Sparse Neuron Ablation Triggers Catastrophic Collapse of the Language Core in Large Vision-Language Models

Cen Lu [1 2] **Yung-Chen Tang** [1 2] **Andrea Cavallaro** [1]

Abstract

Large Vision-Language Models (LVLMs) have shown impressive multimodal understanding capabilities, yet the structures that sustain their functionality remain poorly understood from a mechanistic interpretability standpoint. We propose Consistently Activated Neurons (CAN), a progressive neuron ablation method to identify critical neurons whose removal triggers catastrophic collapse, and use it to investigate structural vulnerabilities in representative 7B LVLMs. Experiments reveal that catastrophic collapse can be triggered by ablating as few as four neurons in `LLaVA-1.5-7b-hf` and a few thousand in `InstructBLIP-vicuna-7b`, both representing a small fraction of model parameters. Notably, critical neurons are predominantly localized in the language model, particularly in its down-projection layer, rather than in the vision components. We also observe a consistent two-stage collapse pattern: initial expressive degradation followed by sudden, complete collapse. These findings reveal that LVLM functionality depends on a sparse subset of neurons concentrated in the language backbone, offering mechanistic insights into how their functionality is structured and where these models are most vulnerable.

1. Introduction

Large Vision-Language Models (LVLMs) (e.g. LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023)) integrate visual perception with language to generate image descriptions or answer (visual) questions. Since these models are increasingly integrated into real-world applications (Ye & Tang, 2025; Zhou et al., 2024), assessing their robustness limits becomes important.

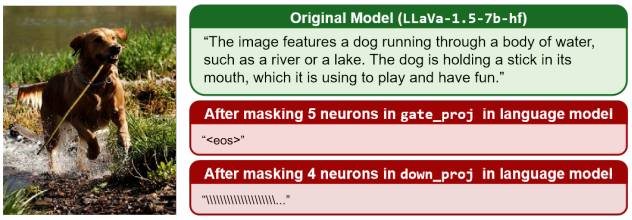


Figure 1. Given the prompt `Describe the object in this image`, the original `LLaVA-1.5-7b-hf` generates correct outputs (green box), but masking only 4–5 neurons in the language model causes catastrophic collapse (red boxes).

Neuron-level analyses revealed that transformer models contain task-specific or language-specific neurons: certain neurons activate predominantly when processing syntax-related tasks, while others respond more strongly to specific languages (AlKhamissi et al., 2024; Tang et al., 2024). Recent research on Large Language Models (LLMs) has also found *super weights* that are decisive for model's functionality (Qin et al., 2025; Yu et al., 2025). A tiny fraction of neurons can be crucial for safety mechanisms in LLMs: safety neurons have been identified that are responsible for model alignment and refusal behaviors, such as rejecting malicious prompts (Chen et al., 2024; Zhao et al., 2025). Inspired by neuroscience, researchers have developed interventional methods, such as neuron ablation (AlKhamissi et al., 2024), that introduce *virtual lesions* to identify critical components (Li & Janson, 2024; Meyes et al., 2019). Recent work has extended mechanistic understanding to multimodal settings, such as identifying domain-specific neurons specialized for medical images or document understanding (Huo et al., 2024). Manipulating these domain-specific neurons causes only modest performance changes on these domains (Huo et al., 2024). However, existing works have largely focused on unimodal LLMs and overlooked the robustness limits of LVLMs from a mechanistic interpretability standpoint, for example the identification of subsets of neural units whose ablation triggers severe performance degradation or complete catastrophic collapse. This problem is important given the deployment of LVLMs

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland [2]Idiap Research Institute, Martigny, Switzerland. Correspondence to: Cen Lu <cen.lu@epfl.ch>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

in safety-critical applications such as autonomous driving (Zhou et al., 2024) and medical diagnosis (Li et al., 2025b).

To characterize these structural vulnerabilities, we define *catastrophic collapse* as a failure state in which the model's output distribution over its vocabulary becomes severely distorted, preventing the generation of any meaningful text across any valid input. Ablating a tiny subset of neurons (as few as four in `LLaVA-1.5-7b-hf`) in the language model can trigger catastrophic collapse, leading to degeneration such as repetitive or empty outputs. We refer to such neurons as *critical neurons*. While the number of neurons required varies across object categories and architectures, the underlying cause is consistent: removing only a tiny fraction of neurons suffices to collapse the tested models (Figure 1).

In summary, our main contributions are:

- We propose the Consistently Activated Neurons (CAN) method as an empirical tool for identifying candidate critical neurons in 7B LVLMs across different images. Analyzing `LLaVA-1.5-7b-hf` and `InstructBLIP-vicuna-7b`, we generate component-specific rankings across vision encoder, cross-modal alignment module, and language model, and object-specific rankings, revealing neurons with the strongest activation patterns.
- Leveraging CAN, we identify a very small subset of neurons that are disproportionately important for the model's output. For example, within our evaluation scope, masking only four neurons in `LLaVA-1.5-7b-hf` suffices to collapse the model, showing structural vulnerability at the 7B parameter scale.
- We localize critical neurons to the language model's feed-forward networks, showing that masking language-model neurons causes catastrophic collapse while vision-component ablations yield only degradation in visual understanding. This contrast reveals the language backbone as the primary structural bottleneck in LVLMs, providing a mechanistic interpretability perspective on multimodal vulnerability.

