This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Structural Graph Probing of Vision–Language Models

Haoyu He*† Yue Zhuo*‡ Yu Zheng✉‡ Qi R. Wang✉†
†Northeastern University ‡Massachusetts Institute of Technology
{he.haoyu1, q.wang}@northeastern.edu {joyzhuo, yu_zheng}@mit.edu

Abstract

Vision–language models (VLMs) achieve strong multimodal performance, yet how computation is organized across populations of neurons remains poorly understood. In this work, we study VLMs through the lens of neural topology, representing each layer as a within-layer correlation graph derived from neuron–neuron co-activations. This view allows us to ask whether population-level structure is behaviorally meaningful, how it changes across modalities and depth, and whether it identifies causally influential internal components under intervention. We show that correlation topology carries recoverable behavioral signal; moreover, cross-modal structure progressively consolidates with depth around a compact set of recurrent hub neurons, whose targeted perturbation substantially alters model output. Neural topology thus emerges as a meaningful intermediate scale for VLM interpretability: richer than local attribution, more tractable than full circuit recovery, and empirically tied to multimodal behavior. Code is publicly available at https://github.com/he-h/vlm-graph-probing.

1. Introduction

Vision–language models (VLMs) have advanced rapidly as general-purpose multimodal systems [3, 5, 25, 36, 51], yet how their competence is organized internally remains incompletely understood [13, 40]. A central unresolved problem is how multimodal computation is organized within the network: how visual evidence and linguistic context are coordinated across layers, how intermediate computation is distributed across populations of units, and which internal structures ultimately govern model behavior. Without an account at that level, interpretability remains largely descriptive, useful for identifying salient inputs or components but limited in its ability to explain how multimodal reasoning is internally structured [20, 46].

Existing analyses of VLMs have primarily emphasized local explanatory signals, including attention patterns, saliency maps, patch attribution, and component-level inspection [1, 9, 33, 38, 44, 46]. While these approaches are valuable for identifying influential inputs or localized mechanisms, they are less well suited to characterizing the large-scale organization through which multimodal behavior is realized. This limitation is especially consequential in transformer-based VLMs, where computation is distributed across large populations of interacting units rather than concentrated in a small number of isolated pathways [3, 25, 42]. More broadly, both neuroscience and mechanistic interpretability point to a common lesson: complex computation often becomes most intelligible at the level of structured populations, interaction patterns, and hub-like organization, rather than at the level of individual units in isolation [8, 12, 15, 19, 34, 39]. For VLMs, this suggests that population-level interaction structure is not merely an auxiliary diagnostic, but a meaningful level of analysis in its own right [24, 49, 50].

In this work, we study VLMs through the lens of *neural topology*: the layerwise structure of neuron correlation induced during multimodal inference. Given an image–question pair, we record hidden activations, construct within-layer neuron correlation graphs, and use these graphs as a central object of analysis [50]. This topology-centered perspective makes it possible to examine multimodal transformers at the level of population organization: to assess how strongly model behavior is reflected in internal interaction structure, to trace how cross-modal organization evolves across depth, and to identify structurally important neurons whose perturbation alters model predictions. Across multiple VLM families, we find that graph-based probes recover substantial signal about model outputs and hallucination behavior; that modality-specific topology reveals systematic changes in cross-modal organization across layers; and that perturbing topology-defined hub neurons significantly changes model predictions.

Taken together, these results establish neural topology as a meaningful perspective on vision–language model interpretability. Rather than treating multimodal interpretability solely as the problem of identifying salient tokens or im-

*Equal contribution.

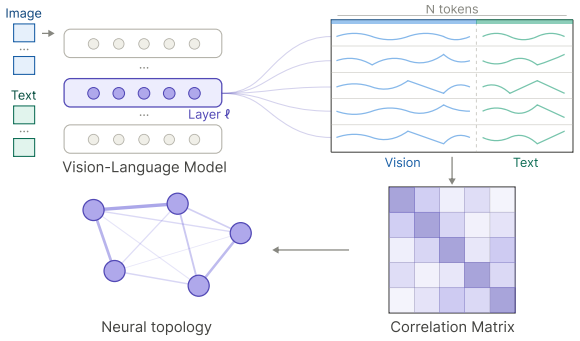

Figure 1. **Overview of neural topology construction.**

age regions, our findings suggest that the organization of neuron correlations itself carries consequential information about model behavior. Under this view, *behavioral predictability, multimodal structure, and intervention* are best understood not as separate analyses, but as complementary perspectives on the same underlying computational organization. This, in turn, points toward a broader approach to VLM interpretability centered on computation as an organized population-level system rather than as a collection of local attribution effects.

2. Neural Topology

We analyze vision-language models at the level of *population structure*. Rather than treating multimodal reasoning as the output of isolated tokens, heads, or neurons, we view each transformer layer as a distributed computational state whose organization is expressed through the correlations among its units. This choice is motivated by findings that functionally relevant organization is often expressed through distributed representational subspaces and heterogeneous network topology, including hub-dominated architectures, rather than through individual units in isolation [14, 35]. In this work, we operationalize that perspective through *neural correlation topology*, a layerwise description of neuron–neuron correlation structure during multimodal inference.

