BEYOND THE CURRENT ANSWER: VISUAL CONTENT TRANSFER INTO VLM TEXT STATES

Anonymous authors
Paper under double-blind review

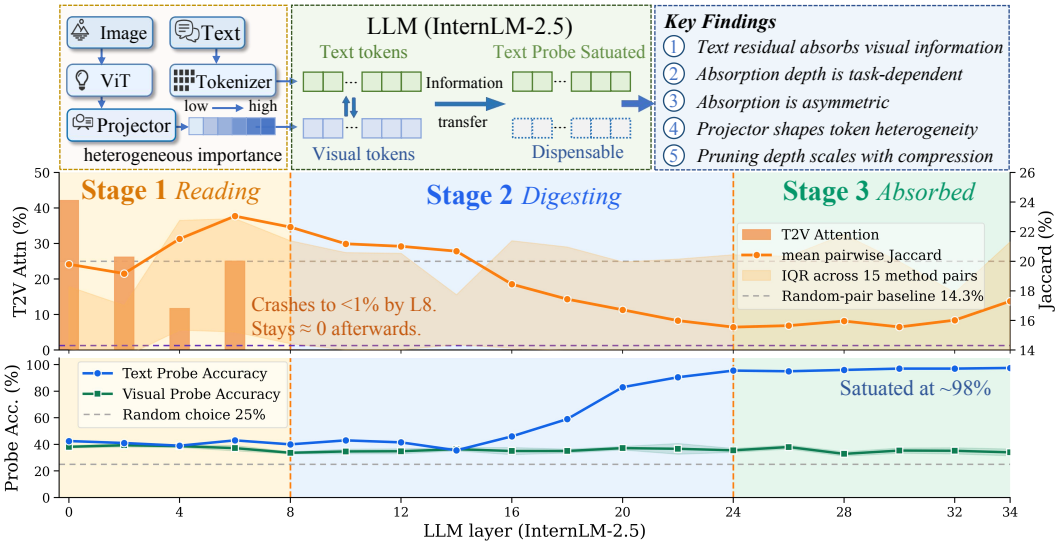

Figure 1: **Original exploratory overview on InternVL/ScienceQA.** Attention, token-set overlap, and answer probes motivate studying visual-to-text information flow. The stage labels and numbered findings show the original framing, not universal conclusions; the 25% line is an illustrative guide, not a calibrated chance baseline. Here, answer-matched donors and carrier patches test bounded content transfer, while symmetric recovery independently measures deletion sufficiency.

ABSTRACT

When a vision–language model tolerates removal of visual states, has it retained image content or only computed the current answer? We separate three questions: which states control the answer, whether text states contain information beyond that answer, and whether visual states remain necessary. Our central test pairs real images with identical questions and annotated answers but different unasked object categories. Patching donor visual activations early versus late changes which image's unasked content is decodable from non-final text states. Across seven tested checkpoints and 500 disjoint pairs per checkpoint, the paired interactions are positive, including a model with no sustained visual–text deletion gap. Complementary opposite-answer patches reveal a depth-dependent shift of causal control from visual to non-final text states. Symmetric zero-out on 1,000 ScienceQA questions yields a positive first-recovery gap in six of seven checkpoints, but sustained ordering is architecture-dependent; open-ended TextVQA reverses the first-recovery ordering in two models. Position-controlled probes expose a substantial causal-order confound. These results support transfer of image-derived semantic content beyond the current answer into text residuals. They do not establish lossless image storage or unrestricted re-queryability: a visual-KV-blocked second-question test largely fails. Content transfer and visual-token dispensability are distinct properties.

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.

1 INTRODUCTION

Decoder-only vision–language models (VLMs) process projected image features and text in a shared token sequence. Visual-token pruning exploits the observation that some of these image positions can be removed with little change in task performance (Chen et al., 2024a; Yang et al., 2025; Zhang et al., 2025a). This observation leaves a mechanistic ambiguity. The model may have extracted only the fact needed for its current answer, or it may have written a broader set of image-derived variables into text states. Both accounts predict accurate answers after sufficiently late visual perturbations. Neither low attention to images nor answer decodability at the final text position distinguishes them.

We study this ambiguity by separating *content*, *carrier*, and *deletion sufficiency*. Content asks which image-derived variables can be recovered from text states. Carrier asks which positions can causally steer an answer. Deletion sufficiency asks whether perturbing selected positions changes task performance. We use the term *visual-to-text absorption* only for the bounded empirical phenomenon supported by the first two measurements: semantic content beyond the queried answer is recoverable from non-final text residuals, and source-conditioned causal control changes carrier with depth. This does not mean that the text stream contains a lossless image or that the model can answer arbitrary later questions from it.

Our primary experiment holds the question and its annotated answer fixed while changing unasked facts. Each target image is paired with a disjoint donor image containing at least one donor-only and one target-only object category. We replace the target's visual residual rows with the donor's real activations, either early or at a later readout block. A probe trained only on native target folds is then applied unchanged to the non-final text states. This tests whether the representation follows *which image was presented*, despite matching the annotated answer. It avoids zero and synthetic-noise activations, although hybrid donor/target states need not be fully in distribution.

A separate same-question, opposite-answer experiment patches visual and text carriers at the same block input. Early visual patches control the generated answer; middle-layer non-final text patches become more effective. These behavioral interventions complement the content probe without assuming that a visual position and a final text position have equal causal access to the question. We explicitly test that assumption with changes in prompt order, multiple text positions, several visual pooling rules, and matched blank images.

Finally, symmetric visual- and text-side zero-out measures where each intervention permits recovery. Its strongest conceptual contribution is the ability to reveal *both* orderings: visual-side recovery often comes first on ScienceQA, whereas text-side recovery comes first on open-ended TextVQA. Larger samples and sustained-crossing checks show that no universal ordering holds. A matched Bunny checkpoint comparison likewise separates content transfer, present in both variants, from complete visual-zeroing recovery, observed only in one.

Our empirical contribution pairs an answer-matched content counterfactual with symmetric carrier/deletion analysis. Content transfer, deletion tolerance, and later access can dissociate, requiring distinct measurements rather than a single "absorption layer."

2 MEASUREMENTS AND EXPERIMENTAL DESIGN

2.1 POSITIONS, INTERVENTIONS, AND DISTINCT ESTIMANDS

Let $H_\ell^{\text{in}}, H_\ell^{\text{out}} \in \mathbb{R}^{B\times N\times d}$ denote residual states at the input and output of decoder block ℓ, with **zero-based** indices. V denotes expanded visual-token rows. T^- denotes nonvisual rows strictly after the last visual row, excluding the final nonvisual row; this is our primary text carrier and content readout. F is the final nonvisual row of the prefill sequence, before any answer token is generated. A row replacement modifies a clone of the residual tensor and leaves the sequence length, attention mask, positional IDs, and all other rows unchanged. It applies to prefill, not each generated token.

We distinguish the following operations. **Real-activation content and carrier patches** use decoder-block *input pre-hooks*. **Symmetric recovery zero-out** uses decoder-block *output hooks*, after attention, MLP, and residual updates and before block $\ell + 1$. The separate zero-out *content-survival* control uses a block-input pre-hook and reads after that block, matching the content assay. Thus an

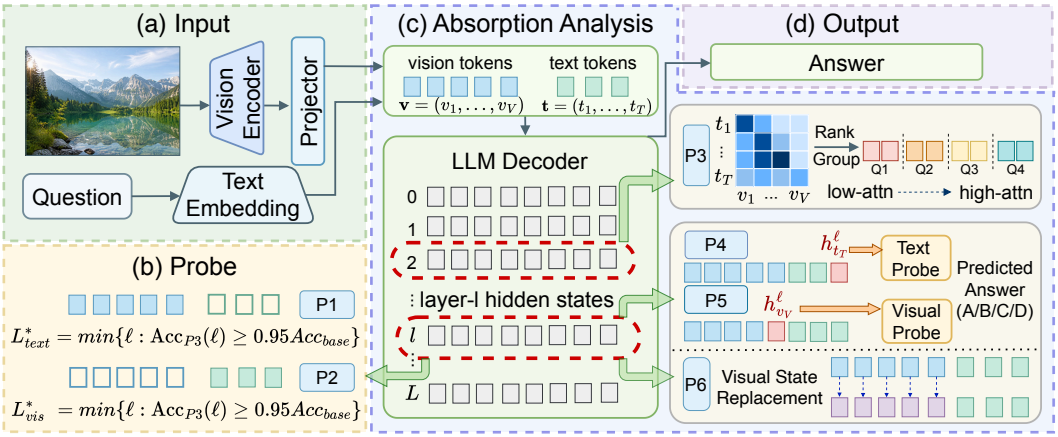

Figure 2: **Original protocol schematic.** P1/P2 denote text/visual zero-out; their recovery scores are A_{T_z}/A_V as defined below. P3 ranks attention quartiles; P4/P5 probe text/visual states; P6 replaces visual activations. The answer-matched content and opposite-answer carrier tests extend this layout. Table 1 specifies current masks, hooks, and evidential roles.

Table 1: **What each assay identifies.** Content is read after the designated block. Recovery assays score generated answers. T_z preserves special/template tokens and the last three prefill positions; it differs from T^-. Sample counts are per checkpoint unless indicated.

