One Scene Code, Two Consumers: Causal Tests of Where Task Differences Live in a Unified Multimodal Model

Tong Zhang[*]
Fudan University

Yun Peng
Fudan University
yunpeng4@sigsoft.org

Tao Xie[†]
Peking University
taoxie@pku.edu.cn

Abstract

Unified multimodal models such as Harmon route image understanding and image generation through one visual encoder with bit-identical weights. Because the encoder never sees the question or the generation regime, its task-blindness is a property of the architecture, not a discovery—so we use it as a controlled testbed for two questions that weight sharing leaves open: is the shared representation *causally interchangeable* between tasks, and where do task differences actually come from? Three causal results give a single answer: the encoder writes an *editable scene code*, and task differences emerge from how each downstream consumer recovers and uses that code. (1) Substituting the rows at a visual fact's token positions with rows computed from a content-flipped scene moves the answer 63–66% of the way to the donor value (random positions: 3–11%), and rows computed by the generation pass repair deleted understanding computation at 12 of 20 layers (71–99%) while wrong-content donors *anti*-rescue; full-grid cross-pass substitution is established in the und→gen direction. (2) The iterative generation consumer re-derives *deleted* evidence from spatial backups: on 98 real COCO scenes, deleting one annotated object instance while others survive lets generation recover .82 of the answer signal vs. .54 when every instance is deleted, while the one-shot understanding reader collapses to .05 under full removal—and with the object fully gone, both consumers seeing the same input, generation still retains +.48 more at every layer, so the surplus is iterative consumption, not the prior. Yet generation faithfully consumes *corrupted* evidence (0.74 → 0.07)—robustness is consumer × evidence-redundancy, not path-intrinsic. (3) Task specialization emerges progressively in the task-conditioned LLM trunk (cross-pass CKA 0.997 → 0.421; substitutability lost by mid-trunk; the divergence replicates at 0.5B), and on Show-o, which mixes task context from layer 0, a causal cross-pass swap collapses at every layer even where cross-pass CKA reads 0.995—similarity is not interchangeability. The mechanism: the passes share one representational *geometry* in rotated *bases*. A single frozen orthogonal map, fit on held-out natural images, causally repairs 99/93/75% of Show-o's swap collapse at layers 0/2/4, while wrong-target rotations and low-rank maps recover ≈0; only deep, where CKA itself collapses, does the geometry diverge. Interchangeability tracks where the architecture injects task context—and what injection first changes is the basis, not the geometry. We state the evidence perimeter explicitly: one model family probed causally, synthetic generation-side stimuli, and direction- and scale-scoped substitution claims.

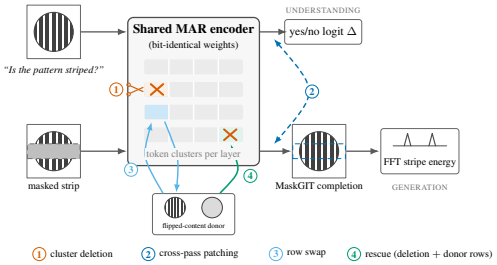

Figure 1: **Method overview.** One shared MAR encoder (bit-identical weights) processes the same image through both task paths: *understanding* (image + question → yes/no logit difference) and *generation* (masked strip → MaskGIT completion → closed-form FFT stripe-energy readout). Matched interventions—cluster deletion, cross-pass patching, row swap from a content-flipped donor, and deletion-plus-rescue—are scored against size-matched random-token nulls through both readouts.

1 Introduction

Unified multimodal models fold image understanding and image generation into a single network, and the tidy engineering story invites a scientific

[*] Work done at Fudan University.
[†] Corresponding author.

reading: a model that reuses the same weights for two tasks must be reusing the same representation, and the representation must serve both tasks in the same way. Harmon (Wu et al., 2025b) is a clean instance: one masked-autoregressive (MAR) visual encoder (Li et al., 2024) is used, with *bit-identical* parameters, on the understanding path (image in, text out) and the generation path (text in, image out).

An obvious objection should be dealt with first. In Harmon, the encoder is *task-blind by construction*: the question text never reaches it, and no regime token distinguishes the two passes, so the encoder's representations of the same image under the two tasks are guaranteed to be near-identical by architecture alone. Measuring high cross-task similarity inside the encoder is therefore not a finding; it is a designed property of the testbed. But that is precisely what makes the testbed useful. If the encoder is task-blind by construction, why should its representations differ? Precisely because sharing is guaranteed at the weight level, this architecture isolates the real questions: is the shared representation *causally interchangeable*—can content computed under one task be read, written, and used to repair the other?—and where do task differences actually come from, if not from the shared encoder?

This paper answers both questions with matched causal interventions pushed through both task paths on the same images, and the answers compose into one claim:

> *The unified model shares an editable scene code; task differences emerge from how each downstream consumer recovers and uses that code.*

Three results support the claim, each scoped to the evidence that establishes it.

(i) The shared code is causally interchangeable and repairable across tasks (§3). Substituting the understanding-pass rows at a fact's token positions with rows from a content-flipped donor scene moves the yes/no answer 63–66% of the way to the donor's value at every encoder layer (random positions: 3–11%; self-swap: bit-exact zero). Rows computed by the *generation* pass repair deleted understanding computation at 12 of 20 layers (71–99%), while content-flipped donors *anti*-rescue; full-grid cross-pass substitution is harmless at every layer und→gen (the reverse lacks a usable negative control and is not claimed).

(ii) Task differences arise from consumer × evidence-redundancy (§4). The identical single-layer deletion abolishes the understanding readout (z to -39.1) while never moving generation on periodic stimuli ($|z_{\text{gen}}| \leq 1.1$)—but the robustness belongs to the evidence, not the pathway. On 98 real COCO scenes, deleting one annotated instance while others survive lets generation recover .82 of the answer signal, vs. .54 when every instance is deleted; the one-shot reader recovers .63 from the identical deletion early in the encoder and collapses to .05 under full removal. With the object fully gone both consumers see the same input, yet generation retains $+.48$ more at every layer: the iterative consumer re-derives *missing* evidence from backups; no consumer resists *wrong* evidence (corruption: $0.74 \rightarrow 0.07$).

(iii) Specialization begins as a rotation of a shared geometry, not a divergence of it (§5). Once visual tokens mix with task context in the LLM trunk, cross-pass similarity declines gradually (CKA $0.997 \rightarrow 0.421$) and substitutability is lost by mid-trunk. On Show-o (Xie et al., 2025), which mixes task context from layer 0, the causal swap collapses at layer 0 (-6.3) *while* CKA reads 0.995. The mechanism: CKA sees rotation-invariant *geometry*; a consumer reads a fixed *basis*. A single frozen orthogonal map, fit on held-out natural images, causally repairs $99/93/75\%$ of the collapse at layers 0/2/4; wrong-target rotations and low-rank maps recover $\approx$0; only deep, where CKA itself collapses, does the geometry part. The task-injection point sets where the basis rotates: late in Harmon, at layer 0 in Show-o, moot in Janus-Pro (Chen et al., 2025b), whose decoupled encoders sit below the different-image floor everywhere.

A methodological discipline underwrites every null result: a readout at its floor silently mimics dissociation, so every readout carries positive controls in both directions (§2). The Limitations section collects all scope bounds.

2 Setup and Method

Model. We study Harmon (Wu et al., 2025b) at two scales: 1.5B (MAR-Huge encoder, L=20 blocks, H=16 heads, Qwen2.5-1.5B trunk) and 0.5B (MAR-Base, L=12, H=12, Qwen2.5-0.5B), public checkpoints, inference only. Images (512^2) are encoded by a KL-16 VAE (Rombach et al., 2022) into a 32×32 latent grid (N=1024 image tokens). The encoder weights are bit-identical across

the two task paths, and—crucially for the design—the encoder receives no question text and no regime token: it is task-blind by construction. All interventions operate inside the shared encoder (or, in §5, the trunk); task heads are untouched.

Stimuli. Procedurally generated scenes with closed-form ground truth: 8 striped-disc stimuli (main sweep), composite five-factor scenes and compact-object variants (swap/rescue/band), plus two-object scenes, a capability suite, text-to-image prompts, and real COCO photographs (POPE and the instance-drop experiment; Appendix D). Procedural control gives the understanding probe an unambiguous yes/no answer and the generation probe a closed-form FFT/Gabor stripe-energy target, so both readouts reference the same physical quantity.

Decompose, intervene, read out through both paths. Per layer we head-average and symmetrize the self-attention map and partition tokens by spectral clustering (k=4, fixed before any intervention; cluster stability ARI 0.87–0.97 across all layers and scales, k-robustness in Appendix A). Interventions target one cluster (or an oracle/factor token set) with one of three operators—pre-softmax attention masking (ATTN-MASK), per-position mean-ablation of the block output (MEAN-ABLATE), or zeroing (ZERO)—always compared against *size-matched random-token nulls* resampled per intervention, so "selective" means beyond what deleting the same amount of anything would do. The same intervention on the same image is pushed through *both* task paths: understanding is read out as the yes/no logit difference Δ_{und}; generation completes a masked center strip of the disc (stripe context visible on both sides) and is scored by the closed-form stripe energy inside the completed strip, Δ_{gen}. Both readouts are reported as z-scores against the matched null; a readout counts as moved at $|z| \geq 2$ in the direction of harm. Substitution protocols (swap, cross-pass patch, rescue) overwrite block-output rows at chosen token positions with donor rows, with bitwise self-patch checks that return exactly 0 in every cell.

Readout-validation discipline. A readout at its floor silently mimics dissociation: interventions cannot move it downward, so "no generation effect" is indistinguishable from "the readout cannot see generation." Our first generation readout (full-disc completion from a gray background) failed exactly this way—clean completions could not reproduce the stripes, and MEAN-ABLATE, injecting corpus-mean rows from other striped stimuli, *restored* them ($0.20 \rightarrow 0.76$): an "ablation" behaving as a treatment. Every readout in this paper therefore carries positive controls in both directions: the clean baseline must approach the source-image ceiling (center-strip design: median 0.725 vs. ceiling 0.775 at 1.5B), and the readout must register real damage (wrong-image donor rows: $0.74 \rightarrow 0.07$). Readouts that fail validation are excluded, not reported—including the 0.5B model on composite scenes, which is why those experiments are single-scale.

Clustering, interventions, nulls, and both readouts are deterministic given a seed; the generation score is closed-form (no learned judge). Code and stimulus generators will be released.

3 The Shared Code Is Causally Interchangeable and Repairable

Task-blindness guarantees that the encoder *computes* nearly the same states on both passes (cross-pass linear CKA is 0.998–1.000 at every layer of both scales, against a 0.62–0.72 different-image baseline; Appendix B, Table 8). What architecture does not guarantee is that this content behaves as one *code*: token-local, value-bearing, and readable/writable by either task. Three substitution protocols test exactly that.

3.1 Swap: rows at a fact's positions carry its value

The design is pure substitution on the intact understanding pass: nothing is deleted. Stimuli are composite five-factor scenes; each factor's donor is a same-geometry scene in which *only* that factor's content is flipped. At one layer at a time we overwrite the block-output rows at a designated position set with the donor's rows at the same positions, and score the *toward-fraction*: how far the yes/no logit difference moves from the clean answer toward the donor scene's own clean answer. Over 12 seeds $\times$ 4 factors $\times$ all 20 layers at 1.5B (Figure 2): oracle-position swaps (the tokens over the factor's content) reach 0.628–0.660 at every layer; size-matched random positions reach only 0.03–0.10; the bitwise self-swap control is exactly 0 in every cell. The answer follows the value written at the fact's token positions—at every depth. (Full per-factor grid: Appendix Table 9; bootstrap 95% CIs for all headline quantities: Appendix Ta-

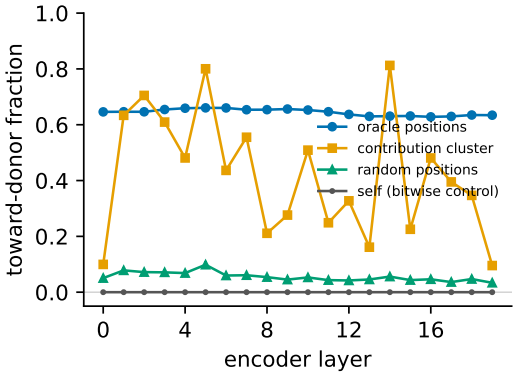

Figure 2: **Swap: swapping rows swaps the answer (1.5B).** Toward-fraction by layer for swaps at the factor's oracle positions, its contribution-cluster positions, and size-matched random positions (12 seeds × 4 factors); the bitwise self-swap control is exactly 0 everywhere. Oracle swaps move the answer 63–66% of the way to the donor's value at every layer.

ble 13.)

3.2 Cross-pass substitution (und→gen)

Can one task consume rows computed by the other? Overwriting the generation pass's visible-token rows with same-image understanding-pass rows—at every MaskGIT step, at every layer—never degrades the completion at either scale (change in stripe energy $+0.068$ at 1.5B, $+0.18$ at 0.5B, flat across depth), while a wrong-image donor collapses it (-0.57) (Figure 3; full grid: Appendix Table 7). If the two passes' computations became task-specific at any encoder depth, the cross curve would have to leave the harmless band from that depth on; it never does.

Two scope notes, stated as claims boundaries rather than caveats in passing. First, the reverse direction (gen→und) is *not claimed*: its patchable positions exclude the disc center where the stripe evidence sits, so even the wrong-image negative control moves the logit difference by at most 0.36 logits against a clean baseline of 3.6—with a control that weak, a null cross effect is uninformative. Full-grid cross-pass substitutability is established in the und→gen direction only. Second, position-shuffled same-image rows are also harmless here, as expected for a spatially periodic stimulus whose token content is position-generic up to phase; the shuffle control does not test position-specificity on these stimuli.

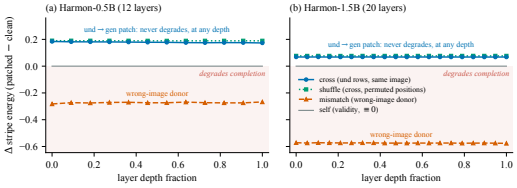

Figure 3: **Cross-pass patching (und→gen): interchangeable content at every encoder depth.** Change in completed-strip stripe energy when block-L outputs of all visible tokens are overwritten at every MaskGIT step, by depth fraction, at (a) 0.5B and (b) 1.5B. Same-image understanding-pass rows (solid) never degrade completion; a wrong-image donor (dashed) collapses it. Self-patch control (gray) is exactly 0.

3.3 Rescue: one pass's rows repair the other

The strongest form of interchangeability is repair. On the composite scenes at 1.5B, deleting a factor's cluster rows and simultaneously patching donor rows into the deleted positions is scored as $\text{rescue}\% = (\text{rescued} - \text{deleted})/(\text{clean} - \text{deleted})$, with three donors separating content from position: *cross* (same image, other pass), *wrong* (content-flipped scene), and *shuffle* (cross rows at permuted positions).

