A Two-Stage Cross-Attention Circuit for Spatial-Relation Binding in Small Diffusion Transformers

Juliana Li [1] **Binxu Wang** [1]

Abstract

We mechanistically resolve how a small PixArt-mini DiT (6 layers, 36 cross-attention heads, T5-XXL conditioning), trained in a controlled object-relation testbed, converts a textual spatial relation into image layout, identifying a two-stage cross-attention circuit at individual-head granularity. Two early-layer heads $\{L_0H_0, L_1H_2\}$ implement *spatial routing*: their QK circuits convert each relation token into an 8×8 directional attention pattern. Zeroing L_0H_0 alone causes $+48$ pp damage to spatial accuracy; the pair causes $+64$ pp (super-additive $I=+0.086$), and among five consensus-ranked candidates only L_1H_2 shows a super-additive interaction with L_0H_0. A second *object-binding* stage is distributed across Layer 2: near-zero effect alone, but strongly super-additive on color and shape when co-ablated with Layer 0; ablation scored on spatial accuracy alone would miss this downstream stage. The circuit emerges via a sharp phase transition, with behavioral accuracy rising $\sim$200 epochs after the projection-magnitude onset; multi-seed replication confirms the sharp transition phenomenon and early-layer routing structure while specific head indices permute. Our results show that head-level causal circuit discovery extends to DiTs, that multi-metric ablation is needed to detect distributed downstream stages, and which of these components should be expected to survive in production-scale DiTs.

1 Introduction

Diffusion transformers (Peebles & Xie, 2023; Chen et al., 2024b), building on diffusion (Ho et al., 2020) and latent-diffusion (Rombach et al., 2022) foundations, underlie many leading text-to-image systems, but their mechanistic analysis has so far been attribution-based (Helbling et al., 2025; Tang et al., 2023) or block-granular for editing and unlearning (Zarei et al., 2025; Basu et al., 2024; Orgad et al., 2023). By contrast, language-model interpretability routinely localizes computations at head-level resolution (Wang et al., 2023; Olsson et al., 2022; Conmy et al., 2023). We localize a specific generative computation to individual DiT cross-attention *heads* and combine head-level ablation, downstream-reader discovery, and checkpoint-resolved emergence tracking. On the diffusion-transformer side we are aware of one recent precedent (Wang et al., 2026), which we build on directly.[1]

We study spatial-relation binding in a PixArt-mini variant (6 layers, 36 cross-attention heads) conditioned on frozen T5 embeddings and trained on $\sim$ 10,000 prompt-image pairs sampled from 144 synthetic two-object prompts spanning eight relations and six color-shape object types. Within this scope the computation factorises into two stages: a *spatial-routing* stage carried by the head pair $\{L_0H_0, L_1H_2\}$, whose QK circuits convert the relation token into a directional attention bias, and a *distributed object-binding* stage carried by Layer 2, which binds color and shape to position and is invisible to ablation scored on spatial accuracy alone; this pitfall is likely to recur in other multi-attribute circuit-discovery settings.

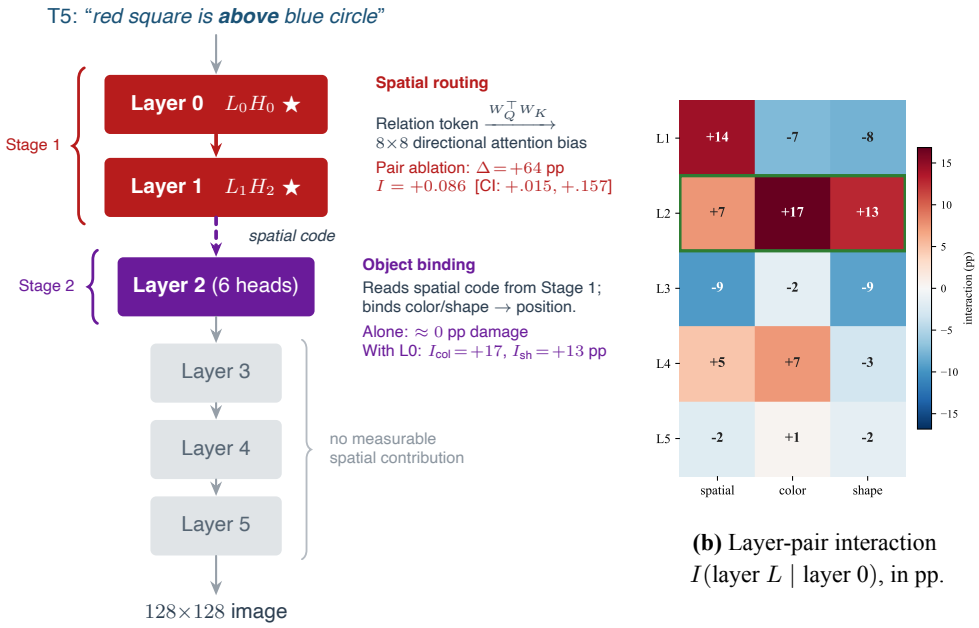
Four independent lines of evidence support the circuit: weight-only spatial-ramp alignment with permutation nulls (discovery), zero-ablation against norm-matched controls and a four-method downstream-reader consensus (necessity), clean-to-corrupt activation patching (sufficiency), and checkpoint-resolved emergence tracking (Nanda et al., 2023; Olsson et al., 2022; Power et al., 2022), which shows the routing stage forming in a sharp phase transition that replicates across three additional training seeds (Table 3).

Scope. Our claims are limited to synthetic prompts, a small model, and four training seeds; we do not expect specific head identities or layer numbers to transfer directly to larger DiTs (Section 4.6).

[1]Harvard University, Cambridge, MA, USA. Correspondence to: Juliana Li <julianali@college.harvard.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Code and reproduction scripts: `https://github.com/lijuliana/Diffusion-Relation-Circuit`.



(a) Two-stage circuit mapped onto the 6-layer DiT.

(b) Layer-pair interaction $I(\text{layer } L \mid \text{layer } 0)$, in pp.

Figure 1. (a) Two-stage cross-attention spatial-binding circuit. Stage 1 (red, layers 0–1): heads $\{L_0H_0, L_1H_2\}$ route the relation token to an 8×8 spatial bias. Stage 2 (purple, layer 2): six heads jointly bind color/shape to position; near-zero alone, super-additive with Stage 1. *(b)* Layer-pair interaction $I(\text{layer } L \mid \text{layer } 0)$ in pp; Layer 2 is super-additive on color/shape but mild on spatial. Quantitative details in Section 4.4.

2 Related Work

DiT interpretability. The direct antecedent is Wang et al. (2026), which showed that small DiTs trained on the same object-relation task implement spatial-relation binding via a sparse cross-attention circuit, which becomes more distributed with pretrained T5 than random embeddings. To our knowledge, head-level circuit identification on DiTs is currently limited to this single line of work. Other DiT interpretability operates at coarser granularity: Helbling et al. (2025) use linear probes on attention outputs to derive segmentation masks; Zarei et al. (2025) localize stored concepts to entire DiT *blocks* for unlearning. Both establish that DiT attention outputs carry interpretable structure, but neither identifies individual heads. UNet-era cross-attention attribution methods (DAAM (Tang et al., 2023), Prompt-to-Prompt (Hertz et al., 2023)) are attribution-based rather than causal, and target a different architecture. We adopt the Wang et al. testbed but use T5 (the harder regime), trace circuit formation across training, and add causal-necessity tests, consensus-based downstream-reader search, and weight-level QK-OV ranking.

LM circuits. Our weight-level analysis follows Elhage et al. (2021): attention is treated as the bilinear-form pair $(W_QW_K^\top, W_OW_V^\top)$, and the QK-OV product across heads predicts functional read-from relations. This formalism underlies head-level circuit work in language models, including the IOI circuit (Wang et al., 2023), induction heads (Olsson et al., 2022), and the automated circuit-discovery procedure of Conmy et al. (2023). We complement weight-level evidence with causal interventions: zero-ablation and clean-to-corrupt activation patching at the per-head output, following the broader causal-mediation programme (Meng et al., 2022; Geiger et al., 2021) and the practical-patching guidance of Heimersheim & Nanda (2024) and Zhang & Nanda (2024). To our knowledge, this combination has not previously been applied to cross-attention in diffusion models. Cross-attention has, however, been used as a control surface for spatial layout in UNet-based diffusion (Chefer et al., 2023; Hertz et al., 2023); our analysis identifies the specific heads in a DiT that play this role.