Assay	Intervention / readout	Data and rule	Evidential role
Answer-matched donor	V input patch; mean T^- output	COCO, 500 disjoint pairs; fixed native probe	Unasked-content transfer
Opposite-answer carrier	V, T^-, or F input patch	POPE, 500 pairs (Smol: 100); source adoption	Behavioral carrier location
Symmetric recovery	Zero V or T_z at output	ScienceQA/TextVQA, 1,000; first/sustained 95%	Deletion sufficiency
Unasked-content survival	Zero V at input; mean T^- output	COCO, 2,000 images; fixed native probe	Corroborating content retention
Position/order probes	Five text positions; four visual pools	ScienceQA, 1,000; matched blanks	Localization, not modality isolation
Stratified real donor	V output patch	ScienceQA 500; POPE 1,000; same/opposite label	Answer-sharing control

identically numbered layer across assays need not denote the same intervention boundary. Appendix A gives the exact operations.

2.2 Answer-matched counterfactual for unasked content

We select 500 target/donor pairs from COCO object annotations (Lin et al., 2014). Within a checkpoint, all 1,000 images are distinct and the target/donor sets are disjoint. Each pair has a byte-identical category question and the same annotated yes/no answer, but differs in unasked categories in *both* directions. Labels fix the pairs before probe fitting; subsequent geometry screening enforces row alignment without selecting for the desired probe outcome. The annotated answer is held fixed, rather than requiring identical generated answers. Pair sets can differ between architectures; the Bunny comparison uses exactly the same pairs.

We run target-native, donor-native, early donor-visual patch, late donor-visual patch, and target-self sham conditions. The text feature is the mean over T^- at the output of the designated readout block. In five folds, we project features to 256 dimensions with a fixed data-independent Gaussian map, fit a scaler and ridge probe ($\alpha = 10$) on the 400 native training targets, and apply the same fitted transformation and probe to every condition for the 100 held-out pairs. In this primary analysis,

donor and intervention features never fit the probe. The queried category is excluded from evaluation separately for each item.

For category c, retain only pairs for which c is unasked and target/donor labels disagree. With a fixed set of condition scores, compute average precision (AP) against target labels and, separately, donor labels. Macro-average over categories having both labels represented. The primary preference and interaction are

$$P_k = \mathrm{mAP}(s_k, y^{\mathrm{target}}) - \mathrm{mAP}(s_k, y^{\mathrm{donor}}), \tag{1}$$

$$I = P_{\mathrm{late}} - P_{\mathrm{early}}. \tag{2}$$

Positive I denotes a targetward change, not positive absolute late preference or a fraction of image information preserved. With target prevalence π_c on discordant cells, constant scores yield $P_c = 2\pi_c - 1$; zero is therefore not a general no-information baseline. We report native endpoints and a sensitivity that gives the two labels equal total weight within each category. Intervals use 5,000 paired resamples, conditional on fitted probes and folds.

The early block is L0 except for Qwen3-VL, where L3 follows all three DeepStack decoder reinjections. The later blocks were chosen from preceding recovery measurements; they are fixed within the donor assay, not fitted to maximize I. SmolVLM2 is evaluated at both L14 and L22. Fail-closed checks require aligned visual rows, a single effective patch/readout at the intended hook, exact insertion, no target/donor image overlap, and exact target-self-sham features.

2.3 Causal carrier patches and symmetric recovery

For carrier localization, source and target images have the same POPE question (Li et al., 2023) but opposite answers, and both native runs must answer correctly. Each of InternVL3.5 and Qwen2.5 uses 500 balanced-direction pairs. At each tested block input, we replace V, T^-, or F with source activations and measure *source-answer adoption*. The primary contrast is adoption under T^- minus adoption under V. Depth and hook are matched; attention access and causal history are not. Question-cluster bootstrap preserves pairing across layers and carriers (72/74 clusters for InternVL/Qwen). Target-self and full-source controls verify that the patch operation has the intended semantics.

For symmetric recovery, T_z comprises eligible nonvisual content rows, preserving registered special/template tokens and the final three prefill positions. Thus "symmetric" describes the matched operation and block boundary, not equal token counts or removal of all text. Let $A_m(\ell)$ be the mean task score after zeroing carrier $m \in \{V, T_z\}$ and A_m^{native} its matched baseline. On the tested grid $\mathcal{G}_m$, define

$$\ell_m^{95} = \min\{\ell \in \mathcal{G}_m : A_m(\ell) \geq .95 A_m^{\mathrm{native}}\}, \tag{3}$$

$$\ell_m^{95,\mathrm{sust}} = \min\{\ell \in \mathcal{G}_m : \forall j \in \mathcal{G}_m,\ j \geq \ell,\ A_m(j) \geq .95 A_m^{\mathrm{native}}\}. \tag{4}$$

The gap is $\Delta^{95} = \ell_{T_z}^{95} - \ell_V^{95}$; positive means visual-side recovery occurs earlier. A missing crossing is undefined, not the last layer. "Sustained" only covers tested grid points. Zeroing residual rows does not remove their sequence positions or explicitly erase all visual KV routes.

2.4 Evaluation scope and uncertainty

The ScienceQA (Lu et al., 2022) recovery cohort contains InternVL3.5-8B, InternVL3-14B, LLaVA-1.5-7B, LLaVA-NeXT-8B, Qwen2.5-VL-7B, Phi-3.5-Vision, and SmolVLM2-2B. Its 1,000 questions form 913 decoded-pixel-identical image clusters. Qwen3-VL-4B is an additional recovery replication. The *seven-checkpoint donor cohort replaces LLaVA-NeXT with Qwen3*; we do not pool these denominators. TextVQA (Singh et al., 2019) uses 1,000 exactly paired questions from 641 images and its native fractional answer score. Recovery inference resamples image clusters, retains all questions in each cluster, and recomputes baselines and crossings in each of 5,000 paired draws. Intervals measure cluster-sampling uncertainty, not uncertainty in hook choice, thresholds, checkpoints, or layer grids. Bootstrap sign frequencies are not p-values.

3 Content beyond the current answer enters text states

The answer-matched experiment yields positive late-minus-early interactions in all seven checkpoints (Figure 3; Table 2), ranging from $+.407$ for LLaVA-1.5 to $+.744$ for SmolVLM2 at L22. All 5,000

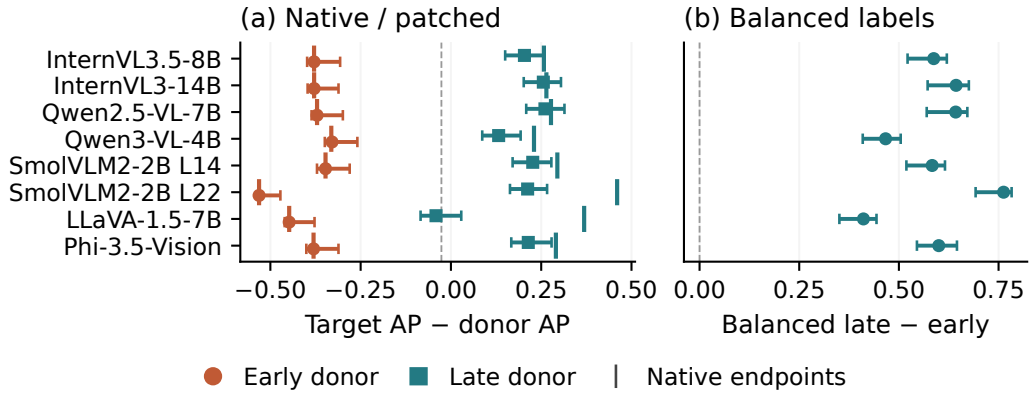

Figure 3: **Donor endpoints and class-balance sensitivity.** Queries and annotated answers match within 500 disjoint pairs per run. Left: early/late patches (95% intervals), native donor/target endpoints (orange/teal ticks), and the constant-score baseline $-.0263$ (dashed). Right: interactions remain positive after assigning equal total mass to both labels per category. LLaVA still lacks a positive late endpoint. Smol readouts are separate fixed runs.

Table 2: **Complete answer-matched donor results.** P is target-label AP minus donor-label AP; $I = P_{\text{late}} - P_{\text{early}}$. All entries use 500 disjoint pairs, five folds, and 5,000 paired bootstrap draws. Early/late are zero-based block-input indices; both conditions are read at the late block's output. The two Smol readouts are separate sensitivity runs within one checkpoint.

Checkpoint	Early/late	P_{early}	P_{late}	I	95% CI for I
InternVL3.5-8B	0/18	-0.379	+0.204	+0.583	[0.507, 0.605]
InternVL3-14B	0/27	-0.379	+0.256	+0.635	[0.556, 0.661]
Qwen2.5-VL-7B	0/17	-0.371	+0.260	+0.631	[0.553, 0.658]
Qwen3-VL-4B	3/18	-0.329	+0.132	+0.461	[0.397, 0.494]
SmolVLM2-2B	0/14	-0.347	+0.227	+0.574	[0.501, 0.600]
SmolVLM2-2B	0/22	-0.531	+0.212	+0.744	[0.672, 0.765]
LLaVA-1.5-7B	0/16	-0.448	-0.041	+0.407	[0.344, 0.436]
Phi-3.5-Vision	0/18	-0.380	+0.214	+0.595	[0.531, 0.630]

bootstrap interactions are positive in each run. Native endpoints distinguish donor-content readout from early probe degradation. The combined result is inconsistent with encoding only the matched annotated answer; the primary cohort does not hold generated answers or confidence fixed.

Early donor vision changes the image-specific content recoverable from non-final text, whereas late replacement preserves more target-specific content (Table 2). All 500 pairs per run pass operation audits, including exactly matching self-sham features.