Across all 20 layers (12 seeds), generation-pass rows repair deleted understanding computation at 12 of 20 layers (cross donor 71–99%; flipped donor negative or near zero at 11 of them, to -40%)—the *anti*-rescue is the signature that the readout follows the donor's *value*, not merely its energy. The und→gen direction repairs at 7 of 20 layers (up to 115%). Under the strict pre-registered four-gate criterion (deletion factor-selective *and* rescue content-specific, both directions), the verdict passes at a minority of layers—layer 15 in both directions, layers 4 and 17 gen→und—so we state the bidirectional claim at that strength and no more. One honest note: the shuffle donor rescues about as well as the cross donor (e.g. 87% vs. 88% at layer 17), so the rescue is content-specific but not position-specific within the deleted set, consistent with the swap and patching shuffle results. (Per-layer grid and gate verdicts: Appendix Table 10.)

Takeaway. The two tasks read and write one editable scene code: the same fact, written as a value at the same token positions, movable by substitution (swap), consumable across tasks (patching, und→gen), and sufficient to repair the other task's deleted computation (rescue, strongest gen→und). Composite-scene evidence is single-scale (1.5B;

§2).

4 Consumer Asymmetry: Robustness Is Consumer × Evidence

If both tasks share one code, where do task differences come from? The dual-path sweep crosses every encoder layer at both scales with two operators (MEAN-ABLATE, ZERO) and two token sets (targeted cluster, discriminative oracle), each cell scored on 8 stripe stimuli against 5 size-matched random-token nulls in the $(z_{\text{und}}, z_{\text{gen}})$ plane, with dissociation pre-registered as a first-class outcome.

The one-sided map. On the periodic stripe stimuli the causal map is strikingly one-sided (Figure 5; Table 5): at 1.5B, 15 of 40 cluster cells abolish the understanding readout (to $z_{\text{und}} = -39.1$) while *not one* cell of any kind moves generation ($|z_{\text{gen}}| \leq 1.1$ everywhere); at 0.5B the map is qualitatively identical (3/24 understanding-only; zero co-causal or generation-only cells at either scale; full grid: Appendix Table 3). The generation readout is positively controlled in both directions (§2), so the silence is real robustness, not a blind readout.

But the robustness is conditional, and corruption defeats it. Two further experiments locate the robustness in the *evidence*, not the pathway. First, corrupted content—wrong-image rows rather than missing rows—destroys generation at every layer ($0.74 \rightarrow 0.07$): the iterative consumer faithfully consumes whatever is written in the code. Second, the band-width experiment deletes each factor's cluster rows over contiguous bands of $w \in \{1, 2, 3, 4, 6\}$ layers on compact-object composite scenes, where the deleted rows carry the *sole* visible evidence for the completion (right-half completion; the factor content sits in the left half). There, generation is as single-layer-fragile as understanding: cluster width-1 normalized drop 0.74 ± 0.73 for generation vs. 0.88 ± 0.43 for understanding, both flat out to six-layer bands, random bands ≈ 0 (Appendix Table 4). Both paths cross 50% of their range already at width 1.

Mechanism: iterative consumption exploits redundancy. The resolution of the apparent contradiction is the spatial redundancy of the visible evidence. The stripe stimuli are periodic and the pattern remains visible on both sides of the completed strip; at every one of the 32 MaskGIT steps the evolving canvas is re-encoded through all layers, so content deleted at one step is re-derived from redundant visible evidence at the next. The compact-object scenes offer no redundant copy, and generation breaks at a single layer. Understanding reads once, through a fragile, layer-local readout that does not exploit redundancy *even when it is present* (stripes, z_{und} to -39.1). And corruption defeats both consumers in every regime. Deletion-robustness is therefore a property of consumer × evidence-redundancy—iterative consumption exploits redundancy that a single pass cannot—not an intrinsic property of the generation pathway.

The decisive test: instance-drop on natural scenes. The stripes-vs-composite contrast confounds redundancy with the stimulus family, and a first synthetic paired design proved a weak instrument (its half-deletion left the intact half of the evidence as an accidental backup, so the no-backup condition never existed) and is discarded. Natural scenes supply the redundancy axis for free: an object annotated with *multiple instances* carries its evidence in physically separate places, a single-instance object does not, and neither fact is designed into the stimulus. For n=98 validated (image, object) POPE items (gt yes, clean signed logit ≥ 1; 56 multi-instance, 42 single-instance), we delete annotated instance footprints *in full—drop-one* (the largest instance; the others remain as a genuine backup) vs. *drop-all* (no backup)—and feed the same deleted token set to both consumers: generation masks it as the MaskGIT completion region and regenerates over 32 steps before re-reading; understanding zeroes the rows and reads once (Figure 4; full grid Appendix Table 6). Three paired contrasts, each significant: *(i) the backup is real and used*—generation recovers .82 when one instance survives vs. .54 when none does ($\Delta = +.29$ $[+.16, +.41]$); *(ii) only the iterative consumer exploits it fully*—on the identical drop-one deletion, generation beats one-shot understanding by $+.20$ $[+.07, +.32]$ at $L4$; *(iii) and the re-derivation is not prior hallucination*—with the object fully removed, both consumers see the same input and the same prior, yet generation retains $+.48$ $[+.34, +.63]$ more, at every probed layer. Internal controls: understanding's drop-all collapse (.05 multi, .02 single) certifies the deletion; random-region nulls retain near-fully for both consumers; and single-instance drop-all sits below multi-instance (.36 vs. .54)—even full removal is buffered by scene context only where the scene offers it. One asymmetry fades with depth: understanding's drop-one recov-

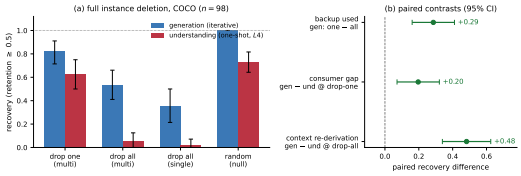

Figure 4: **Instance-drop redundancy (Harmon-1.5B, COCO, n=98).** (a) Recovery after *full* deletion of annotated instances, by consumer; bootstrap 95% CIs. (b) Paired contrasts, all significant: the spatial backup is used (+.29), only the iterative consumer exploits it fully (+.20), and the surplus survives full removal with identical input to both consumers (+.48)—re-derivation, not prior. Full grid: Appendix Table 6.

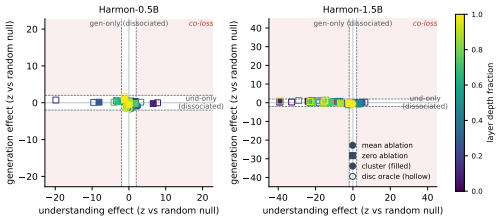

Figure 5: **Dual-path sweep.** Every (layer, operator, token-set) cell at both scales in the $(z_{\text{und}}, z_{\text{gen}})$ plane; dashed lines mark $|z|$=2. All hits lie on the understanding axis; no cell is co-causal or generation-only ($|z_{\text{gen}}| \leq 1.1$ across cluster cells). Representative layers: Table 5 (appendix).

ery rises .63 → .82 from $L4$ to $L15$ while drop-all stays ≤.18, so late *encoder* layers already integrate neighboring-instance evidence—the one-shot reader is redundancy-blind early, not everywhere.

Generality, briefly. Where the readouts validate, the understanding-side map generalizes: on two-object scenes, 38/40 cells are object-selective (z_{und} to −51.2; the other object's size-matched cluster inert); on real COCO photographs (POPE), object-cluster MEAN-ABLATE flips the answer on 47–54% of clean-correct items (oracle 67–74%) vs. 1–9% random and ≤ 3.5% other-object controls, replicating at 0.5B (Table 19). The generation-side twin is a null: regenerating 30 COCO images with the object's cluster rows deleted leaves OWLv2 detectability at the random-band null in every cell ($|z| \leq 0.35$; Table 20), though regeneration itself halves detector confidence, so the claim rests on the matched contrast. On the model's own text-to-image outputs, mid-generation cluster deletion flips self-consistency answers in 3–4% of cases (vs. ∼2% random; n=1534). Protocols and caveats: Appendices D and E.

Auxiliary finding: content-addressed, not routing-addressed. Severing all pre-softmax attention edges into the targeted cluster (ATTN-MASK) barely moves either readout at any layer, while MEAN-ABLATE/ZERO of the same cluster's block-output rows reach $z_{\text{und}} = -39.1$; a knockout atlas over all 464 heads (both scales, both passes) finds none privileged—deleting *all* heads at a layer moves understanding by ≈0.3 logits, an order of magnitude below block-output zeroing. The scene code is addressed by what it is, not where it is routed: the attention map here is stable and clusterable, yet causally idle—a sharp caution for attention-centric interpretability. Full atlas: Appendix C.

5 The Sharing Boundary: Specialization Emerges in the Trunk

Sections 3 and 4 are no-fork results for the shared encoder. But the encoder's output rows enter the LLM trunk (Qwen2.5, 28 layers at 1.5B), where visual tokens first mix with task context—question text on one pass, generation regime on the other; if sharing is coextensive with task-blindness, specialization should begin exactly here, progressively rather than switch-like.

Progressive divergence, loss of substitutability. We ran the capture-and-substitute logic one stage downstream, on the trunk's visual-token hidden states at shared visible positions (12 stripe seeds, 1.5B). Correlationally, cross-pass linear CKA (Kornblith et al., 2019) declines monotonically through the trunk—0.997 (L0) → 0.421 (L27)—where the encoder held 0.998–1.000 flat; the different-image baseline declines in parallel (0.78 → 0.35), never crossed, so the divergence is graded, not a metric depth artifact (Figure 6a); the decline replicates at 0.5B (0.997 → 0.551). Causally, gen→und substitution at a single trunk layer shows partial interchangeability early (substitutability index $1 - d_{\text{cross}}/d_{\text{mism}}$: 0.84 at L0) that is lost by mid-trunk, where the cross donor hurts nearly as much as a wrong-image donor (L9: −0.47 vs. −0.71 logits; L15: −0.56 vs. −0.58; index 0.34 → 0.04; Figure 6b). A shuffle control: same-image rows at *permuted* positions track the wrong-image donor (L9: −0.72)—the trunk code is position-bound. Late in the trunk both donors go to zero together: the readout stops depending on the visual rows. Bounds: donor rows carry a RoPE position mismatch (the position-matched cross-vs-

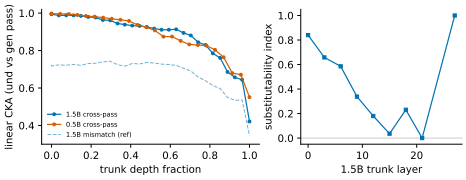

Figure 6: **Task specialization emerges progressively in the trunk (1.5B).** (a) Cross-pass CKA on visual-token hidden states declines monotonically ($0.997 \rightarrow 0.421$; different-image baseline dashed, never crossed) where the encoder held 0.998–1.000 flat. (b) Causal substitution (gen→und): cross donors much less harmful than wrong-image donors early; gap closed by mid-trunk; both near-null late as the readout leaves the visual rows. Self-patch control exactly 0.

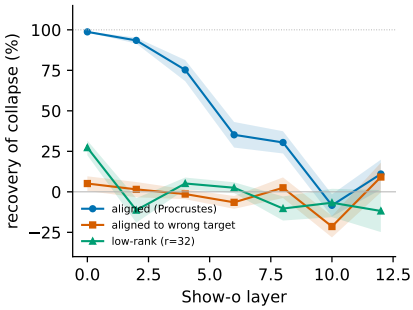

Figure 7: **A fixed orthogonal map causally repairs the cross-pass swap (Show-o, natural images).** Recovery of the same-content cross-swap collapse by layer: Procrustes-aligned donor rows (blue) vs. the same map fit to wrong target images (orange) and the best rank-32 linear map (green); item-level bootstrap 95% bands, n=53. Alignment recovers $99/93/75\%$ at L0/2/4; both controls sit at $\approx$0.

wrong *contrast* survives), and trunk substitutability is gen→und, stripe stimuli, 1.5B only.

Cross-architecture contrast, made causal. Is the divergence profile isolation structure, or an accident of Harmon? On Show-o (Xie et al., 2025), a 1.3B model of the opposite design (one shared transformer, text and image tokens *joint from layer 0*), the same CKA protocol declines from the first layer ($0.995 \rightarrow 0.354$; Figure 8)—where Harmon's trunk begins. A causal mirror of the swap (gen-pass rows into the und pass at all 1024 grid positions; self-swap and random-position controls; Table 11) finds *no layer at which same-content substitution is benign*: even same-image donors collapse the readout (-6.3 at L0, below the -5.1 scrambled-position null, $z = -9.1$), because Show-o writes layout and text-prefix context into the image rows from L0. Where Harmon's same-content swaps are near-no-ops everywhere, Show-o's are benign nowhere—yet at L0 the causal failure coexists with CKA 0.995: similarity overstates sharing.

Why similarity overstates sharing: one geometry, rotated bases. CKA is invariant to orthogonal transformations—it measures a representation's *geometry*, not its coordinates—while a downstream consumer reads rows in a fixed *basis*. If the two passes encode the same geometry in rotated bases, CKA reads high while a raw swap fails, and supplying the rotation should repair the swap. It does: one orthogonal map R per layer (Procrustes, fit gen→und on held-out natural images, model frozen, disjoint fit/test/control splits) causally recovers $99/93/75\%$ of Show-o's cross-swap collapse at L0/L2/L4 (POPE-yes COCO photographs, per-image object questions; n=53 over three independent splits plus a validity-floor sensitivity run; Figure 7). The recovery is specific to the correct isometry: R fit to *different* target images recovers $\approx$0, and the best rank-k linear maps fail—a rotation, not a projection, and not just any rotation. Recovery fades through L6–L8 (35–30%); only deep, where CKA collapses ($\rightarrow 0.08$) and no rotation reduces the Procrustes residual, does the *geometry* itself diverge. On Harmon, whose task-blind encoder keeps the passes near one basis (raw trunk deviations only 0.4–0.6 logits), alignment recovers 71–95% at layers 9–15, direction-consistent at small range. Instrument lesson: on procedural stripes the wrong-target control *also* recovered ($\approx 90\%$)—low-dimensional stimulus geometry cannot certify alignment specificity; natural images were required.

Negative control: Janus-Pro. Janus-Pro-1B (Chen et al., 2025b) decouples visual encoding (SigLIP-L for understanding, LlamaGen VQ-16 for generation) beneath a shared 24-layer trunk; the mirrored protocol yields cross-pass CKA of 0.21–0.31, flat across all 24 layers—*below* either pass's different-image baseline everywhere (Table 12): sharing absent by construction measures as absent, and no causal swap is needed below this floor. The three architectures span the sharing axis under one protocol (Figure 8): Harmon (isolated encoder) holds 0.998–1.000, interchangeable everywhere; Show-o (joint from L0) declines $0.995 \rightarrow 0.354$, interchangeable nowhere; Janus-Pro (decoupled) flat-low from L0. Causal interchangeability tracks where task context enters, not weight sharing per se.