2.1. Neuron Correlation Topology

Given an image I and text prompt T, a frozen VLM produces hidden activations at each transformer layer ℓ. We denote the hidden representation at layer ℓ as $H^{(\ell)} \in \mathbb{R}^{d\times N}$, where N is the number of multimodal tokens and d is the hidden dimension. Each row $H^{(\ell)}_{i,:}$ therefore represents the response of neuron i across all tokens in the joint multimodal context. We use these hidden states only to infer correlation structure; the downstream analysis module never receives the activation values themselves.

We represent each layer as a weighted graph

$$G^{(\ell)} = (V, E, W^{(\ell)}), \tag{1}$$

where each node in V corresponds to a neuron, so that $|V| = d$, and $E = V \times V$ denotes the complete set of neuron pairs. The edge weight $W^{(\ell)}_{ij}$ measures the functional coupling between neurons i and j, defined here as the Pearson correlation between their activation profiles across tokens:

$$W^{(\ell)}_{ij} = \mathrm{corr}\left(H^{(\ell)}_{i,:}, H^{(\ell)}_{j,:}\right) = \frac{\left(H^{(\ell)}_{i,:} - \bar{H}^{(\ell)}_{i,:}\right)^{\top}\left(H^{(\ell)}_{j,:} - \bar{H}^{(\ell)}_{j,:}\right)}{\left\|H^{(\ell)}_{i,:} - \bar{H}^{(\ell)}_{i,:}\right\|\left\|H^{(\ell)}_{j,:} - \bar{H}^{(\ell)}_{j,:}\right\|}. \tag{2}$$

This construction yields a layerwise correlation graph in which edge weights reflect how similarly pairs of neurons respond within the same inference pass, shown in Figure 1. In other words, neural topology captures the organization of co-activation structure within each layer, a correlation-based view of within-layer population structure rather than a literal wiring diagram of the model.

2.2. Vision, Text, and Multimodal Topology

To examine how internal structure differs across modalities, we construct three related topologies at each layer. The multimodal graph $G^{(\ell)}$ is computed from the full hidden state obtained under joint image–text inference. To extract modality-specific structure, we use the same multimodal forward pass and partition the hidden states by token type, using positional indices to separate visual and textual tokens. Let $H^{(\ell)}_{\mathrm{vis}}$ and $H^{(\ell)}_{\mathrm{text}}$ denote the subsets of hidden activations associated with vision and text tokens, respectively. We then construct

$$G^{(\ell)}_{\mathrm{vis}} = \mathrm{GraphCorr}\left(H^{(\ell)}_{\mathrm{vis}}\right), G^{(\ell)}_{\mathrm{text}} = \mathrm{GraphCorr}\left(H^{(\ell)}_{\mathrm{text}}\right), \tag{3}$$

using the same correlation-based procedure as in the multimodal case. Because these graphs are derived from different token subsets within the same forward pass, differences among $G^{(\ell)}$, $G^{(\ell)}_{\mathrm{vis}}$, and $G^{(\ell)}_{\mathrm{text}}$ reflect how correlation structure specializes to visual tokens, textual tokens, and their joint multimodal context.

2.3. Neural Topology Representation

To analyze these graphs without directly exposing hidden activation magnitudes or token semantics, we represent each neuron using a learnable one-hot node identity embedding. Each neuron is assigned a unique basis vector, which is projected by a trainable embedding layer into a low-dimensional feature space. The resulting node features preserve neuron identity while keeping the analysis centered on correlation structure rather than raw hidden-state content. A graph convolutional network (GCN) operates over the correlation graph to produce topology-dependent

Table 1. **Graph probing across multimodal benchmarks.** Accuracy and F1 of linear vs. graph-based probes trained on layer-wise neuron-correlation topology for TDIUC, CLEVR, MMMU, MMMU-Pro, BLINK, and EMMA. Best result in each model–dataset pair is in **bold**. Graph-based probes outperform linear baselines in most settings, indicating that population-level connectivity carries task-relevant information. See Table 5 for Precision/Recall.

Dataset	InternVL3-1B				Qwen2.5-VL-3B				LLaVA-1.5-7B			
	Linear		GCN		Linear		GCN		Linear		GCN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
TDIUC	0.884	0.846	**0.965**	**0.959**	0.943	0.930	**0.976**	**0.970**	**0.971**	**0.966**	0.954	0.947
CLEVR	0.980	0.980	**0.993**	**0.993**	0.920	0.919	**0.963**	**0.963**	0.602	0.563	**0.679**	**0.676**
MMMU	0.293	**0.253**	**0.321**	**0.253**	0.293	0.211	**0.336**	**0.335**	**0.314**	**0.300**	0.279	0.162
MMMU-Pro	0.309	0.226	**0.359**	**0.260**	0.286	0.164	**0.318**	**0.290**	0.291	0.171	**0.314**	**0.305**
BLINK	0.549	0.530	**0.592**	**0.589**	0.544	0.543	**0.565**	**0.564**	**0.647**	**0.646**	0.592	0.591
EMMA	0.343	0.195	**0.380**	**0.268**	0.329	0.146	**0.343**	**0.288**	0.307	**0.228**	**0.329**	0.218

node representations:

$$Z^{(\ell)} = \mathrm{GCN}\Big(W^{(\ell)}, X\Big) = \sigma\Big(D^{-\frac{1}{2}} W^{(\ell)} D^{-\frac{1}{2}} X W_g\Big), \tag{4}$$

where X denotes the node embedding matrix, D is the degree matrix of $W^{(\ell)}$, W_g is a learnable weight matrix, and $\sigma(\cdot)$ is a nonlinear activation function. Since the GCN receives only graph structure and node identities, the resulting representation reflects how neurons are organized within the layer rather than what individual neurons encode.

