$ | AUC | $R^2$ | AUC |
| GQA | 0.982 | 1.000 | 0.954 | 1.000 | 0.962 | 1.000 | 0.973 | 1.000 |
| MMStar | 0.901 | 0.999 | 0.809 | 0.997 | 0.856 | 0.999 | 0.810 | 0.997 |
| POPE | 0.982 | 1.000 | 0.958 | 1.000 | 0.965 | 1.000 | 0.974 | 1.000 |
| HallusionBench | 0.918 | 1.000 | 0.816 | 1.000 | 0.819 | 0.997 | 0.883 | 1.000 |
| MathVista | 0.861 | 0.998 | 0.761 | 1.000 | 0.791 | 0.996 | 0.776 | 0.999 |

Decodability dissociates from transfer: every model decodes the imposed level ($R^2$ 0.76–0.98, AUROC $\geq$ 0.996), and the model that fails to transfer is not the one that decodes worst: all three non-transferring cells are Idefics3's, which ranks second or third on every benchmark. The axis is low-pass, not corruption-general: a third operator (Gaussian noise) breaks transfer cleanly on MathVista in all four directions despite a usable accuracy range, so the supported claim is a low-pass accessibility axis, already present at the vision–language connector (a connector-token baseline matches or exceeds the later prompt state).

The decisive negative is item-level: predicting the item-level accuracy loss $\mathrm{Acc}(\mathrm{clean}) - \mathrm{Acc}(\mathrm{degraded})$ after removing per-condition means (five benchmarks, 250 base items, 3,750 degraded views, 16 paths per view) is null on all five (best connector residual: GQA $R^2 = 0.031$, CI $[-0.089, 0.094]$). The null is not a weak probe: a capacity sweep on the sibling residual-accuracy target does find signal, but it decomposes into a question-difficulty prior (it survives blanking the image) or into perceiving the answer itself (on POPE the ground-truth answer decodes from the sighted state at AUROC 0.979 versus 0.591 blind). Predicting correctness by perceiving the answer is not self-knowledge.

### F.4 Matched-Verifier Comparison

If hidden states carried privileged evidence about the model's own errors, a verifier reading them should beat one reading only observable evidence when everything else is matched. Reasoning traces on GQA decompose into OBSERVE and INFER steps supervised by factorized counterfactuals built from executable ground truth (scene graphs and programs): each correct path yields a same-style, same-length quadruple {OK, V (one visual fact falsified, valid inference preserved), R (facts kept, one logical operator flipped), VR (both)}, with text-identical image-swap pairs and image-identical inference-swap pairs, answer correctness and error position balanced, and splits

image-disjoint. The corpus runs to hundreds of thousands of quadruples across query, multi-step and five-family factorial generations, with schema, executor, image-existence and checksum audits at zero errors. The *policy-latent* verifier reads the answering model's internal states (sighted replay, blank replay, their difference), and the *output-only* verifier reads the image, step text and observable probability features, with the same training data, splits, capacity and multimodal access. On the controlled factorial corpus (20,000 paths, 75,220 steps) the two are indistinguishable at near-ceiling (grounding AUROC difference $+0.00040$, CI $[-0.00162, +0.00317]$) while both clearly beat text- and probability-only baselines: parity, not a floor effect. On $6{,}400$ natural on-policy paths, scored with the controlled-set checkpoint without retraining or re-thresholding, the policy latent is significantly *worse* at step-level grounding diagnosis ($-0.1011$, $[-0.1247, -0.0811]$), tied on reasoning diagnosis and best-of-$N$ selection, and equal-budget typed repair shows no latent advantage ($+0.0025$, CI crossing zero, Table 26). Both verifiers route $\sim 88\%$ of natural paths to "no repair needed", so the repair comparison is underpowered for equivalence and is reported as such.