3 Methods

Model and data. We study a PixArt-mini (Chen et al., 2024b) variant: L=6 transformer blocks, H=6 cross-attention heads/block (36 total), d=384, d_{head}=64, with frozen T5-XXL conditioning (Raffel et al., 2020) and a learned $4096 \to 384$ caption projection W_{cap}. It is trained for 4000 epochs on ∼10,000 synthetic prompt-image pairs per epoch, drawn from 144 prompts of the form "[color$_1$] [shape$_1$] is [relation] [color$_2$] [shape$_2$]" (12 ordered object-attribute templates × 12 phrase strings over 8 cardinal

and diagonal relations; full combinatorics in Appendix A). Held-out sets test unseen color×shape combinations and relation paraphrases. Evaluation uses post-hoc OpenCV detection for separately reported spatial, color, and shape accuracies, with 1000-replicate prompt-bootstrap 95% CIs throughout; per-metric criteria and thresholds (strict vs. loose) are in Appendix A.

Candidate head discovery. We rank the 36 heads by an offline weight-only score: project the per-relation T5 effect vector $\mathbf{p}_r = W_{\text{cap}}\mathbf{v}_r \in \mathbb{R}^{384}$ through each head's QK circuit and evaluate the resulting per-position score $\mathbf{q}_{l,h,r} = \mathbf{P}(W_Q^{(l,h)})^\top W_K^{(l,h)}\mathbf{p}_r$ on a coarse 8×8 position grid (the 16×16 latent's sinusoidal position embedding $\mathbf{P}$ evaluated at 8×8 resolution, giving 64 scores; the pattern is smooth at the patch scale, so the coarse grid captures its directional structure without 256-way noise). We then measure its alignment to a canonical spatial-ramp basis $\{\mathbf{R}_r\}$ (e.g. *above*: top-heavy ramp), reporting two scores per head:

$$|p|(l,h) = \max_r \frac{|\mathbf{q}_{l,h,r}^\top \mathbf{R}_r|}{\|\mathbf{R}_r\|},$$

$$|\cos|(l,h) = \max_r \frac{|\mathbf{q}_{l,h,r}^\top \mathbf{R}_r|}{\|\mathbf{q}_{l,h,r}\|\,\|\mathbf{R}_r\|}.$$

$|p|$ captures both directional alignment and write strength; $|\cos|$ is direction-only. We rule out chance with a 200-permutation null over relation labels and a shape-identity negative control (Appendix B).

Discovery metrics use QK patterns only; the OV/output path enters only in the per-head output ablations (Section 4.3, Appendix C) and activation patching, where we test what each head injects into the residual stream.

Causal ablation. For each ablation condition we zero the target head's value-weighted attention output, $\text{softmax}(QK^\top/\sqrt{d_{\text{head}}})V$, before head concatenation and the shared output projection W_O, at every denoising step on both CFG branches, and regenerate 10 images/prompt. We adopt a single sign convention throughout: *damage* $\Delta\text{Acc}(l,h) = \text{Acc}_{\text{baseline}} - \text{Acc}_{\text{ablated}}$, so positive values mean worse performance after the intervention. We compare each head's damage to a *norm-matched random-head control* (five non-shortlist heads of comparable OV Frobenius norm) to absorb the off-distribution effect of zeroing. Implementation details (tensor shapes, scheduler/CFG settings, evaluator) are in Appendix C.

Downstream-reader search. QK-OV composition (Elhage et al., 2021) gives a clustered ranking (top-5 Frobenius scores 18.2–11.5) that does not cleanly identify the on-path reader. We add three signals and take the unweighted mean rank: offline QK-pattern alignment with the spatial ramp (M1), path-patching recovery loss (M2), in-vivo activation correlation with L_0H_0 (M3), and per-head spatial-ramp correlation of in-vivo head-output patterns (M4). Verification then runs single-head *pair-ablation* (super-additive $I(L_0H_0, h_d) = \Delta(\text{pair}) - \Delta(L_0H_0) - \Delta(h_d)$ on every metric), *layer-pair ablation* (zero all six heads of layer L jointly with layer 0 to catch distributed stages), *triple ablation* relative to the source pair, and clean-to-corrupt activation patching at the per-head output. Per-method definitions and tensor-level details are in Appendix D.

Table 1. Top heads by spatial-ramp projection magnitude $|p|$ at the final checkpoint, computed from each head's QK product (no OV terms). The $|\cos|$ shelf is not discriminative; $|p|$ separates L_0H_0 and L_1H_2 from the next tier.

Head	$\lvert\cos\rvert$	Projection magnitude $\lvert p\rvert$
L_0H_0	0.98	78.2
L_1H_2	0.99	38.4
L_0H_2	0.64	19.3
L_1H_0	0.79	16.4
L_2H_2	0.92	16.1
L_2H_1	0.88	15.7
L_0H_5	0.52	14.9
L_1H_1	0.72	14.6
median (36 heads)	0.69	4.6

Training-time emergence. We re-evaluate alignment scan, variance partition, and accuracy at eleven checkpoints (epochs 100, 250, 500, 600, 700, 750, 800, 900, 1000, 2000, 4000), with density concentrated in 500–1000 where a pilot scan showed rapid change.

4 Results and Discussion

Results follow the methods order: each subsection states a quantitative finding and its implication for the next analysis. Fig. 1 summarizes the resulting circuit, which the next subsections build piece by piece.

Throughout, the QK product determines the relation-conditioned attention pattern, while the OV/output analysis measures its contribution to the residual stream.

4.1 Head discovery and statistical significance

The projection magnitude $|p|$ (the ℓ_2 norm of the head's QK-induced spatial pattern projected onto the spatial-ramp basis, combining directional alignment with signal strength) cleanly separates the routing heads: L_0H_0=78.2, L_1H_2=38.4, and the third-ranked L_0H_2=19.3 (a 2.0× gap), against a 36-head median of 4.6. Direction-only $|\cos|$ does not: it ranks the same top-2 but with only a 0.06 gap between the second- and third-ranked heads (Table 1).

Both L_0H_0 and L_1H_2 exceed all 200 shuffled-relation null values (empirical $p < 0.005$; null max 8.3), and a shape-identity negative control gives $|p| < 3$ for every shortlisted head, ruling out a generic-high-variance artefact. The signal is concentrated in layers 0–1; the only mid-layer

Table 2. Per-checkpoint L_0H_0 alignment metrics and end-task loose accuracy. The discriminative metric is $|p|$ (projection magnitude), not $|\cos|$: $|\cos|$ is high from initialization, but $|p|$ rises $\sim 1000\times$ during training.

Epoch	L_0H_0 $\lvert\cos\rvert$	L_0H_0 $\lvert p\rvert$	Loose acc.
100	0.85	0.05	0.00
250	0.78	0.09	0.00
500	0.69	0.10	0.00
600	0.80	0.20	0.00
700	0.77	0.26	0.06
750	0.76	0.66	0.11
800	0.74	1.54	0.16
900	0.73	4.79	0.20
1000	0.74	9.54	0.38
2000	0.75	39.87	0.78
4000	0.73	47.69	0.81

heads exceeding $|p| = 15$ are L_2H_2 and L_2H_1. Across 1000 bootstrap resamples, L_0H_0 and L_1H_2 are always the top two heads, and top-eight rankings are stable (Kendall-$\tau = 0.93$).

A variance partition of the caption projections confirms that this sparsity is non-trivial: the shape and color factors each have larger marginal R^2 than the relation factor (0.28 and 0.15 for the two shapes vs. 0.11 for the relation, against a total R^2=0.61; Appendix B, Table 8), yet the relation factor concentrates into L_0H_0 and L_1H_2. Relation binding is thus implemented by a sparse subset of heads even where the caption-projection geometry alone would not pick them out.

4.2 Temporal emergence is a sharp phase transition in projection magnitude

Tracking $|\cos|$, projection magnitude $|p|$, and loose held-out accuracy across the eleven checkpoints (Table 2, Fig. 4) exposes a decoupling: $|\cos|$ is high from initialization and never trends (it stays in $[0.69, 0.85]$), whereas $|p|$ crosses 1 near epoch 800 and rises sharply through epoch 1000, while behavioral accuracy rises more gradually over the same window, reaching 0.38 only by epoch 1000 and its plateau ($\sim$ 0.8) later still.[2] This connects to the grokking/phase-transition literature (Nanda et al., 2023) and cautions against $|\cos|$ alone as a circuit-formation indicator: only $|p|$, which folds in write strength, tracks behavioral emergence.