Two exceptions constrain interpretation. LLaVA's late preference is $-.041$ $[-.084, .029]$: its positive interaction identifies a targetward *change*, not a positive late endpoint. Smol shows transfer at both L14 and L22 despite a zero sustained ScienceQA recovery gap. A null deletion-order gap therefore does not imply absent content transfer. The measured content is annotated object presence, not all visual attributes, spatial relations, or pixels.

Native endpoints and prevalence controls. All runs evaluate 75 eligible categories, with target prevalence .20–.83 and constant-score macro preference $-.0263$. Negative early preference alone does not identify donor transfer. Every L0 patch instead exactly matches donor-native features and scores; Qwen3's L3 preference is $-.329$ versus donor-native $-.331$. InternVL donor-native target/donor AP is $.408/.787$, versus target-native $.722/.465$. Equal-label-mass AP retains positive interactions in all seven checkpoints ($+.411$–$+.763$), with intervals excluding zero. Refitting over

Table 3: **Unasked content survives late visual zeroing.** Macro-AP on 2,000 unique images per checkpoint. Δ is the observed late-minus-early difference; intervals use paired bootstrap. All conditions share a native-trained probe.

Readout	Native	Early	Late	Δ	95% CI
InternVL L18	0.3831	0.0511	0.3356	+0.2845	[0.2780, 0.3055]
Qwen L17	0.3463	0.0447	0.3388	+0.2941	[0.2873, 0.3161]
Smol L14	0.3295	0.0452	0.1913	+0.1461	[0.1392, 0.1625]
Smol L22	0.6482	0.0502	0.4040	+0.3539	[0.3430, 0.3756]

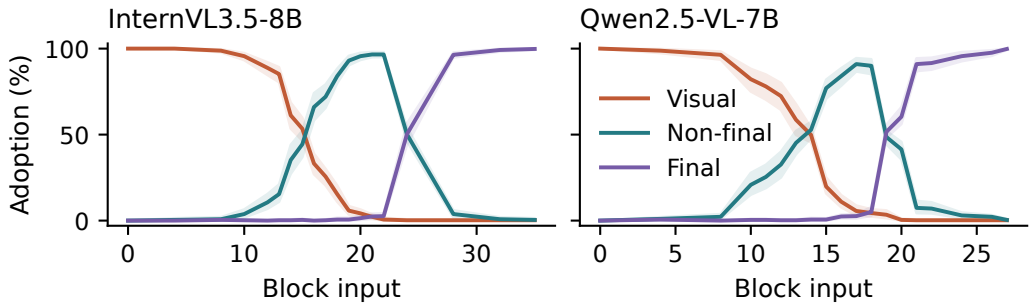

Figure 4: **Behavioral carrier handoff with real activations.** Same-question, opposite-answer patches at the same block input; 500 balanced pairs per checkpoint. Shading denotes paired question-cluster 95% intervals. Non-final text overtakes visual states in the middle layers; the final prefill position dominates later. Carrier histories and available attention context remain different.

five split/projection seeds and two native training sources (target or donor) gives 100 positive interactions across ten readout/checkpoint runs, minimum $+.379$. This changes probe training, not patch direction. Fixed-support and independent train/test-refitting sensitivities also preserve positive interactions (Appendix B).

Matching actual native answers. Greedy native generation on the same 500 pairs leaves 444 InternVL and 422 Qwen pairs with both answers correct. Their fixed-probe interactions remain $+.579$ $[.507, .609]$ and $+.641$ $[.561, .677]$. Every new native text feature exactly matches its cached counterpart. This sensitivity selects on native answers alone, not intervention outcomes; it matches decoded yes/no labels, not confidence (Appendix B).

Independent content-survival control. On 2,000 unique COCO images, a fixed native-trained probe predicts unasked categories from T^- (Table 3). Late zeroing retains 87.6/97.8% of native AP for InternVL/Qwen, not of the whole image; both native-minus-late intervals exclude zero. Smol replicates substantial but incomplete survival. This assay corroborates the real-donor result (Appendix D).

4 Causal control shifts between carriers

At L0, source-answer adoption is 100/0% under visual/non-final-text patches in both principal checkpoints (Figure 4). By the middle layers, non-final text dominates (Table 4); both answer directions agree. A separate 100-pair Smol assay changes from 100/0% at L0 to 6/94% at L14. Text therefore becomes a functional carrier of source-conditioned control, beyond a passive correlate at the final position.

At the last tested block, final-only patches exceed non-final text by 99.4 points in both principal checkpoints, motivating a separate F carrier. The 500-pair sets reuse images across questions: InternVL/Qwen contain 444/417 unique images. A sensitivity retaining image-disjoint pairs in manifest order gives 175/166 pairs and middle-layer text advantages of 67.4 [57.7,76.2] / 84.3

Table 4: **Middle-layer carrier handoff.** Source-answer adoption (%) under visual (V), non-final (T^-), and final-text (F) patches. Each model uses 500 pairs; text-lead intervals resample 72/74 question clusters for InternVL/Qwen.

Checkpoint	Block	V	T^-	F	Text lead [95% CI]
InternVL3.5-8B	18	15.4	84.0	0.6	+68.6 [61.3, 75.5]
Qwen2.5-VL-7B	17	5.6	91.0	2.6	+85.4 [79.0, 90.5]

Table 5: **ScienceQA recovery, 1,000 questions / 913 image clusters.** V/T lists visual/text recovery blocks; intervals are for the text-minus-visual gap. First-crossing intervals are positive in 6/7 checkpoints; sustained intervals are positive in only 3/7. Both LLaVA intervals include zero under sustained recovery, Phi reverses, and Smol is equal.

	First recovery		Sustained recovery	
Checkpoint	V/T	Gap 95% CI	V/T	Gap 95% CI
InternVL3.5-8B	18/24	[6, 6]	18/24	[6, 6]
InternVL3-14B	27/36	[9, 9]	27/36	[7, 9]
LLaVA-1.5-7B	16/18	[2, 2]	16/18	[-1, 2]
LLaVA-NeXT-8B	14/17	[3, 7]	26/17	[-9, 3]
Qwen2.5-VL-7B	17/22	[4, 5]	17/22	[4, 5]
Phi-3.5-Vision	15/20	[4, 5]	29/20	[-9, -9]
SmolVLM2-2B	17/17	[0, 3]	22/22	[0, 0]

[77.6,90.4] points. Appendix C reports the subset selection and question-cluster inference. All effects remain conditional on correct, aligned native pairs.

Controls for the replacement operation. Self-replacement of all, visual, or primary text rows preserves the target answer; replacing all rows with source activations reproduces the source answer. Each control passes all 3,000 audited cells per principal model (six designated blocks, 500 pairs), with 250 pairs per answer direction. Question-cluster resampling keeps carrier outcomes paired. These checks validate execution without equating causal histories; Appendix C reports crossover estimates and Smol controls.

Content residence does not guarantee later access. A stronger test appends a different question after a native image/question prefix and a shared gold first answer, then blocks original visual KV access for every second-question query at every layer. Qwen2.5 achieves 50.1% across 1,000 observations from 500 eligible pairs; only 1/500 pairs have both answers correct. Thus decodable semantic variables do not guarantee successful new answers under this protocol. Appendix D specifies cache retention, masking, eligibility, and decoding. This negative result does not establish impossibility for other readouts or query interfaces.

5 Content transfer and deletion sufficiency dissociate

The enlarged ScienceQA evaluation supports a common but non-universal first-crossing asymmetry (Table 5). Sustained recovery exposes behavior that a single threshold crossing hides. LLaVA-NeXT's visual crossing changes from L14 to L26, and Phi's from L15 to L29, following late dips. Smol's first and sustained point gaps are both zero. Qwen3, excluded from the original seven-model denominator, gives L18/L28 and gap CI [10,11] under both definitions. These bootstrap-gap intervals differ from the exploratory $n = 200$ Holm–McNemar tests reported in Appendix F.

Open-ended OCR reverses the direction. Table 6 reports both models' recovery points and gap intervals. InternVL's first point gap is -13 layers (bootstrap median -12); its L31 text dip delays sustained text recovery to L32. Qwen reverses under both definitions. At L20, its visual/text scores

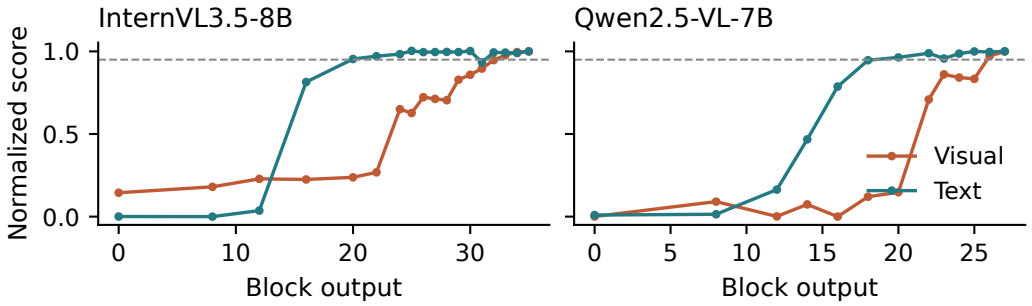

Figure 5: **Open-ended OCR reverses first-recovery ordering.** Native fractional TextVQA scores normalized by each matched baseline, 1,000 questions per model. The dashed line is 95% recovery. Visual zero-out remains damaging after text-side recovery; the late InternVL text dip also shows why first and sustained definitions must be distinguished.

Table 6: **TextVQA reverses first-recovery ordering.** 1,000 questions / 641 images per model, with native fractional scoring. V/T denotes recovery blocks; 95% intervals are for the text-minus-visual gap, using paired image-cluster bootstrap.