To obtain a fixed-dimensional structural signature for each layer, we aggregate the node representations with complementary global pooling operators:

$$h^{(\ell)} = \mathrm{Concat}\Big(\mathrm{Mean}\Big(Z^{(\ell)}\Big), \mathrm{Max}\Big(Z^{(\ell)}\Big)\Big). \tag{5}$$

The mean-pooled term captures the overall correlation tendency of the layer, while the max-pooled term preserves salient high-response structure. Each transformer layer therefore yields a graph-level signature $h^{(\ell)}$ that can be used to study behavioral predictability, multimodal structure, and intervention targets in a common representation space.

3. Predictability

A central question is whether neural topology encodes behaviorally meaningful information, rather than merely reflecting incidental patterns of co-activation. We address this question by asking whether layerwise correlation graphs support reliable prediction of model behavior across grounded reasoning, semantic recognition, and hallucination classification tasks. This section serves as the first empirical test of our framework: rather than claiming mechanism from probe performance alone, we use predictability to establish that neural topology is a structured and behaviorally informative representation, thereby motivating the more detailed structural and causal analyses that follow.

Experimental Setup. We evaluate three representative VLMs: InternVL3-1B [52], Qwen2.5-VL-3B [6], and LLaVA-1.5-7B [29]. In the main text, we focus on CLEVR [21], TDIUC [22], and MHaluBench [10], which respectively test numerical grounding, semantic recognition, and multimodal hallucination. Broader results on MMMU [47], MMMU-Pro [48], BLINK [16], and EMMA [18] are reported in Table 1. For each dataset, we split examples into 80% training and 20% test sets, and train both a linear probe and a GCN probe on each layer's graph representation. Full dataset descriptions, model details, and training settings are provided in Section 8.

3.1. Behavioral Predictability

We use CLEVR object counting to probe quantitative reasoning, and color recognition on both CLEVR and TDIUC alongside TDIUC sports classification to probe semantic understanding. These tasks allow us to test whether graph-derived representations of internal activity encode information that is predictive of model behavior.

Probing Performance. Table 1 reports out-of-sample probing accuracy for linear and GCN probes across CLEVR and TDIUC, together with broader results on MMMU, MMMU-Pro, BLINK, and EMMA. Overall, graph-based probes outperform linear baselines on most model–dataset pairs, with the clearest gains on CLEVR counting and TDIUC. The strongest improvements appear on CLEVR, where the GCN probe improves over the linear baseline by 7.7% on LLaVA, 4.3% on Qwen2.5-VL, and 1.3% on InternVL3. Broader multimodal benchmarks exhibit a more mixed pattern, suggesting that topology is especially informative on grounded tasks with tighter alignment between internal multimodal organization and target outputs.

These results indicate that modeling relational structure within the graph yields additional predictive value beyond a linear readout of the same graph-derived representation.

Table 2. **Regression on object counting from neural topology.** Linear and graph-based probes on CLEVR counting. Graph-based probes improve regression performance across all three VLMs.

Model	Probe	MSE ↓	R^2 ↑	Pearson ↑
InternVL3-1B	Linear	0.020	0.996	0.998
	GCN	**0.007**	**0.999**	**0.999**
Qwen2.5-VL-3B	Linear	0.081	0.985	0.992
	GCN	**0.038**	**0.993**	**0.996**
LLaVA-1.5-7B	Linear	0.605	0.884	0.949
	GCN	**0.379**	**0.928**	**0.963**

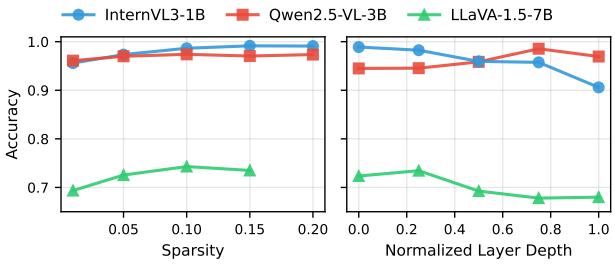

Figure 2. **Sparsity robustness and depth dependence of graph probing.** Left: probing accuracy as a function of graph sparsity (top 1%–20% of neuron correlations retained). Right: probing accuracy across normalized layer depth. Accuracy is stable across sparsity levels, while depth-wise peak predictiveness differs by architecture.

Notably, all results are obtained on sparse correlation graphs with density at most 0.2, showing that high predictive performance does not require dense connectivity.

