Structural Graph Probing of Vision–Language Models

Haoyu He[*♮] **Yue Zhuo**[*†] **Yu Zheng**[✉†] **Ryan Qi Wang**[✉♮]

[♮]Northeastern University, Boston, MA 02115 USA; [†]Massachusetts Institute of Technology, Cambridge, MA 02139 USA

CVPR JUNE 3-7, 2026 DENVER COLORADO

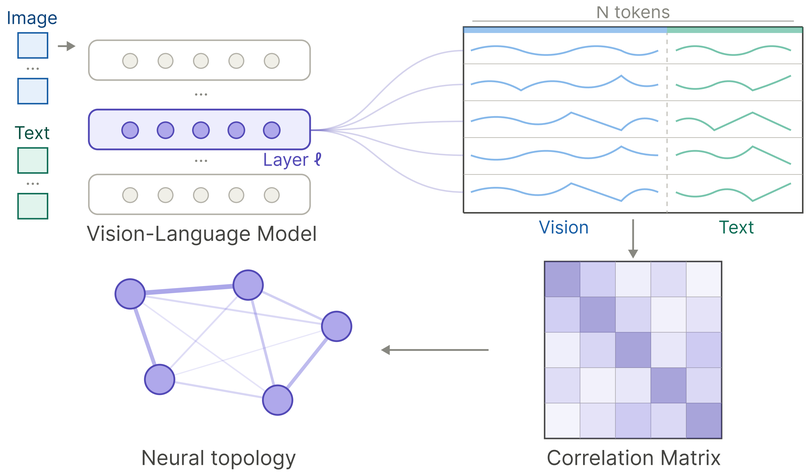
The Big Idea

Question. How does information actually *flow* inside a VLM?

Our move. Don't read activations – read *topology*. Each transformer layer = a neuron–neuron correlation graph. Analyse VLMs as *population-level structure*.

- **Topology alone** predicts VLM behaviour – no activation magnitudes, no tokens.
- **Mid-layer hubs** mediate vision ↔ text fusion as depth grows.
- **Hubs are causal**: tiny perturbations crash accuracy.

Framework



Hidden states → neuron correlation matrix → sparse graph → GCN probe. *Model-agnostic*: every transformer layer becomes a graph.

Method in 3 Steps

1. Build. Correlate every neuron pair, per layer:

$$W_{ij}^{(\ell)} = \mathrm{corr}\big(H_{i,:}^{(\ell)}, H_{j,:}^{(\ell)}\big), \quad H^{(\ell)} \in \mathbb{R}^{d\times N}.$$

2. Sparsify. Keep top-$k\%$ ($\leq 20\%$) edges – robust to pruning.

3. Probe topology only. Learnable neuron identity X (*no activations leak in*); GCN → small classifier:

$$Z^{(\ell)} = \sigma\big(D^{-\frac{1}{2}} W^{(\ell)} D^{-\frac{1}{2}} X\, W_g\big),$$
$$h^{(\ell)} = [\mathrm{Mean}(Z^{(\ell)}); \mathrm{Max}(Z^{(\ell)})].$$

Takeaways

- **Topology** > **activations** for predicting VLM behaviour.
- **Fusion is depth-localised**, mediated by mid-layer hubs.
- **Hubs are causal** control points for multimodal reasoning.

Result 1: Topology Predicts Behaviour

Setup. Run task. Build per-layer graph. Probe topology only.

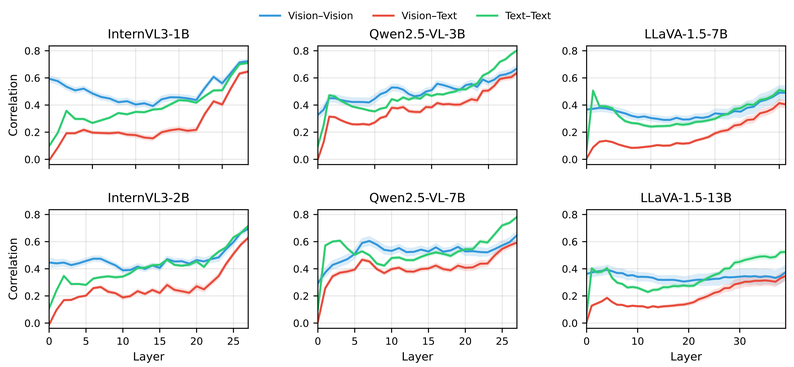
Outcome. Topology probes beat activation-based linear probes on grounded reasoning + compositional counting:

Model	Task	Linear	GCN	Δ
InternVL3-1B	TDIUC	0.884	**0.965**	**+8.1**
	CLEVR	0.980	**0.993**	**+1.3**
Qwen2.5-VL-3B	TDIUC	0.943	**0.976**	**+3.3**
	CLEVR	0.920	**0.963**	**+4.3**
LLaVA-1.5-7B	CLEVR	0.602	**0.679**	**+7.7**

Bonus – hallucinations (MHaluBench, F1): text baselines ≤ 0.66 vs. GCN 0.79–0.91 across all three VLMs.