		First recovery		Sustained recovery	
Checkpoint	Native	V/T	Gap CI	V/T	Gap CI
InternVL3.5-8B	0.6507	33/20	[-13, -9]	33/32	[-13, 0]
Qwen2.5-VL-7B	0.8700	26/20	[-8, -4]	26/20	[-8, -2]

are .1283/.8383 (14.7/96.4% of baseline). First-gap intervals remain negative at 90% and 98% thresholds (Appendix F), confirming task-dependent deletion tolerance.

Late recovery under opposite-answer donors. We explicitly distinguish random, same-output-label, and different-output-label visual donors. On 500 ScienceQA items per run, all three strategies recover at L14 in Qwen and at L18 in InternVL. The cleaner semantic POPE comparison uses 1,000 aligned targets, unique different-image donors, and same versus strictly opposite yes/no labels. Both conditions first and sustainably recover at L18, with CI [18,18]. L18 accuracies are 86.9/87.0%; their paired difference is $+0.1$ points with image-cluster CI $[-0.9, 1.1]$. Holm-adjusted McNemar $p = 1.0$ supports no detected difference, not statistical equivalence. All 20,000 operation-level checks pass. Appendix G discloses and corrects the original deterministic-offset pairing description.

6 Position and Architecture Constrain the Interpretation

Probes localize information but do not isolate modality. In the standard templates, image tokens precede the question. A final text token can attend to the question and image prefix; visual tokens cannot attend to the later question. On 1,000 InternVL ScienceQA items, L24 probe accuracies at first/25%/middle/75%/last text positions are 39.5/36.3/46.5/64.0/95.7% (Appendix E, Figure 6). This is a large positional gradient. Question-first prompts increase absolute visual decodability, and we validate that order separately on 200 generation items (90.5% versus 96.0% standard accuracy for InternVL; 87.0% versus 91.5% for Qwen).

Matched blanks change the interpretation again. Under question-first order, InternVL's largest all-layer real-minus-blank visual gain is 2.7 points across last/mean/max pooling, or 2.0 with learned attention pooling; final text gains 23.2 points at L24. Qwen's visual last/mean/max/learned maxima are 12.8/8.0/3.7/-1.3 points, versus a 16.3-point final-text maximum. Architecture and causal position both matter. We retain these probes as localization diagnostics and use the donor and carrier assays for the transfer claim. We do not use a fixed "25% chance" baseline for variable-choice tasks.

The Bunny pair changes deletion behavior, not transfer on/off. Bunny v1.0/v1.1 (He et al., 2024) share the backbone/encoder family and reported training data, with different aggregation designs. Identical 500-pair L0/L12 donor comparisons give interactions $+.512$ $[.438, .538]$ / $+.565$ $[.493, .598]$, with paired difference $+.0528$ $[.0024, .1139]$. The paired comparison uses identical bootstrap multiplicities for both checkpoints. A separate 250/250 train/test split reverses the point difference to $-.0297$, with 95% refitting-resampling range $[-.0674, .1032]$. This stability range does not replace the primary confidence interval (Appendix B): both variants transfer unasked content, but their relative ordering is not robust. In exploratory ScienceQA deletion ($n = 200$), baselines are 80.0/80.5%, yet only v1.0 recovers from visual zeroing. This tested pair links aggregation/checkpoint differences to transfer profile and deletion sufficiency. Separate training prevents isolating one parameter's effect or claiming that absorption is universally switchable.

Reading the dissociations together. Smol and Bunny show content transfer despite a null recovery gap or absent complete visual-zeroing recovery. TextVQA shows that visual states can remain necessary for another task; the second-question test separates probe access from native decoding. These dissociations require distinct content, carrier, and task-performance measurements. A positive donor interaction alone does not identify a safe pruning schedule, which requires task-specific validation.

7 Related work

Visual-token pruning methods such as FastV (Chen et al., 2024a) and VisionZip (Yang et al., 2025) reduce inference cost by identifying dispensable image features. Work on redundancy and relevance studies task-dependent information flow and truncation (Zhang et al., 2025a). Our purpose is to distinguish the explanations compatible with late visual tolerance; we do not propose a new pruning algorithm or claim a speedup from zeroing rows.

Multimodal causal tracing (Basu et al., 2024) and visual processing analyses (Neo et al., 2024; Yu & Lee, 2025) motivate studying internal positions. Zhang et al. (2025b) use attention knockout to identify general image-to-question transfer, later task-relevant transfer, and question-to-final-position flow. Our contribution is not the discovery of that staged pathway: real-state replacement adds a matched-annotated-answer counterfactual for unasked facts, while symmetric recovery tests whether transfer entails deletion tolerance. Kaduri et al. (2025) show that query tokens can carry global image information and support descriptive generation with direct image access blocked. Their compressed re-prompting setting and our strict visual-KV isolation retain different contexts and use different tasks; our negative reuse result is protocol-specific. Probes and representation comparisons remain descriptive tools (Belrose et al., 2023; Kornblith et al., 2019); neither alone identifies a causal carrier.

8 Limitations and conclusion

The measured content consists of annotated object-presence variables and can include category correlations; it is not a complete image representation. Donor selection, geometric alignment, and correct-baseline screening in the carrier assay restrict the populations to which the results apply. Real donor rows avoid synthetic zero activations, but the resulting hybrid computation can still differ from native execution. Probes provide a particular readout; primary bootstrap intervals condition on training folds and fitted weights. Matching hook depth does not equate causal histories. Recovery depends on preserved text positions, tested grids, thresholds, and task scoring. Architecture comparisons are observational outside the limited Bunny pair, whose separately trained checkpoints do not eliminate every training confound. The study concerns the tested decoder-only VLMs and does not establish the same behavior in recurrent or cross-attention-only multimodal decoders. Broader claims require attribute/relation interventions and tests of alternative query interfaces under matched retained context.

Our evidence supports a bounded form of visual-to-text absorption: unasked image-derived semantic variables are decodable and causally transferable into non-final text states, while source-conditioned answer control shifts toward text with depth. It does not support universal visual-before-text deletion ordering, lossless image storage, or unrestricted re-queryability. Treating content transfer, deletion sufficiency, and behavioral access separately makes both the positive evidence and the counterexamples interpretable, with each inference tied to its measured intervention and readout.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, interpretation of saved results, implementation checks, and preparation of statistical-analysis, native-answer-screening, figure, and build code. These uses include methodological feedback, probe refitting from cached features, and sensitivity analyses from frozen predictions. Native-answer screening was additionally executed for the robustness checks described in the paper. The authors are responsible for the final text, claims, code, and results, including material prepared with AI assistance.

ETHICS STATEMENT

This study analyzes existing VLM checkpoints on established image datasets. It does not introduce a deployment system or a new collection of human-subject data. The measured redundancy should not be treated as a guarantee of safe token removal on untested inputs or tasks. We distinguish annotated semantic content from complete image understanding, report negative results, and disclose the correction to the earlier distractor-pairing description in Appendix G.

REPRODUCIBILITY STATEMENT

Section 2 defines interventions, carrier masks, estimands, and resampling units. Appendices A–H specify implementation, pairing, probing, grids, controls, and source results. Tables and figures are generated from saved outputs with source checksums and audit checks. The statistical supplement reproduces donor and carrier metrics; the inference supplement supplies intervention sources, adapters, portable manifests, environment/checkpoint records and exact projected inputs for independent probe refitting. Historical weight revisions that were not recorded remain explicitly unknown.

REFERENCES

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *NeurIPS*, 2024.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

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 *ECCV*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *CVPR*, pp. 14549–14558, 2025. URL https://arxiv.org/abs/2411.17491.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pp. 8317–8326, 2019. URL https://arxiv.org/abs/1904.08920.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*, 2025.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. *arXiv preprint arXiv:2508.20279*, 2025.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *CVPR*, pp. 19781–19791, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

A IMPLEMENTATION AND INTERVENTION BOUNDARIES

Residual indexing. Visual rows are resolved after architecture-specific expansion of image placeholders, not by assuming that the tokenizer's unexpanded input length equals the decoder length. The residual tensor has shape $[B, N, d]$. In content/carrier assays, source and target must have matching visual layouts, sequence geometry, and text-token identities at the relevant positions. The patch copies all selected rows. Source states are captured from a native source run at the same decoder block boundary used for insertion in the target. A one-time prefill guard prevents reapplying the intervention during autoregressive decoding.

```
# decoder-block input:  real donor patch
H = target_input.clone()
H[:, rows, :]  = source_input[:, rows, :]
return H   # all other inputs unchanged
# decoder-block output:  symmetric zero-out
H = block_output.hidden.clone()
H[:, rows, :]  = 0
return replace_hidden(block_output, H)
# content readout after recovery/readout block
z = H_out[:, post_image_nonfinal_text, :].mean(dim=1)
```

The implementation preserves tuple/dataclass output structure where required. Zero-out changes residual values, not token count. It does not itself remove visual keys/values from every attention cache. In the separate unasked-content survival assay, both early and late zeroing occur at *block input*, before the selected block computes the content readout; this must not be confused with the block-output recovery curves.

Three different text masks. The content probe and primary carrier patch use T^-: all nonvisual positions after the last visual position, excluding the last nonvisual token. This excludes the final answer-generating prefill state, not an already generated answer. F contains that final prefill state alone. Recovery uses the script's `question_only` mode: nonvisual, non-special positions eligible under the model adapter, excluding the final three sequence positions. The mode name does not guarantee that every retained role delimiter is a single registered special token, nor does it denote a linguistically parsed question span. It is a reproducible token-ID/mask operation. These masks have different purposes and must not be substituted for one another.