The Phase II onset (first epoch at which $|p| > 1$) is $E^\star \approx 800$ in the original run (seed 43).

Multi-seed replication. To test whether the phase transition and early-layer routing structure are robust to training stochasticity, we retrained the model from scratch on 3 additional seeds (44, 45, 46) using a spec-matched regenerated dataset (identical prompt distribution, independently sampled placements) and re-ran the full alignment scan at each checkpoint. Table 3 summarizes the results. (The replication covers the routing stage only; the Layer-2 binding ablations have not been re-run per seed.)

Table 3. Multi-seed replication: phase-transition onset $E^\star$ and top-2 heads by final-checkpoint $|p|$. All seeds exhibit a sharp phase transition; the routing stage consistently lands on early-layer heads, but the specific (l, h) indices permute.

Seed	$E^\star$	Top-1	Top-2	Both L0-L1?
43 (original)	≈800	L_0H_0	L_1H_2	✓
44	600	L_0H_2	L_2H_3	×
45	600	L_0H_3	L_0H_4	✓
46	600	L_1H_5	L_0H_4	✓

Three results emerge. First, every seed exhibits a *sharp* transition, and its onset is identical across the three replication seeds ($E^\star$=600); the original run's later $E^\star$≈800 likely reflects the original dataset's slightly different placement statistics, so we claim a reproducible transition phenomenon within a 600–800 window rather than a common epoch. Second, *routing structure replicates but identity permutes*: in 2/3 of replication seeds (and the original) the top-2 heads lie in layers 0–1, confirming the early-layer spatial-routing stage is not a single-run accident; however, the specific head index within a layer changes across seeds, consistent with the permutation symmetry of random initialization (Chughtai et al., 2023). Third, the final-checkpoint max $|p|$ is comparable across seeds (24–48), indicating comparable routing strength regardless of which head realizes it.

4.3 Causal necessity via zero-head ablation

Single-head zero-ablation (Table 4, Fig. 5; qualitative samples in Fig. 6) reveals asymmetric necessity: L_0H_0 is individually necessary (loose accuracy $0.84 \to 0.35$), whereas L_1H_2 has little effect alone (+6 pp) but interacts super-additively with L_0H_0, the pair dropping accuracy to 0.20 ($I = +0.086$, 95% CI $[+0.015, +0.157]$; Section 4.4). This holds despite comparable QK-side alignment (L_1H_2: $|\cos| = 0.99$, $|p| = 38.4$), and reflects *redundancy*: with L_0H_0 intact, its OV pathway routes enough of the spatial signal that removing L_1H_2 alone barely degrades the model. The norm-matched controls are five non-shortlist cross-attention heads with OV Frobenius norm within ±15% of L_0H_0; matching on output norm rules out the possibility that the effect is merely caused by removing a high-norm cross-attention output rather than a functionally specific one. Their ablation effects are indistinguishable from zero (−0.3 pp mean, 1.0 pp SD); L_0H_0's +48 pp effect is more than 40 control SDs outside this range.

Layer-resolved ablation sharpens the localization: zeroing all six heads in Layer 0 causes +31 pp damage to loose accuracy, while zeroing any other entire layer changes ac-

[2] The emergence sweep used an earlier alignment pipeline, so its absolute $|p|$ values are offset from the final-checkpoint scan of Section 4.1; the sharp rise and its timing are unaffected.

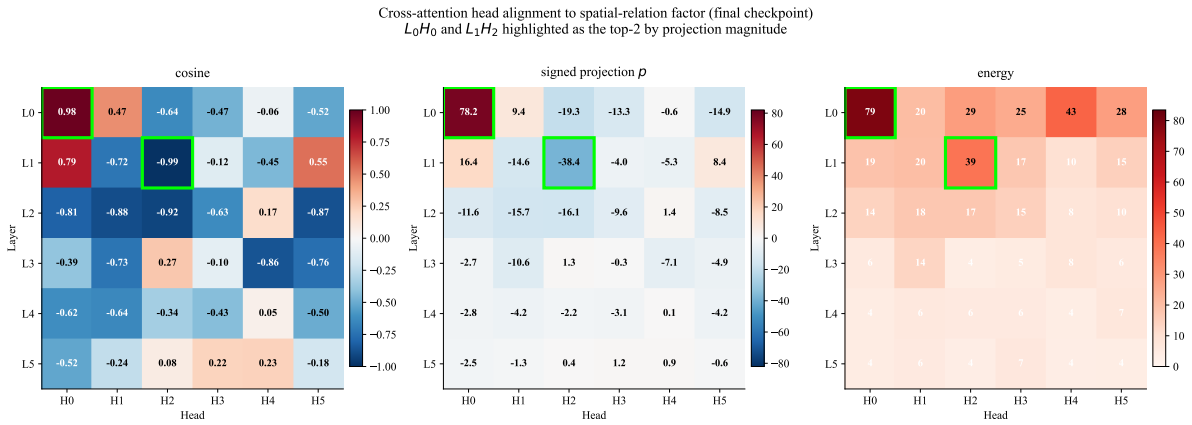


Figure 2. Per-head alignment of the QK-induced spatial attention pattern against the canonical spatial-ramp basis at the final checkpoint. Three metrics are reported: *signed* cosine, *signed* projection p, and (non-negative) energy of the pattern. L_0H_0 and L_1H_2 (lime boxes) are the two outliers: L_0H_0's QK pattern aligns with the canonical ramp direction (deep red, $p{=}+78.2$), L_1H_2's aligns with the negated ramp (deep blue, $p{=}-38.4$). The opposite signs are not a contradiction: each head's downstream contribution passes through W_OW_V, which absorbs the sign, so both heads inject the same spatial signal into the residual stream. $|\cos|$ ranks L_0H_0/L_1H_2 on top but with only a 0.06 gap to the next-best head (L_2H_2, $|\cos| = 0.92$), versus a $2.0\times$ gap by $|p|$.

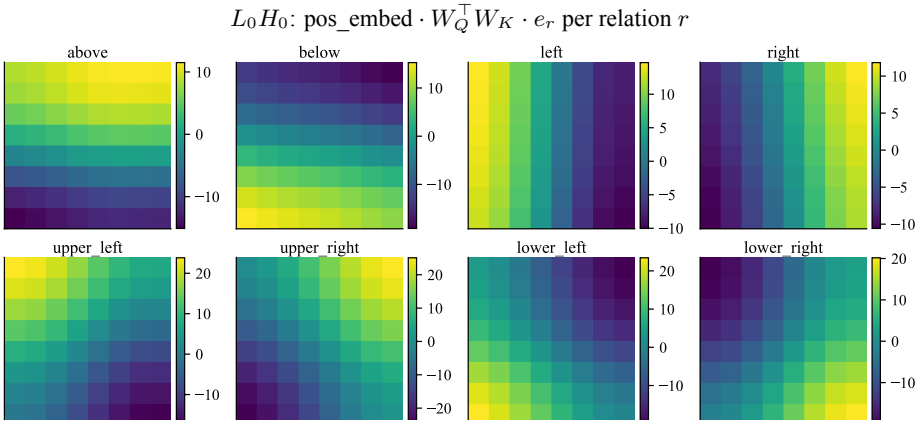


Figure 3. Per-relation QK-attention pattern of L_0H_0. Each panel shows the 8×8 image-grid attention score pos_embed $\cdot\, W_Q^\top W_K \cdot e_r$ for a relation r, where e_r is the mean-centered relation effect vector. Each panel is itself an 8×8 ramp aligned with r (*above*: score concentrated on top rows; *lower_right*: bottom-right corner). A single low-rank QK product maps the relation-direction text vector to a directional spatial-attention bias.

curacy by less than ±5 pp (Appendix G, Fig. 8). The cross-attention spatial signal is concentrated in Layer 0, and within Layer 0 in L_0H_0.

Relation-resolved effects are asymmetric: L_0H_0 ablation hits diagonals (*upper_right, lower_right, lower_left*) at +46–50 pp vs. +10–20 pp for the four cardinals, consistent with diagonals needing both an x- and y-axis bias, so each L_0H_0 output position carries more relation-relevant information. L_1H_2 alone has no measurable effect (+3 pp ±4) on any relation. Per-relation bars in Appendix H, Fig. 11.