Beyond classification, Table 2 evaluates CLEVR object counting as a regression problem. Across all three models, graph-based probes reduce MSE and improve both R^2 and Pearson correlation, indicating that the probe captures genuine functional dependence between topology and target counts rather than overfitting to dataset idiosyncrasies. This shows that the advantage of topology extends beyond discrete label prediction to finer-grained numerical estimation. Taken together, these results establish that neural topology contains recoverable signal about both semantic and quantitative task behavior.

Sparsity and Graph Construction. A practical challenge in graph probing for VLMs is scale: each layer contains thousands of neurons, producing correlation graphs with tens of millions of possible edges. Fully connected graphs are computationally prohibitive and can obscure the strongest structural relations. To address this, a sparse construction strategy retains only the top-k fraction of edges with the largest pairwise correlations, yielding a weighted graph that remains tractable while preserving dominant dependencies.

To evaluate the effect of this design choice, we sweep sparsity levels from 0.01 to 0.20. As shown in Figure 2(a), probe accuracy remains largely stable across this range, with only marginal gains as additional weaker edges are included. This suggests that, under our probing setup, the most predictive structural signal is already concentrated in the strongest correlations. From a practical perspective, this justifies sparse graph construction as an efficient approximation; from an analytical perspective, it indicates that task-relevant topology can be recovered without modeling the full dense correlation graph.

Layerwise Predictability. To examine how predictive topology varies with depth, probe accuracy is evaluated across layers of each VLM in Figure 2(b). Qwen2.5-VL-3B exhibits a clear mid-to-late peak, reaching its highest accuracy around layer 27 before slightly declining near the final layer. In contrast, LLaVA-1.5-7B and InternVL3-1B show flatter or gradually declining trends across depth.

These results suggest that the depth at which topology is most behaviorally informative differs across architectures. Rather than drawing a strong functional conclusion from probe accuracy alone, these layerwise differences serve as motivation for the structural analysis in the next section, where we examine how multimodal correlation patterns evolve across depth.

3.2. Hallucination Detection

We test whether neural topology can distinguish hallucinating from non-hallucinating responses on MHaluBench. A binary classifier is trained on graph representations extracted from each model to predict hallucination status. As text-only controls, we construct two simple baselines using `word2vec` [32, 53]: the mean embedding of each question–answer prompt. Separate linear probes are trained on these features to estimate how much hallucination is predictable from shallow textual statistics alone.

As shown in Table 3, graph-based probes consistently outperform these text-only baselines across all three models, indicating that hallucination status is associated with structural information in neuron–neuron correlation graphs beyond simple lexical cues. This result is best viewed as evidence of informativeness rather than as a competitive hallucination-detection system: the main takeaway is that topological representations capture signals related to whether a response is grounded or hallucinatory.

4. Multimodal Structure and Alignment

Beyond predictability, a key question is what kind of multimodal organization neural topology reflects. We examine

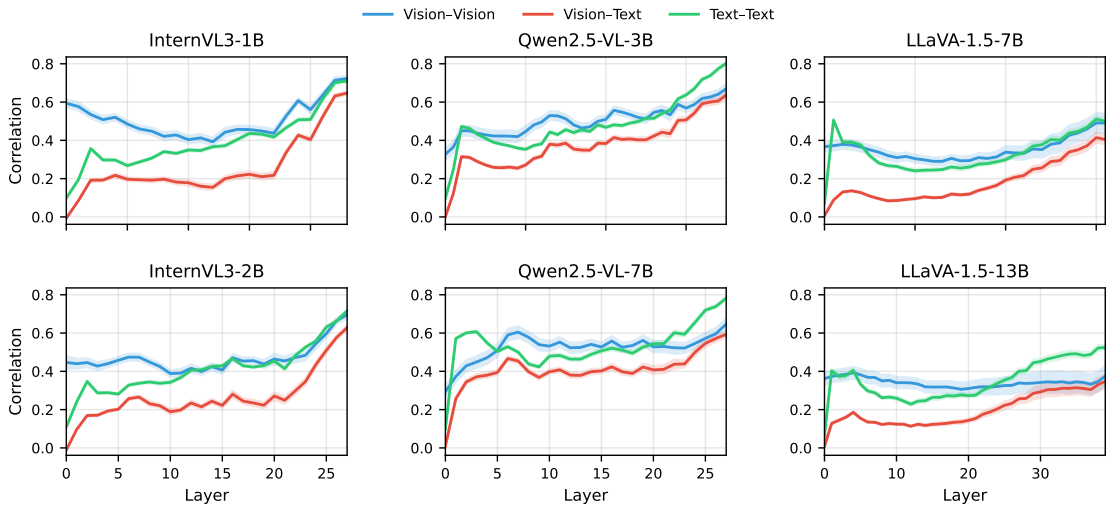

Figure 3. **Token-level cross-modal correlation dynamics across depth.** Layer-wise token–token correlations for Vision–Vision, Vision–Text, and Text–Text pairs on TDIUC (mean ± std) across multiple VLM families and scales. Vision–Text correlations increase with depth, consistent with progressively stronger multimodal integration in later layers.