Checkpoints and multimodal entry. The recovery set covers InternVL3.5-8B and InternVL3-14B; LLaVA-1.5-7B and LLaVA-NeXT-8B; Qwen2.5-VL-7B; Phi-3.5-Vision; and SmolVLM2-2B. The donor set substitutes Qwen3-VL-4B for LLaVA-NeXT. The additional matched checkpoint pair is Bunny-Llama-3-8B v1.0/v1.1. These are checkpoint-level comparisons, not seven independent architectural design interventions. The model families build on visual instruction tuning and multimodal alignment (Liu et al., 2024; Chen et al., 2024b; Wang et al., 2024; Marafioti et al., 2025; He et al., 2024).

For Qwen3, three visual embeddings are reinjected at decoder blocks 0, 1, and 2; the early donor patch is at block-input L3, after these additions. Its vision-encoder feature indexes (5, 11, 17) are not decoder reinjection indexes. SmolVLM2 provides a one-shot visual-entry comparison without those decoder reinjections. The strict-donor runs use fixed image geometry (square letterboxing to 448 pixels) to permit complete aligned row substitution; conclusions are conditional on that preprocessing.

Token order. All recovery/donor models use their native image-first templates. The table abstracts over special tokens and image expansion; it is not a literal serialization of each chat template.

Checkpoint family	Standard causal order	Question-first probe control
InternVL3/3.5	Prefix, image, question, assistant prefix	InternVL3.5 only
LLaVA-1.5/NeXT	Prefix, image, question, assistant prefix	Not evaluated
Qwen2.5/3-VL	Prefix, image, question, assistant prefix	Qwen2.5 only
Phi-3.5-Vision	Prefix, image, question, assistant prefix	Not evaluated
SmolVLM2	Prefix, image, question, assistant prefix	Not evaluated
Bunny v1.0/v1.1	Prefix, image, question, assistant prefix	Not evaluated

The question-first control places the question before the image block. It changes the causal context and possibly template familiarity, so its generation-validity check is reported separately from probe accuracy. It does not make every position's causal history identical.

B STRICT DONOR CONSTRUCTION AND INFERENCE

COCO category-presence labels define the queried category and the unasked labels. Pair construction requires different images, a matching queried answer, and at least one target-only and one donor-only unasked category. A focal category in each direction is fixed before folds for secondary score diagnostics; the primary AP interaction uses *all* eligible discordant unasked cells, not only those two focal categories. Geometry screening rejects incompatible layouts rather than truncating the patch to a partial overlap. The number of rows patched is verified live.

The 500 target images and 500 donor images are globally disjoint within each checkpoint run, and no target or donor repeats within its set. Five shuffled folds (seed 42) contain 400 training targets and 100 held-out pairs each. The Gaussian projection is fixed independently of labels; scaling and ridge fitting use native training targets only. The ridge scores are ranking scores, not calibrated probabilities. All probe evaluation excludes the category actually asked by each prompt. Training fits the category vector on native targets; the masking concerns the reported unasked evaluation.

The AP comparison uses the *same* score vector against target and donor labels. A category is eligible if its discordant subset contains both labels. Each paired bootstrap draw applies the same multiplicities to early, late, and native conditions. AP is recomputed with tie-aware weighted ranking, then macro-averaged over categories with support in that draw. Consequently percentile intervals and bootstrap means can be asymmetric around the observed point estimate. The intervals do not refit the ridge model and do not quantify probe-training randomness or generalization to all possible pair-selection schemes.

Complete point estimates, early/late block indices, and interaction intervals appear in main-text Table 2. The two Smol readouts are separate fixed sensitivity runs within a single checkpoint, not additional architectures or a matched-pair estimate of the effect of changing readout depth.

Every run passes the complete-pair audit (500 expected, 500 completed), live row-alignment and exact-insertion checks, one-readout checks, and native/self-sham identity. An all-positive set of bootstrap draws is a resampling sign frequency, not a permutation-test significance level. The interaction and native endpoints support content beyond the matched annotated answer; generated answers and confidence are not matched. They do not imply exclusive storage in text or independent grounding of every correlated category.

Absolute endpoints and the AP baseline. On an eligible category's discordant subset, donor labels complement target labels. Constant scores have target AP π_c and donor AP $1-\pi_c$, yielding preference $2\pi_c - 1$. All ten runs share 75 eligible-category support profiles; target prevalence ranges from .20 to .833, with macro constant-score preference $-.02634$. This is an analytic constant-score reference, not an exact finite-sample permutation null. Subtracting it from both early and late leaves the primary interaction unchanged. Table 7 exposes both APs for every condition, rather than interpreting a negative preference alone as donor alignment. At L0 the early-patch features and scores exactly match donor-native values in every run. Qwen3's L3 intervention is not identical to donor-native; its respective preferences are $-.3293$ and $-.3313$. The supplementary archive provides category supports and all held-out scores.

Table 7: **Absolute discordant-category endpoints.** AP uses the same held-out score vector against target or donor labels; P is their difference. Native conditions use the same readout block and target-native-trained probe as interventions.

Checkpoint	Condition	Target AP	Donor AP	P
Bunny v1.0	Target native	0.720	0.463	+0.257
Bunny v1.0	Donor native	0.438	0.746	-0.308
Bunny v1.0	Early patch	0.438	0.746	-0.308
Bunny v1.0	Late patch	0.686	0.482	+0.204
Bunny v1.1	Target native	0.734	0.454	+0.280
Bunny v1.1	Donor native	0.418	0.767	-0.349
Bunny v1.1	Early patch	0.418	0.767	-0.349
Bunny v1.1	Late patch	0.691	0.475	+0.216
InternVL3-14B	Target native	0.722	0.458	+0.264
InternVL3-14B	Donor native	0.408	0.787	-0.379
InternVL3-14B	Early patch	0.408	0.787	-0.379
InternVL3-14B	Late patch	0.716	0.461	+0.256
InternVL3.5-8B	Target native	0.722	0.465	+0.257
InternVL3.5-8B	Donor native	0.408	0.787	-0.379
InternVL3.5-8B	Early patch	0.408	0.787	-0.379
InternVL3.5-8B	Late patch	0.690	0.487	+0.204
LLaVA-1.5-7B	Target native	0.792	0.423	+0.369
LLaVA-1.5-7B	Donor native	0.382	0.830	-0.448
LLaVA-1.5-7B	Early patch	0.382	0.830	-0.448
LLaVA-1.5-7B	Late patch	0.553	0.594	-0.041
Phi-3.5-Vision	Target native	0.738	0.447	+0.291
Phi-3.5-Vision	Donor native	0.410	0.791	-0.380
Phi-3.5-Vision	Early patch	0.410	0.791	-0.380
Phi-3.5-Vision	Late patch	0.688	0.474	+0.214
Qwen2.5-VL-7B	Target native	0.731	0.454	+0.277
Qwen2.5-VL-7B	Donor native	0.409	0.780	-0.371
Qwen2.5-VL-7B	Early patch	0.409	0.780	-0.371
Qwen2.5-VL-7B	Late patch	0.721	0.461	+0.260
Qwen3-VL-4B	Target native	0.702	0.472	+0.230
Qwen3-VL-4B	Donor native	0.425	0.757	-0.331
Qwen3-VL-4B	Early patch	0.426	0.755	-0.329
Qwen3-VL-4B	Late patch	0.642	0.510	+0.132
SmolVLM2-2B L14	Target native	0.741	0.446	+0.295
SmolVLM2-2B L14	Donor native	0.425	0.772	-0.347
SmolVLM2-2B L14	Early patch	0.425	0.772	-0.347
SmolVLM2-2B L14	Late patch	0.699	0.473	+0.227
SmolVLM2-2B L22	Target native	0.853	0.393	+0.460
SmolVLM2-2B L22	Donor native	0.359	0.890	-0.531
SmolVLM2-2B L22	Early patch	0.359	0.890	-0.531
SmolVLM2-2B L22	Late patch	0.690	0.477	+0.212

Equal-label-mass sensitivity. For each eligible category and paired bootstrap draw, assign every target-positive cell weight $1/n_+$ and every target-negative cell weight $1/n_-$. The same weights also balance the complementary donor labels. Compute tie-aware weighted AP against each label vector and their difference, then macro-average over categories with both labels present in that draw. Constant scores now have preference zero. This sensitivity uses the frozen predictions without refitting any probe; 5,000 paired draws use seed 42. All ten interaction intervals remain positive (Table 8). For LLaVA, the balanced late endpoint is $-.025$, so the sensitivity does not establish positive late alignment. Its original late-minus-constant preference is $-.015$ with CI $[-.061, .041]$.

Table 8: **Class-balanced AP preference.** Equal total mass is assigned to positive and negative labels within each eligible category and draw. Original-prevalence estimates remain the primary results.