2. Related work

Neuron ablation is a powerful technique for understanding the internal mechanisms and functional specialization within neural networks (AlKhamissi et al., 2024). For example, the *neural erosion* method (Alexos et al., 2024) applies synaptic pruning and neuron deactivation in LLMs, showing that LLMs lose cognitive abilities in a hierarchical manner: mathematical reasoning first, followed by linguistic capabilities. By replacing neuron activations with their mean values across samples, researchers have identified specialized entropy neurons and token frequency neurons that impact model confidence rather than direct predictions, revealing calibration circuits across architectures (Stolfo et al., 2024). Additionally, recent work applies neuron ablation to evaluate role-playing prompts in medical LLMs, showing that different role assignments activate similar groups of neurons and affect only linguistic features rather than core reasoning (Liang et al., 2025). These studies focus on identifying functional neurons and overlook the ablation thresholds that may lead to catastrophic collapse.

While traditional neuron ablation identifies neurons critical for general model performance, *safety neuron identification* targets neurons responsible for safety alignment, ensuring models reject malicious prompts and generate safe responses. Contrastive methods to identify safety neurons in LLMs showed that these neurons are both sparse and effective, with only 5% of neurons responsible for 90% of safety performance (Chen et al., 2024). Safety neurons constitute fewer than 1% of all parameters and reside in self-attention layers (Zhao et al., 2025). Complementing neuron-level analyses, Zhou et al. (Zhou et al., 2025) showed that ablating safety-related attention heads degrades safety capabilities. While these safety-focused methods rely on contrastive activation patterns between aligned and unaligned models, we develop a neuroscience-inspired approach to identify critical neurons in LVLMs through activation-gradient ranking to investigate failure risks, as illustrated in Section 4.

3. Problem formulation

Let $\mathcal{M}$ be an LVLM that processes a visual input $\mathbf{v}$ and a textual prompt $\mathbf{t}$ to generate $\mathcal{M}(\mathbf{v}, \mathbf{t})$, the output response. Let $\mathcal{N} = \{n_{\ell,j} \mid \ell \in \mathcal{L}, j \in \mathcal{J}_\ell\}$ denote the set of neurons across all layers $\mathcal{L}$ and indices $\mathcal{J}_\ell$, where ℓ is the layer index and j is the position index. Given a dataset $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{t})\}_{i=1}^{M}$ containing M vision-language pairs with a fixed textual prompt, our objective is to identify a critical neuron subset $\mathcal{C}^* \subset \mathcal{N}$, with $|\mathcal{C}^*| \ll |\mathcal{N}|$ that satisfies:

$$\mathcal{C}^* = \arg\min_{\mathcal{C} \subset \mathcal{N}} |\mathcal{C}| \quad \text{s.t.} \quad \frac{1}{M} \sum_{i=1}^{M} \Delta_{\text{perf}}(\mathcal{M}|\mathcal{C}, \mathbf{v}_i, \mathbf{t}) \geq \tau, \tag{1}$$

where $\Delta_{\text{perf}}(\mathcal{M}|\mathcal{C}, \mathbf{v}_i, \mathbf{t})$ quantifies performance degradation of model $\mathcal{M}$ when neurons in $\mathcal{C}$ are ablated on input $(\mathbf{v}_i, \mathbf{t})$, and τ is a threshold indicating catastrophic failure. We define τ through joint criteria based on perplexity (Bahl et al., 1983) and CLIP score (Hessel et al., 2021), as described in Section 5.

4. CAN: Consistently Activated Neurons

To identify critical neurons in LVLMs, we draw inspiration from lesion-function analyses in neuroscience (Bates