Bootstrap CIs from 1000-replicate prompt resamples on this set place $\Delta\text{Acc}_{\text{loose}}(L_0H_0) = +0.48$ (95% CI $[+0.41, +0.55]$) and $\Delta\text{Acc}_{\text{loose}}(L_1H_2) = +0.06$ (95% CI $[+0.02, +0.10]$). Across the full 36-head sweep on the held-out set, the 34 non-source heads have mean $\Delta\text{Acc}_{\text{loose}} = -1.6$ pp (SD 2.2 pp); L_0H_0's +52 pp held-out effect lies $\sim$24 SDs outside that distribution.

The L_0H_0 effect persists on unseen color×shape pairs ($\Delta = +0.51$, $n{=}12$) and synonym-paraphrased relations ($\Delta = +0.19$, $n{=}7$), with both bootstrap CIs excluding zero, but shrinks on syntax-shifted variants where baseline accuracy is already degraded; see Appendix H (and Fig. 9 for held-out CIs).

Ablating all 36 cross-attention heads collapses spatial accuracy to 2% while much object appearance survives (Table 4); we analyze this residual self-attention prior, and the source pair's overlapping OV directions, in Appendix G.

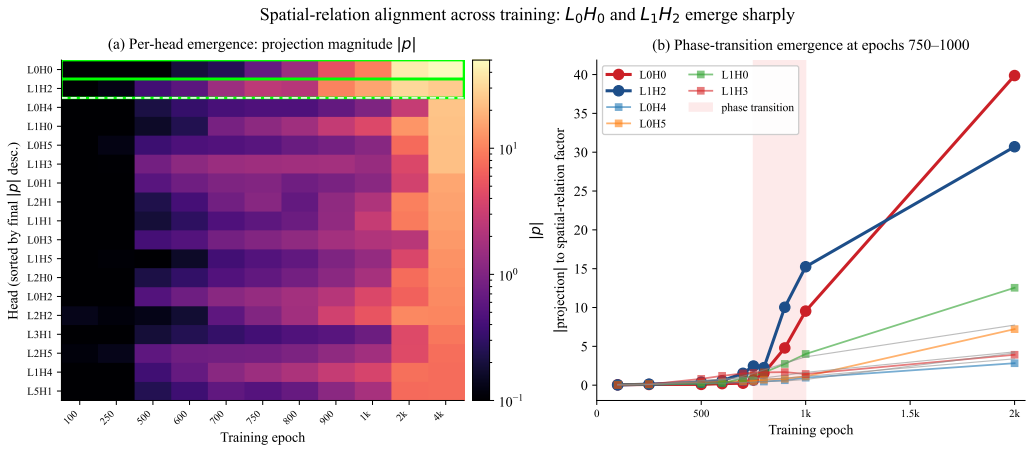


Figure 4. (a) Per-head emergence of projection magnitude $|p|$ across the top 18 of 36 cross-attention heads at 11 training checkpoints, sorted by final-checkpoint $|p|$ (log color scale); the bottom 18 heads stay near $|p| \approx 1$ throughout and are omitted. Lime boxes and the dashed line mark the two routing heads L_0H_0 and L_1H_2, which form the brightest band; only the top heads reach $|p| > 10$. (b) Projection magnitude for the top-6 heads over training (epochs 0–2000), with the phase-transition window (epochs 750–1000) shaded. L_0H_0 and L_1H_2 (red, blue) rise by $\sim 1000\times$ over the full training run ($0.05 \rightarrow 47.7$ for L_0H_0), with the sharp lift-off concentrated in the shaded window ($0.66 \rightarrow 9.5$). Other top-ranked layer-0/1 heads (grey) acquire only modest projection.

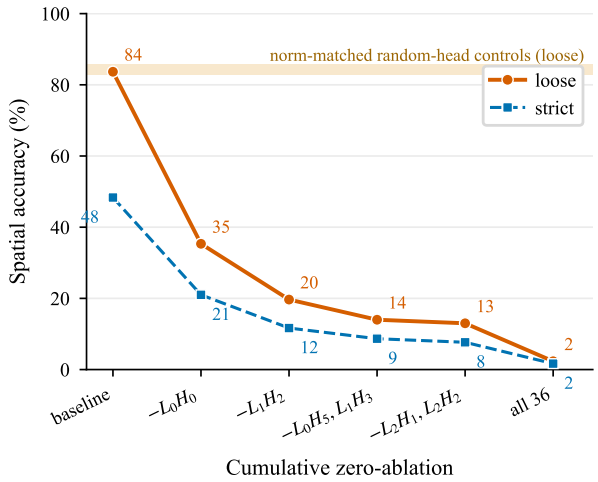


Figure 5. Cumulative multi-head ablation on the 30-prompt training-template set, loose (red) and strict (blue) spatial accuracy: baseline 84% $\rightarrow$ 35% on L_0H_0 alone, $\rightarrow$ 20% with both source heads zeroed, $\rightarrow$ 13% at top-6, $\rightarrow$ 2% at all-36. Shaded band: accuracy range of the five norm-matched random-head controls.

4.4 From source heads to the downstream binding stage

QK-OV composition alone does not cleanly identify a single downstream reader: the top-5 mid-layer heads have clustered Frobenius scores (18.2–11.5). The four-method consensus (QK-OV, path-patching, activation correlation, spatial-ramp projection; Appendix D, Table 9) yields top-5 $\{L_1H_2, L_0H_5, L_1H_3, L_2H_1, L_2H_2\}$; L_2H_3 (rank-1 by QK-OV) falls to consensus rank 21.5 once the other signals are included.

Table 4. Cumulative zero-head ablation, loose spatial accuracy on the same 30-prompt training-template evaluation set used by pair-ablation, n_{img}=10 each. Random control: mean of 5 non-source heads (std 0.011). Top-k *consensus* = source pair plus next $k-2$ heads from the cross-method consensus ranking ($L_0H_5, L_1H_3, L_2H_1, L_2H_2$).

Condition	Loose acc.	Δacc.
Baseline	0.84	–
Ablate L_0H_0	0.35	+0.48
Ablate L_1H_2	0.78	+0.06
Norm-matched control (mean of 5)	0.84	+0.00
Ablate $\{L_0H_0, L_1H_2\}$	0.20	+0.64
Ablate top-4 (consensus)	0.14	+0.70
Ablate top-6 (consensus)	0.13	+0.71
Ablate all 36 cross-attn heads	0.02	+0.81

Pair-ablation. For each consensus-top-5 candidate we run the four-condition pair-ablation (baseline, L_0H_0 only, h only, $\{L_0H_0, h\}$ together) on the same 30-prompt set as Table 4 (n_{img}=10). The result (Table 5) is single-headed: only $\{L_0H_0, L_1H_2\}$ has a super-additive interaction beyond resampling noise ($I = +0.086$, prompt-bootstrap 95% CI $[+0.015, +0.157]$, 1000 resamples, n_{prompt}=23 after dropping prompts with baseline accuracy at floor). The four other candidates have $|I| \leq 0.013$ with CIs containing zero.

Activation patching. Patching L_0H_0's clean head-output into a relation-flipped corrupt forward pass recovers $\sim$83% of the loose-accuracy gap (clean 0.88 $\rightarrow$ corrupt 0.00 $\rightarrow$ patched 0.71, n=8; Fig. 7), a converse causal test indicating that L_0H_0's output contains information sufficient to restore much of the relation-dependent layout. Patching L_1H_2 recovers 0% despite its positive pair-ablation interaction; we attribute this to a combination of redundancy

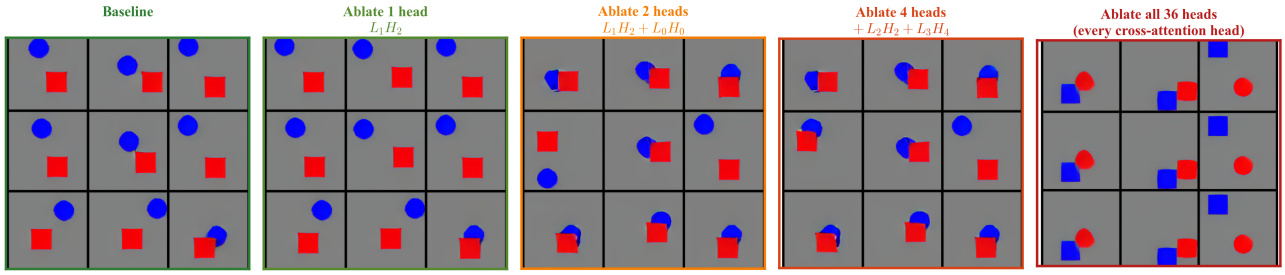


Figure 6. Qualitative cumulative ablation, fixed prompts and seed (epoch 4000, guidance 4.5). Same three relation prompts and same noise seed across all five conditions. *Baseline*: clean two-object layouts. *Ablate* L_1H_2 *alone*: visually indistinguishable from baseline. *Ablate* $\{L_0H_0, L_1H_2\}$: objects begin merging in the center. *Ablate* 4 *heads*: heavier overlap. *Ablate all* 36 *cross-attention heads*: layout is fully destroyed but coherent shapes survive.