Table 26: Matched policy-latent versus output-only verifier comparison (Qwen3-VL-8B). The two AUROC rows and the typed-repair cell are differences, policy minus output-only. The path-level row gives the two levels themselves, policy then output-only. A dash marks an arm that does not exist in that setting.

| Metric | Controlled | Natural paths |
|---|---|---|
| Grounding AUROC | $+0.00040$ | $-0.1011$ |
| Reasoning AUROC | $-0.00002$ | $-0.0208$ |
| Path-level score | F1 $0.9397/0.9410$ | BoN $0.6100/0.6038$ |
| Typed repair | – | $+0.0025$, CI $\ni 0$ |

### F.5 A Terminated Behavioral Direction

A counting task that varied connectedness (connected components against item counts) produced directionally consistent negative shifts across four stimulus redesigns, but each version failed a neutrality control (off-target gray lines themselves shifted counts), and the independent confirmation of the derived scaffold-correction effect came out directionally consistent yet below both magnitude criteria ($S = -0.0560$ against $-0.07$, accuracy gain $+0.0321$ against $+0.05$). No activation patch was run and no criterion was relaxed. The direction is kept in the record as what fixed criteria do to a real but sub-threshold effect.

## G Validity Checks That Rejected Something

A check that never rejects anything demonstrates nothing. Table 27 lists every case in the program in which a capability criterion, an instrument check, a guard or a scored prediction rejected a model, a task size, an instrument, a run or a sentence, with what it removed and where the record is. Three rules govern their reading: (i) a rejected model or task size is never read as a mechanism result, (ii) an instrument that fails its own validity check is never read as a negative finding, and (iii) a correction made after a number was seen is disclosed as such. Nothing in the main text depends on a reader accepting this appendix on trust.

Table 27: What the validity checks caught. "Removed" names the claim, model, run or wording the check took out of the paper.

| Check | What it caught | Removed |
|---|---|---|
| *Capability criteria* | | |
| Behavioral screen, scale | Qwen3-VL-4B fails two cells (worst accuracy $0.9423$, pair consistency $0.8993$), and 2B fails everything including parse (down to $0.6250$). The failure is monotonic in scale. | Both models from causal work. Nothing about their mechanisms is concluded. |

*continued on the next page*

*Table 27, continued.*

| Check | What it caught | Removed |
|---|---|---|
| Behavioral screen, third family | Idefics3-8B-Llama3 (32 layers) parses 1.0000 on all 9,216 items but misses accuracy on every cell (worst 0.7474, pair consistency 0.5807). Ovis1.6-Gemma2-9B (42 layers) emits chain-of-thought and never reaches a letter within 8 tokens. | Both as third-family candidates. The absolute-versus-relative depth question waited for gemma-3. |
| Behavioral screen, size 16 | Vertical accuracy 0.5677 [0.5356, 0.5998], consistency 0.4575, flip 0.6675, worst stratum 0.3750. Not a backend accident (128-cell HF recheck 0.96875 agreement) and not a manifest defect. | Every size-16 mechanism arm. Scope bounded at eight entries. |
| Neutrality control, topology task | A counting task varied connectedness. Every redesign failed a neutrality or capability control (off-target gray lines shifted counts), and the derived correction missed both magnitude criteria on independent confirmation ($S = -0.0560$ vs. $-0.07$). | The task family. No patch was ever run. |
| *Interface agreement* | | |
| vLLM–HF agreement, Idefics3 | 83.38% item-level agreement. On a vLLM-selected subset the HF argmax contradicted the model's own answer on 4–5 of 64 pairs, leaving the estimand undefined. Re-deriving the subset from HF gives 0.8919, below the 0.90 floor by 0.0081. | An exploratory scan tier, with two scan outputs kept on disk, never analyzed. Agreement $\geq 0.98$ became a precondition for any new checkpoint, and the tier's near-tautological consistency clause is recorded as a wording defect. |
| Exact argmax, real-chart sweep | $221/224 = 0.9866$ donors classified correctly in the A–D set under the logit interface although all three failing pairs (`car-ecd-ecd10k-004647`, `-005456` and `-006770`, with argmax-minus-truth logit $+3.00$, $+0.25$ and $+1.00$) generate correctly under the deployment backend (reproduced in three reruns, the first shown in Fig. 6). | The committee and binding arms on real charts (Qwen3-VL). The layer-24 crossing is diagnostic only. |
| Agreement criterion, real-chart steering | $124/128 = 0.9688$ against 0.98. The four disagreements are near-ties on the bf16 grid. | The vLLM-selected population: the study's population is the HF forward that measures the intervention ($1{,}382/1{,}400$ agreement), and the bridge is retired program-wide. |
| *Instrument checks* | | |
| Binding-ID positive control (i) | A structurally analogous MCQ control (`A) star B) hexagon ...`) lets the model bind by position and cannot certify the instrument. The substituted control failed. | The first estimator. The control the protocol specified was built (prose-separated bindings, names crossed with slots). |
| Binding-ID positive control (ii) | The faithful estimator, certified on the published task, failed its visual positive control against a 20% criterion inherited from position-basis patching. Context freezing costs the reference operation $3.2\times$ (26.05% $\rightarrow$ 8.11%). | That criterion. Criteria were re-fixed as fractions of the same-run reference and confirmed on held-out pairs and an untouched family. "Ceiling" was later narrowed to "reference": the reference operation is one-sided and was exceeded $1.70\times$ in one family. |
| Window-restricted ratio | A windowed row-identity intervention clears a 0.40 fraction at 1.04 only because the window handicaps the reference $6\times$ and the primary $1.7\times$. The absolute effect falls from 1.84% to 1.09%. | The windowed ratio as a criterion. |

