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

Juliana Li & Binxu Wang · Kempner Institute at Harvard University · ICML 2026 Mech Interp

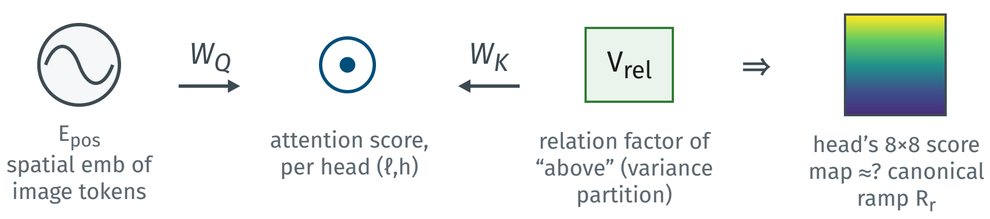
Motivation & Literature Gap

- Text-to-image DiTs can follow *"red square is **above** blue circle"*, but **how a relation word becomes image layout** has never been resolved mechanistically.
- Existing DiT interpretability stops short of heads: attribution maps (DAAM, ConceptAttention) or whole blocks (unlearning). **No individual head has been tied to a named generative computation.**
- **We ask: which cross-attention heads turn spatial phrases like "above" into layout, and can we prove it causally?**

Setup & Methods

- **Model:** PixArt-mini DiT, 6 layers × 6 cross-attention heads (**36 total**), d = 384.
- **Task:** ~1000 synthetic prompts "[colour] [shape] is [relation] [colour] [shape]", 8 spatial relations, 128×128 images.
- **Evaluation:** OpenCV detector scores disjoint **spatial / colour / shape** accuracies; prompt-bootstrap 95% CIs on every number.

1. **Discover** - project each relation vector through every head's QK circuit; score alignment to canonical directional ramps; 200-permutation null:



2. **Verify** - zero-head ablation vs norm-matched controls; pair, triple & layer ablation; activation patching.
3. **Find readers** - 4-method consensus: QK–OV composition, path patching, activation correlation, ramp projection.
4. **Trace** - re-run the full analysis at 11 training checkpoints.

$$|p|(l,h) = max_r \, |q_{l,h,r}{}^{\top} R_r| \, / \, \|R_r\|$$

projection of a head's QK write onto the canonical spatial ramp R_r; captures direction **and** write strength.

Key Results

*T5 prompt: "red square is **above** blue circle"*

↓

Layer 0 — **L0H0 ★**

↓

Layer 1 — **L1H2 ★**

↓

Layer 2 — all 6 heads

↓

Layers 3 – 5

↓

Generated 128×128 image

Stage 1 - Spatial routing: decide *where*

L0H0 + L1H2 read the relation token and write an **8×8 directional attention bias** ("above" → mass in the top rows) into the residual stream. Pair ablation: **−64 pp**, super-additive I = +0.086.

spatial code, via residual stream

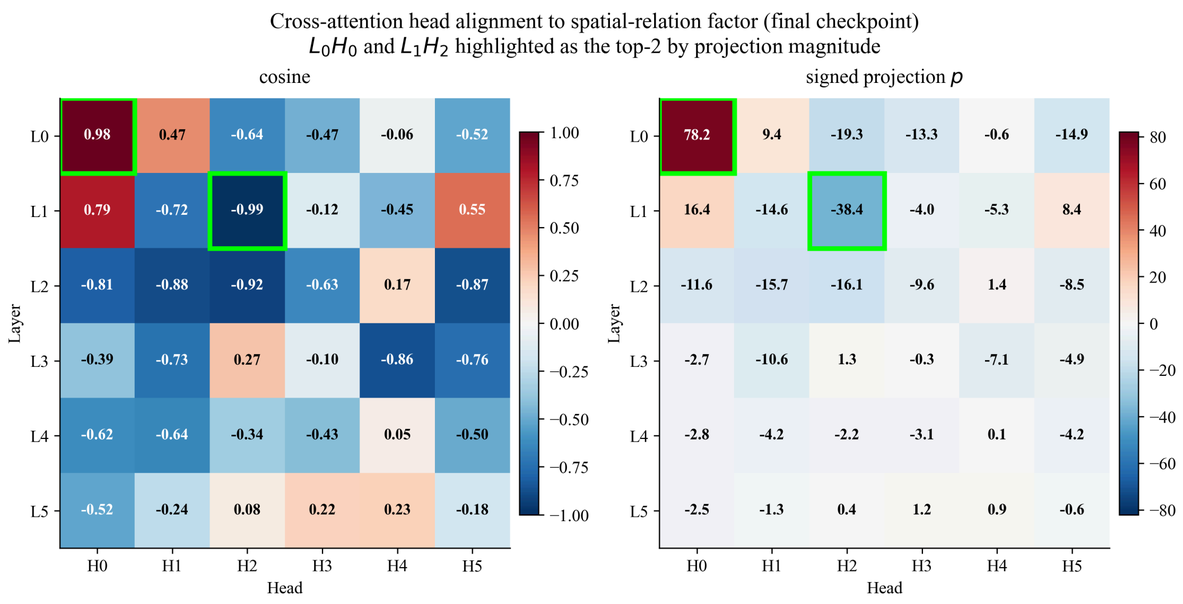
Stage 2 - Object binding: fill in *what*