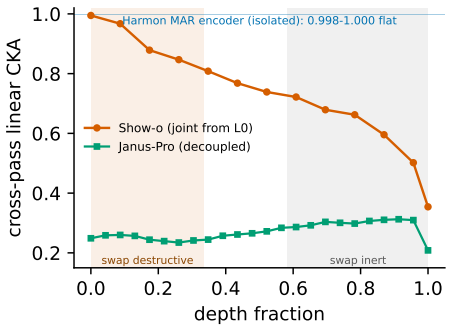

Figure 8: **Cross-architecture spectrum.** Cross-pass CKA by relative depth: Harmon's task-blind encoder flat at 0.998–1.000; Show-o (joint from L0) declining from layer 0 (0.995 $\rightarrow$ 0.354); Janus-Pro (decoupled) flat-low at 0.21–0.31, below its different-image baselines (Table 12). Shading: Show-o swap regimes (Table 11), destructive early, inert late.

Takeaway. The two tasks share one scene *geometry*; what differs is the *basis* each pass reads it in, and the basis rotates where the architecture injects task context—immediately in Show-o, only downstream in Harmon. True specialization is deeper still, where the geometry itself parts. The weights that are shared are the ones that never see the task.

6 Related Work

Unified multimodal models. The design space spans a sharing axis: fully shared backbones (Chameleon Chameleon Team, 2024; Emu3 Wang et al., 2024; MMaDA Yang et al., 2025), single transformers mixing autoregressive text with (discrete-)diffusion images (Show-o Xie et al., 2025; Transfusion Zhou et al., 2025), partially decoupled designs (Janus/Janus-Pro Wu et al., 2025a; Chen et al., 2025b; BAGEL Deng et al., 2025), and decoupled hybrids (BLIP3-o Chen et al., 2025a). Kang et al. (2026) show behaviorally that the choice matters, without a mechanistic account of *what* is shared; Harmon (Wu et al., 2025b), whose bit-identical MAR encoder (Li et al., 2024) is isolated from the task-conditioned trunk, is the axis's cleanest causal testbed, and our results supply that account.

Localization granularity. Circuit discovery localizes computation in *routing structures*—edges and mover heads—via necessity-based ablation (Conmy et al., 2023; Yao et al., 2024); neuron-level work localizes signals in MLP channels (Dai et al., 2022; Wang et al., 2022; Leng and Xiong, 2024; Tang et al., 2024); Meng et al. (2022) localize factual associations at the last subject-token position. We extend token-anchored localization to vision—visual facts are content-addressed at object-region token positions while attention routing is causally idle—suggesting routing-axis concentration is a property of sparse text retrieval, not of transformers per se; circuit-discovery pipelines should test both axes.

Activation-space editing and its boundary. Hase et al. (2023) showed causal tracing and weight editing answer different questions; we avoid that invalidated inference by localizing and editing at the same site, in activation space, with datapoint-level controls (cf. activation patching and causal mediation Vig et al., 2020; Meng et al., 2022; CKA Kornblith et al., 2019 on the correlational side).

Cross-task sharing. Function vectors (Todd et al., 2024) carry task identity through privileged heads; modular circuit reuse (He et al., 2025) matches components across tasks partly correlationally. Our cross-task rescue is the directly causal counterpart at representation level, and it *locates* the boundary where sharing ends rather than assuming it. In neuroscience, primate prefrontal cortex represents related tasks in shared subspaces engaged task-specifically (Tafazoli et al., 2025; Kumar et al., 2024; Goldstein et al., 2025); our result is that picture with an interventional stamp—substitution and repair, not alignment.

7 Conclusion

In a unified model whose visual encoder is task-blind by construction, we answered the two questions weight sharing leaves open. The encoder writes one *editable scene code*: substitution moves the answer predictably, either task can consume rows computed by the other, and generation-pass rows repair deleted understanding computation. Task differences emerge downstream: in *consumption*—the iterative consumer re-derives deleted evidence from the image's backups, where the one-shot reader collapses and corrupted content defeats both—and in the *task-conditioned trunk*, where what is lost first is coordinates, not content: a frozen orthogonal map repairs the cross-pass swap where raw substitution collapses. Unified models share one scene geometry; task differences come from the basis each consumer reads it in and how each consumer uses the code—both set by where the architecture injects task context.

Limitations

Evidence scope on the generation side. The generation-side causal evidence originated on synthetic stimuli—8 procedurally generated stripe scenes (the redundant-evidence regime) plus compact-object composite scenes (the unique-evidence regime)—chosen because they admit a closed-form generation readout. The redundancy mechanism itself, however, no longer rests on them: the instance-drop experiment (Table 6) establishes the backup effect, the consumer asymmetry, and the context re-derivation surplus on 98 real COCO scenes with full deletion, gt-yes-only items, and paired bootstrap CIs. Its own bounds are stated there: generation's clean-regeneration baseline loses signal on about half the items (making the generation recoveries conservative lower bounds), the completion-then-re-read readout is one readout, not a panel, and the und-side depth structure (one-shot redundancy-blindness fading by $L15$) means the consumer contrast is layer-scoped. An earlier synthetic paired design (within-factor redundant/unique members) was discarded in audit—its half-deletion and mixed ground-truth directions disqualify its magnitudes. The understanding-side map replicates on real COCO photographs (POPE: object-token ablation flips 50–70% of answers vs. 2–6% controls; Appendix E), and the regeneration benchmark adds a generation-side real-image null (cluster-band deletion leaves OWLv2 detectability at the random-band null, $|z| \leq 0.35$; Table 20)—the latter bounded by a compressed dynamic range, regeneration itself halving detector confidence (source $0.36 \rightarrow \approx 0.17$).

Directionality and layer coverage of substitution claims. Full-grid cross-pass substitutability is established in the und→gen direction only; the gen→und full-grid control has a dynamic range of 0.36 logits against a 3.6 baseline and supports no claim. The rescue result is strongest gen→und (12/20 layers); the strict bidirectional four-gate criterion passes at a minority of layers. No claim in this paper should be read as "activations are interchangeable at every layer in both directions."

Scale and model coverage. The composite-scene experiments (swap, rescue, band-width) are single-scale: the 0.5B model fails readout validation on those synthetic scenes and is excluded under the discipline of §2—though on real COCO photographs 0.5B is behaviorally valid (clean POPE accuracy 77.2%) and the real-image cluster causality replicates there (Table 19), so both the trunk divergence and the real-image causal map are now two-scale results. The trunk divergence replicates at 0.5B (CKA $0.997 \rightarrow 0.551$), but the trunk substitutability analysis is underpowered at that scale and remains one direction, quantified at 1.5B, stripe stimuli. The Show-o comparison now includes a causal cross-pass swap, but on 6 stimuli of one stimulus family, in the understanding direction only, and with the positional/text-context entanglement of the two passes as the measured property—it establishes the interchangeability contrast, not transfer of the full protocol. The decoupled-encoder negative control is no longer future work: Janus-Pro-1B's cross-pass CKA sits below the different-image floor at every shared-trunk layer (§5; Table 12), but that evidence is correlational (CKA) by design, its generation pass is teacher-forced, and extending the causal battery (deletion, swap, rescue) to decoupled architectures remains open.

Statistical and design caveats. Hit classification used $|z| \geq 2$ without multiple-comparison correction; headline cells sit at $|z| \approx 10$–39 where any correction is immaterial, but near-threshold cells should be read accordingly. The shuffle control is uninformative on spatially periodic stimuli at the encoder, so encoder position-specificity is untested there (in the trunk it is informative, and shows a position-bound code); rescue is content-specific but not position-specific within the deleted set. The clustering uses a fixed $k{=}4$ (k-stability and k-sweeps in Appendix A). The band-width experiment's seed accumulation is uneven across start layers (12 seeds at starts $\{2, 5\}$, 9–10 at the others). Finally, the encoder-level CKA of ≈ 1 is partly an architectural necessity (task-blindness by construction) and is never offered as a finding; the findings are the causal interchangeability, the consumer-by-evidence mechanism, and the progressive trunk fork.

References

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. 2025a. BLIP3-o: A family of fully open

unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Spotlight.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8493–8502. ArXiv:2104.08696.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*. BAGEL.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, and 3 others. 2025. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1):10529. Preprint arXiv:2310.07106.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 61125–61152.

Yinhan He, Wendy Zheng, Yaochen Dong, Yushun Zhu, Chen Chen, and Jundong Li. 2025. Towards global-level mechanistic interpretability: A perspective of modular circuits of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 22865–22880. PMLR.

Jiwon Kang, Heeji Yoon, Jaewoo Jung, Jaewon Min, Minkyeong Jeon, Biyeon Hwang, Sangwon Jung, and Seungryong Kim. 2026. Transferability between understanding and generation in unified multimodal models. In *European Conference on Computer Vision (ECCV)*. ArXiv:2607.04423.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15:5523.

Yongqi Leng and Deyi Xiong. 2024. Towards understanding multi-task learning (generalization) of LLMs via detecting and exploring task-specific neurons. *arXiv preprint arXiv:2407.06488*.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, volume 37. ArXiv:2406.11838.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

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

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2112.10752; introduces the KL-regularized latent VAE used by LDM.

Sina Tafazoli, Flora M. Bouchacourt, Adel Ardalan, Nikola T. Markov, Motoaki Uchimura, Marcelo G. Mattar, Nathaniel D. Daw, and Timothy J. Buschman. 2025. Building compositional tasks with shared neural subspaces. *Nature*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2402.16438; ACL Anthology 2024.acl-long.309.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11132–11152.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2025a. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2410.13848.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. 2025b. Harmonizing visual representations for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. ArXiv:2503.21979.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv:2408.12528.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38. ArXiv:2505.15809.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. In *Advances in Neural Information Processing Systems*, volume 37.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2025. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*. ArXiv:2408.11039.

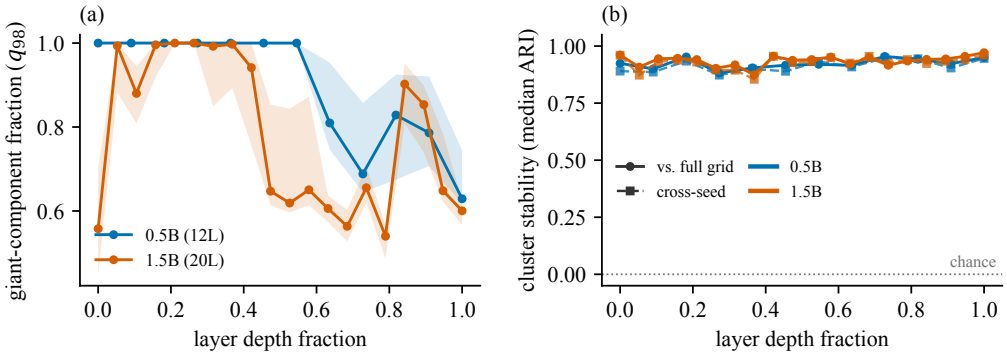

Figure 9: **Gate A.** Cluster stability (ARI) and spatial coherence across layers at both scales, and attention-graph fragmentation in the last third of depth.

Stimulus	k	ARI_{seed}	Overlap	Jaccard	$\text{ARI}^{\text{obj}}_{k=6}$	$\text{ARI}^{\text{set}}_{k=6}$
stripes	2	0.70±0.17	0.90±0.05	0.90±0.05	0.01±0.03	0.41±0.23
	3	0.70±0.16	0.97±0.03	0.77±0.20	0.39±0.35	0.57±0.34
	4	0.72±0.17	0.99±0.02	0.69±0.21	0.58±0.34	0.67±0.37
	6	0.71±0.15	1.00±0.01	0.57±0.18	—	—
	8	0.71±0.14	1.00±0.01	0.43±0.05	0.64±0.33	0.48±0.45
	10	0.68±0.13	1.00±0.00	0.41±0.06	0.58±0.30	0.39±0.42
two objects	2	0.43±0.39	0.45±0.27	0.45±0.27	0.22±0.42	0.41±0.30
	3	0.67±0.18	0.67±0.16	0.67±0.16	0.22±0.41	0.67±0.15
	4	0.58±0.24	0.76±0.25	0.60±0.19	0.72±0.39	0.70±0.38
	6	0.67±0.18	0.84±0.18	0.64±0.13	—	—
	8	0.71±0.14	0.93±0.09	0.58±0.11	0.55±0.42	0.75±0.21
	10	0.66±0.13	0.96±0.07	0.55±0.08	0.41±0.39	0.66±0.24

Table 1: Robustness of communication-contribution spectral clustering to the number of clusters k (Harmon-1.5B, pooled over encoder layers $\{2, 5, 8, 11, 14, 17\}$ and 6 stimulus seeds; mean±sd). ARI_{seed}: cross-seed partition stability; Overlap/Jaccard: object-cluster match to the striped-disc tokens; $\text{ARI}^{\text{obj}}_{k=6}$: label agreement with the k=6 partition restricted to object tokens; $\text{ARI}^{\text{set}}_{k=6}$: persistence of the object cluster as a token set relative to k=6.

A Gate details and clustering robustness

Gate A: cluster stability and coherence. Before any intervention we pre-registered three feasibility gates, all of which passed. Re-running the spectral clustering (Ng et al., 2002) on resampled attention estimates leaves the partition essentially unchanged: adjusted Rand index 0.87–0.97 across all layers at both scales, with spatial coherence ≈ 0.71 (clusters group neighboring image tokens rather than scattering across the grid). The attention affinity graph fragments in the last third of encoder depth (Figure 9).

Sensitivity to the number of clusters k. The choice $k=4$ was fixed before any intervention was run; Table 1 sweeps $k \in \{2, 3, 4, 6, 8, 10\}$ on the single-object and two-object stimuli at 1.5B. Cross-seed partition stability is flat in k ($\text{ARI}_{\text{seed}} \approx 0.7$ for every $k \geq 3$), and the object cluster survives as a token set (overlap with the object ≥ 0.97 for all $k \geq 3$ on single-object scenes), so the intervened token sets are not artifacts of $k=4$.

Gate B: per-layer selectivity, both scales, all three operators. Table 2 reports the null-calibrated selectivity by layer and operator at the gate stage, for both the targeted cluster and the discovery oracle, at both scales—including the pre-softmax ATTN-MASK operator that the routing contrast (Appendix C) shows to be near-null while the output-level operators are selective. The one ATTN-MASK cell that moves (+4.1/+4.7 at 0.5B layer 3) moves in the harmless direction.

		mean		zero		attn-mask	
Scale	L	z_{clus}	z_{orac}	z_{clus}	z_{orac}	z_{clus}	z_{orac}
1.5B	5	-4.55	-4.71	5.32	7.11	-0.05	0.49
	10	-2.60	-1.65	-2.71	-10.38	-0.09	-1.15
	15	-2.20	-2.56	-2.17	-14.11	0.33	-0.37
	18	-3.32	-5.02	-2.26	-24.72	-0.37	0.04
0.5B	3	0.33	0.05	-6.52	-7.69	4.12	4.66
	6	-0.75	-1.79	0.14	3.46	0.47	1.48
	9	-1.47	-2.26	-0.05	-1.15	0.69	1.28
	10	-0.74	-1.67	0.08	0.27	-0.23	-0.76

Table 2: Gate B (causal handle validation), full results including the attention-mask op: effect sizes of the communication-cluster (z_{clus}) and discovery-oracle (z_{orac}) ablation deltas against the size-matched random-set null (striped stimuli), per scale, layer, and ablation op.