*continued on the next page*

*Table 27, continued.*

| Check | What it caught | Removed |
|---|---|---|
| Non-collinearity, size 8 | The slot-permuted control is collinear with the correct delta (max $\lvert\cos\rvert$ 0.8379 / 0.8195 against $< 0.75$), reproduced exactly on separate GPUs. | Any binding conclusion at size 8, in either direction. |
| Probe convergence | Standardizing before PCA but not after left 119 of $\sim$570 logistic fits at the iteration cap. A non-converged probe underestimates decodability, i.e. biases toward the convenient verdict. | The run. The re-run reports zero non-converged folds. |
| Simulated null for an equivalence bound | A "label absent at every interface" bound of 0.30, calibrated for one interface but applied as an intersection over 38, was cleared by pure noise in only 1 of 5 seeds. | The bound moved to 0.33 before any real feature was probed (disclosed as a loosening), and zero of 38 interfaces fell below it. |
| Shifted control geometry | The two families' probe verdicts disagreed only because a shifted control lands inside the legend-row band in one layout and outside it in the other. A fixed-grid map that flattened by construction is kept as an analysis error. | Any specificity claim in either direction. The specificity rule, which tested significance without a magnitude floor, is recorded as defective and was not repaired after the fact. |
| *Guards and scoring corrections* | | |
| Absolute-cosine gate | Two runners implemented "not collinear" as max absolute cosine $< 0.5$, which flags the anti-correlated pattern that validated the control (mean cosines $-0.268$ / $-0.202$). One family was wrongly blocked on a $-0.5374$ slot difference. | The premature invalid labels. The gate reads max positive cosine $< 0.50$, corrected before any causal forward. |
| Analyzer stricter than its criterion | A flip-calibration cell with clean baselines $63/64$ and $64/64$ was rejected by an analyzer demanding exact 1.00, which the protocol never required. It re-scores at accuracy 0.9922, consistency 0.9844. | The over-strict rule, before any configuration was selected. Grid-completeness guards added. |
| "Already donor" is not a flip | The first flip scorer counted a donor-letter output as a flip even when the baseline already emitted it. | That scoring. A flip requires the token to change from baseline and equal the predicted letter. No selection changed. |
| Geometry guard, gemma-3 shard | A control-equalization rule could expand a control mask into the primary region on the 48-px grid. The guard stopped it at record 15. | Fourteen records, quarantined. Equalization restricted to positions outside primary and changed regions. |
| Hook placement | Spurious recovery at layers 0–2 (DeepStack features are added after those layers, the hook fired before) and a final-layer mismatch (post-norm state vs. pre-norm hook). | Both checks replaced by stronger deterministic ones. No primary estimand touched. |
| *Audits and scored predictions* | | |
| Analysis-code audit | Four gate-completeness gaps in eleven scripts (an unwired supplementary null, an ungated ceiling check, a silent `defaultdict`, and a coverage guard skipping a mid-list `nan`). | Nothing: all analyses re-run, zero of ten verdicts changed, six regression tests added. |
| Purpose retracted | The difficulty ladder's mechanism-under-failure purpose: its errors are upstream of the mechanism, the estimand is undefined on failed pairs, and the subset would be selected on the dependent variable. | That purpose. The color control stands. |
| Concession clause | Cross-family transport fixed in advance the sentence to withdraw under each outcome, and the outcome required withdrawing "abstract". | "Abstract row identity" became "layout-specific row identity" throughout. |