Table 5. Pair-ablation interaction scores at the final checkpoint, source $= L_0H_0$. Loose-spatial accuracy. $\Delta\text{Acc}(L_0H_0) = +0.483$ for all rows (same source-only condition). Interaction $I = \Delta\text{Acc}(\text{pair}) - \Delta\text{Acc}(L_0H_0) - \Delta\text{Acc}(\text{cand})$; positive = super-additive. I and its 1000-resample bootstrap 95% CI are computed on the n_{prompt}=23 subset that excludes prompts with baseline accuracy at floor, so I does not exactly equal the difference of the (full-set) Δ columns.

h	$\Delta(h)$	$\Delta(\text{pair})$	I	95% CI
L_1H_2	$+0.060$	$+0.640$	$+0.086$	$[+0.015, +0.157]$
L_0H_5	-0.007	$+0.480$	$+0.012$	$[-0.052, +0.080]$
L_1H_3	$+0.003$	$+0.500$	$+0.008$	$[-0.026, +0.043]$
L_2H_1	$+0.027$	$+0.500$	-0.013	$[-0.035, +0.013]$
L_2H_2	$+0.003$	$+0.487$	-0.002	$[-0.039, +0.033]$

with L_0H_0 and patch-site sensitivity (full discussion in Appendix E).

Among the individually tested candidates, the spatial-only pair-ablation thus finds no additional super-additive interaction above noise, and the effect persists for the source pair on the unseen color×shape set ($I = +0.084$, CI $[+0.05, +0.12]$). This spatial-metric null does not, however, exclude distributed or identity-binding readers: single-head tests can miss distributed stages, and the color and shape metrics detect binding errors far more directly than spatial accuracy, which registers identity only indirectly through its object-identification step (Appendix A). We therefore re-ran the interventions scored on color and shape (per-head table in Appendix F), and on a layer-resolved variant.

Layer-resolved pair-ablation. For each layer $L \in \{1, \ldots, 5\}$ we zero all six heads of layer 0 and additionally all six heads of layer L. The interaction $I(\text{layer } L \mid \text{layer } 0) = \Delta(\text{both}) - \Delta(\text{layer } 0) - \Delta(\text{layer } L \text{ alone})$ on three metrics is shown in Fig. 1(b) and Table 6.

Layer 2 stands out: zeroing it on top of layer 0 produces $+16.8$ pp super-additive damage on color and $+12.6$ pp on shape, while layer 2 zeroed alone has essentially no effect ($\leq$ 1 pp on every metric): the signature of a stage acting downstream of layer 0 on the color/shape-to-position assignment. The load-bearing point is not the exact localization to layer 2, but that this stage appears only when color and shape are scored separately from spatial accuracy.

Table 6. Layer-pair interaction $I(\text{layer } L \mid \text{layer } 0)$ in pp; positive = super-additive. Same 19-prompt training-template set.

L	I_{spatial}	I_{color}	I_{shape}
1	$+14.2$	-7.4	-7.9
2	$\mathbf{+7.4}$	$\mathbf{+16.8}$	$\mathbf{+12.6}$
3	-9.5	-1.6	-8.9
4	$+4.7$	$+7.4$	-3.2
5	-2.1	$+0.5$	-1.6

Table 7. Triple-ablation interaction $I(h \mid \text{src})$ in pp where src $= \{L_0H_0, L_1H_2\}$. Sorted by mean of color/shape interactions. The first six rows have super-additive color interactions $\geq +6$ pp; the bottom two (L_0H_1, L_1H_5) are weaker controls.

Head h	I_{spatial}	I_{color}	I_{shape}
L_0H_5	$+6.8$	$\mathbf{+7.9}$	$\mathbf{+5.8}$
L_2H_2	$+5.8$	$\mathbf{+10.0}$	$+1.1$
L_2H_1	$+4.2$	$\mathbf{+7.4}$	$+2.6$
L_0H_4	$+5.8$	$+6.3$	$+3.2$
L_1H_3	$+5.3$	$+5.8$	$+1.1$
L_0H_2	$+3.7$	$+1.6$	$+5.3$
L_0H_1	$+2.6$	$+3.2$	-0.5
L_1H_5	$+2.1$	-3.2	$+0.0$

Single-head triple ablation. At single-head resolution, we ran the triple-ablation $I(h \mid \{L_0H_0, L_1H_2\}) = \Delta(\text{src}+h) - \Delta(\text{src}) - \Delta(h \text{ alone})$ for the consensus candidates plus four heads with the largest single-head color or shape damage on the held-out set ($L_0H_4, L_0H_1, L_1H_5, L_0H_2$). Several candidates are super-additive on *binding* accuracy when ablated jointly with the source pair (Table 7): L_0H_5 on color ($+7.9$) and shape ($+5.8$); L_2H_2 on color ($+10.0$); L_2H_1 on color ($+7.4$); L_0H_4 on color ($+6.3$). Spatial-resolved interactions are smaller ($+2$–$+7$ pp) and roughly half the binding-resolved interactions, consistent with these heads contributing primarily to object binding rather than spatial routing.

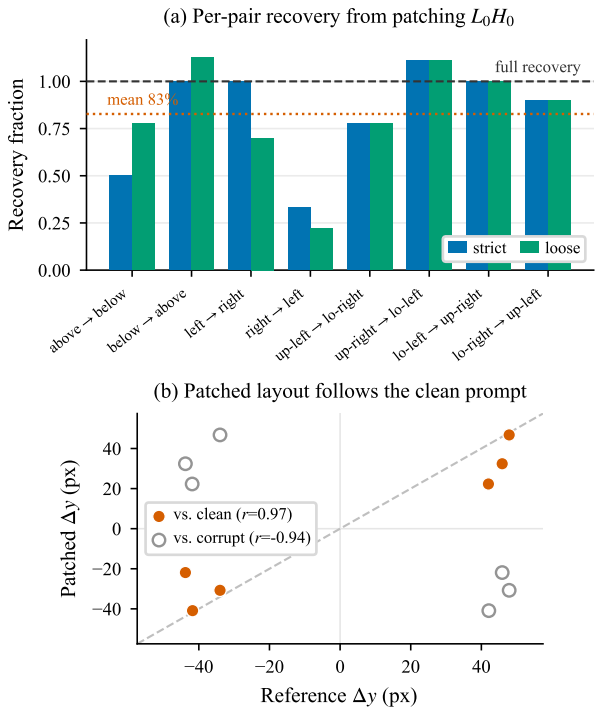


Figure 7. Cross-prompt activation patching at L_0H_0. *(a)* Per-relation-flip recovery fraction on strict and loose metrics; mean $\sim 83\%$. *(b)* Patched layout's Δy tracks the clean prompt's Δy along the identity line (filled, r=0.97) and anti-correlates with the corrupt prompt's (open, r=−0.94), confirming causal control over placement direction.

Why the original pair-ablation missed this. No layer-2 head individually damages binding by more than 1–3 pp (max $\Delta_{\text{shape}} = 4.7$ pp at L_2H_3): the binding stage is *distributed* across the layer, so single-head pair-ablation registers nothing above bootstrap noise. This matches Wang et al. (2026)'s observation that DiTs trained with pretrained T5 (vs. random embeddings) recruit a broader head pool: here, Layer 2. Bar-chart visualization of the triple-ablation interactions is in Appendix I, Fig. 12.

4.5 Prospects for scaling to production DiTs

What transfers directly. PixArt-style models (Chen et al., 2024b;a) keep exactly the interface we study, per-block cross-attention from image tokens to frozen-T5 tokens, so the full discovery pipeline (QK spatial-ramp projection, per-head ablation and patching) ports unchanged; only the head search space grows.[3] PixArt-α-XL is thus the natural first test at scale.