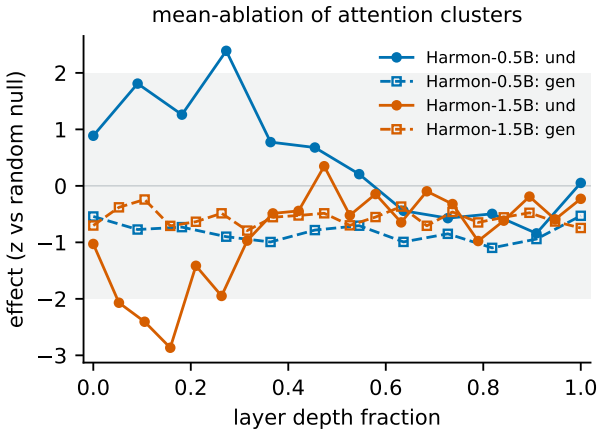

Figure 10: **Depth profile (MEAN-ABLATE cluster cells).** Null-calibrated z_{und} (solid) and z_{gen} (dashed) by layer depth at both scales; shaded band marks $|z| \leq 2$. z_{gen} stays within the null band at every layer of both scales.

B Full grids for the main protocols

Dual-path sweep (§4). Table 3 reports every cell of the dual-path sweep—all layers, both operators (MEAN-ABLATE/ZERO), both token sets, both scales—with z_{und}, z_{gen}, and the selectivity class per cell. No cell at either scale is generation-only or co-causal. The depth profile of the in-distribution MEAN-ABLATE operator is shown in Figure 10.

Band-width (deletion-depth) fragility (§4). Table 4 reports the full width × path × condition grid: normalized drop for cluster-band vs. size-matched random-band zeroing at widths $w \in \{1, 2, 3, 4, 6\}$, on both paths, at 1.5B. Both cluster curves are flat in width and already past 50% of the clean-to-floor range at $w = 1$; random bands never leave the noise band.

Dual-path sweep, representative layers (§4). Table 5 lists representative layers of the sweep plotted in Figure 5; the full grid is Table 3.

Instance-drop redundancy grid (§4). Table 6 reports the full instance-drop grid behind Figure 4: recovery by consumer, deletion arm, and instance multiplicity, with the paired contrasts and both disclosed bounds (the lossy clean-regeneration baseline; the depth profile of the one-shot reader).

Cross-pass activation patching (§3). Table 7 reports the full patching grid per scale, direction, and layer: the bitwise self-patch exactness check, the same-image cross-pass patch, and the mismatched-image and token-shuffled donor controls. The gen→und rows are included for completeness; as discussed in §3.2, that direction's negative control is too weak to support any claim.

	1.5B						0.5B					
	cluster			oracle			cluster			oracle		
L	z_u	z_g	cls	z_u	z_g	cls	z_u	z_g	cls	z_u	z_g	cls
op = mean												
0	-1.03	-0.70	–	-2.08	-0.95	U	0.89	-0.54	–	1.94	-0.55	–
1	-2.07	-0.38	U	-2.35	-0.38	U	1.81	-0.77	–	1.58	-0.79	–
2	-2.41	-0.24	U	-1.95	-0.25	–	1.26	-0.73	–	0.66	-1.36	–
3	-2.87	-0.71	U	-2.34	-0.60	U	2.39	-0.90	–	1.58	-0.88	–
4	-1.42	-0.63	–	-2.00	-0.70	U	0.77	-0.99	–	0.77	-0.97	–
5	-1.95	-0.49	–	-1.95	-0.50	–	0.68	-0.78	–	0.86	-1.52	–
6	-0.97	-0.79	–	-1.28	-0.80	–	0.21	-0.71	–	0.38	-1.71	–
7	-0.49	-0.55	–	-1.40	-0.63	–	-0.44	-0.99	–	0.16	-0.98	–
8	-0.44	-0.52	–	-1.10	-0.50	–	-0.57	-0.85	–	-0.04	-1.57	–
9	0.35	-0.49	–	-0.67	-0.61	–	-0.50	-1.10	–	-0.67	-1.08	–
10	-0.52	-0.70	–	-1.17	-0.74	–	-0.84	-0.94	–	-0.84	-1.63	–
11	-0.14	-0.55	–	-1.41	-0.67	–	0.05	-0.53	–	0.05	-1.67	–
12	-0.65	-0.37	–	-0.82	-0.36	–	–	–	–	–	–	–
13	-0.10	-0.71	–	-0.58	-0.70	–	–	–	–	–	–	–
14	-0.32	-0.46	–	-0.32	-0.47	–	–	–	–	–	–	–
15	-0.98	-0.65	–	0.04	-0.59	–	–	–	–	–	–	–
16	-0.62	-0.56	–	-0.46	-0.60	–	–	–	–	–	–	–
17	-0.19	-0.48	–	-0.35	-0.59	–	–	–	–	–	–	–
18	-0.59	-0.63	–	-1.19	-0.62	–	–	–	–	–	–	–
19	-0.23	-0.75	–	-0.37	-0.73	–	–	–	–	–	–	–
op = zero												
0	-1.11	-0.45	–	-8.14	0.42	U	-0.48	0.42	–	0.54	0.44	–
1	-24.00	0.80	U	-32.79	0.37	U	6.59	-0.24	–	7.78	-0.01	–
2	-39.10	0.40	U	-39.10	0.44	U	1.78	0.33	–	-19.83	0.72	U
3	0.98	0.06	–	-21.99	1.09	U	-8.12	0.15	U	-9.57	0.09	U
4	-3.36	-0.24	U	-21.84	0.95	U	-1.03	0.33	–	-0.53	0.32	–
5	4.86	0.28	–	6.42	0.28	–	-2.48	0.45	U	-3.40	0.55	U
6	3.92	-0.13	–	-7.03	0.70	U	0.20	0.19	–	3.22	0.14	–
7	-2.25	-1.10	U	-18.09	-0.02	U	-2.90	0.41	U	-3.22	0.33	U
8	3.67	-0.62	–	-18.41	1.03	U	1.68	-0.21	–	-4.04	0.11	U
9	2.49	-0.54	–	-7.70	0.76	U	-0.07	0.46	–	-1.20	0.56	–
10	-2.66	-0.71	U	-10.30	0.61	U	-0.02	-0.12	–	0.16	0.02	–
11	-12.40	0.27	U	-29.05	1.13	U	-1.13	0.87	–	-1.28	1.68	–
12	-22.97	-0.14	U	-21.94	0.16	U	–	–	–	–	–	–
13	-0.54	0.00	–	-16.26	1.25	U	–	–	–	–	–	–
14	-11.91	0.04	U	-21.37	0.96	U	–	–	–	–	–	–
15	-2.42	-0.21	U	-14.51	1.41	U	–	–	–	–	–	–
16	-15.59	-0.48	U	-22.26	0.53	U	–	–	–	–	–	–
17	-6.71	-0.17	U	-38.99	0.92	U	–	–	–	–	–	–
18	-2.15	0.03	U	-23.06	0.64	U	–	–	–	–	–	–
19	-1.95	-0.63	–	-15.14	0.91	U	–	–	–	–	–	–

Table 3: Dual-path v2 sweep, full results. Per layer, ablation op, and token set (communication cluster vs. discovery oracle): z-scores of the understanding (z_u) and generation (z_g) readout deltas against the size-matched random-set null, and the resulting selectivity class (U = und-only, G = gen-only, S = shared, – = neither). Left block: Harmon-1.5B (20 layers); right block: Harmon-0.5B (12 layers).

		Understanding		Generation		
Width w	Layers deleted	cluster	random	cluster	random	n
1	L	0.88±0.43	0.08±0.21	0.74±0.73	0.02±0.43	136
2	$L..L{+}1$	0.89±0.43	0.10±0.18	0.72±0.65	0.21±0.55	136
3	$L..L{+}2$	0.75±0.36	0.03±0.06	0.63±0.52	−0.05±0.33	176
4	$L..L{+}3$	0.77±0.37	0.06±0.09	0.73±0.71	0.06±0.47	116
6	$L..L{+}5$	0.88±0.53	0.06±0.10	0.73±0.61	0.04±0.41	116

Table 4: Band-width (deletion-depth) fragility on Harmon-1.5B, composite factor scenes (texture and color factors), start layers $L \in \{2, 5, 8, 11, 14, 17\}$ with 12 seeds at starts $\{2, 5\}$ and 9–10 at the others. Per cell: normalized drop (mean±sd), the fraction of the clean-to-floor range removed, $(\text{clean} - \text{deleted})/(\text{clean} - \text{floor})$, where the floor zeroes *all* grid rows over the same band and the deletion zeroes the target factor's contribution-cluster rows (*cluster*) or a size-matched random token set drawn once per (seed, layer, factor) and reused across widths (*random*) at every band layer on both paths. Width is the effective band width after end-clipping at the last encoder block. Both cluster curves are flat in width and cross 50% of the range already at $w = 1$: on these compact-object scenes, generation is as single-layer-fragile as understanding (Section 4). n per cell counts (seed, start-layer, factor) records; seed accumulation is uneven across start layers (12 at starts $\{2, 5\}$, 9–10 elsewhere), hence the varying n.

		Cluster ZERO		Oracle ZERO	
Scale	L	z_{und}	z_{gen}	z_{und}	z_{gen}
1.5B	1	−24.0	+0.80	−32.8	+0.37
	2	−39.1	+0.40	−39.1	+0.44
	12	−23.0	−0.14	−21.9	+0.16
	17	−6.7	−0.17	−39.0	+0.92
0.5B	3	−8.1	+0.15	−9.6	+0.09
	7	−2.9	+0.41	−3.2	+0.33

Table 5: Dual-path sweep, representative layers (stripe stimuli; all listed cells understanding-only). Full grid in Table 3.

Encoder representational-similarity map. Table 8 reports the per-layer cross-pass linear CKA on the shared visible tokens with the different-image control, and the within-/cross-pass LOSO ridge-probe transfer accuracies, at both scales (Figure 11). Because the encoder is task-blind by construction (§1), the flat 0.998–1.000 profile is an architectural expectation confirmed, not a finding; it is reported as the calibration backdrop against which the trunk divergence of §5 is read. The probe is at ceiling everywhere and hence uninformative; the CKA panel, with its calibrated different-image baseline, carries the evidence.

Factor swap (§3.1). Table 9 reports the full swap grid on the composite scenes at 1.5B: toward-fraction per layer and factor for the communication-cluster, discovery-oracle, and size-matched random swaps with understanding-pass donor rows, plus the oracle swap with generation-step-0 donor rows. The step-0 variant is much weaker (0.09–0.12), but the attenuation is built in: step-0 rows exist only for visible tokens ($\sim$47% of positions) and the masked strip is precisely where the factor evidence sits—itself consistent with the iterative-consumption picture of §4.

Rescue (§3.3). Table 10 reports the circuit-reuse experiment at 1.5B per layer: deletion factor-selectivity (target vs. unrelated factors), rescue fractions in both directions for the cross-pass donor against the content-flipped (wrong) and within-cluster shuffled donors, and the capture-only reuse cosine, which rises from 0.18 at layer 0 to 0.85 at layer 17, peaking where the strict gates pass. Under the pre-registered four-gate criterion (deletion factor-selective: target drop $\geq 20\%$ of the clean range, unrelated $< 5\%$; rescue content-specific: cross $\geq 50\%$, wrong $< 10\%$), the scale-up passes at layer 15 in both directions and at layers 4 and 17 gen→und; the relaxed repair criterion passes at 12/20 layers gen→und and 7/20 und→gen.

Show-o causal cross-pass swap (§5). Table 11 reports the full layer sweep for the causal cross-pass row swap on Show-o: shift in the understanding yes/no logit difference for the same-image, content-flipped,

Deletion	Consumer	Multi-inst.	Single-inst.
drop one	gen	**.82** [.71, .91]	—
	und ($L4$)	.63 [.50, .75]	—
drop all	gen	.54 [.41, .66]	.36 [.21, .50]
	und ($L4$)	.05 [.00, .13]	.02 [.00, .07]
random (null)	gen	1.00 [1.00, 1.00]	
	und ($L4$)	.74 [.64, .82]	
Paired contrasts (multi-instance, bootstrap 95% CI):			
backup: gen, one − all		+.29 [+.16, +.41]	
consumer: gen − und @ one		+.20 [+.07, +.32]	
re-derivation: gen − und @ all		+.48 [+.34, +.63]	

Table 6: **Instance-drop redundancy on natural scenes (Harmon-1.5B, COCO).** Per validated (image, object) item (n=98; gt yes; clean signed yes/no logit ≥ 1; 56 objects with ≥ 2 COCO instances, 42 with one), annotated instance footprints are deleted *in full*: *drop-one* removes the largest instance and leaves the others (a genuine spatial backup survives), *drop-all* removes every instance (no backup); 3 size-matched random regions per item are the null. The same token set is consumed two ways: *gen* masks it as the MaskGIT completion region (32-step iterative regeneration, then re-read), *und* zeroes the encoder rows and reads once. Cells report recovery (retention ≥ 0.5 of clean) with item-bootstrap 95% CIs. Understanding's drop-one recovery rises with depth ($.63/.73/.82$ at $L4/10/15$; drop-all stays $\leq .18$), so the $L4$ consumer gap is the conservative one; the drop-all contrast is significant at every layer ($+.36$ to $+.48$). Generation's clean-regeneration baseline itself loses signal (45/98 items retain $\geq .5$ after clean completion of the same region), so gen recoveries are conservative lower bounds.

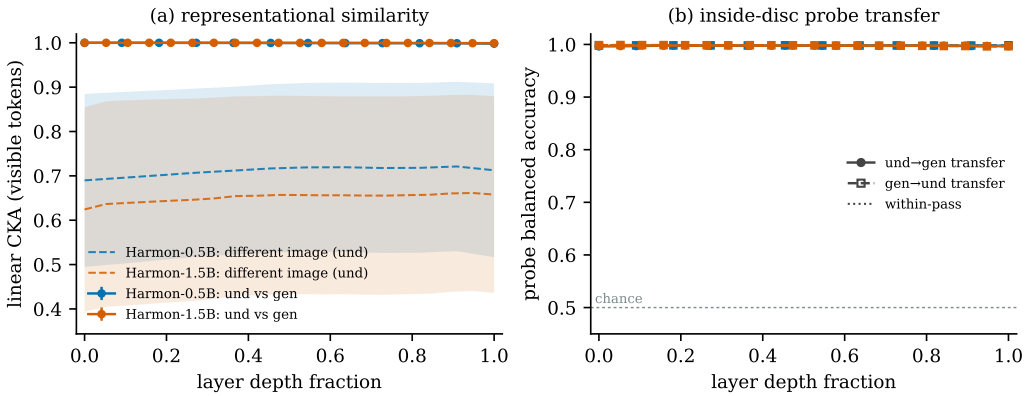

Figure 11: **Encoder-level representational identity between the passes.** (a) Cross-pass linear CKA on shared visible tokens by depth fraction at both scales (solid), against a different-image baseline (dashed): 0.998–1.000 everywhere vs. 0.62–0.72. (b) LOSO probe transfer, at ceiling and uninformative.

and random-position conditions, with the bit-exact self-swap validity control. The frequency-mismatched donor (not tabulated) behaves like the content-flipped donor early (-6.41 at L0) and goes inert on the same schedule. Show-o's two passes place the image block at different absolute positions (understanding 2–1025, generation 130–1153) under different text prefixes; the swap maps grid index to grid index and deliberately does not correct that difference—it is the architectural property under test.