Table 3. **Hallucination detection from neural topology on MHaluBench.** Accuracy of a graph-based probe versus two text-only baselines (mean word2vec embedding and token length) for binary hallucination classification. Graph-based probes outperform both baselines across all models.

Method	InternVL3-1B	Qwen2.5-VL-3B	LLaVA-1.5-7B
Emb. Avg.	0.664	0.654	0.649
Length	0.500	0.633	0.642
GCN	**0.789**	**0.910**	**0.908**

the hidden states at three complementary levels: token-level correlation dynamics, neuron-level persistence of structural roles, and graph-level alignment across modality conditions. Together, these analyses characterize how visual and linguistic information become coupled, stabilized, and organized within the internal topology of VLMs.

4.1. Cross-Modal Correlation Dynamics

We begin with a token-level view of multimodal interaction. Let $H^{(\ell)} \in \mathbb{R}^{d\times N}$ denote the hidden activations at layer ℓ, where d is the hidden dimension and N is the number of multimodal tokens. To quantify token–token dependencies, we compute the Pearson correlation of the transposed activations $C^{(\ell)}_{\text{tok}} = \text{corr}\big(H^{(\ell)\top}\big)$. For vision and text token sets $\mathcal{V}$ and $\mathcal{T}$, we define intra- and inter-modality coupling as

$$\mu^{(\ell)}_{\text{VV}} = \frac{1}{|\mathcal{V}|(|\mathcal{V}|-1)} \sum_{i,j\in\mathcal{V}} C^{(\ell)}_{\text{tok}}[i,j],$$
$$\mu^{(\ell)}_{\text{TT}} = \frac{1}{|\mathcal{T}|(|\mathcal{T}|-1)} \sum_{i,j\in\mathcal{T}} C^{(\ell)}_{\text{tok}}[i,j], \tag{6}$$
$$\mu^{(\ell)}_{\text{VT}} = \frac{1}{|\mathcal{V}|\,|\mathcal{T}|} \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{T}} C^{(\ell)}_{\text{tok}}[i,j].$$

As shown in Figure 3, both vision–text and text–text correlations increase with depth, whereas vision–vision correlations remain comparatively flat. This pattern is consistent with progressively stronger multimodal integration in later layers, although the statistic itself is descriptive rather than mechanistic. In decoder-style VLMs, this asymmetry may reflect the role of visual tokens as conditioning inputs that increasingly shape the language-side representation as depth increases.

4.2. Structural Hub Stability

A natural follow-up is whether topology identifies structural roles that persist across inputs. For each correlation graph $W^{(\ell)}$, we define the degree of neuron i as

$$d^{(\ell)}_i = \sum_j |W^{(\ell)}_{ij}|. \tag{7}$$

Neurons within the top-$k\%$ of $d^{(\ell)}_i$ are treated as *hub neurons*. For a set of samples $\mathcal{S}$ with corresponding hub sets

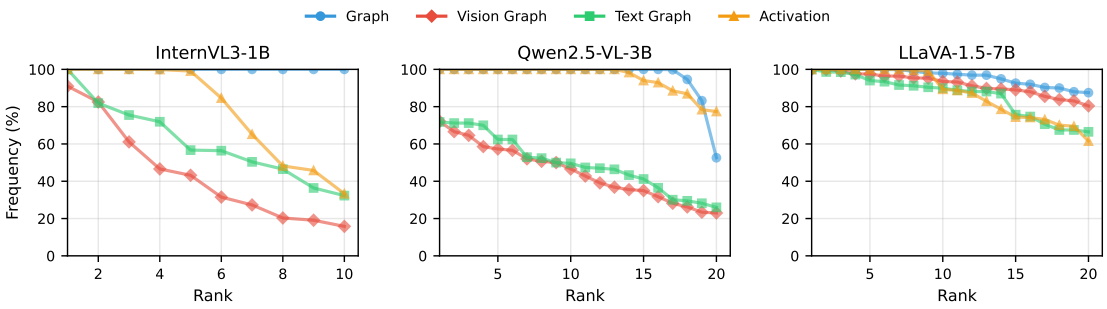

Figure 4. **Cross-sample stability of hub definitions.** Recurrence of top 1% hubs across samples on TDIUC for graph-wide, modality-specific, and activation-based hub definitions. Graph-derived hubs are the most stable, indicating that structurally central neurons occupy more persistent roles than alternative hub candidates.

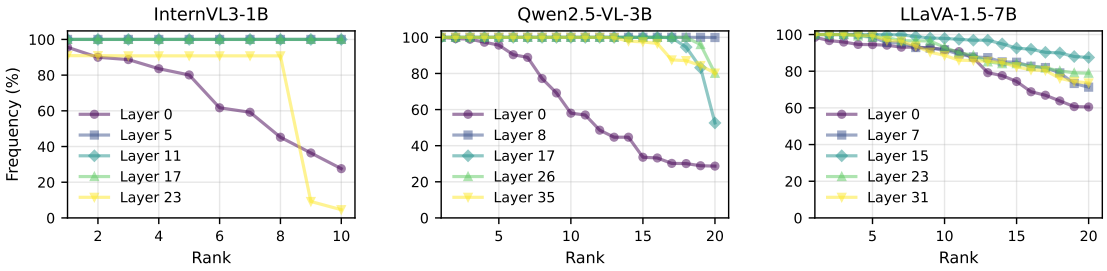

Figure 5. **Layer-wise stability of graph hubs.** Recurrence of top 1% graph-derived hubs across samples at different depths on TDIUC. Intermediate layers show the strongest hub stability, suggesting the most persistent population-level organization emerges in mid-depth representations.