*continued on the next page*

*Table 27, continued.*

| Check | What it caught | Removed |
|---|---|---|
| Predictions scored | The second-model gate was predicted to fail and passed at 1.0000 (a stale cross-reference in the prediction text is preserved verbatim and corrected in place). The Qwen2.5-VL row-identity magnitude was predicted smaller and came out 11–29× larger. A constant-freezing-cost prediction missed (1.9× vs. 3.2×). The gemma-3 query-arm prediction was wrong in the model's favor. The layer-24 deflation, the real-chart gate band and the steering null were predicted correctly. | Nothing. The misses are kept beside the hits. |

## H Extended Limitations and Outlook

**New versus replication.** Middle-depth assembly of the answer symbol by sparse attention heads is reported for text-only models (Lieberum et al., 2023; Wiegreffe et al., 2025) at closely matching relative depths, and our destination side replicates that in a VLM. That visual information leaves the image tokens early for the readout position is the flow the visual-token literature describes (Sec. 2). The source side is new: a content-only visual key, no addressable binding slot at its position, a layout-bound binding code, and a separation dated to the vision encoder rather than inferred from the language model: a value installed at the image tokens, certified in place, and bypassed. So is the handle: a train-free injection that decides which legend entry the model reads.

**Limitations.** The key null (no position-basis intervention at the key moves the answer) is relative to that instrument. A learned fixed-rank subspace at the key (Geiger et al., 2024) leaves it intact on Qwen2.5-VL-7B (Appendix C.8), within that class (linear, one layer at a time, five layers, one checkpoint). An edge-level path analysis (Goldowsky-Dill et al., 2023; Syed et al., 2024) could still overturn it, while patching every position the label reaches leaves the null intact, holds when the donor value is re-imposed at every layer so that it cannot be overwritten, and locates the carrier elsewhere (Sec. 4). The chain closes on one checkpoint: Qwen2.5-VL-7B carries the probe, the certified install, the readback and the emitted-answer control. Qwen3-VL-8B carries the same first half and gemma-3 the certificate. Behavioral control does not transfer across checkpoints (Sec. 7). Recovery is normalized by each model's own margin span, so magnitudes do not compare across checkpoints, and the equivalence bound is an absolute number applied to a model-relative quantity. Behavioral criteria and mechanism forwards use separate inference stacks. Agreement is measured and gated, and every causal measurement is within-harness against a baseline recomputed there, so backend disagreement cannot manufacture a recovery, only make a selection unrepresentative, which is why the real-chart diagnosis operates through one interface (Appendix E.5). The task is operation-specified retrieval on four-entry legends: an abstract prompt scores 0.43 on the same images (Appendix C.5), and the binding instrument is valid through four entries only.

**Conclusion.** Does the model's answer run through what the probe reads? Where it can be tested by intervention, no: a decodable, verifiably installable key value the encoder wrote is overwritten rather than read. Over natural benchmarks, where use can only be bounded by prediction, the pattern is the same. What the answer does run through (the legend rows at the merger, a readout committee, a layout-specific row identity) was found by intervention, replicates structurally across three model families, and on Qwen2.5-VL decides what the model says. On that one checkpoint the whole sequence is measured. The scope is operation-specified retrieval on three checkpoints: head and layer indices are model-specific, causal binding evidence is Qwen-only, and real charts support the value locus and carry a readable row identity that the model's errors bypass.

**Outlook.** The open directions are validity and use, not scale: (i) a non-collinear binding instrument beyond four entries, (ii) a real-chart readout criterion applied through one interface, (iii) an edge-level analysis at the key, and (iv) whether the row identity is *necessary* and not only sufficient. Its obvious control does not exist: matched on removed energy, an in-span control overlaps the identity, an orthogonal one cannot reach its energy, and we report no necessity result in either direction. The row identity is a handle on what the model reads and, on real charts, a diagnostic: the failures it

locates lie in the value-side read that follows a correct identity, which is where a repair would have to act and where the readout committee of Sec. 5 is the natural target.