Janus-Pro decoupled-encoder negative control (§5). Table 12 reports the full per-layer grid behind the Janus-Pro-1B negative control: linear CKA per shared-trunk layer between understanding-pass (SigLIP-L) and generation-pass (LlamaGen VQ-16) activations at the 576 image-token positions, with the within-pass different-image baselines. The protocol mirrors the Show-o replication—same striped-disc stimulus family (384px, 6 stimuli)—and the 24×24 SigLIP patch grid and VQ code grid match exactly, so positions correspond one-to-one. The generation rows come from a causal teacher-forced forward pass, so the two passes are internally symmetric in this respect; the evidence is correlational by design—cross-pass similarity below the different-image floor at every layer certifies absence without a causal swap.

	1.5B					0.5B				
L	d_{self}	d_{cross}	d_{mism}	d_{shuf}	compat	d_{self}	d_{cross}	d_{mism}	d_{shuf}	compat
direction = gen→und										
0	0.000	0.000	-0.281	-0.281	1.0000	0.000	-0.000	0.359	0.031	1.0000
1	0.000	0.016	-0.203	-0.234	0.9230	0.000	0.000	0.328	0.047	1.0000
2	0.000	-0.016	-0.250	-0.203	0.9376	0.000	-0.031	0.297	-0.016	0.8946
3	0.000	-0.031	-0.203	-0.203	0.8462	0.000	0.016	0.219	-0.000	0.9287
4	0.000	0.000	-0.219	-0.219	0.9999	0.000	0.031	0.250	0.031	0.8749
5	0.000	0.031	-0.234	-0.172	0.8666	0.000	-0.016	0.203	-0.000	0.9229
6	0.000	0.047	-0.266	-0.203	0.8234	0.000	-0.016	0.219	0.047	0.9285
7	0.000	0.047	-0.219	-0.141	0.7857	0.000	-0.031	0.141	-0.000	0.7774
8	0.000	0.016	-0.234	-0.156	0.9333	0.000	-0.016	0.031	0.031	0.4990
9	0.000	-0.031	-0.234	-0.187	0.8668	0.000	-0.016	0.047	0.047	0.6659
10	0.000	0.000	-0.203	-0.141	0.9999	0.000	0.016	0.063	0.062	0.7499
11	0.000	0.047	-0.188	-0.172	0.7499	0.000	-0.016	0.063	0.047	0.7504
12	0.000	0.000	-0.156	-0.219	0.9999	–	–	–	–	–
13	0.000	0.000	-0.094	-0.187	0.9998	–	–	–	–	–
14	0.000	0.016	-0.188	-0.187	0.9165	–	–	–	–	–
15	0.000	0.047	-0.125	-0.203	0.6249	–	–	–	–	–
16	0.000	0.016	-0.125	-0.219	0.8748	–	–	–	–	–
17	0.000	0.031	-0.109	-0.156	0.7142	–	–	–	–	–
18	0.000	0.047	-0.125	-0.094	0.6249	–	–	–	–	–
19	0.000	0.000	-0.156	-0.156	0.9999	–	–	–	–	–
direction = und→gen										
0	0.000	0.069	-0.572	0.077	0.8801	0.000	0.184	-0.282	0.190	0.3470
1	0.000	0.068	-0.574	0.076	0.8810	0.000	0.182	-0.273	0.190	0.3337
2	0.000	0.069	-0.572	0.077	0.8801	0.000	0.181	-0.275	0.190	0.3399
3	0.000	0.068	-0.573	0.076	0.8810	0.000	0.181	-0.271	0.190	0.3325
4	0.000	0.068	-0.574	0.076	0.8810	0.000	0.180	-0.270	0.190	0.3337
5	0.000	0.068	-0.573	0.076	0.8807	0.000	0.178	-0.274	0.189	0.3482
6	0.000	0.069	-0.573	0.077	0.8799	0.000	0.178	-0.273	0.189	0.3494
7	0.000	0.069	-0.573	0.076	0.8804	0.000	0.176	-0.268	0.189	0.3451
8	0.000	0.068	-0.575	0.076	0.8814	0.000	0.175	-0.273	0.189	0.3592
9	0.000	0.068	-0.575	0.077	0.8813	0.000	0.176	-0.274	0.189	0.3587
10	0.000	0.068	-0.575	0.077	0.8815	0.000	0.176	-0.274	0.189	0.3600
11	0.000	0.068	-0.574	0.076	0.8810	0.000	0.173	-0.267	0.189	0.3515
12	0.000	0.068	-0.575	0.077	0.8811	–	–	–	–	–
13	0.000	0.068	-0.574	0.076	0.8809	–	–	–	–	–
14	0.000	0.068	-0.574	0.077	0.8812	–	–	–	–	–
15	0.000	0.068	-0.576	0.076	0.8816	–	–	–	–	–
16	0.000	0.068	-0.575	0.077	0.8816	–	–	–	–	–
17	0.000	0.068	-0.574	0.077	0.8813	–	–	–	–	–
18	0.000	0.068	-0.575	0.077	0.8817	–	–	–	–	–
19	0.000	0.068	-0.577	0.077	0.8815	–	–	–	–	–

Table 7: Cross-path activation patching, full results per scale, direction, and layer. d_{self}: self-patch deviation (exactness check); d_{cross}: same-image cross-pass patch; d_{mism}: mismatched-image donor; d_{shuf}: token-shuffled donor; compat $= 1 - |d_{\text{cross}}|/(|d_{\text{cross}}| + |d_{\text{mism}}|)$-style patch-compatibility index.

	1.5B						0.5B					
L	$\mathrm{CKA}_{\times}$	$\mathrm{CKA}_{\mathrm{diff}}$	$b_{u\to u}$	$b_{g\to g}$	$b_{u\to g}$	$b_{g\to u}$	$\mathrm{CKA}_{\times}$	$\mathrm{CKA}_{\mathrm{diff}}$	$b_{u\to u}$	$b_{g\to g}$	$b_{u\to g}$	$b_{g\to u}$
0	1.000	0.624	0.998	0.998	0.998	0.999	1.000	0.690	0.998	0.998	0.996	0.998
1	1.000	0.637	0.999	0.999	0.997	0.999	1.000	0.695	0.998	0.998	0.997	0.998
2	1.000	0.639	0.998	0.998	0.998	0.999	1.000	0.701	0.998	0.998	0.998	0.998
3	1.000	0.641	0.998	0.998	0.998	0.999	1.000	0.707	0.998	0.998	0.998	0.998
4	1.000	0.644	0.998	0.998	0.998	0.999	1.000	0.712	0.998	0.998	0.998	0.998
5	1.000	0.646	0.998	0.998	0.998	0.998	1.000	0.717	0.998	0.998	0.998	0.998
6	1.000	0.649	0.998	0.998	0.998	0.998	0.999	0.719	0.998	0.998	0.998	0.998
7	1.000	0.654	0.998	0.998	0.998	0.999	0.999	0.719	0.998	0.998	0.998	0.998
8	1.000	0.655	0.998	0.997	0.998	0.998	0.999	0.718	0.998	0.998	0.998	0.998
9	1.000	0.657	0.998	0.997	0.998	0.998	0.999	0.718	0.998	0.998	0.998	0.998
10	1.000	0.657	0.997	0.997	0.998	0.998	0.999	0.721	0.998	0.997	0.997	0.998
11	1.000	0.656	0.997	0.997	0.998	0.998	0.998	0.713	0.998	0.997	0.997	0.998
12	1.000	0.656	0.997	0.997	0.997	0.998	–	–	–	–	–	–
13	1.000	0.656	0.997	0.998	0.997	0.998	–	–	–	–	–	–
14	1.000	0.655	0.997	0.998	0.997	0.998	–	–	–	–	–	–
15	1.000	0.657	0.997	0.997	0.998	0.997	–	–	–	–	–	–
16	1.000	0.657	0.997	0.997	0.998	0.997	–	–	–	–	–	–
17	0.999	0.660	0.997	0.996	0.998	0.997	–	–	–	–	–	–
18	0.999	0.661	0.997	0.997	0.997	0.996	–	–	–	–	–	–
19	0.999	0.658	0.996	0.996	0.997	0.996	–	–	–	–	–	–

Table 8: Divergence map, full per-layer results (means over 8 stimuli). $\mathrm{CKA}_{\times}$: linear CKA between und- and gen-pass representations of the same image; $\mathrm{CKA}_{\mathrm{diff}}$: different-image control; $b_{a\to b}$: balanced accuracy of a linear probe trained on pass a and evaluated on pass b (u = understanding, g = generation).

	color				relation				shape				texture			
L	clus	orac	rand	g0-orac	clus	orac	rand	g0-orac	clus	orac	rand	g0-orac	clus	orac	rand	g0-orac
0	0.38	0.51	0.17	0.21	-0.00	0.91	0.02	0.08	0.01	0.29	0.01	0.13	0.00	0.88	-0.00	0.09
1	0.39	0.52	0.17	0.21	0.48	0.91	0.05	0.08	0.74	0.28	0.10	0.14	0.93	0.88	-0.01	0.09
2	0.87	0.53	0.18	0.20	0.43	0.91	0.05	0.09	0.67	0.27	0.07	0.14	0.85	0.88	-0.01	0.09
3	0.59	0.55	0.16	0.20	0.27	0.90	0.04	0.08	0.74	0.28	0.09	0.13	0.84	0.89	-0.00	0.09
4	0.31	0.55	0.16	0.20	0.24	0.92	0.05	0.09	0.52	0.27	0.07	0.12	0.85	0.90	-0.01	0.09
5	0.95	0.55	0.16	0.19	0.34	0.92	0.05	0.09	0.93	0.26	0.19	0.13	0.98	0.91	-0.01	0.09
6	0.27	0.56	0.15	0.19	0.17	0.93	0.04	0.09	0.35	0.23	0.06	0.11	0.96	0.92	-0.01	0.08
7	0.59	0.56	0.14	0.19	0.33	0.93	0.05	0.08	0.36	0.20	0.06	0.10	0.94	0.93	-0.00	0.08
8	0.64	0.56	0.15	0.19	0.10	0.93	0.04	0.08	0.09	0.20	0.03	0.09	0.01	0.93	-0.01	0.08
9	0.86	0.58	0.14	0.19	0.07	0.92	0.03	0.09	0.03	0.19	0.03	0.09	0.14	0.93	-0.01	0.07
10	0.89	0.59	0.14	0.19	0.30	0.93	0.04	0.09	0.24	0.16	0.03	0.09	0.62	0.93	-0.00	0.07
11	0.90	0.59	0.15	0.19	0.03	0.92	0.02	0.09	0.05	0.14	0.00	0.08	0.01	0.93	-0.00	0.07
12	0.81	0.59	0.13	0.19	0.14	0.92	0.04	0.09	0.07	0.12	0.01	0.06	0.30	0.92	-0.00	0.07
13	0.47	0.59	0.14	0.19	0.12	0.91	0.03	0.09	0.06	0.10	0.01	0.05	0.00	0.92	-0.00	0.07
14	0.86	0.60	0.13	0.19	0.40	0.92	0.05	0.10	1.00	0.09	0.05	0.05	0.99	0.92	-0.01	0.07
15	0.85	0.62	0.14	0.17	0.06	0.91	0.03	0.10	0.01	0.08	0.01	0.05	-0.01	0.91	-0.00	0.06
16	0.33	0.60	0.11	0.18	0.16	0.92	0.05	0.10	0.46	0.09	0.03	0.05	0.98	0.91	-0.01	0.06
17	0.53	0.61	0.11	0.17	0.32	0.92	0.03	0.10	0.01	0.09	0.00	0.05	0.73	0.90	-0.00	0.06
18	0.70	0.63	0.14	0.16	0.19	0.92	0.03	0.09	0.03	0.09	0.02	0.06	0.47	0.91	-0.01	0.05
19	0.31	0.62	0.11	0.16	0.05	0.93	0.02	0.10	0.02	0.09	0.00	0.06	-0.00	0.90	-0.00	0.04

Table 9: Factor-swap test on Harmon-1.5B, full results: fraction of the donor-ward readout shift (toward-frac) per layer and factor, for the communication-cluster swap (clus), discovery-oracle swap (orac), and size-matched random swap (rand) with understanding-pass donor rows, plus the oracle swap with generation-step-0 donor rows (g0-orac). Self-swap validity condition is 0.00 everywhere and omitted.

	dissociation (und drop)		rescue g2u			rescue u2g			
L	target	unrelated	cross	wrong	shuf	cross	wrong	shuf	Reuse(C)
0	0.10±0.11 (48)	0.04±0.04 (48)	0.78±0.27 (18)	-0.29	0.76	0.78±0.73 (11)	0.07	0.69	0.18±0.52
1	0.71±0.32 (48)	0.16±0.17 (48)	0.56±0.33 (48)	-0.24	0.37	1.00±0.25 (29)	0.66	0.86	0.72±0.19
2	0.77±0.22 (48)	0.07±0.04 (48)	0.95±0.27 (48)	-0.40	0.82	0.96±0.30 (33)	0.48	0.86	0.68±0.48
3	0.73±0.39 (48)	0.06±0.04 (48)	0.86±0.15 (48)	-0.23	0.69	0.93±0.43 (27)	0.14	0.81	0.65±0.49
4	0.52±0.42 (48)	0.04±0.03 (48)	0.98±0.27 (36)	-0.40	0.93	0.99±0.18 (34)	0.49	0.96	0.66±0.47
5	0.80±0.38 (48)	0.37±0.34 (48)	0.43±0.39 (41)	-0.15	0.31	0.95±0.36 (36)	0.37	0.91	0.80±0.23
6	0.52±0.34 (48)	0.29±0.15 (48)	0.71±0.26 (44)	0.10	0.65	0.95±0.24 (24)	0.44	0.90	0.41±0.44
7	0.56±0.24 (48)	0.05±0.03 (48)	0.74±0.24 (47)	-0.31	0.73	0.92±0.43 (33)	0.52	0.80	0.78±0.35
8	0.21±0.24 (48)	0.12±0.06 (48)	1.20±0.24 (28)	0.48	1.04	1.01±0.10 (19)	-0.06	0.93	0.53±0.49
9	0.49±0.56 (48)	0.09±0.09 (48)	0.77±0.22 (29)	0.19	0.69	1.08±0.29 (24)	0.12	0.88	0.48±0.37
10	0.53±0.29 (48)	0.08±0.07 (48)	0.77±0.15 (47)	-0.28	0.64	1.00±0.42 (29)	0.16	0.90	0.42±0.67
11	0.75±0.41 (48)	0.07±0.07 (48)	0.89±0.15 (44)	0.36	0.79	1.00±0.31 (24)	-0.33	0.98	0.79±0.31
12	0.50±0.37 (48)	0.11±0.06 (48)	0.79±0.24 (41)	-0.07	0.74	0.87±0.49 (27)	-0.07	0.84	0.61±0.52
13	0.12±0.13 (48)	0.04±0.03 (48)	0.90±0.14 (23)	0.15	0.86	0.92±0.44 (21)	-0.33	1.02	0.45±0.42
14	1.23±0.59 (48)	0.14±0.08 (48)	0.86±0.12 (48)	-0.06	0.52	0.98±0.16 (36)	0.44	0.92	0.82±0.34
15	0.24±0.23 (48)	0.04±0.04 (48)	0.83±0.19 (33)	0.05	0.77	0.99±0.33 (23)	-0.40	0.90	0.62±0.48
16	0.68±0.52 (48)	0.09±0.06 (48)	0.92±0.14 (36)	-0.05	0.73	0.95±0.31 (36)	0.45	0.89	0.66±0.47
17	0.65±0.27 (48)	0.03±0.03 (48)	0.90±0.08 (47)	-0.19	0.90	0.87±0.43 (24)	0.11	0.87	0.85±0.15
18	1.26±0.94 (48)	0.09±0.07 (48)	0.83±0.15 (45)	0.08	0.83	0.89±0.47 (22)	-0.03	0.88	0.73±0.24
19	0.20±0.23 (48)	0.07±0.06 (48)	0.80±0.22 (23)	0.20	0.75	0.82±0.57 (22)	0.05	0.87	0.65±0.28