What may change. Joint-attention (MMDiT) models such as Stable Diffusion 3 (Esser et al., 2024) and FLUX (Black Forest Labs, 2024) fuse text and image tokens in a single self-attention, so text-to-image routing is no longer segregated into dedicated heads: the relation-to-position map is still well-defined but its natural unit may be a low-rank *subspace* rather than a head. Web-scale data also dilutes the task pressure that plausibly drives our routing sparsity, consistent with the coarser block-level localization reported at scale (Zarei et al., 2025; Basu et al., 2024). We thus expect the underlying function (relation token → directional spatial bias) to persist (Hertz et al., 2023; Chefer et al., 2023) but spread across more heads, as Wang et al. (2026) already observe when moving from random to pretrained T5 conditioning.

[3] PixArt-α-XL has $28 \times 16 = 448$ cross-attention heads vs. our 36.

4.6 Limitations

Training-seed variability: the detailed ablation and patching analysis is from seed 43; the 3-seed replication (Table 3) confirms the phase transition and early-layer routing are robust, but the specific head indices permute. The replication covers only the routing stage (alignment scan); re-running the per-seed ablations, both spatial ΔAcc on the new routing heads and the Layer-2 color/shape interactions that establish Stage 2, remains open. Head-level circuits in small models are known to vary across retraining (Chughtai et al., 2023), and self-repair (McGrath et al., 2023) may redistribute function; the head indices $\{L_0H_0, L_1H_2\}$ are thus "the two early-layer routing heads of seed 43," not universal identifiers. **Scale:** the model has 36 cross-attention heads; production DiTs have hundreds, or a different conditioning interface altogether (Section 4.5). **Distributed binding:** no single layer-2 head shows > 3 pp damage; finer localization needs denser combinatorial ablation. **Time-aggregated:** step-resolved ablation across the noise schedule is left for future work.

5 Conclusion

Within this controlled testbed on a small PixArt-mini DiT, we resolve spatial-relation binding into a two-stage cross-attention circuit at head resolution: an early-layer routing pair that translates the relation token into a directional spatial bias, and a distributed Layer-2 stage that binds object identity to those positions. Multi-seed replication shows the sharp phase transition and early-layer routing localization are robust to training stochasticity, with the specific head indices permuting across seeds. The main methodological lesson is that spatial accuracy alone misses the downstream binding stage; multi-metric interventions are therefore essential when circuits jointly manipulate position and object identity, in DiTs and plausibly in language-model circuits that span factual, stylistic, and structural axes. More broadly, the head-level circuit vocabulary developed for language models transfers to generative image models once cross-attention is read as the text-to-image routing surface.

References

Basu, S., Zhao, N., Morariu, V., Feizi, S., and Manocha, D. Localizing and editing knowledge in text-to-image generative models. In *International Conference on Learning Representations (ICLR)*, 2024.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: attention-based semantic guidance for text-to-image diffusion models. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *European Conference on Computer Vision (ECCV)*, 2024a.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024b.

Chughtai, B., Chan, L., and Nanda, N. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning (ICML)*, 2023.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. arXiv:2404.15255, 2024.

Helbling, A. et al. ConceptAttention: Diffusion transformers learn highly interpretable features. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.04320.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

McGrath, T., Rahtz, M., Kramár, J., Mikulik, V., and Legg, S. The hydra effect: Emergent self-repair in language model computations. arXiv:2307.15771, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. Transformer Circuits Thread, 2022.

Orgad, H., Kawar, B., and Belinkov, Y. Editing implicit assumptions in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv:2201.02177, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F. What the DAAM: Interpreting stable diffusion using cross attention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Wang, B., Fan, J., and Pan, X. Circuit mechanisms for spatial relation generation in diffusion transformers. arXiv:2601.06338, Kempner Institute, Harvard University, 2026.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Zarei, A., Basu, S., et al. Localizing knowledge in diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.18832.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: metrics and methods, 2024.

A Experimental setup details

Architecture. Cross-attention has the conventional form

$$\text{Attn} = \text{softmax}\left(\frac{XW_Q(EW_K)^\top}{\sqrt{d_{\text{head}}}}\right) EW_V, \tag{1}$$

with $X \in \mathbb{R}^{256\times 384}$ (spatial latent tokens on a 16×16 grid), $E \in \mathbb{R}^{L_{\text{tok}}\times 4096}$ (T5 embeddings, $L_{\text{tok}} \approx 15$), d=384, d_{head}=64, H=6 heads/layer, L=6 layers. Training: 4000 epochs with EMA, CFG conditioning-dropout rate 0.1, $\sim$10,000 prompt-image pairs per epoch (40 steps $\times$ 256 batch).

Prompt combinatorics. The 144 unique prompts arise as 2×1 ordered color pairs $\times$ 3×2 ordered shape pairs $=$ 12 object-pair templates, times 12 unique phrase strings: the 8 included relations (cardinal and diagonal; *in front of*/*behind* excluded) expand into 12 strings with 1–3 paraphrases each (e.g. *above* = *above*/*above and to the left of*/*above and to the right of*), giving $12 \times 12 = 144$. No augmentations are applied.

Held-out sets. (i) *Unseen color×shape*: n=12 prompts using color/shape combinations not seen during training. (ii) *Synonym-paraphrased relations*: n=7 prompts substituting synonyms (*on top of*, *underneath*, *above and to the left of*). (iii) Harder OOD: see Appendix H.

OpenCV evaluator. For each 128×128 image, we threshold to a binary mask, find external contours via `findContours(RETR_EXTERNAL)`, classify each contour by mean RGB inside the contour (within margin 25 of canonical red/blue) and by polygon-segment count from `approxPolyDP` ($3 \to$ triangle, $4 \to$ square, $\geq 7 \to$ circle). Each metric then has its own success criterion. *Color* (resp. *shape*) requires that some detected object matches each prompted color (resp. shape). *Spatial* metrics require that each prompted (color, shape) pair matches exactly one detected object; the two matched centroids' offset $(\Delta x, \Delta y)$ is then scored against the prompted relation. The *strict* spatial metric categorizes the offset into one of the 8 relations (cardinal if the off-axis offset is within 5 px, else diagonal) and requires an exact category match; the *loose* metric only requires the correct sign(s) with magnitude ≥ 8 px on each prompted axis (e.g. *upper_right*: $\Delta x \geq 8$ and $\Delta y \leq -8$). The two metrics are not nested: a generation offset that is diagonal in category but prompted as *above* can pass loose while failing strict. Because the spatial metrics locate objects by their prompted color and shape, a spatially-correct layout with wrong object identities scores zero on the spatial metrics through the identification step; color and shape accuracies remain informative for such images.

Inference settings. DPM-Solver multi-step scheduler at 14 inference steps, CFG scale 4.5, 10 images per prompt per condition. Pilot single-step ablations produced <2 pp change, so we report all-step ablation throughout.

B Candidate-head discovery details

Spatial-ramp basis. The basis consists of D=8 relation directions $\mathcal{D} = \{$above, below, left, right, upper_left, upper_right, lower_left, lower_right$\}$. For each r with unit vector $(d_y^{(r)}, d_x^{(r)}) \in \{-1, 0, +1\}^2$, the ramp $\mathbf{R}_r \in \mathbb{R}^{64}$ over the 8×8 patch grid is $[\mathbf{R}_r]_{(i,j)} = d_y^{(r)}\tilde{y}_i + d_x^{(r)}\tilde{x}_j$ with $\tilde{y}_i, \tilde{x}_j \in \text{linspace}(-1, 1, 8)$, mean-centered. The latent has 16×16 patches; we evaluate the QK score on an 8×8 grid (the sinusoidal position embedding sampled at half resolution, 64 positions) because the relation-conditioned pattern varies smoothly across neighbouring patches, so the coarser grid preserves its directional structure while halving the per-axis noise. Results are qualitatively unchanged at the full 16×16 resolution.