Punchline. *Wiring beats firing* – connectivity already encodes how a VLM solves the task.

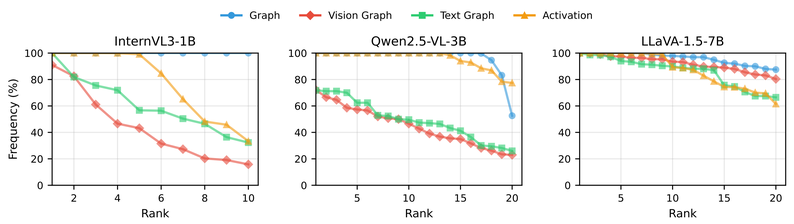
Result 2: Fusion Grows With Depth



What we measure. Mean vision–text token correlation per layer.

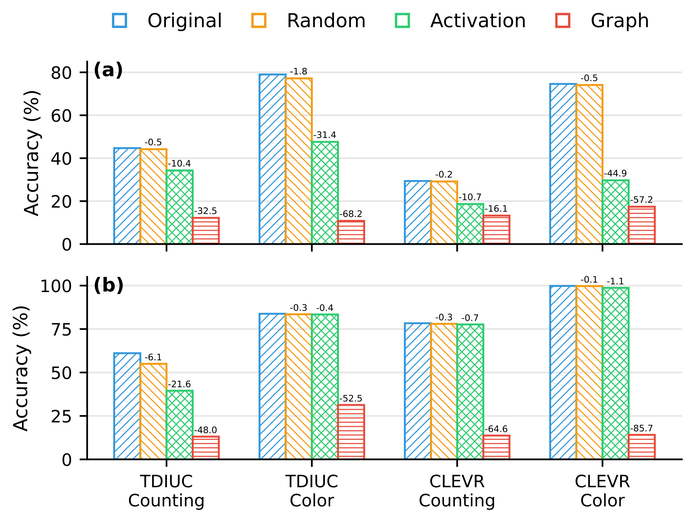
What we see. V–T and T–T couplings ↑ with depth, V–V stays flat ⇒ vision is *absorbed into the language stream*, consistently across model sizes.

Result 3: Mid-Layer Hubs Are Stable



Hub strength $d_i^{(\ell)} = \sum_j |W_{ij}^{(\ell)}|$; keep top-$1\%$ per sample. **Graph-defined hubs recur most consistently** across samples (vs. activation- or modality-specific hubs), peaking at *mid-depth* – exactly where fusion is strongest.

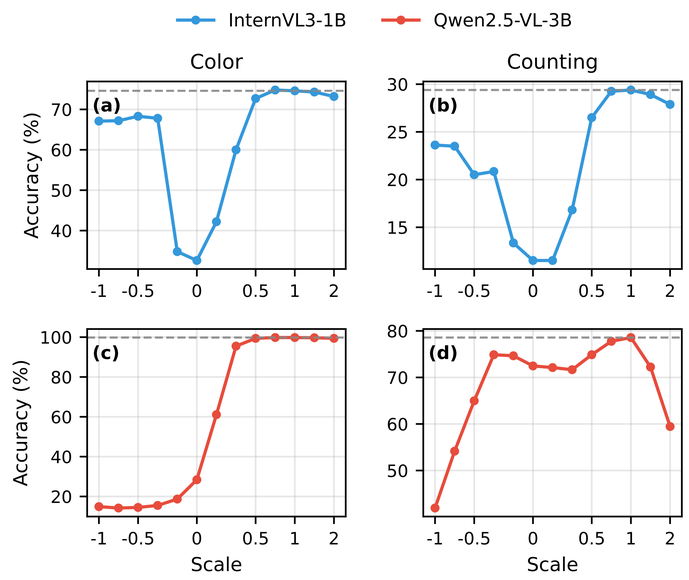
Causal Test 1: Topology Picks Critical Neurons



Test. Remove top-1% neurons per sample, ranked by Graph degree vs. Activation magnitude vs. Random.

Outcome. **Graph-degree ablation drops accuracy hardest** on TDIUC + CLEVR – topology, not magnitude, picks the behaviourally critical units.

Causal Test 2: Hubs Are the Control Knobs



Test. Scale a single hub in InternVL3-1B (n62 @ L11) or five hubs in Qwen2.5-VL-3B (L0); leave the rest of the model untouched.

Outcome. Tiny perturbations crash accuracy *symmetrically* under amplify / suppress – hubs sit on a *sharp ridge* of multimodal reasoning.