Table 10: Circuit-reuse main experiment on Harmon-1.5B (recomputed per layer from the raw run records). Dissociation: normalized understanding drop of the target factor vs. mean |drop| of unrelated factors after ablating the factor cluster (mean±sd, n in parentheses). Rescue: fraction of the deleted effect recovered, $(\text{rescued} - \text{deleted})/(\text{clean} - \text{deleted})$, for cross-pass donors vs. content-flipped (wrong) and within-cluster shuffled donors, in both directions (g2u: gen rows into the und pass; u2g: und rows into the gen band). Reuse(C): cosine between the two passes' per-factor drop vectors.

		cross-pass donor		
L	self	same	wrong (toward)	rand-pos
0	0.000	−6.30	−6.55 (1.12)	−5.07
2	0.000	−6.37	−6.47 (1.10)	−5.30
4	0.000	−6.48	−6.54 (1.12)	−5.24
6	0.000	−6.32	−6.44 (1.10)	−5.40
8	0.000	−5.14	−5.46 (0.93)	−4.34
10	0.000	−2.70	−3.23 (0.55)	−2.47
12	0.000	−0.60	−1.15 (0.20)	−0.53
14	0.000	+0.06	−0.37 (0.06)	−0.10
16	0.000	+0.59	+0.36 (−0.06)	+0.27
18	0.000	+0.91	+0.65 (−0.11)	+0.38
20	0.000	+0.56	+0.52 (−0.09)	+0.39
22	0.000	+0.37	+0.39 (−0.07)	+0.38

Table 11: Show-o causal cross-pass row swap (6 stimuli, understanding direction, every second trunk layer). Cell values are the change in the understanding yes/no logit difference vs. the clean run (clean baseline $+4.3$) when generation-pass residual rows at all 1024 image-grid positions are substituted into the understanding pass at layer L: bitwise self-swap control (self; exactly 0 at every layer), same-image cross-pass donor (same), content-flipped donor (wrong; in parentheses the fraction of the shift toward the donor's answer, >1 denoting overshoot), and the random-position control (rand-pos; 3 draws per stimulus, $n = 18$). Wrong-content donors flip the answer to the donor's on every stimulus through L8; the same-content donor is more destructive than the position-scrambled null through L10 ($z \leq -2.8$) and no layer exists at which it is benign; from L14 all image-row interventions are inert (readout consolidated outside the image rows). The swap maps grid index to grid index and deliberately preserves the two passes' positional/text-prefix difference—the architectural property under test.

L	$\mathrm{CKA}_{\times}$	$\mathrm{CKA}_{\mathrm{diff}}^{\mathrm{und}}$	$\mathrm{CKA}_{\mathrm{diff}}^{\mathrm{gen}}$
0	0.249 ± 0.012	0.689 ± 0.146	0.622 ± 0.180
1	0.259 ± 0.012	0.696 ± 0.145	0.661 ± 0.194
2	0.260 ± 0.012	0.709 ± 0.144	0.686 ± 0.198
3	0.257 ± 0.011	0.727 ± 0.141	0.705 ± 0.197
4	0.244 ± 0.010	0.743 ± 0.140	0.713 ± 0.207
5	0.239 ± 0.011	0.753 ± 0.138	0.720 ± 0.210
6	0.235 ± 0.010	0.761 ± 0.138	0.726 ± 0.208
7	0.241 ± 0.008	0.756 ± 0.138	0.732 ± 0.209
8	0.244 ± 0.007	0.759 ± 0.138	0.745 ± 0.204
9	0.257 ± 0.007	0.750 ± 0.139	0.753 ± 0.201
10	0.261 ± 0.005	0.748 ± 0.139	0.765 ± 0.202
11	0.265 ± 0.006	0.745 ± 0.139	0.772 ± 0.200
12	0.272 ± 0.006	0.734 ± 0.140	0.767 ± 0.204
13	0.284 ± 0.007	0.718 ± 0.141	0.786 ± 0.205
14	0.286 ± 0.007	0.709 ± 0.142	0.786 ± 0.202
15	0.292 ± 0.008	0.703 ± 0.142	0.791 ± 0.200
16	0.304 ± 0.007	0.683 ± 0.144	0.784 ± 0.202
17	0.301 ± 0.006	0.664 ± 0.145	0.783 ± 0.200
18	0.298 ± 0.006	0.646 ± 0.145	0.786 ± 0.196
19	0.307 ± 0.006	0.629 ± 0.147	0.781 ± 0.199
20	0.311 ± 0.005	0.615 ± 0.148	0.780 ± 0.199
21	0.313 ± 0.005	0.598 ± 0.148	0.782 ± 0.194
22	0.310 ± 0.008	0.573 ± 0.151	0.781 ± 0.191
23	0.208 ± 0.032	0.401 ± 0.145	0.656 ± 0.294

Table 12: Janus-Pro-1B decoupled-encoder negative control (mean $\pm$ sd over 6 striped-disc stimuli, 384px; protocol mirrored from the Show-o replication). $\mathrm{CKA}_{\times}$: linear CKA per shared-trunk layer between understanding-pass (SigLIP-L) and generation-pass (LlamaGen VQ-16, teacher-forced) activations of the *same* image at the 576 image-token positions (the 24×24 SigLIP patch grid and VQ code grid match exactly); $\mathrm{CKA}_{\mathrm{diff}}^{\mathrm{und}}$/$\mathrm{CKA}_{\mathrm{diff}}^{\mathrm{gen}}$: within-pass different-image baselines. Cross-pass CKA (0.208–0.313) sits *below* both baselines at every layer: the two passes disagree about the same image more than one pass disagrees about different images, so sharing absent by construction measures as absent. The generation rows come from a causal teacher-forced forward (both passes internally symmetric in this respect); the evidence is correlational (CKA), matching the negative-control role—no causal swap is needed once similarity is below the content-variation floor.

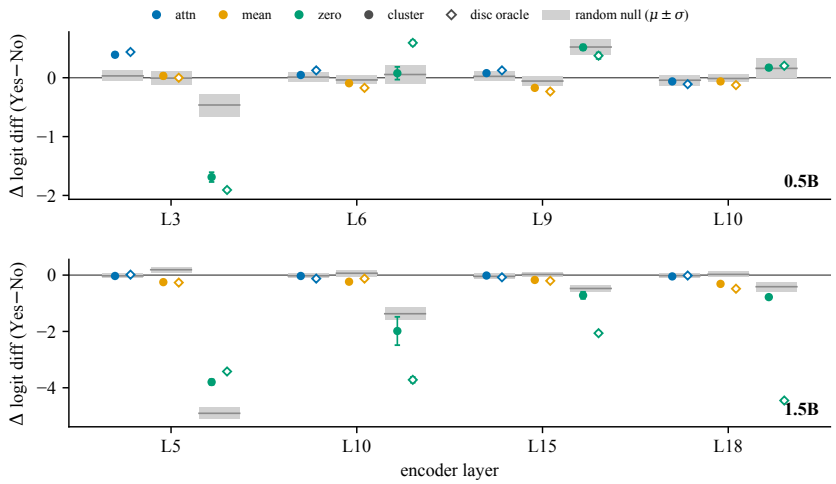

Figure 12: **Operator contrast (Gate B).** Null-calibrated selectivity by layer and operator on the understanding readout: output-level MEAN-ABLATE/ZERO are selective while pre-softmax ATTN-MASK is near-null—the stripe signal survives attention re-routing and travels through value and residual pathways.

Bootstrap confidence intervals. Table 13 reports stimulus-level bootstrap 95% confidence intervals (10,000 resamples over the stimulus/seed/item axis) for the headline swap, trunk-patching, band-width, and POPE quantities. The band-width 50%-crossing for generation on color is right-censored ($[1.03, \infty)$; 22% of resamples never cross within the measured widths), consistent with the redundancy-conditional reading in §4; POPE intervals are item-level over the 57 clean-correct object-present items.

C Head-knockout atlas and routing contrast

The operator contrast (Figure 12) holds the token set fixed and varies the stage of the layer's computation: pre-softmax ATTN-MASK removes the cluster from the attention routing while leaving its token states in the residual stream, and is near-null everywhere, while output-level MEAN-ABLATE/ZERO on the identical tokens are strongly selective. The head atlas (Figure 13; Table 14) varies the routing unit: each head's output-projection slice is zeroed for all tokens at one layer, propagated through both passes, on 4 stimuli, scored relative to sibling heads with a robust z (median/MAD), with an all-heads-at-a-layer knockout as the collective reference. The strongest single-head effect is $|z| \approx 6$; single-axis flag rates (8–12%) sit at or below the $\approx 15\%$ rate that a synthetic-noise calibration of the same pipeline returns under pure noise; and even deleting all heads at a layer moves the understanding logit difference by at most ≈ 0.3 logits, an order of magnitude below cluster-level block-output zeroing at the same layers. Three caveats: the atlas uses 4 stimuli; the z-scores are relative within a layer, so uniformly important heads would score flat (the all-heads reference addresses this, and it too is small); and the knockout removes only the head's output-projection contribution. One boundary: the operator profile is capability-dependent—count questions respond to MEAN-ABLATE while ZERO stays near null (Appendix D)—so operator choice is part of the hypothesis, not a nuisance parameter.

D Generality experiments

Two-object dissociation. On single-object stimuli the object cluster is nearly degenerate (it covers most of the informative region), so we generate scenes with a striped disc *and* a plain square in disjoint, side-randomized slots. The partition then supplies a natural size-matched control: the *other* object's cluster. At 1.5B, 38 of 40 (layer, operator) cells are object-selective on understanding (deleting the striped object's cluster reaches $z_{\text{und}} = -51.2$ at layer 17 ZERO, while the plain square's cluster at the same layer is harmless); at 0.5B, 17 of 24 cells replicate the direction. Generation never moves at either scale. The two non-selective 1.5B cells are cluster-assignment misses, not absences of signal (the disc oracle stays at -15.5 and -20.3). Full grid: Table 15.

quantity	subset	point	95% CI	n
Row-swap (shift toward donor, `toward_frac`)				
oracle row	gen0/pooled	0.111	[0.100, 0.121]	12
oracle row	gen0/best L2	0.132	[0.120, 0.145]	12
oracle row	gen0/L2	0.132	[0.119, 0.145]	12
oracle row	gen0/L10	0.108	[0.096, 0.120]	12
oracle row	gen0/L17	0.094	[0.085, 0.103]	12
random position	gen0/pooled	0.028	[0.025, 0.030]	12
random position	gen0/best L1	0.053	[0.045, 0.061]	12
random position	gen0/L2	0.045	[0.036, 0.052]	12
random position	gen0/L10	0.032	[0.025, 0.040]	12
random position	gen0/L17	0.022	[0.017, 0.028]	12
oracle row	und/pooled	0.645	[0.602, 0.686]	12
oracle row	und/best L5	0.660	[0.612, 0.708]	12
oracle row	und/L2	0.646	[0.600, 0.692]	12
oracle row	und/L10	0.653	[0.606, 0.696]	12
oracle row	und/L17	0.630	[0.587, 0.672]	12
random position	und/pooled	0.055	[0.052, 0.058]	12
random position	und/best L5	0.098	[0.084, 0.118]	12
random position	und/L2	0.072	[0.063, 0.080]	12
random position	und/L10	0.053	[0.045, 0.062]	12
random position	und/L17	0.036	[0.030, 0.043]	12
LLM-trunk cross-pass patching				
Δ cross	15b/L9	-0.469	[-0.510, -0.427]	12
Δ mismatch	15b/L9	-0.708	[-0.750, -0.667]	12
Δ cross	15b/L12	-0.427	[-0.490, -0.365]	12
Δ mismatch	15b/L12	-0.521	[-0.583, -0.448]	12
Δ cross	15b/L15	-0.562	[-0.615, -0.510]	12
Δ mismatch	15b/L15	-0.583	[-0.625, -0.542]	12
CKA cross	15b/first(L-1)	0.997	[0.996, 0.997]	12
CKA mismatch	15b/first(L-1)	0.724	[0.629, 0.808]	12
CKA cross	15b/mid(L13)	0.926	[0.922, 0.930]	12
CKA mismatch	15b/mid(L13)	0.737	[0.673, 0.794]	12
CKA cross	15b/last(L27)	0.421	[0.402, 0.441]	12
CKA mismatch	15b/last(L27)	0.345	[0.313, 0.377]	12
Δ cross	05b/L6	-0.177	[-0.240, -0.115]	12
Δ mismatch	05b/L6	-0.198	[-0.240, -0.156]	12
Δ cross	05b/L9	-0.146	[-0.198, -0.094]	12
Δ mismatch	05b/L9	-0.188	[-0.250, -0.125]	12
Δ cross	05b/L12	-0.052	[-0.104, 0.000]	12
Δ mismatch	05b/L12	-0.042	[-0.083, 0.000]	12
CKA cross	05b/first(L-1)	0.997	[0.996, 0.997]	12
CKA mismatch	05b/first(L-1)	0.693	[0.591, 0.787]	12
CKA cross	05b/mid(L11)	0.924	[0.921, 0.928]	12
CKA mismatch	05b/mid(L11)	0.744	[0.690, 0.794]	12
CKA cross	05b/last(L23)	0.551	[0.511, 0.587]	12
CKA mismatch	05b/last(L23)	0.349	[0.295, 0.401]	12
Band-width deletion (50%-crossing width, cluster)				
w_{50}	gen/color	1.651	[1.028, ∞]	9
w_{50}	gen/texture	1.000	[1.000, 1.000]	9
w_{50}	und/color	1.000	[1.000, 1.000]	9
w_{50}	und/texture	1.000	[1.000, 1.000]	9
POPE adversarial (COCO; flip rate, clean-correct *yes*)				
flip rate	L5/mean/cluster	0.526	[0.404, 0.649]	57
flip rate	L5/mean/oracle	0.737	[0.614, 0.842]	57
flip rate	L5/mean/clus\orac	0.107	[0.036, 0.193]	57
flip rate	L5/mean/rand	0.094	[0.035, 0.164]	57
flip rate	L5/zero/cluster	0.825	[0.719, 0.912]	57
flip rate	L5/zero/oracle	0.719	[0.597, 0.825]	57
flip rate	L5/zero/clus\orac	0.357	[0.232, 0.482]	57
flip rate	L5/zero/rand	0.661	[0.538, 0.772]	57
flip rate	L10/mean/cluster	0.544	[0.421, 0.667]	57
flip rate	L10/mean/oracle	0.702	[0.579, 0.825]	57
flip rate	L10/mean/clus\orac	0.078	[0.019, 0.160]	57
flip rate	L10/mean/rand	0.035	[0.000, 0.082]	57
flip rate	L10/zero/cluster	0.491	[0.351, 0.614]	57
flip rate	L10/zero/oracle	0.719	[0.597, 0.825]	57
flip rate	L10/zero/clus\orac	0.098	[0.020, 0.188]	57
flip rate	L10/zero/rand	0.088	[0.029, 0.158]	57
flip rate	L15/mean/cluster	0.474	[0.351, 0.614]	57
flip rate	L15/mean/oracle	0.667	[0.544, 0.789]	57
flip rate	L15/mean/clus\orac	0.098	[0.020, 0.188]	57
flip rate	L15/mean/rand	0.053	[0.012, 0.099]	57
flip rate	L15/zero/cluster	0.509	[0.386, 0.632]	57
flip rate	L15/zero/oracle	0.649	[0.526, 0.772]	57