Checkpoint	Early	Late	Interaction	95% CI
Bunny v1.0	-0.294	+0.226	+0.520	[0.451, 0.551]
Bunny v1.1	-0.334	+0.244	+0.578	[0.512, 0.615]
InternVL3-14B	-0.363	+0.280	+0.643	[0.572, 0.675]
InternVL3.5-8B	-0.366	+0.221	+0.587	[0.522, 0.620]
LLaVA-1.5-7B	-0.436	-0.025	+0.411	[0.351, 0.444]
Phi-3.5-Vision	-0.362	+0.238	+0.600	[0.545, 0.646]
Qwen2.5-VL-7B	-0.359	+0.283	+0.642	[0.570, 0.672]
Qwen3-VL-4B	-0.312	+0.155	+0.467	[0.409, 0.505]
SmolVLM2-2B L14	-0.333	+0.250	+0.583	[0.519, 0.616]
SmolVLM2-2B L22	-0.524	+0.238	+0.763	[0.692, 0.783]

Probe refitting and native training-source control. Using the frozen features, we repeat the five-fold pipeline with seeds 42–46, jointly changing the shuffled pair folds and the Gaussian projection (seed plus 1009). For each seed, one probe is trained on native target features/labels and another on native donor features/labels, with the same 256-dimensional projection, training-only scaler, and ridge $\alpha = 10$. Holding out a pair excludes both of its images from training. Each fitted probe is applied unchanged to target-native, donor-native, and both intervention conditions. No intervention feature or held-out label fits the scaler or probe. Seed 42 target-trained predictions reproduce the original cache exactly in all ten runs.

All $10 \times 5 \times 2 = 100$ interaction point estimates remain positive under both original-prevalence and equal-label-mass AP. Table 9 reports the original-prevalence range across seeds; a minimum/maximum is a descriptive range, not a confidence interval. For the seed 42 donor-trained control, all ten conditional 95% intervals from 5,000 paired draws exclude zero (balanced AP does too). These intervals still condition on that fitted probe. The exercise samples fitting variability but is not a nested bootstrap over probe training, and the donor-trained control does not reverse the cached target-receives-donor intervention. It addresses dependence on the probe's native training population, rather than supplying a reciprocal causal experiment.

Table 9: **Five-seed probe-refit sensitivity.** Interactions for target-native or donor-native training; each bracketed seed range is a minimum/maximum, not an interval estimate. The last two columns use only seed 42 with donor-native training.

Checkpoint	Target seed range	Donor seed range	I_{42}	Conditional 95% CI
Bunny v1.0	[0.512, 0.565]	[0.521, 0.561]	+0.521	[0.456, 0.557]
Bunny v1.1	[0.546, 0.580]	[0.546, 0.597]	+0.546	[0.482, 0.581]
InternVL3-14B	[0.591, 0.654]	[0.593, 0.668]	+0.647	[0.580, 0.683]
InternVL3.5-8B	[0.560, 0.610]	[0.590, 0.627]	+0.590	[0.528, 0.625]
LLaVA-1.5-7B	[0.379, 0.445]	[0.381, 0.448]	+0.407	[0.343, 0.436]
Phi-3.5-Vision	[0.512, 0.595]	[0.564, 0.615]	+0.596	[0.518, 0.621]
Qwen2.5-VL-7B	[0.595, 0.631]	[0.613, 0.668]	+0.613	[0.537, 0.637]
Qwen3-VL-4B	[0.461, 0.486]	[0.486, 0.503]	+0.503	[0.429, 0.528]
SmolVLM2-2B L14	[0.467, 0.574]	[0.518, 0.562]	+0.524	[0.461, 0.562]
SmolVLM2-2B L22	[0.713, 0.762]	[0.749, 0.790]	+0.790	[0.708, 0.799]

Fixed category-support sensitivity. Fifteen of the 75 eligible categories have fewer than five discordant examples of one label; thirty have fewer than ten. We therefore fix category sets using minimum per-label support of 5 or 10 in the original 500 pairs, leaving 60 or 45 categories. Each set remains fixed throughout 5,000 paired bootstrap draws (seed 42). If a draw contains only one

label for a category, its late-minus-early AP interaction is zero: AP is 1 for the all-positive label and 0 for the all-negative label under either ranking. This convention preserves the fixed macro-average denominator rather than dropping categories. Table 10 reports the original-prevalence metric; the supplement also supplies equal-label-mass results. All intervals remain positive under both metrics. These are support-restricted sensitivities, not replacements for the primary estimand.

Table 10: **Fixed category support.** Minimum examples per label, within the discordant unasked cells. Category sets remain fixed across paired resampling.

	At least 5 (60 categories)		At least 10 (45 categories)	
Checkpoint	I	95% interval	I	95% interval
Bunny v1.0	+0.541	[0.468, 0.565]	+0.555	[0.487, 0.584]
Bunny v1.1	+0.601	[0.528, 0.627]	+0.616	[0.547, 0.649]
InternVL3-14B	+0.652	[0.572, 0.673]	+0.673	[0.596, 0.697]
InternVL3.5-8B	+0.628	[0.556, 0.648]	+0.652	[0.583, 0.678]
LLaVA-1.5-7B	+0.429	[0.369, 0.457]	+0.476	[0.418, 0.505]
Phi-3.5-Vision	+0.647	[0.579, 0.669]	+0.671	[0.603, 0.693]
Qwen2.5-VL-7B	+0.654	[0.576, 0.678]	+0.666	[0.592, 0.688]
Qwen3-VL-4B	+0.487	[0.424, 0.515]	+0.534	[0.468, 0.560]
SmolVLM2-2B L14	+0.588	[0.519, 0.613]	+0.610	[0.543, 0.636]
SmolVLM2-2B L22	+0.772	[0.697, 0.782]	+0.803	[0.734, 0.814]

Resampling probe training and independent evaluation pairs. To assess fitting uncertainty beyond frozen predictions, we split the 500 pairs once into 250 training and 250 evaluation pairs (permutation seed 1729); neither image of any evaluation pair enters training. Projection seed 1051 and dimension 256 remain fixed. A training-only scaler and target-native ridge probe ($\alpha = 10$, exact Cholesky solver) are fit on the training set and applied unchanged to early and late evaluation features. For each of 1,000 draws (seed 2027), training pairs and evaluation pairs are independently sampled with replacement, and both scaler and ridge are refit. The eligible-category set is fixed from the evaluation labels, using the same zero-interaction convention for absent labels as above. Both raw and equal-label-mass metrics are retained.

Table 11 gives the unresampled holdout estimate and the central 95% of refitted/resampled interactions. The latter is a stability range, not a substitute confidence interval for the primary five-fold estimate. Duplicate training pairs reduce the number of distinct training images and can shift the resampling distribution below the unresampled estimate. This analysis explicitly varies probe training and test sampling while still conditioning on the projection, fixed split, selected image population, and VLM features. It neither estimates uncertainty over all designs nor resolves category correlations. All 10,000 refitted/resampled interactions remain positive under both metrics. For Bunny, however, the unresampled v1.1-minus-v1.0 holdout difference is $-.0297$, and its paired resampling range is $[-.0674, .1032]$ (644/1,000 draws positive). Thus the primary fixed-probe ordering is not invariant to training/evaluation design; both checkpoints exhibit transfer, but their relative strength is less robust. Exact projected inputs, split indices, training resamples and test multiplicities are supplied for independent CPU refitting.

Native generated-answer matching. We additionally run native target and donor generation for all 500 original pairs in InternVL3.5 and Qwen2.5. Prompts, image preprocessing, fast-processor settings, and row layouts match the saved experiment. Generation is greedy with at most four new tokens; a case-insensitive leading yes/no word defines the decoded label, otherwise the answer is invalid. All 2,000 responses parse successfully. The original non-final-text readout is captured during this generation, with exact feature equality to the saved cache for every image (maximum absolute error zero). This checks compatibility before combining new answer labels with frozen intervention predictions.

InternVL target/donor native correctness is 479/460 out of 500; Qwen is 460/454. Our primary sensitivity retains only pairs where both are correct, giving 444/422 pairs. A secondary sensitivity

Table 11: **Training and evaluation resampling.** A separate 250/250 holdout sensitivity. C is the fixed number of eligible evaluation categories; the range covers the middle 95% of 1,000 independently resampled/refitted runs.

Checkpoint	C	Holdout I	95% resampling range	Positive draws
Bunny v1.0	71	+0.452	[0.237, 0.389]	1000/1000
Bunny v1.1	71	+0.422	[0.262, 0.412]	1000/1000
InternVL3-14B	71	+0.547	[0.330, 0.486]	1000/1000
InternVL3.5-8B	71	+0.482	[0.298, 0.453]	1000/1000
LLaVA-1.5-7B	71	+0.363	[0.213, 0.343]	1000/1000
Phi-3.5-Vision	71	+0.517	[0.300, 0.452]	1000/1000
Qwen2.5-VL-7B	71	+0.574	[0.329, 0.483]	1000/1000
Qwen3-VL-4B	71	+0.404	[0.211, 0.355]	1000/1000
SmolVLM2-2B L14	71	+0.495	[0.274, 0.425]	1000/1000
SmolVLM2-2B L22	71	+0.683	[0.457, 0.608]	1000/1000

retains equal valid generated labels even if both are wrong, giving 449/430 pairs. Selection uses native answers only, never intervention outcomes, unasked probe scores, or a refitted decoder. We then recompute discordant-category AP from the original held-out predictions on each subset, with 5,000 paired pair-bootstrap draws. Eligible-category support is recomputed within the subset and each draw. Table 12 reports original-prevalence interactions; all four intervals remain positive under equal-label-mass AP as well. These conditional analyses restrict the evaluated population and leave model confidence unmatched. They do not replace the original annotated-answer-matched estimand or establish equivalence of internal decision states.

Table 12: **Matching actual generated yes/no answers.** Native answer labels select the subset; the original probe remains fixed. n is the number of pairs and C the number of eligible categories. Confidence intervals condition on the selected subset and fitted probe.