Effect-vector decomposition. For each prompt embedding $\mathbf{e}_i \in \mathbb{R}^{4096}$ we run an additive-effects regression over factors $\mathcal{F} = \{\text{rel}, \text{color}_1, \text{color}_2, \text{shape}_1, \text{shape}_2, \text{interactions}\}$:

$$\mathbf{e}_i = \boldsymbol{\mu} + \sum_{f\in\mathcal{F}} \alpha_{f,i}\,\mathbf{v}_f + \boldsymbol{\epsilon}_i. \tag{2}$$

Pushing $\mathbf{v}_{\text{rel}}$ through the caption projection gives $\mathbf{p}_{\text{rel}} = W_{\text{cap}}\mathbf{v}_{\text{rel}} \in \mathbb{R}^{384}$. The variance partition reports $R_f^2 = 1 - \text{SS}_{\text{resid, w/o } f}/\text{SS}_{\text{total}}$; the marginal R^2 per factor is in Table 8. Object properties dominate the raw caption-projection variance (expected, since each prompt has two colored and two shaped objects and only one relation), while the head-resolved view (Fig. 2) concentrates the spatial-relation factor into L_0H_0's low-rank QK-pattern direction ($|p| = 78.2$) and disperses object factors across many heads.

Permutation null. 200 shuffled-relation refits give a null distribution with max $|p| = 8.3$ and mean 4.1 ± 1.6. L_0H_0's observed $|p| = 78.2$ exceeds all 200 null values (empirical $p < 0.005$, the resolution limit of 200 permutations), sitting

Table 8. Marginal R^2 of caption-projection variance at the final checkpoint, by factor (one prompt = two objects × {color, shape} + relation, so color/shape have two regressors each).

Factor	R^2_{marginal}
Total (all factors)	0.61
Shape (object 2)	0.28
Shape (object 1)	0.15
Spatial relationship	0.11
Color (object 1)	0.11
Color (object 2)	0.11

~ 46 null SDs from the null mean. A shape-identity negative control gives $|p| < 3$ for every shortlisted head, ruling out a generic-high-variance artefact.

C Causal ablation details

Zero-head ablation. With the per-head attention output tensor of shape $(B, H, P, d_{\text{head}})$ where B is the CFG-doubled batch (B=2), H=6 heads, P=256 image patches, d_{head}=64, we set tensor$[:, h_{\text{target}}, :, :] = 0$ at every denoising step and on both CFG branches. The per-head output is zeroed *before* concatenation and projection through W_O.

Norm-matched random control. Five non-shortlist heads of comparable OV Frobenius norm. Mean ablation effect -0.003 ± 0.010 (accuracy fraction, i.e. -0.3 ± 1.0 pp); L_0H_0's $+0.48$ effect lies more than 40 control SDs from this mean.

D Downstream-reader search details

Method definitions.

- M1 (**offline QK-pattern alignment**): rank by | cos | between the candidate head's QK-induced spatial pattern and the spatial-ramp basis.
- M2 (**path-patching recovery loss**): from a clean / relation-flipped pair, patch L_0H_0 from clean (restoring attention) and additionally re-zero h_d; the recovery loss measures h_d's on-path causal contribution.
- M3 (**in-vivo activation correlation**): capture every head's per-step head-output for 30 prompts, compute per-prompt Pearson correlation between L_0H_0 and h_d, average across prompts.
- M4 (**per-head spatial-ramp projection**): per-position ℓ_2 norm of h_d's head-output (averaged over time and head_dim) correlated with the canonical 8×8 ramp R_r for the prompt's relation; report mean | cos | across prompts.

For each method we rank the 35 non-L_0H_0 heads. Path-patching does not sweep L_0, so L_0 entries get the column median. Consensus is the unweighted mean rank.

Table 9. Cross-method consensus ranking. Each cell is the head's rank in that method (1 = best); consensus is the mean rank.

Cand.	M1	M2	M3	M4	Cons.
L_1H_2	2	11	7	2	5.50
L_0H_5	5	11	3	6	6.25
L_1H_3	6	1	10	17	8.50
L_2H_1	8	3	12	14	9.25
L_2H_2	14	3	20	1	9.50
L_2H_3	19	32	23	12	21.50
L_3H_1	15	3	16	30	16.00
L_4H_2	26	3	14	21	16.00

Activation-patching protocol. Patches are applied to h_d's value-weighted attention output (before head concatenation and W_O), on the conditional half of the CFG batch only, at all 14 denoising steps. Eight relation-flip pairs: above↔below, left↔right, and the four diagonals.

E Reconciling the patching null with pair-ablation

L_1H_2 has a positive pair-ablation interaction ($I = +0.086$, CI excludes zero) but a 0% activation-patching recovery. Three explanations are consistent with this asymmetry:

- **Redundancy.** L_1H_2 writes a partially-redundant copy of L_0H_0's output; in the corrupt forward pass L_0H_0 already injects the flipped pattern into the residual, so patching L_1H_2 from clean adds the same direction without overriding it. Joint ablation eliminates both copies at once.
- **Interaction-only contribution.** If L_1H_2 contributes only when the source-pair output exceeds a threshold, single-head patching cannot trigger it. The pair-ablation interaction is positive because L_1H_2's contribution becomes detectable only when L_0H_0 is removed.
- **Patching-site mismatch.** We patch the value-weighted attention output (before concatenation and W_O); a more sensitive site might be the residual stream after the layer's output projection.

Discriminating among these requires denser patching experiments (residual-stream patching, scaled patching, partial-replacement patching), left for future work.

F Metric-resolved pair-ablation table

Table 10. Metric-resolved pair-ablation. Δ = baseline − ablated; positive = damage. Same 30-prompt training-template set used in Table 4.

Condition	Δ_{rel}	Δ_{col}	Δ_{sh}	Δ_{bind}
L_0H_0 alone	$+0.483$	$+0.147$	$+0.263$	$+0.273$
$\{L_0H_0, L_1H_2\}$	$+0.640$	$+0.180$	$+0.270$	$+0.310$
$\{L_0H_0, L_0H_5\}$	$+0.480$	$+0.190$	$+0.267$	$+0.293$
$\{L_0H_0, L_1H_3\}$	$+0.500$	$+0.210$	$+0.283$	$+0.297$
$\{L_0H_0, L_2H_1\}$	$+0.500$	$+0.173$	$+0.283$	$+0.300$
$\{L_0H_0, L_2H_2\}$	$+0.487$	$+0.140$	$+0.260$	$+0.267$

G Self-attention prior under full cross-attention ablation

Ablating all 36 cross-attention heads collapses spatial accuracy to 2% but preserves $\sim$90% color and $\sim$30% shape (Table 4, last row). This surviving generation comes from self-attention, trained during CFG-unconditional steps ($p(\text{img})$, no text) to produce coherent shaped objects without any text conditioning. Relational information enters only via cross-attention, so removing it destroys the spatial signal but leaves the self-attention object prior intact. The two source heads project onto overlapping ramp directions (cosine $\rho \approx 0.7$ between their OV writes) but jointly cover more of the basis than either alone; the parallel redundancy is consistent with CFG dropout favoring such backup configurations.

Layer-resolved ablation. Zeroing all six heads of a single layer (Fig. 8) localizes the spatial signal: Layer 0 ablation causes $+31$ pp damage on loose spatial accuracy, while every other layer changes accuracy by less than ± 5 pp. Color and shape accuracies drop only modestly in Layer 0 (object identity is largely preserved by self-attention).

H Harder-OOD prompt variants

We built three harder OOD variants and re-encoded 24 prompts via T5-XXL: *distractor* (filler clause appended: "red square is above blue circle, in a quiet scene"), *extra adjectives* ("the small red square is above the large blue circle"), and *syntax-fronted* ("Above the blue circle is the red square"). The vanilla effect ($+50$ pp damage, baseline 0.83) reproduces the in-distribution result. On all three harder variants the model's baseline drops to 0.38–0.42 before any ablation. L_0H_0 ablation produces additional damage on every variant ($+7$ to $+14$ pp), with relative drops $|\Delta|$/baseline of 17% (distractor), 37% (extra adjectives), and 22% (syntax-fronted), versus 61% on vanilla. The circuit is not invariant to prompt syntax; its absolute effect is largest on the training distribution. This bounds the generalization claim and identifies prompt-syntax robustness as a separate axis.