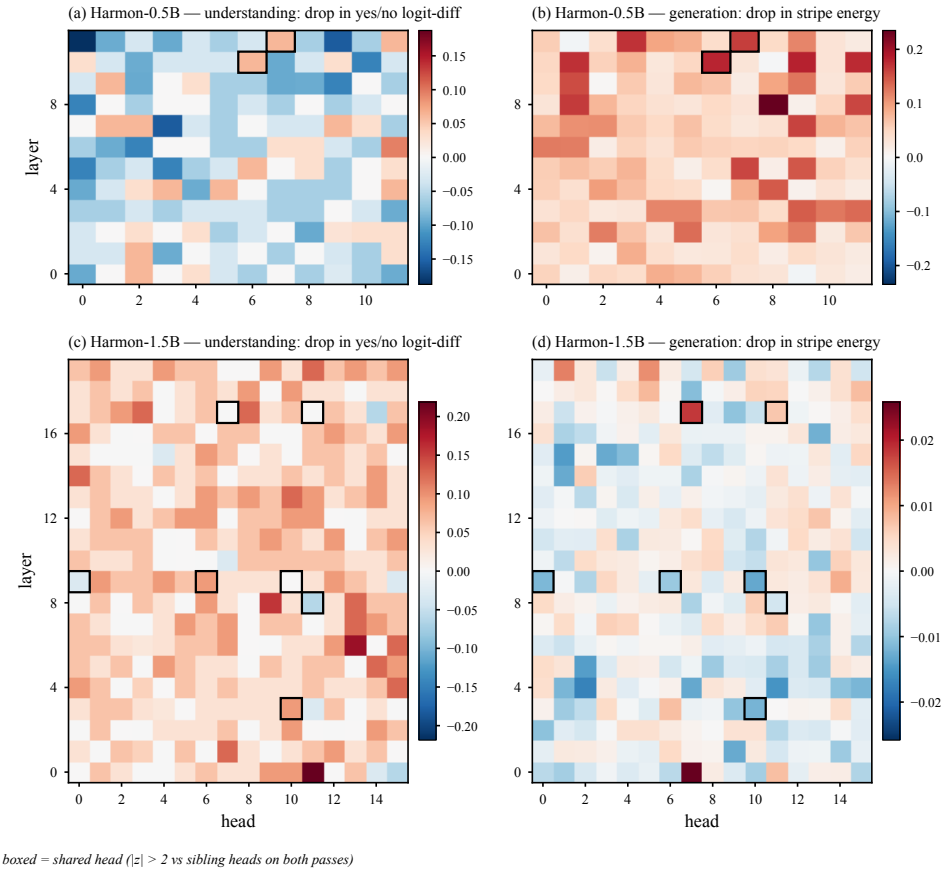

Figure 13: **Head-level knockout atlas: no privileged heads.** Layers × heads heatmaps of per-head knockout effects through both passes: drop in the understanding logit difference (a, c) and in the generation stripe energy (b, d), at 0.5B (a, b) and 1.5B (c, d). Boxes mark shared flags. Effects are small and dispersed everywhere.

	1.5B							0.5B						
L	sh	und	gen	in	$\max \lvert z \rvert$	Δ_u^{all}	Δ_g^{all}	sh	und	gen	in	$\max \lvert z \rvert$	Δ_u^{all}	Δ_g^{all}
0	0	2	1	13	4.45	-0.31	0.24	0	0	1	11	2.04	-0.03	0.17
1	0	1	2	13	3.31	0.03	-0.01	0	1	1	10	3.67	-0.19	0.10
2	0	0	0	16	1.94	0.00	0.00	0	1	0	11	2.02	-0.00	0.12
3	1	1	3	11	3.56	0.03	-0.01	0	2	0	10	2.02	0.19	0.18
4	0	1	0	15	2.60	-0.06	-0.01	0	0	1	11	2.54	0.09	0.15
5	0	0	0	16	1.35	-0.19	-0.01	0	3	2	7	4.19	0.09	0.10
6	0	2	0	14	5.06	-0.19	-0.01	0	1	4	7	5.78	0.03	0.06
7	0	0	1	15	2.18	-0.06	0.00	0	1	1	10	2.70	0.06	0.16
8	1	2	2	11	3.65	-0.13	0.00	0	1	3	8	6.32	0.03	0.16
9	3	4	1	8	3.72	-0.16	0.00	0	2	0	10	2.69	0.28	0.07
10	0	4	4	8	4.06	-0.06	0.00	1	1	3	7	3.13	0.12	0.17
11	0	0	0	16	1.53	0.06	-0.00	1	3	1	7	2.70	0.22	0.15
12	0	0	1	15	2.90	0.00	0.00	–	–	–	–	–	–	–
13	0	0	1	15	2.35	-0.06	0.00	–	–	–	–	–	–	–
14	0	1	4	11	6.36	-0.12	-0.01	–	–	–	–	–	–	–
15	0	0	3	13	3.14	0.03	0.00	–	–	–	–	–	–	–
16	0	0	2	14	3.62	-0.12	-0.02	–	–	–	–	–	–	–
17	2	5	3	6	6.71	-0.03	0.00	–	–	–	–	–	–	–
18	0	2	3	11	4.22	-0.03	-0.00	–	–	–	–	–	–	–
19	0	0	0	16	1.84	-0.22	-0.01	–	–	–	–	–	–	–

Table 14: Head-level co-loss atlas, per scale and layer: number of heads classified as shared (sh), und-only (und), gen-only (gen), and inert (in); largest per-head effect size $\max |z|$ over both readouts; and the all-heads reference deltas Δ_u^{all}, Δ_g^{all} (ablating every head of the layer at once) for the understanding and generation readouts.

	1.5B							0.5B						
L	z^u_{str}	z^u_{oth}	z^u_{orac}	z^u_{sq}	z^g_{str}	z^g_{oth}	sel	z^u_{str}	z^u_{oth}	z^u_{orac}	z^u_{sq}	z^g_{str}	z^g_{oth}	sel
op = mean														
0	-3.82	-1.09	-13.00	-1.52	-0.42	0.36	Y	-3.65	2.04	-4.08	1.67	-1.82	1.44	Y
1	-7.16	0.72	-8.33	-1.41	-0.77	0.06	Y	-5.18	3.17	-5.53	2.32	-1.72	1.69	Y
2	-8.83	-1.34	-10.96	-1.23	-0.79	0.08	Y	-7.34	2.21	-7.64	2.91	-1.95	0.68	Y
3	-5.94	-0.89	-6.92	-1.62	-0.86	0.20	Y	-6.51	3.02	-8.58	2.55	-1.73	0.78	Y
4	-9.28	-1.01	-10.02	-1.13	-0.80	0.53	Y	-4.31	3.78	-11.28	3.78	-1.80	1.02	Y
5	-7.48	-0.61	-7.60	-0.98	-0.57	0.66	Y	-4.56	4.62	-10.11	4.47	-1.31	1.00	Y
6	-10.10	0.49	-9.98	0.00	-0.64	0.46	Y	-2.94	3.40	-5.75	3.52	-2.19	0.33	Y
7	-13.29	1.20	-11.84	0.37	-0.56	1.17	Y	-3.20	3.51	-5.63	3.80	-1.76	0.51	Y
8	-16.97	1.06	-15.45	1.21	-0.94	0.25	Y	-2.07	4.88	-5.45	4.17	-1.44	0.96	Y
9	-15.03	2.48	-10.79	1.06	-0.68	-0.05	Y	-3.69	4.17	-4.33	3.85	-1.90	0.52	Y
10	-16.04	0.91	-13.89	1.03	-0.91	0.66	Y	-2.44	3.49	-2.60	3.49	-1.67	1.13	Y
11	-3.56	1.34	-14.35	0.33	-0.58	0.79	Y	-1.05	2.57	-1.05	3.66	-1.81	0.90	N
12	-20.50	2.46	-13.50	1.11	-0.81	0.52	Y	–	–	–	–	–	–	–
13	-1.69	0.99	-15.53	0.99	-0.55	0.56	N	–	–	–	–	–	–	–
14	-9.55	1.06	-6.07	0.66	-0.95	1.56	Y	–	–	–	–	–	–	–
15	-3.41	1.03	-12.85	0.79	-0.74	0.43	Y	–	–	–	–	–	–	–
16	-16.39	0.86	-12.11	0.44	-0.83	0.76	Y	–	–	–	–	–	–	–
17	-14.79	1.15	-16.87	0.60	-0.93	0.13	Y	–	–	–	–	–	–	–
18	-2.95	0.69	-15.74	0.69	-0.71	0.91	Y	–	–	–	–	–	–	–
19	-0.24	0.77	-20.29	0.59	-0.71	0.41	N	–	–	–	–	–	–	–
op = zero														
0	-4.32	1.77	-19.28	1.02	-0.09	0.69	Y	-1.73	0.55	-7.26	1.46	0.11	0.70	N
1	-19.29	0.95	-11.88	1.28	0.40	0.07	Y	-1.15	2.64	-0.05	3.36	0.13	0.13	N
2	-15.90	-1.48	-22.49	-1.29	0.45	0.29	Y	-7.89	2.60	-6.61	2.82	-0.42	0.33	Y
3	-17.95	0.54	-22.19	0.15	0.38	-0.09	Y	-0.60	1.97	-3.85	-1.42	-0.08	0.67	N
4	-20.59	-1.23	-20.78	-1.23	0.71	0.47	Y	-4.67	2.82	-5.03	2.94	0.05	-0.47	Y
5	-9.64	3.53	-2.20	2.65	0.64	-0.85	Y	-0.46	3.01	-9.63	2.92	-0.21	-0.74	N
6	-9.06	2.35	-7.07	2.60	-0.07	0.12	Y	-6.15	4.47	-3.89	4.67	-0.32	-0.33	Y
7	-13.71	0.94	-16.09	0.85	0.42	0.13	Y	-6.75	3.69	-5.78	3.99	0.39	0.07	Y
8	-6.73	-1.16	-6.25	-1.21	0.47	0.46	Y	-1.16	1.30	-8.34	1.30	-0.69	-0.40	N
9	-25.51	-0.74	-20.81	-0.15	0.22	-0.16	Y	-3.26	1.57	-5.72	2.61	-0.45	0.31	Y
10	-21.60	2.05	-29.70	1.73	0.06	-0.19	Y	-2.30	1.57	0.55	2.89	-0.40	-0.38	Y
11	-20.57	2.27	-40.67	1.51	-0.23	0.18	Y	2.76	5.89	2.45	6.10	0.07	-0.05	N
12	-26.81	0.66	-19.48	3.17	0.50	-0.09	Y	–	–	–	–	–	–	–
13	-11.63	0.36	-44.23	-0.12	0.20	-0.06	Y	–	–	–	–	–	–	–
14	-6.60	-0.05	-5.68	0.05	0.94	0.20	Y	–	–	–	–	–	–	–
15	-22.40	3.05	-59.32	2.35	0.20	0.05	Y	–	–	–	–	–	–	–
16	-38.80	0.41	-39.52	0.27	0.47	0.16	Y	–	–	–	–	–	–	–
17	-51.20	1.73	-57.03	1.86	0.10	0.05	Y	–	–	–	–	–	–	–
18	-15.33	4.23	-60.35	2.70	-0.37	0.41	Y	–	–	–	–	–	–	–
19	-4.13	-0.82	-50.29	-0.56	-0.06	0.10	Y	–	–	–	–	–	–	–

Table 15: Multi-object selectivity sweep, full results per scale, layer, and op. z^u: understanding-readout effect sizes when ablating the striped object's cluster (str), the other object's cluster (oth), the discovery oracle (orac), or the square object's cluster (sq); z^g: generation-readout effect sizes for the striped/other clusters; sel = Y when the understanding effect is selective for the striped object.

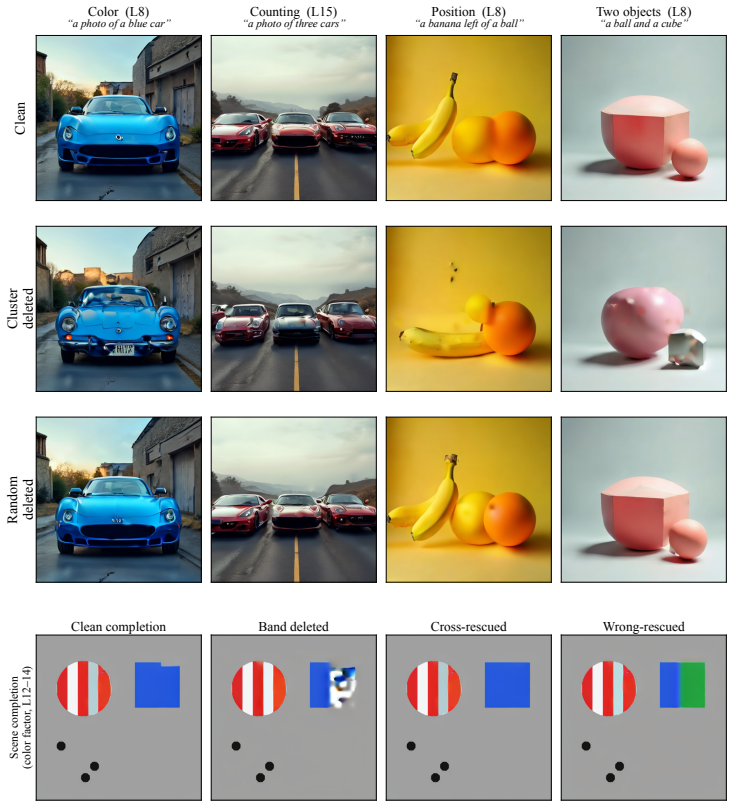

Figure 14: **Qualitative panel.** *Rows 1–3:* Harmon-1.5B text-to-image generations for four GenEval-style prompt categories under three conditions: clean, prompt-relevant cluster deleted at every MaskGIT step, and size-matched random deletion. Cluster deletion corrupts the prompt-relevant content while random deletion is nearly indistinguishable from clean. *Row 4:* factor-scene rescue (color factor, band L12–14): deletion corrupts the completion, understanding-pass donor rows restore it, and a content-flipped donor does not.