Checkpoint	Subset	n	C	I	95% CI
InternVL3.5-8B	Both correct	444	75	+0.579	[0.507, 0.609]
InternVL3.5-8B	Same valid answer	449	75	+0.578	[0.503, 0.607]
Qwen2.5-VL-7B	Both correct	422	75	+0.641	[0.561, 0.677]
Qwen2.5-VL-7B	Same valid answer	430	75	+0.639	[0.561, 0.674]

C BEHAVIORAL CARRIER LOCALIZATION AND CONTROLS

The source-specific carrier assay uses a different pairing criterion from the strict donor: questions match and answers are *opposite*. Native source and target predictions must each be correct. Yes-to-no and no-to-yes directions each contribute 250 of the 500 principal-model pairs. The intervention copies the corresponding native source rows into a target run, preserving unpatched target states. Unlike the content assay, the outcome is the model's generated source answer, not a probe score.

Checkpoint	Pairs	Question clusters	Block	Text lead (pp)	95% CI
InternVL3.5-8B	500	72	18	68.6	[61.3, 75.5]
Qwen2.5-VL-7B	500	74	17	85.4	[79.0, 90.5]

The four audited controls are target-all self replacement, target-visual self replacement, target-primary-text self replacement, and full-source replacement. The first three preserve the native target answer and the fourth adopts the native source answer. Six designated control blocks times 500 pairs give

3,000 audited cells per control and principal model; this count does not imply controls were run at every plotted layer. Primary-model curves are paired within question clusters; the 500 pairs should not be interpreted as 500 independent question templates. The smaller Smol replication has 100 pairs and 39 question clusters, with non-final-text advantage 88 points [83.3,94.4] at L14.

Measured handoff begins at InternVL L16 (CI [16,16]) and Qwen L14 ([14,15]) under the saved crossover analysis. This is a source-adoption crossover on the tested grid, distinct from any recovery threshold. Equal block depth and input-hook placement allow an operational comparison of carriers, but cannot remove their naturally different attention contexts. We therefore describe an intervention-specific shift of causal control and do not infer a modality-intrinsic information capacity.

Image reuse and a disjoint-subset sensitivity. Question-cluster intervals account for repeated templates, but not image reuse across different templates. Decoded RGB pixel hashes identify 444/417 unique images among the InternVL/Qwen 500-pair sets; 217/205 images appear in both source and target roles. We therefore select an image-disjoint subset by scanning the saved manifest order and retaining a pair only when neither image has appeared in any retained pair. Selection uses image identities, never intervention outcomes. It retains 175/166 pairs, with direction counts 88/87 and 80/86. We resample their 59/55 whole question clusters in 5,000 draws (seed 42), keeping carrier outcomes paired. Table 13 reports the resulting middle-layer contrasts, not a new crossover estimate. Smol independently retains 72 pairs (35/37 directions). All intervals stay positive. This subset sensitivity changes the sample composition and does not replace the full-sample curves or quantify every source of dependence.

Table 13: **Carrier image audit and disjoint-subset check.** S, D count unique source/target images in the full set; overlap counts images used in both roles. n, q are pairs/question clusters retained without any image reuse. Text lead is at L18/L17/L14 for InternVL/Qwen/Smol.

Checkpoint	S	D	Overlap	n	q	Lead (pp)	95% CI
InternVL3.5-8B	327	334	217	175	59	67.4	[57.7, 76.2]
Qwen2.5-VL-7B	310	312	205	166	55	84.3	[77.6, 90.4]
SmolVLM2-2B	89	90	13	72	32	87.5	[80.9, 95.3]

D UNASKED-CATEGORY SURVIVAL AND THE BOUNDARY OF REUSE

The zero-out content assay uses 2,000 unique images per checkpoint, five folds, and a common native-trained probe applied unchanged to all conditions. Each item queries one category, which is excluded from its evaluation. A block-input visual-zero intervention is placed early or at the fixed readout block. We then pool non-final text at that block's output. Late perturbation therefore precedes the readout rather than leaving the recorded activation untouched.

Table 14: **Unasked-category macro-AP under visual zeroing.** The observed difference is computed from the original sample AP values. The bootstrap mean is reported separately because it differs from that point estimate for this nonlinear metric. All intervals use 5,000 paired draws.

Readout	Native	Early	Late	Point gap	Bootstrap mean	Gap 95% CI
InternVL L18	0.3831	0.0511	0.3356	+0.2845	+0.2916	[0.2780, 0.3055]
Qwen L17	0.3463	0.0447	0.3388	+0.2941	+0.3015	[0.2873, 0.3161]
Smol L14	0.3295	0.0452	0.1913	+0.1461	+0.1509	[0.1392, 0.1625]
Smol L22	0.6482	0.0502	0.4040	+0.3539	+0.3590	[0.3430, 0.3756]

Native-minus-late intervals are [.0409,.0552] for InternVL and [.0022,.0118] for Qwen. Thus neither result establishes lossless preservation even for the measured label set. Smol's late/native AP ratios are 58.1% at L14 and 62.3% at L22. These ratios are not estimates of mutual information or of the fraction of pixels represented. Qwen3 supplies a further survival control with early block L3 after reinjection; its bootstrap mean late-minus-early AP is +.2002 [.1880,.2130].

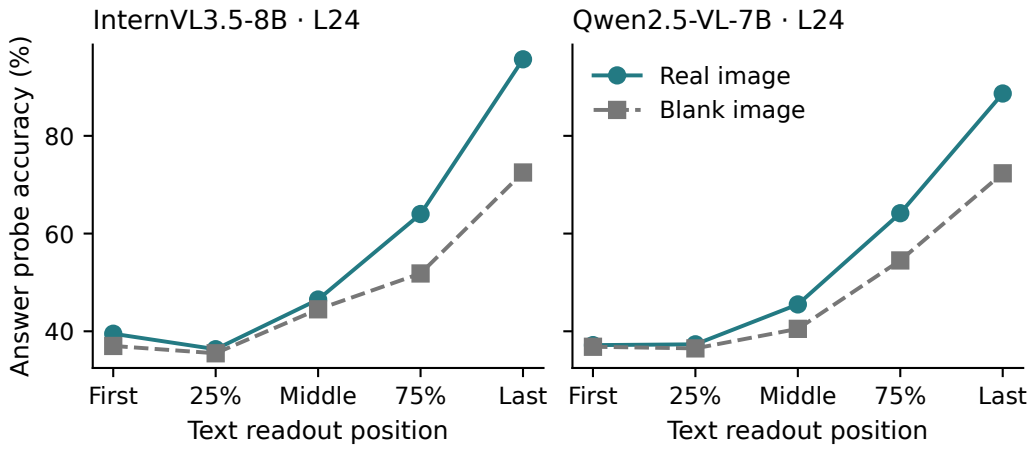

Figure 6: **Text position changes answer decodability.** Real/blank-image comparisons at L24, using 1,000 ScienceQA items per checkpoint and three held-out splits. The strong position gradient and substantial blank-image accuracy show why last-text versus visual probing cannot isolate a modality effect. These are localization measurements, separate from the real-activation carrier interventions.

Visual isolation for a second question. The Qwen2.5 protocol serializes a native chat prefix as user: image and Q_1; assistant: gold A_1 ("Yes." or "No."); then appends user: Q_2 and the assistant generation prefix. Pairs share Q_1, its annotated answer, and Q_2, but have opposite Q_2 labels. Native generated Q_1 and full-vision Q_2 answers must both be correct on both sides before a pair is eligible; the inserted A_1 is nevertheless the shared gold answer. Four strata balance the Q_1 answer and the direction of the Q_2 label contrast, with 125 pairs each.

Both conditions prefill the same prefix with native visual access, retaining all prefix KV entries, including image, question, and gold-answer positions. The second question is then evaluated as a cache continuation. In the isolated condition, an additive causal attention mask sets columns corresponding to original image-token KV positions to the dtype's minimum finite value, for *every* Q_2 prompt query and every subsequent generated-token query, at every decoder layer. Other causal accesses remain available. The unblocked control uses the same tokenization, cache positions, and M-RoPE reference. No cache entries are deleted, compacted, or reindexed. Prefix fingerprints, mask dimensions, and actual attention to blocked visual columns are audited (tolerance 10^{-7}).

Native Q_1 screening and both Q_2 conditions use greedy decoding with at most four new tokens. Yes/no parsing first accepts a leading yes/no, then an unambiguous occurrence of either word. Blocking visual access yields 501/1,000 correct responses and only 1/500 jointly correct pairs. The selected native baseline is 100% by construction, not an unfiltered task-accuracy estimate. The test retains rich first-turn text context but changes direct visual access; it does not rule out trained readouts or different querying interfaces. Auxiliary retrieval and transplantation assays are not needed for this conclusion.

E Position-controlled probing details

InternVL3.5 and Qwen2.5 each use 1,000 ScienceQA items with three held-out split seeds (42, 43, 44). Fixed probes use standardized features and logistic regression; the same train/test split is used for corresponding real/blank conditions. Text readouts are the first, quarter, middle, three-quarter, and final eligible text positions. Visual features use last-token, mean, maximum, and learned attention pooling. The learned pool uses a fixed 64-dimensional projection, a trainable pooling query, and a linear classifier (389 parameters in the reported setup), with validation selection inside the training partition. The probe tests and the 200-item generation-validity check are distinct evaluations.

Figure 6 shows the five-position real/blank comparison at L24. It describes accessible information at different causal positions and is not a modality-controlled comparison.