# I Side Study: Base-Color Attraction in Geometry Judgments

This study shares the program's minimal-pair and patching methodology and nothing else: different stimuli, a different task, a different claim, and a single checkpoint (Qwen3-VL-8B). No claim in the main text depends on it. Its cross-model failures are part of the capability record in Appendix G.

Texture–shape conflict is classic for CNNs (Geirhos et al., 2019). The causal VLM version uses rendered stimuli whose ground truth is the scene graph. Two-image panels ask whether reference and candidate objects share the same 3D shape. The key pair reuses the identical wrong mesh, pose, camera, background and environment lighting (HDRI) while changing only the candidate's physically-based-rendering (PBR) material between shared and non-shared. With *material-only* the condition sharing only the material (different mesh) and *neither* sharing nothing, the estimand is $D = \mathrm{Acc}(\mathrm{neither}) - \mathrm{Acc}(\mathrm{material\text{-}only})$. Both conditions have correct answer DIFFERENT, so a positive $D$ is an attraction toward SAME created purely by the shared material. Six design iterations eliminated label validity, prompt semantics, presentation order, view matching and mesh difficulty as confounds. The first fully controlled evaluation (72 audited Objaverse meshes (Deitke et al., 2023), unseen PBRs) gives $D = 0.344$ $[0.253, 0.434]$. An independent new-pair, new-PBR, new-pixel replication reproduces the different-mesh attraction, $D_{\mathrm{sym}} = 0.115$ $[0.088, 0.142]$, but not the same-mesh component, so the broad "shared PBR damages geometry" claim is rejected. Rendering the same pairs with single material channels attributes the effect to base color: full PBR $D = 0.128$, base color only $0.164$, microgeometry (normal/roughness) only $0.056$, uniform clay $0$, with base color exceeding microgeometry by $0.090$ $[0.049, 0.133]$: color and texture suffice, and complete material structure is not necessary. Patching the candidate-object visual tokens localizes the effect: the final merger alone recovers $78.4\%$ [70.2, 85.5] on discovery and $78.8\%$ [70.8, 85.4] on zero-overlap confirmation pairs, with left-region, cross-unit and layer-misalignment controls passing and individual DeepStack taps near zero. Smaller Qwen3-VL scales and seven 7B–9B alternatives fail same/different or response-format capability controls, so the mechanism is one-checkpoint. Those are invalid cross-model tests, not negative replications.

# J Models and Benchmarks

Table 28 lists every checkpoint the paper reports and where each is used, and Table 29 lists the benchmarks and data sources. The legend stimuli are generated for this paper (Appendix A).

**Checkpoints.** All weights are the public releases, run without fine-tuning: `Qwen/Qwen2.5-VL-7B-Instruct` (Bai et al., 2025), `Qwen/Qwen3-VL-8B-Instruct`, `-4B-Instruct` and `-2B-Instruct` (Qwen Team, 2025), `google/gemma-3-12b-it` (Gemma Team, 2025), `HuggingFaceM4/Idefics3-8B-Llama3` (HuggingFaceM4, 2024), `AIDC-AI/Ovis1.6-Gemma2-9B` (AIDC-AI, 2024), `llava-hf/llava-onevision-qwen2-7b-ov-hf` (Li et al., 2024), `OpenGVLab/InternVL3-8B` (Zhu et al., 2025), `OpenGVLab/InternVL3_5-8B` (Wang et al., 2025), `AIDC-AI/Ovis2.5-9B` (Lu et al., 2025) and `allenai/Molmo-7B-D-0924` (Deitke et al., 2025).

**Visual tokens on the legend canvas.** The three checkpoints that pass the legend task's behavioral screen see the $768\times768$ canvas as different grids. Qwen2.5-VL-7B resizes it to $756$ px and emits $729$ merged tokens ($27\times27$, $28$ px each), so symbol boundaries do not fall on token boundaries. Qwen3-VL-8B emits $576$ ($24\times24$, $32$ px each), and three DeepStack taps add the features of encoder blocks $8$, $16$ and $24$ to the first three language-model layers (Appendix B.2). gemma-3-12b-it emits $256$ ($16\times16$ after pooling), with pan-and-scan off.