Capability suite (understanding readout only). We repeat the object-cluster ablation protocol on shape, count, and relation questions with procedurally controlled ground truth (Table 16). Object-cluster ablations are selective for all three capabilities, with capability-specific depth and operator profiles: shape responds to late-layer ZERO, count to mid-depth MEAN-ABLATE (inverting the stripe-task operator ordering), relation strongest early-to-mid at 1.5B. The closed-form generation score exists only for stripes, so this suite probes the understanding side only, and each question is interpreted only where its clean yes/no contrast validates the readout: strong for shape at both scales, moderate for count and relation at 1.5B, weak for relation at 0.5B, absent for count at 0.5B—those cells are reported but not interpreted.

Text-to-image self-consistency. On GenEval-style prompts in four categories, the model generates an image and then answers the prompt's own yes/no question about its generation. Deleting the prompt-relevant cluster mid-generation flips the self-consistency answer in $\sim 2\%$ of cases overall, at or near the size-matched random-control rate in every category (Table 17)—the deletion robustness replicates on the generative path's own distribution—even as the same deletion visibly corrupts prompt-relevant content (Figure 14).

E Real-image causal benchmark (POPE)

To test the understanding-side causal map on natural images against a public benchmark, we port the cluster-ablation protocol to the POPE object-hallucination benchmark (Li et al., 2023), adversarial split, over COCO val2017 images (Lin et al., 2014): 200 items, 193 usable (Harmon-1.5B; clean accuracy 74.6%). For each clean-correct ground-truth-*yes* item we compute contribution clusters ($k = 6$) at encoder layers 5/10/15/17 and ablate five token sets: the cluster maximally overlapping the queried object's COCO

Scale	Cap	L	op	d_{clus}	d_{orac}	μ_{null}	σ_{null}	z_{clus}	z_{orac}	n
1.5B	count	5	mean	-0.48	-0.23	0.11	0.19	-3.09	-1.77	6
1.5B		5	zero	-0.94	-0.17	-0.06	0.38	-2.32	-0.28	6
1.5B		10	mean	-0.42	-0.12	0.04	0.07	-6.21	-2.21	6
1.5B		10	zero	-0.33	-0.38	-0.21	0.15	-0.84	-1.15	6
1.5B		15	mean	0.06	-0.10	0.05	0.09	0.21	-1.69	6
1.5B		15	zero	-0.35	-0.40	-0.33	0.16	-0.17	-0.44	6
1.5B		18	mean	-0.12	-0.10	0.02	0.07	-2.20	-1.90	6
1.5B		18	zero	0.17	0.48	0.23	0.13	-0.46	1.88	6
1.5B	relation	5	mean	-1.30	-0.28	-0.14	0.10	-12.18	-1.43	5
1.5B		5	zero	-1.15	-0.40	0.18	0.09	-14.16	-6.18	5
1.5B		10	mean	-1.00	-0.08	0.04	0.08	-12.36	-1.37	5
1.5B		10	zero	-1.55	-0.63	-0.51	0.13	-8.08	-0.94	5
1.5B		15	mean	-0.05	-0.08	0.02	0.08	-0.95	-1.27	5
1.5B		15	zero	-0.08	-0.18	0.02	0.08	-1.14	-2.36	5
1.5B		18	mean	0.05	0.02	0.03	0.08	0.18	-0.14	5
1.5B		18	zero	-0.25	-0.15	0.10	0.08	-4.44	-3.21	5
1.5B	shape	5	mean	0.02	-0.06	-0.05	0.12	0.61	-0.07	6
1.5B		5	zero	-2.33	-2.58	-3.24	0.54	1.66	1.20	6
1.5B		10	mean	0.08	-0.19	-0.01	0.08	1.18	-2.15	6
1.5B		10	zero	-2.08	-2.25	-1.11	0.36	-2.69	-3.16	6
1.5B		15	mean	-0.31	-0.10	0.04	0.09	-4.05	-1.67	6
1.5B		15	zero	-1.87	-1.42	-0.67	0.17	-6.91	-4.29	6
1.5B		18	mean	-0.02	0.02	0.05	0.07	-1.07	-0.48	6
1.5B		18	zero	-4.35	-4.35	-3.59	0.21	-3.62	-3.62	6
0.5B	count	3	mean	-0.79	-0.20	0.02	0.11	-7.33	-2.01	7
		3	zero	-1.11	-0.46	-1.16	0.20	0.27	3.53	7
		6	mean	-0.75	-0.41	0.00	0.10	-7.59	-4.16	7
		6	zero	-0.59	-0.16	-0.19	0.31	-1.26	0.10	7
		9	mean	-0.63	-0.38	-0.00	0.09	-6.79	-4.05	7
		9	zero	-0.48	-0.21	-0.01	0.21	-2.26	-0.97	7
		10	mean	-0.32	-0.27	0.01	0.09	-3.86	-3.24	7
		10	zero	-0.68	-1.00	-0.45	0.18	-1.26	-2.99	7
0.5B	relation	3	mean	-0.42	0.12	-0.05	0.07	-5.50	2.48	6
		3	zero	-0.23	0.23	-0.15	0.09	-0.83	4.29	6
		6	mean	0.13	0.10	0.02	0.07	1.56	1.26	6
		6	zero	0.12	0.39	0.45	0.12	-2.76	-0.47	6
		9	mean	0.10	0.04	0.02	0.07	1.07	0.23	6
		9	zero	-0.04	0.69	-0.33	0.10	2.97	10.37	6
		10	mean	0.15	-0.02	0.04	0.08	1.38	-0.72	6
		10	zero	-0.36	-0.32	-0.34	0.07	-0.29	0.27	6
0.5B	shape	3	mean	-0.34	-0.16	-0.19	0.13	-1.08	0.24	7
		3	zero	0.37	0.07	-1.17	0.16	9.77	7.84	7
		6	mean	-0.14	-0.09	0.04	0.09	-2.08	-1.48	7
		6	zero	-0.93	-0.42	-0.39	0.19	-2.82	-0.16	7
		9	mean	-0.18	-0.16	-0.05	0.09	-1.37	-1.18	7
		9	zero	-1.16	-1.36	-1.44	0.20	1.37	0.39	7
		10	mean	-0.23	-0.32	-0.08	0.07	-2.11	-3.31	7
		10	zero	-0.43	-0.54	-0.03	0.11	-3.78	-4.79	7

Table 16: Capability-specific ablations (count / relation / shape), full results: readout deltas for the communication cluster and discovery oracle, the size-matched random null (mean μ, sd σ), and the resulting effect sizes, per scale, capability, layer, and op.

	self-consistency acc.		target flips (latent)		target flips (pixel)	
Category	latent	pixel	cluster	rand	cluster	rand
colors	0.979	0.979	0.007	0.014	0.007	0.007
counting	0.760	0.760	0.042	0.028	0.069	0.035
position	0.983	0.983	0.000	0.000	0.000	0.000
two-object	0.986	0.986	0.042	0.021	0.042	0.021

Table 17: Text-to-image self-consistency benchmark. Left: clean cycle-consistency accuracy (generate, then answer the prompt's question about the own generation) with latent- and pixel-space re-encoding. Right: rate at which ablating the prompt-relevant communication cluster flips the self-consistency answer on target items, vs. the size-matched random control (averaged over intervention layers, weighted by item count).

segmentation, that oracle token set itself, the cluster *minus* the oracle tokens, the largest other-object cluster, and size-matched random sets. Under MEAN-ABLATE—the in-distribution operator on real images—the object cluster flips 47–54% of answers and the oracle 67–74% at every probed layer, while the other-object cluster stays at $\leq 3.5\%$ and randoms at 1–9% (Table 18). ZERO at layer 5 is non-specific (other-object 73.7%, random 66.1%): zeroing early-layer activations on a natural image acts as corruption, not deletion, which is why MEAN-ABLATE is the operator of record here.

One accounting note, prompted by a protocol audit: the target cluster is *selected* by its overlap with the oracle, so cluster-vs-random alone is circular. The discipline is the cluster-minus-oracle condition: the cluster's non-object tokens flip only 4–12%, within the band of their own size-matched random control. The cluster's causal power on real images is therefore carried entirely by its object-token coverage: the token-level causal locus is the object itself, and contribution clustering finds it unsupervised—it is not evidence of extra-object causal structure.

		answer flip rate (%)						
L	op	cluster	oracle	clus\orac	other obj.	rand	rand$_{c\setminus o}$	rand (*no*)
5	MEAN-ABLATE	52.6	73.7	10.7	3.5	9.4	7.7	6.5
5	ZERO	82.5	71.9	35.7	73.7	66.1	39.9	1.5
10	MEAN-ABLATE	54.4	70.2	7.8	3.5	3.5	5.9	2.3
10	ZERO	49.1	71.9	9.8	10.5	8.8	11.1	4.6
15	MEAN-ABLATE	47.4	66.7	9.8	1.8	5.3	3.3	1.1
15	ZERO	50.9	64.9	11.8	5.3	5.8	6.5	1.1
17	MEAN-ABLATE	50.9	66.7	3.8	1.8	1.8	1.3	1.5
17	ZERO	57.9	63.2	9.4	1.8	2.3	3.1	2.3

Table 18: Real-image causal benchmark: POPE adversarial split on COCO val2017 (200 items, 193 usable; Harmon-1.5B; contribution clusters, $k=6$). Flip rates are over the 57 clean-correct ground-truth-*yes* items; clean POPE accuracy is 74.6% (yes 61.3%, no 87.0%). Token sets: the contribution cluster with maximal overlap with the queried object (cluster); the COCO-segmentation oracle token set (oracle); the cluster with oracle tokens removed (clus\orac); the largest cluster with zero or minimal oracle overlap (other obj.); and size-matched random sets—rand matched to the cluster size, rand$_{c\setminus o}$ to the cluster\oracle size. The last column is the flip base rate of cluster-sized random ablation on clean-correct ground-truth-*no* items (calibration). ZERO at layer 5 is non-specific (other-object 73.7%, random 66.1%)—out-of-distribution destruction rather than targeted deletion—so MEAN-ABLATE is the operator of record on real images.

A 0.5B replication runs the identical protocol on the same 193 items at depth-matched layers 3/6/9/11 (Table 19). Unlike the synthetic composite scenes, real photographs pass the behavioral gate at 0.5B (clean accuracy 77.2%, 74 clean-correct *yes*-items), and MEAN-ABLATE reproduces the pattern at lower amplitude: layer 3 cluster 40.5%, oracle 47.3%, cluster-minus-oracle 8.5%, random 4.1%, with effects tapering to 21.6–29.7% by layers 9–11—the same depth profile as 1.5B, with ZERO again noisier and MEAN-ABLATE the operator of record.

Generation-side companion: regeneration under cluster-band deletion. POPE exercises the understanding readout only; a companion benchmark ports the generation side to the same distribution. For each of 30 COCO val2017 images the right half is regenerated by 32-step MaskGIT completion

		answer flip rate (%)						
L	op	cluster	oracle	clus\orac	other obj.	rand	rand$_{c\setminus o}$	rand (*no*)
3	MEAN-ABLATE	40.5	47.3	8.5	0.0	4.1	3.3	6.7
3	ZERO	50.0	58.1	23.9	21.6	18.0	15.5	7.1
6	MEAN-ABLATE	32.4	39.2	7.6	2.7	5.4	5.6	1.3
6	ZERO	52.7	47.3	34.8	28.4	25.2	21.7	8.4
9	MEAN-ABLATE	29.7	21.6	9.9	1.4	3.2	2.3	1.3
9	ZERO	28.4	23.0	5.6	2.7	5.4	3.8	6.7
11	MEAN-ABLATE	21.6	16.2	4.5	0.0	3.6	2.5	0.9
11	ZERO	27.0	16.2	3.0	4.1	4.1	3.0	3.1

Table 19: Real-image causal benchmark, 0.5B replication: POPE adversarial split on COCO val2017 (193 usable items; Harmon-0.5B; contribution clusters, $k=6$; encoder layers 3/6/9/11, depth-matched to the 1.5B grid). Flip rates are over the 74 clean-correct ground-truth-*yes* items; clean POPE accuracy is 77.2% (yes 79.6%, no 75.0%), vs. 74.6% at 1.5B. Token sets as in Table 18. Under MEAN-ABLATE, the operator of record, the object cluster flips 21.6–40.5% of answers against 3.2–5.4% cluster-sized randoms, with cluster-minus-oracle at 4.5–9.9% (best selectivity at layer 3: cluster 40.5% vs. random 4.1%), and effects taper with depth (layers 9–11), matching the 1.5B profile. ZERO is noisier at this scale (layer-6 clus\orac 34.8% vs. its size-matched random 21.7%), the early-layer non-specificity caveat of Table 18 again.

(CFG 3.0) while the queried object's contribution-cluster rows are zeroed over a width-3 layer band at start layers $\{4, 10, 15\}$ at every MaskGIT step, against 3 size-matched random bands and the largest unrelated cluster. The readout is OWLv2 detectability of the regenerated object (prompt "a photo of a {category}", maximum box confidence inside the object region); det-suppression is the drop relative to the clean regeneration, z-calibrated against the random-band null. Table 20 reports the grid: no cell is selective (cluster $|z| \leq 0.08$, all cells $|z| \leq 0.35$), while the identical cluster at a single layer flips 47–54% of understanding answers on the same images. The bound: regeneration itself halves in-region detector confidence (source 0.358, all regenerated variants $\approx$0.17), so the null rests on the matched cluster-vs-random contrast at equal baseline.

L	token set	n	det-supp	null $\mu \pm \sigma$	z	selective
4	cluster	30	$+0.014$	$+0.015 \pm 0.105$	-0.01	no
4	other	11	-0.021	$+0.015 \pm 0.105$	-0.34	no
10	cluster	30	-0.010	-0.009 ± 0.119	-0.01	no
10	other	9	$+0.019$	-0.009 ± 0.119	$+0.24$	no
15	cluster	30	-0.003	-0.012 ± 0.107	$+0.08$	no
15	other	9	$+0.023$	-0.012 ± 0.107	$+0.32$	no

Mean in-region confidence by group:
source 0.358; regenerated: clean 0.168, cluster-deleted 0.167, random-band 0.169, other-cluster 0.145.

Table 20: Generation-side real-image benchmark: right-half regeneration of 30 COCO val2017 images (32-step MaskGIT, CFG 3.0) under width-3 band deletion at start layers $L \in \{4, 10, 15\}$ of the contribution-cluster rows covering the queried object (cluster), of the largest unrelated cluster (other; defined on the subset of images with one, hence the smaller n), and of 3 size-matched random bands (the null). Detectability of the regenerated object is scored by OWLv2 (prompt "a photo of a {category}"), as the maximum box confidence inside the object region; det-supp is the drop in that confidence relative to the clean regeneration, and z calibrates it against the random-band null. No cell is selective: cluster cells sit at $|z| \leq 0.08$, all cells at $|z| \leq 0.35$. The bottom panel gives the dynamic-range caveat: regeneration itself halves detector confidence relative to the source image ($0.358 \rightarrow \approx 0.17$ for every regenerated variant), so the absolute range for detecting a further drop is compressed—the null rests on the matched cluster-vs-random contrast at equal baseline, not on absolute detectability.