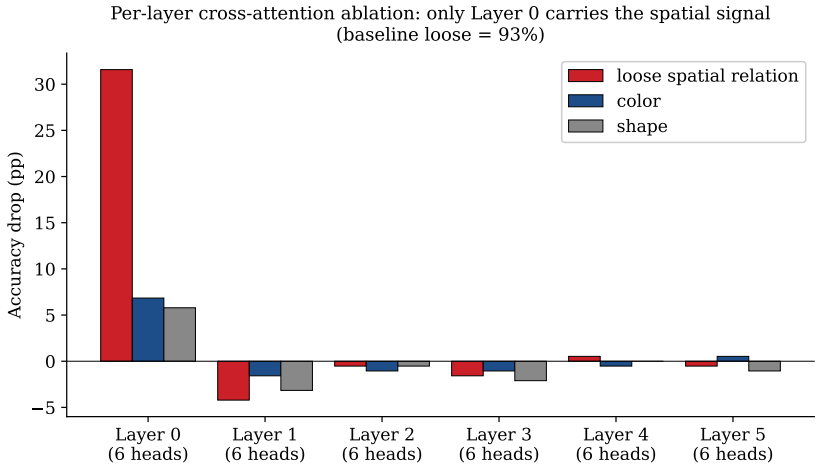


Figure 8. Zero-ablation of every head in a single transformer layer (loose spatial accuracy on the full training distribution of 144 unique prompts; baseline ≈ 0.93, distinct from the 30-prompt training-template subset used by Table 4 and pair-ablation, where baseline is 0.84). Layer 0 ablation causes $+31$ pp damage; layers 1–5 change accuracy by less than ± 5 pp.

Per-relation breakdown. The single-head L_0H_0 ablation effect is also relation-asymmetric (Fig. 11): diagonals (*upper_right, lower_right, lower_left*) take $+46$–50 pp damage, while the four cardinals take $+10$ to $+20$ pp damage. L_1H_2 alone has no measurable effect ($+3$ pp ± 4) on any relation, consistent with the redundancy-with-L_0H_0 picture.

I Triple-ablation bar chart

Companion visualization to Tables 6–7: layer-pair interactions concentrate at Layer 2 on color/shape (panel a), and triple ablations relative to the source pair show consistently larger color than spatial interactions (panel b).

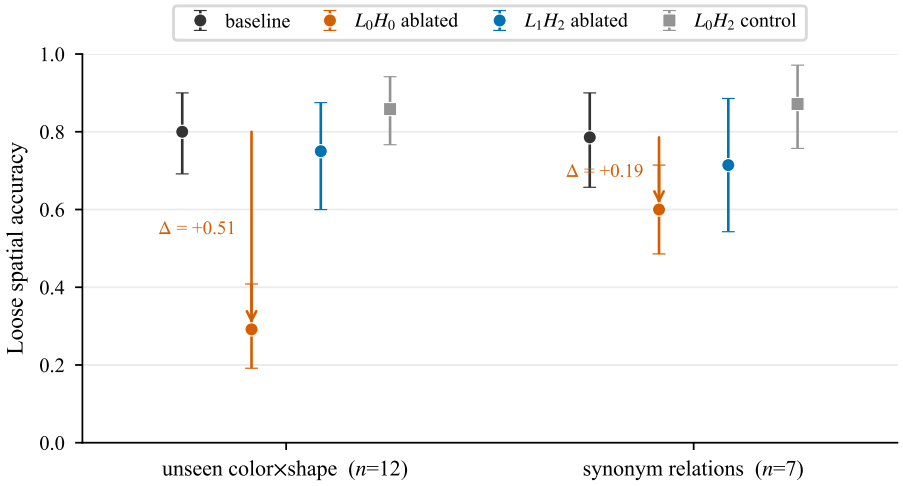


Figure 9. Held-out OOD: loose spatial accuracy with prompt-bootstrap 95% CIs (dot-and-whisker). **Unseen color×shape pairs** (n=12): L_0H_0 ablation drops accuracy from 0.80 to 0.29 ($\Delta = +0.51$, CI $[+0.41, +0.61]$). **Synonym-paraphrased relations** (n=7): from 0.79 to 0.60 ($\Delta = +0.19$, CI $[+0.01, +0.33]$). Both CIs exclude zero. L_1H_2 ablation (blue) and norm-matched control L_0H_2 (grey) are within noise of baseline.

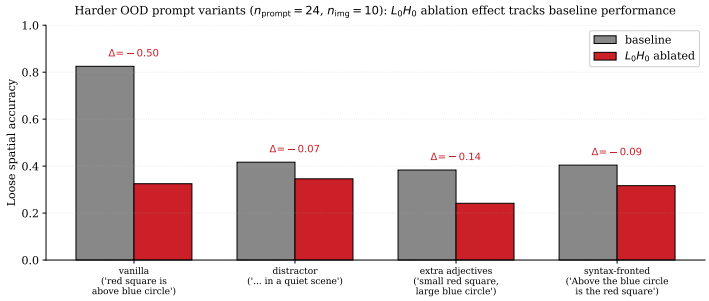


Figure 10. Harder OOD prompt variants. Vanilla template: L_0H_0 ablation causes $+50$ pp damage. The three harder variants degrade the model's baseline accuracy substantially (from 0.83 to 0.38–0.42); L_0H_0's marginal contribution becomes correspondingly smaller in absolute terms ($+7$ to $+14$ pp).

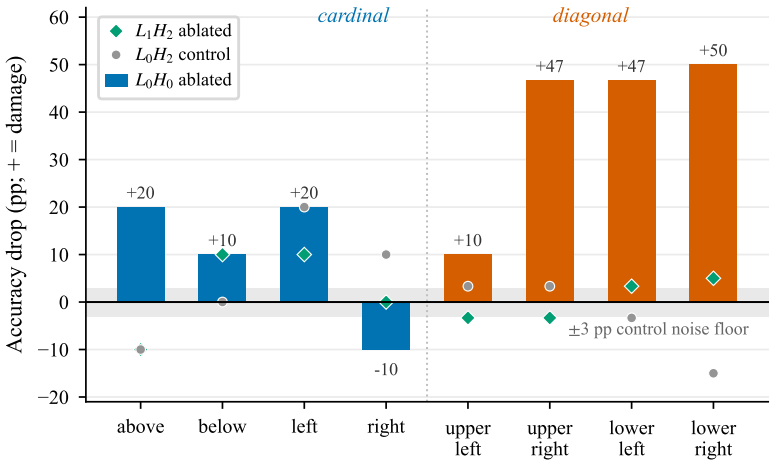


Figure 11. Per-relation ablation effect (bars: L_0H_0; markers: L_1H_2 and the norm-matched control L_0H_2; positive = damage). Diagonals take $+47$–50 pp damage vs. $+10$–20 pp for cardinals, consistent with diagonals needing simultaneous x- and y-axis coding; L_1H_2 and the control stay within the ± 3 pp noise floor (grey band).

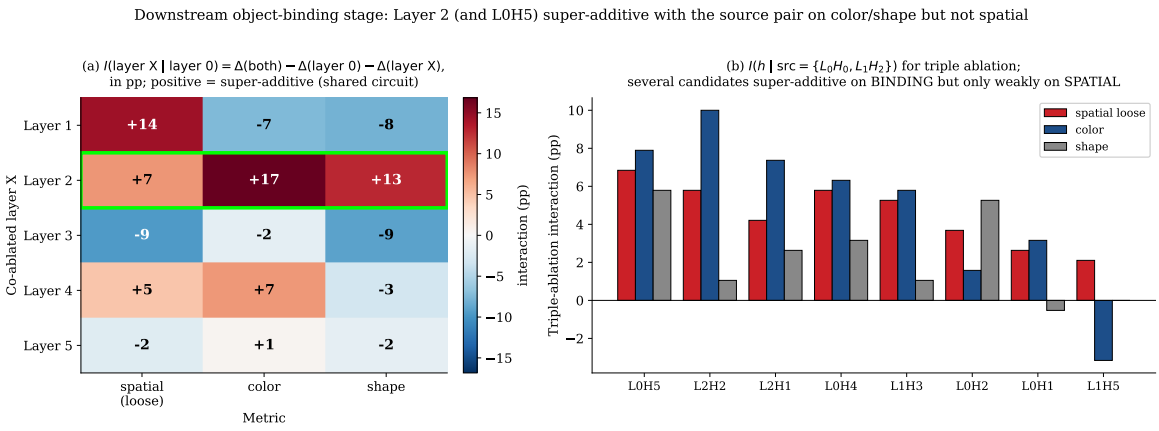


Figure 12. *(a)* Layer-pair interaction with layer 0; layer 2 (lime) shows positive interaction on color and shape, but only a smaller interaction on spatial accuracy. *(b)* Triple-ablation interaction $I(h \mid \{L_0H_0, L_1H_2\})$. Spatial interactions (red bars) are systematically smaller than the corresponding color interactions (blue bars).