$\mathcal{H}_s^{(\ell)}$, we measure cross-sample recurrence as

$$\pi_i^{(\ell)} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1}[i \in \mathcal{H}_s^{(\ell)}]. \tag{8}$$

The distributions in Figure 4 show that graph-defined hubs exhibit substantially greater recurrence across samples than activation-based and modality-specific alternatives. This comparison disentangles three possible sources of hub persistence: full multimodal topology, unimodal subnetwork structure, and activation magnitude. Graph-defined hubs are the most stable, indicating that topology captures a more persistent notion of structural centrality than either alternative.

Stability further peaks in the middle layers (Figure 5), broadly aligning with the region where cross-modal coupling becomes strongest. Rather than taking this as direct evidence of mediation, we interpret it as a compact set of structurally central neurons appears to recur reliably across diverse inputs, suggesting that multimodal processing is organized around persistent topological loci rather than being uniformly distributed across the network.

4.3. Cross-Modal Graph Alignment

The final analysis tests whether modality-specific graphs occupy a shared structural space. For each layer ℓ, we extract graph-level embeddings $h_\Omega^{(\ell)}$ and $h_\Gamma^{(\ell)}$ from the GCN representations under different modality conditions. Rather than performing explicit node matching, we align these graph embeddings contrastively: positive pairs are drawn from the same sample and layer, while negative pairs are drawn from different samples or layers. We train the alignment model using a symmetric InfoNCE objective with cosine similarity and temperature τ:

$$\mathcal{L} = -\frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\log \frac{e^{s(z_{\Omega,i}, z_{\Gamma,i})/\tau}}{\sum_j e^{s(z_{\Omega,i}, z_{\Gamma,j})/\tau}} + \log \frac{e^{s(z_{\Gamma,i}, z_{\Omega,i})/\tau}}{\sum_j e^{s(z_{\Gamma,i}, z_{\Omega,j})/\tau}} \right]. \tag{9}$$

We evaluate alignment using Graph AUC (GAUC) [27], which measures how reliably matched graph representations are ranked above mismatched ones.

As shown in Table 4, multimodal–multimodal matching provides the highest alignment score (0.9598 GAUC), serving as a reference point for near self-alignment within the learned structural space. Matching LLaVA's text and image pathways yields a lower score (0.8188), indicating that unimodal graphs derived from the same multimodal model

Table 4. **Cross-modal graph alignment.** GAUC at layer 6 of LLaVA-1.5-7B for graph embeddings matched across modalities. Higher GAUC indicates stronger structural correspondence.

Ω Modality	Γ Modality	GAUC
Image+Text	Image+Text	0.960
Text	Image	0.819
LLaMA (Text)	LLaVA (Text)	0.680

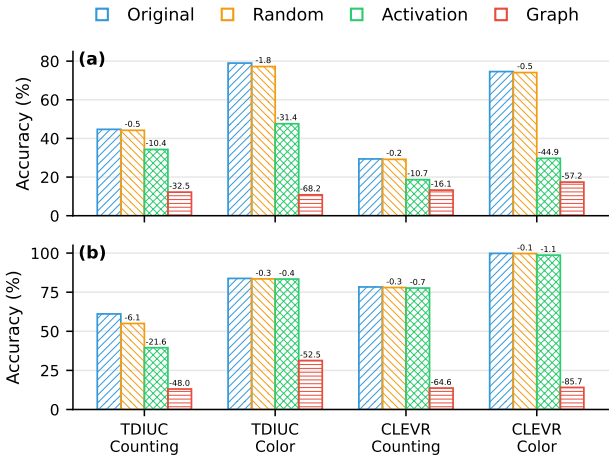

Figure 6. **Ablation by zeroing selected neurons.** Accuracy on TDIUC and CLEVR after zeroing the top 1% of neurons selected per sample by random choice, activation magnitude, or graph degree for InternVL3-1B (a, layer 11) and Qwen2.5-VL-3B (b, layer 0). Zeroing graph-selected neurons yields the largest drop.

remain only partially aligned. Alignment drops further when comparing LLaVA text graphs with those from the unimodal LLaMA backbone [43] (0.6803), suggesting that multimodal finetuning alters the inherited text-side topology in a substantial way.

Within this setting, the results suggest that multimodal training does not collapse visual and linguistic pathways into a single undifferentiated topology; instead, it brings them into partial correspondence while preserving meaningful structural differences.

5. Causal Intervention Analysis

If neural topology captures behaviorally meaningful structure, it should also identify components whose targeted perturbation materially changes model behavior. To test this, we intervene on topology-defined neurons and edges and measure the resulting performance degradation to assess whether structural centrality provides a useful criterion for selecting behaviorally influential loci.