Layer 2 (all 6 heads) reads Stage 1's spatial code and binds **colour / shape → position**. Near-zero alone; with layer 0: **+17 pp colour, +13 pp shape**.

In one sentence: Stage 1 decides where each object goes; Stage 2 puts the right object there.

1. Two heads route spatial relation

- L0H0 projection magnitude **|p| = 47.7** - 1.9× the runner-up L1H2, 8.4× the 36-head median, **≈ 27σ above the permutation null** ($p < 10^{-4}$).
- |cos| is a flat 0.7-0.8 shelf across heads: direction alone doesn't find sources; **direction × write-strength does**.
- Raw caption variance is dominated by shape (R^2 0.28 vs 0.11 for relation), yet the relation factor concentrates into these two heads - **sparse circuitry without sparse geometry.**

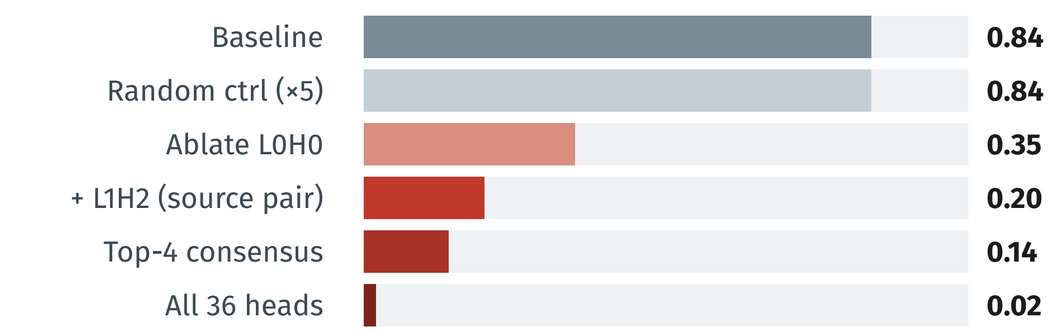


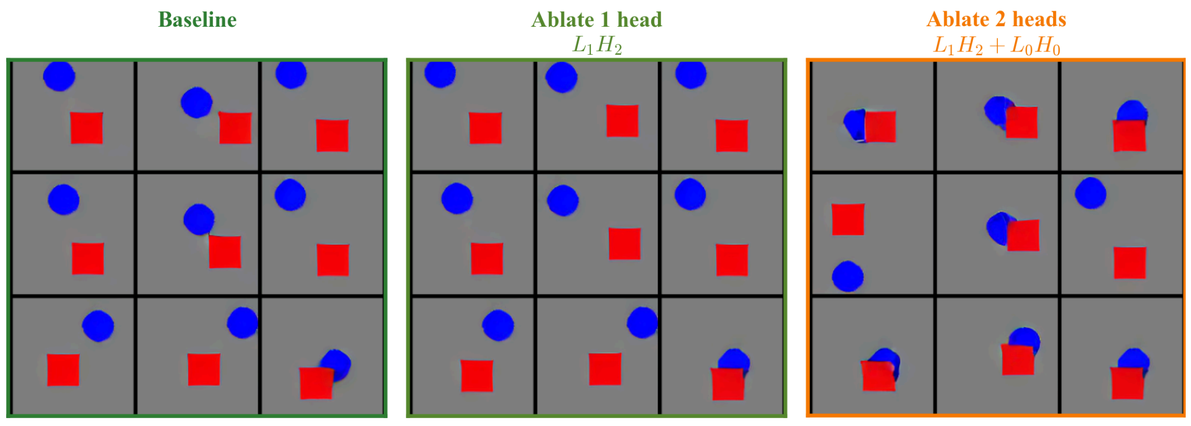
Per-head alignment to the spatial-relation factor (final checkpoint). Lime boxes: L0H0 and L1H2, the top-2 by |p|.