**Benchmarks.** The degradation ladder (Appendix A.5) draws its base items from GQA (Hudson & Manning, 2019), MMStar (Chen et al., 2024b), POPE (Li et al., 2023), HallusionBench (Guan et al., 2024) and MathVista (Lu et al., 2024), and renders each at eleven levels plus a blank floor. The path-selection study (Appendix F.2) samples $16$ reasoning paths on $951$ HallusionBench, $958$ MathVista, $1{,}499$ MMStar and $1{,}500$ POPE questions, with MMBench's English dev split (v1.1) (Liu et al.,

Table 28: Checkpoints reported in the paper. Encoder: family, transformer blocks and patch size. Language model: backbone and number of layers. The Qwen3-VL connectors add DeepStack taps from three encoder blocks. The first three pass the legend task's behavioral screen and carry the mechanism results. The screens are the legend task's (Appendix E.1) and the geometry side study's (Appendix I).

| Model | Vision encoder | Connector | Language model | Used in |
|---|---|---|---|---|
| Qwen2.5-VL-7B | ViT, 32, 14 px | 2×2 patch merger | Qwen2.5-7B, 28 | Secs. 4–7 (primary) |
| Qwen3-VL-8B | ViT, 27, 16 px | 2×2 patch merger | Qwen3-8B, 36 | Secs. 4–7, App. F, I |
| gemma-3-12b-it | SigLIP, 27, 14 px | 4×4 pooling, linear | Gemma 3 12B, 48 | Sec. 7, App. E.2 |
| Qwen3-VL-4B | ViT, 24, 16 px | 2×2 patch merger | Qwen3-4B, 36 | both screens |
| Qwen3-VL-2B | ViT, 24, 16 px | 2×2 patch merger | Qwen3-1.7B, 28 | both screens |
| Idefics3-8B-Llama3 | SigLIP, 27, 14 px | pixel shuffle, linear | Llama 3.1 8B, 32 | App. F, both screens |
| Ovis1.6-Gemma2-9B | SigLIP, 27, 14 px | visual embedding table | Gemma 2 9B, 42 | both screens |
| LLaVA-OneVision-7B | SigLIP, 26, 14 px | two-layer MLP | Qwen2-7B, 28 | App. F, geometry screen |
| InternVL3-8B | InternViT, 24, 14 px | pixel shuffle, MLP | Qwen2.5-7B, 28 | App. F, geometry screen |
| InternVL3.5-8B | InternViT, 24, 14 px | pixel shuffle, MLP | Qwen3-8B, 36 | geometry screen |
| Ovis2.5-9B | SigLIP 2, 27, 16 px | visual embedding table | Qwen3-8B, 36 | geometry screen |
| Molmo-7B-D | CLIP ViT-L, 24, 14 px | attention pooling, MLP | Qwen2-7B, 28 | geometry screen |

2024) as the within-benchmark control. The verifier comparison (Appendix F.4) builds its reasoning paths from the full GQA release, with image-disjoint splits of our own. The real charts are re-rendered from the matplotlib sources that ECDBench and ECD-10k-Images (Yang et al., 2025) ship (Appendix A.4). The geometry side study renders Objaverse (Deitke et al., 2023) meshes with the LAS&T generation scripts (Eppel et al., 2025), VasTexture PBR materials and Poly Haven HDRI lighting.

Table 29: Benchmarks and data sources. Ladder items are base items, each rendered at eleven levels plus a blank floor.

| Source | Split or subset | Answer | Items used | Used in |
|---|---|---|---|---|
| GQA | testdev-balanced, *verify* | yes/no | 250 ladder | App. F.3 |
| MMStar | test | four-way choice | 250 ladder | App. F.2, F.3 |
| POPE | test, all three settings | yes/no | 100 ladder | App. F.2, F.3 |
| HallusionBench | test | yes/no | 100 ladder | App. F.2, F.3 |
| MathVista | testmini, choice and integer | choice, integer | 100 ladder | App. F.2, F.3 |
| MMBench | dev, English, v1.1 | choice | control | App. F.2 |
| GQA | questions v1.2, scene graphs v1.1 | reasoning steps | 20,000+6,400 paths | App. F.4 |
| ECDBench, ECD-10k | shipped matplotlib sources | four-way choice | 350 pairs | Sec. 7, App. E |
| Objaverse | 12 LVIS categories | same/different | 72 meshes | App. I |