In standard order, InternVL last/mean/max visual pooling peaks at 41.2/37.0/40.7%, versus matched-blank peaks 36.5/37.2/37.0%; learned pooling peaks at 40.2%, versus 37.0% blank. Selecting only the last visual token is therefore not the sole cause of low standard-order answer decodability. In question-first order at L24, last/mean/max real scores become 56.5/58.5/62.7%, but blank scores are 53.8/70.7/72.0%. Absolute accuracy alone would hide this change in the control condition.

Question-first validity is defined as retaining at least 90% of native-template generation accuracy. InternVL retains 90.5/96.0% and Qwen retains 87.0/91.5% in the separate 200-item check. Passing that screen does not prove that the reordered template has the same distribution as training. All-layer maxima for visual pools and final text can occur at different layers and should not be read as simultaneous per-layer differences. Probe saturation, if used descriptively, is denoted $\ell^{\mathrm{sat}}_{\mathrm{probe}}$ (within 2% of its own maximum); it is not identified with ℓ^{95}_{V} or a causal absorption point. Logit-lens vocabulary changes likewise illustrate task-conditioned extraction rather than broad image-content preservation.

F Recovery robustness and tested grids

Image-cluster bootstrap resamples whole decoded-pixel clusters in ScienceQA and whole source images in TextVQA. Every sampled cluster contributes all its questions and the same multiplicity is used for visual and text conditions. Each carrier's native baseline is recomputed per draw. The original ScienceQA set has 913 clusters among 1,000 questions; TextVQA has 641 (282 singleton images and 359 images with two questions). We do not smooth away late performance dips. A confidence interval collapsing to a single integer denotes a stable discrete threshold crossing on this grid, not exact knowledge of an architecture-invariant transition layer.

ScienceQA checkpoint	90% first-gap CI	95% first-gap CI	98% first-gap CI
InternVL3.5-8B	[6, 7]	[6, 6]	[5, 6]
InternVL3-14B	[8, 9]	[9, 9]	[7, 9]
LLaVA-1.5-7B	[1, 5]	[2, 2]	[-3, 2]
LLaVA-NeXT-8B	[3, 9]	[3, 7]	[2, 3]
Qwen2.5-VL-7B	[4, 4]	[4, 5]	[5, 7]
Phi-3.5-Vision	[4, 7]	[4, 5]	[3, 4]
SmolVLM2-2B	[0, 0]	[0, 3]	[-1, 5]
Qwen3-VL-4B	[9, 12]	[10, 11]	[11, 11]

Qwen3 is an additional replication and is excluded from the seven-checkpoint denominator in the main text. Its first and sustained visual/text point crossings are L18/L28. The earlier $n = 200$ paired McNemar analysis with Holm correction resolves 5/7 models; that count is exploratory historical evidence, not a significance test for the enlarged bootstrap-gap table. Failure to cross in a resample must remain nonfinite and be disclosed rather than converted into an arbitrary finite last-layer value. All 95% ScienceQA gap draws reported in Table 5 are finite.

TextVQA	First V/T	Median gap	95% CI	Sust. V/T	Median gap	95% CI
InternVL3.5-8B	33/20	-12	[-13, -9]	33/32	-1	[-13, 0]
Qwen2.5-VL-7B	26/20	-6	[-8, -4]	26/20	-6	[-8, -2]

In Qwen TextVQA, first visual/text recovery at 90% is L26/L18 with gap CI $[-8, -8]$, and at 98% is L27/L22 with CI $[-5, -3]$. This reinforces the task reversal beyond a single chosen threshold. The InternVL 95% sustained interval includes zero because of a late text-side dip, despite its negative first-gap interval.

Recorded ScienceQA grids. The following are the actual tested output-hook block indices, not an assumed uniform step size. Paired carrier comparisons use the saved grids; finer untested behavior is not resolved by the bootstrap.

Checkpoint	Carrier	Tested blocks
InternVL3.5-8B	visual	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 34
InternVL3.5-8B	text	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 34
InternVL3-14B	visual	0, 8, 16, 20, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47
InternVL3-14B	text	0, 8, 16, 20, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47
LLaVA-1.5-7B	visual	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
LLaVA-1.5-7B	text	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
LLaVA-NeXT-8B	visual	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
LLaVA-NeXT-8B	text	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
Qwen2.5-VL-7B	visual	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26
Qwen2.5-VL-7B	text	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26
Phi-3.5-Vision	visual	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
Phi-3.5-Vision	text	0, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
SmolVLM2-2B	visual	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23
SmolVLM2-2B	text	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23
Qwen3-VL-4B	visual	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 34
Qwen3-VL-4B	text	0, 8, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 34

G DISTRACTOR-PAIRING CORRECTION AND STRATIFIED REPLICATION

An audit of the earlier $n = 200$ activation-patching implementation found deterministic offset pairing, $i \mapsto (i + N/2) \bmod N$, rather than enforced same-class distractors. In the ScienceQA run, 60 pairs shared an option letter and 140 did not. The correct description is *deterministic real-image distractor*. Sharing an option letter across different multiple-choice questions is not the same as sharing a semantic image class. That experiment cannot be interpreted as a controlled same-class test.

The replacement analysis explicitly stratifies random, same-output-label, and different-output-label real-image donors. The ScienceQA runs have $n = 500$ each: Qwen first/sustained recovery is L14 under all three strategies, with respective same/different/random CIs [14,14], [14,14], and [12,14]. Their L14 accuracies are 89.0/89.2/90.0% against matched baselines of 90.8/90.6/90.6%. All transition-layer pairwise intervals contain zero and Holm-adjusted McNemar $p = 1.0$. InternVL recovers at L18 across the three strategies, with CI [18,20].

The stronger semantic POPE replication uses 1,000 aligned targets, 167 target-image clusters, 887 baseline-correct items, unique donors, and different image content for every pair. Same-answer and strictly opposite-answer donor strategies have identical target ordering and baseline-correctness vectors. Both first and sustained recovery are L18, with item and image-cluster intervals [18,18]. The L18 opposite-minus-same difference is $+.1$ percentage points with cluster interval $[-.9, 1.1]$. No tested layer differs after Holm correction. These are robustness checks against answer sharing, not equivalence tests or guarantees of semantic image dissimilarity.

All 20,000 audited intervention cells execute exactly one intended hook, align all 256 visual rows, insert the source tensor with zero error, and reject a no-op patch. This operation-level audit addresses silent skipped samples and incomplete replacements. Patching is at block *output*, unlike the principal content/carrier input-hook assays.

H BUNNY SCOPE AND SOURCE-RESULT INVENTORY

The exact common-pair strict-donor comparison matches ordered image IDs, questions, answers, unasked labels, focal categories, and probe-fold membership. Both checkpoints use early/late block inputs L0/L12. The bootstrap reuses identical pair multiplicities for both models. Both interactions are positive in every draw, while the v1.1-minus-v1.0 difference is positive in 97.76% of draws, not all draws. The paired-difference interval is [.0024,.1139]. It measures a modest change in the tested transfer statistic, not an on/off transition.

The separate POPE carrier comparison uses 500 common pairs and 73 shared question clusters. At L10 the v1.1-minus-v1.0 visual-adoption difference is +5.6 points [2.929,8.525], and the text-minus-visual relay interaction differs by -7.6 points $[-13.145, -2.495]$. This is a different task, estimand, and resampling unit from the unasked-content comparison. Agreement of pair identities does not imply that serialized manifest files are byte-identical.

Reproduction inventory. Two complementary supplements separate statistical checking from model execution. `analysis_supplement.zip` provides held-out scores, labels, supports, folds, pair IDs, native answers, carrier outcomes, resampling records, checksums and CPU commands. It reproduces endpoints, balanced and fixed-support AP, all 100 fixed-score refit metrics, answer subsets and carrier sensitivities. `inference_supplement.zip` adds the content/carrier/deletion implementations, model adapters, position and second-query assays, portable full pair manifests, raw result records and inspected environment versions. Raw result records are split into two accompanying ZIP parts. Exact projected features enable independent refitting of the 250/250 holdout bootstrap without a VLM; other projections can be regenerated through the supplied inference entry points.

The inference archive records current checkpoint download revisions and etags where recoverable, hashes of config/processor code, full weight-payload SHA256 values, sizes and safetensors headers. This release-time inventory does not retrospectively establish unknown historical revisions. Model weights and dataset images are obtained separately; no claim of independent replication of every historical GPU run follows from publishing the code. Source/export hashes document path anonymization and the model-path configuration overlay; the intervention algorithms are unchanged.

The following script families define the operations: `counterfactual_unasked_donor.py` for fixed answer-matched pairs and shared probes; `source_specific_handoff.py` for carrier replacement; `visual_removal_ablation.py` and `text_removal_ablation.py` for symmetric recovery; `unasked_content_causal_survival.py` for content zeroing; `position_controlled_probing.py` for ordering and pooling; and `activation_patching.py` for stratified donors. Companion statistics scripts recompute cluster bootstrap summaries from per-item outputs.

The compact result inventory includes strict-donor statistics for seven checkpoints, exact paired Bunny statistics, the original-seven ScienceQA cluster summaries plus Qwen3, both TextVQA cluster summaries, the audited POPE patching comparison, and the Smol content-survival readouts. Source hashes, per-pair/per-layer outputs, image/pair manifests, and exact hook audits provide traceability. The figure/table builder checks complete-run status, sham identity, cohort membership, and source hashes before formatting the stored statistics. This is a consistency audit of saved runs, not a claim that model inference has been independently rerun.

The original spatial-quartile, compression-ratio, attention, and CKA explorations are not needed for the bounded claim established here. They should not be interpreted as proving an architectural law, lossless copying, or a universally safe pruning schedule. No new inference-speed claim follows from an intervention that retains sequence length.