2. These two heads are causally necessary

- Zeroing **L0H0 alone**: spatial accuracy **0.84 → 0.35 (-48 pp)**, a **160:1** effect vs norm-matched controls.
- Zeroing the **pair**: → 0.20 (**-64 pp**); super-additive I = +0.086.
- L1H2 alone: -6 pp - a **redundant backup** that matters only once L0H0 fails. Effect persists on held-out OOD prompts (–51 pp).
- Holds **off-distribution**: unseen colour x shape pairs -51 pp; synonym phrasings -19 pp; **diagonal relations** hit hardest (-48 pp vs -15 for cardinals).

Cumulative zero-ablation - loose spatial accuracy





Same prompts & seed: ablating L1H2 alone is indistinguishable from baseline; ablating the pair collapses routing; objects merge at the center.

3. Evidence for a second, attribute-level stage

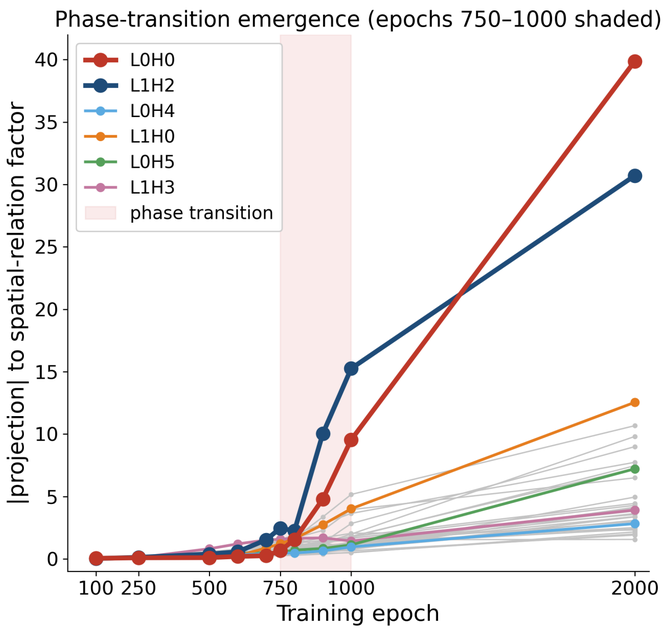
- Ablating layer 2 alone: **≤ 1 pp** on every metric. Ablating it **with layer 0**: colour and shape fall **17 / 13 pp more** than the two single ablations predict.
- A backup spatial router would show extra *spatial* damage. Layer 2's shows up on *attributes*, so it **binds identity to position**, using layer 0's spatial code.
- The stage is distributed across L2: no single L2 head ablation exceeds 3 pp.
- **Caveat:** our metrics are coupled (the detector must match colour/shape before scoring relation), decoupled metrics are ongoing work.

	spatial	colour	shape
L1	+14	−7	−8
L2	+7	**+17**	**+13**
L3	−9	−2	−9
L4	+5	+7	−3
L5	−2	+1	−2

Layer-pair interaction I(layer L | layer 0), in pp. Contrast the rows: **L1**'s extra damage is *spatial* (+14). Only **L2**'s is *color & shape - the second stage* downstream of routing.

Lesson: ablate on every output metric, not just the target one.

4. The circuit emerges in a sharp phase transition



- |p| rises **~1000×** over training (0.05 → 47.7); a **14× jump concentrated in epochs 750–1000**.
- Behavioural accuracy follows **~200 epochs later**; grokking-like lag.
- |cos| is flat from initialisation, a **misleading progress measure**.

|p| for the top-6 heads, 10 checkpoints; transition window shaded.

Takeaways & What's Next

- **Head-level causal circuit discovery transfers from language models to diffusion transformers:** individual heads implement nameable generative computations.
- **Metric-resolved ablation should be standard practice:** single-metric ablation provably misses downstream stages that operate on other metrics (in any multi-attribute generative system, including LLMs).
- **Scope & next steps:** single seed, 36-head model, time-aggregated ablation → multi-seed replication, step-resolved ablation across the noise schedule, dense layer-2 ablation, larger DiTs.