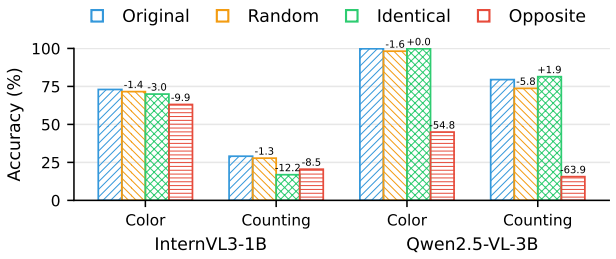

Figure 7. **Edge-level intervention on top-ranked neuron pairs.** For the strongest dataset-level graph-defined neuron pair, one endpoint is replaced by its partner's activation (Identical), the negated partner activation (Opposite), or a random matched-shape vector (Random), and downstream accuracy is measured on color and counting tasks for InternVL3-1B and Qwen2.5-VL-3B. Opposite intervention causes the largest performance drop.

Ablation of Top Neurons. The first question is whether neurons ranked highly by graph structure are more behaviorally important than those selected by simpler criteria. For each sample, the top 1% of neurons are ablated, chosen either by connectivity degree in the correlation graph or by activation magnitude, and the resulting performance on TDIUC and CLEVR is reported in Figure 6.

Across both tasks, ablating degree-ranked neurons produces the largest performance drop, indicating that graph-based ranking identifies neurons whose removal has stronger behavioral consequences than activation-based selection. The depth at which this effect is strongest differs across models: InternVL3-1B shows its largest decline around layer 11, whereas Qwen2.5-VL-3B is most sensitive at layer 0. These differences suggest that behaviorally influential topology is concentrated at different depths in different architectures, though this should not be taken as a definitive localization of multimodal fusion. Overall, structural centrality provides a more behaviorally aligned intervention criterion than activation magnitude alone.

Edge-Level Intervention. Beyond individual neurons, we next test whether strong graph-defined edges encode functionally meaningful pairwise relations, rather than merely linking individually important nodes. For the edge with the highest aggregate degree across the dataset, one endpoint is intervened on while the remainder of the representation is held fixed. Specifically, its activation is replaced with that of its partner on the same edge (IDENTICAL), with the negated partner activation (OPPOSITE), or with a random vector of matched shape (RANDOM).

As shown in Figure 7, these interventions produce a consistent ordering of effects. The IDENTICAL intervention causes little degradation and can even slightly improve performance, suggesting that preserving partner-consistent ac-

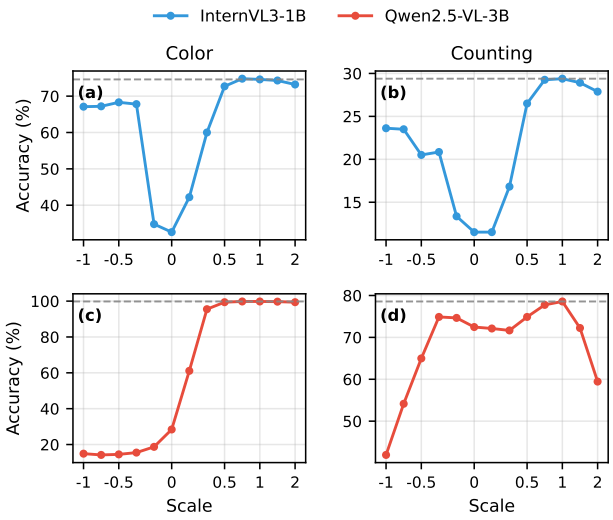

Figure 8. **Direct perturbation of topology-defined hub neurons.** Selected high-degree hub neurons in InternVL3-1B (neuron 62, layer 11) and Qwen2.5-VL-3B (neurons 71, 318, 294, 528, 583, layer 0) are scaled and evaluated on color and counting tasks. Performance degrades under both positive and negative perturbation.

tivity largely maintains the underlying relation. RANDOM replacement causes a moderate decline, whereas OPPOSITE replacement is most destructive, especially in Qwen2.5-VL-3B, where both color and counting performance drop sharply. InternVL3-1B is more robust overall but exhibits the same qualitative ordering. This pattern indicates that the behavioral importance of a strong edge depends not only on the identity of its endpoint neurons, but also on the sign and alignment of their coordinated activity — topology is informative at the level of relations as well as nodes.

Perturbation of Hub Activations. As a final probe of causality, a small set of hub neurons is directly scaled while all other activations are held fixed: one hub in InternVL3-1B (neuron 62, layer 11) and five in Qwen2.5-VL-3B (neurons 71, 318, 294, 528, 583, layer 0), as illustrated in Figure 8. Even small perturbations produce substantial performance degradation, and the effect is approximately symmetric under both amplification and suppression.

This symmetric sensitivity suggests that these hubs operate within a relatively narrow functional range: performance deteriorates not only when their activity is suppressed, but also when it is exaggerated. Across all intervention types, topology-defined hubs consistently occupy behaviorally influential positions. More broadly, neural topology proves useful not only for prediction and structural analysis, but also for identifying localized targets whose perturbation has substantial downstream effects.

6. Related Work

Interpretability in Vision–Language Models. Recent VLMs, including BLIP [25], LLaVA [29], Qwen-VL [6], and InternVL [52], achieve strong multimodal reasoning by combining visual encoders with large language models. Despite success on instruction following and visual question answering, their internal mechanisms remain difficult to interpret [13, 28]. Existing analyses rely on attention flow [1], saliency and gradient-based attribution [38], or interactive visualization [2, 40]. These methods provide local, token-level explanations but limited insight into the global organization of computation across layers. In parallel, mechanistic interpretability in language models has identified sparse features [12], causal mediation pathways [17], and mechanisms for editing factual associations [31]. Our work extends this structural perspective to VLMs by modeling each layer as a neuron–neuron correlation graph and analyzing multimodal computation through its topology.

Neural Topology and Representation Structure. Prior work has studied neural networks through representational similarity [11, 24], neural connectivity patterns [4, 30, 45], and emergent modularity in transformers [49]. Related mechanistic studies have also used causal interventions to test the functional importance of internal components [7, 26, 37]. However, these approaches do not examine the population-level topology of VLM layers or use such structure to analyze multimodal behavior. Our framework introduces a topology-based view of VLMs by treating each layer as a neuron correlation graph and studying it through graph embeddings, modality-specific subgraphs, and causal interventions on structurally prominent neurons.

7. Discussion

The broader implication of this work is a shift in what should count as an explanatory unit for vision–language models. In neuroscience, understanding complex behavior required moving beyond single-neuron selectivity toward population dynamics at an intermediate, mesoscopic scale. Our results suggest an analogous perspective: within-layer co-activation topology is not merely another summary of hidden states, but a window into how computation is organized across structured populations, persistent hubs, and coordinated relations. The value of this perspective is precisely that it sits between local attribution and full circuit recovery: rich enough to expose behaviorally consequential internal organization, yet tractable enough to compare across layers, modalities, and models. Although these graphs are not literal wiring diagrams, they suggest that multimodal reasoning may be more fruitfully understood as an emergent property of organized neural populations than as the sum of independently interpretable components.

Acknowledgments

H.H. and Q.R.W.'s work is supported by the National Science Foundation (NSF) under Grant Nos. 2125326, 2114197, 2228533, and 2402438, as well as by the Northeastern University iSUPER Impact Engine. Any opinions, findings, conclusions, or recommendations expressed in the paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 1, 8

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415, 2022. 8

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[4] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, 2025. 8

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. 1

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 8, 1

[7] Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025. 8

[8] Danielle S. Bassett and Edward T. Bullmore. Small-world brain networks revisited. *The Neuroscientist*, 23(5):499–516, 2017. PMID: 27655008. 1

[9] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. 1

[10] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*, 2024. 3, 1

[11] Yongqiang Chen, Yatao Bian, Bo Han, and James Cheng. How interpretable are interpretable graph neural networks? *arXiv preprint arXiv:2406.07955*, 2024. 8

[12] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. 1, 8

[13] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024. 1, 8

[14] Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7: 1220–1234, 2025. 2

[15] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html. 1

[16] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 3, 1

[17] Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022. 8

[18] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025. 3, 1

[19] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6): 2035–2040, 2009. 1

[20] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024. 1

[21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3, 1

[22] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017. 3, 1

[23] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019. 1, 8

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 8

[26] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023. 8

[27] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects, 2019. 6

[28] Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025. 8

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3, 8, 1

[30] Yiheng Liu, Xiaohui Gao, Haiyang Sun, Bao Ge, Tianming Liu, Junwei Han, and Xintao Hu. Brain-inspired exploration of functional networks and key neurons in large language models. *arXiv preprint arXiv:2502.20408*, 2025. 8

[31] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022. 8

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 4, 1

[33] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 1

[34] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in. 1

[35] Ben Piazza, Dániel L Barabási, André Ferreira Castro, Giulia Menichetti, and Albert-László Barabási. Physical network constraints define the lognormal architecture of the brain's connectome. *bioRxiv*, pages 2025–02, 2025. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 1

[37] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. 8

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 8

[39] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: A structural description of the human brain. *PLOS Computational Biology*, 1(4):null, 2005. 1

[40] Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-interpret: an interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*, 2024. 1, 8

[41] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 1

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019. 1

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 7

[44] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *CoRR*, abs/1905.09418, 2019. 1

[45] Xiongye Xiao, Chenyu Zhou, Heng Ping, Defu Cao, Yaxing Li, Yizhuo Zhou, Shixuan Li, and Paul Bogdan. Exploring neuron interactions and emergence in llms: From the multifractal analysis perspective. *CoRR*, 2024. 8

[46] Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*, 2024. 1

[47] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline

multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024. 3, 1

[48] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025. 3, 1

[49] Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers. *arXiv preprint arXiv:2305.18390*, 2023. 1, 8

[50] Yu Zheng, Yuan Yuan, Yue Zhuo, Yong Li, Gabriel Kreiman, Tomaso Poggio, and Paolo Santi. Probing neural topology of large language models. *arXiv preprint arXiv:2506.01042*, 2025. 1

[51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1

[52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3, 8, 1

[53] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010. 4, 1