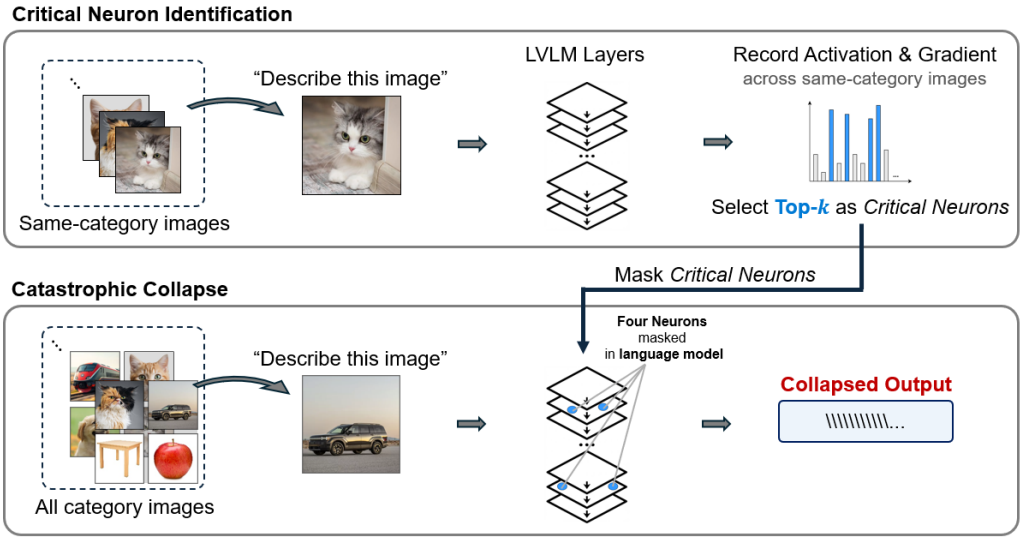


Figure 2. CAN first ranks neurons by importance combining activation and gradient magnitudes. Next, masking the top-k critical neurons causes catastrophic collapse, with the generation of meaningless outputs regardless of the input.

et al., 2003). *Lesion-symptom mapping* establishes functional specificity by associating selective cognitive impairments with localized brain damage (Kimberg et al., 2007; Meyer et al., 2016). Translating this principle to neural networks, we hypothesize that LVLMs contain sparse subsets of neurons that are disproportionately critical for a broad range of tasks (Qin et al., 2025; Yu et al., 2025). Specifically, we employ neuron ablation techniques to validate the importance of identified neurons, aligning with neuroscience practices such as *virtual lesioning* (e.g., Transcranial Magnetic Stimulation (Hallett, 2007)). Furthermore, we use random neuron ablation as a control to mimic the practice of comparing targeted lesions to random ones in neuroscience experiments (Rorden & Karnath, 2004).

Inspired by activation-based saliency (Simonyan et al., 2014) and gradient-based pruning criteria (Lee et al., 2019; Lin et al., 2018), we propose CAN. Unlike these methods, which target performance-preserving pruning in unimodal settings, CAN is designed to identify collapse-triggering neurons across the multimodal components of LVLMs. CAN quantifies neuron importance through *activation magnitude* and *gradient sensitivity*. The activation-based importance contribution focuses on how strongly neurons respond to inputs (forward pass). The gradient-based importance contribution focuses on how changes to neurons would affect the loss, revealing their impact on model predictions (backward pass). The method is explained in Figure 2.

For our localization protocol, we extract activation profiles across all layers during forward propagation. For each neuron at layer ℓ and index j, we record its activation values across all samples in $\mathcal{D}_{\text{cal}}$, the calibration set used for neuron importance scoring:

$$\mathbf{A}_{\ell,j} = \{a_{\ell,j}(\mathbf{v}_i, \mathbf{t}) \mid (\mathbf{v}_i, \mathbf{t}) \in \mathcal{D}_{\text{cal}}\}, \tag{2}$$

where $|\mathbf{A}_{\ell,j}| = M$ and $a_{\ell,j} \in \mathbb{R}$.

To quantify neuron importance, we adopt a scoring mechanism combining activation magnitude with gradient sensitivity, related to criteria in pruning literature (Molchanov et al., 2017). For each neuron at layer ℓ and index j, we compute:

$$\mathcal{I}_{\ell,j} = \frac{1}{M} \sum_{(\mathbf{v}_i,\mathbf{t}) \in \mathcal{D}_{\text{cal}}} \left| \frac{\partial \mathcal{L}(\mathbf{v}_i, \mathbf{t})}{\partial a_{\ell,j}(\mathbf{v}_i, \mathbf{t})} \right|^{\alpha} \cdot |a_{\ell,j}(\mathbf{v}_i, \mathbf{t})|, \tag{3}$$

where $\frac{\partial \mathcal{L}}{\partial a_{\ell,j}}$ represents the gradient sensitivity of the loss with respect to neuron activation, and the hyperparameter $\alpha \in [0, 1]$ controls the balance between the gradient term and the activation magnitude. By selectively combining the magnitude and gradient signals through α, we obtain a flexible measure of neuron criticality across LVLMs.

Next, we rank all neurons by their importance scores in descending order. The top-k neurons are then selected as critical neurons:

$$\mathcal{C}_k = \{n_{\ell,j} : \mathcal{I}_{\ell,j} \text{ ranks in top-}k \text{ globally}\}, \tag{4}$$

where the hyperparameter k is determined through iterative masking by increasing k until a catastrophic performance degradation occurs. This identifies the number of critical neurons k^* that suffices to trigger catastrophic collapse for each specific model and component.

5. Assessing performance degradation

We evaluate performance degradation using perplexity (Bahl et al., 1983), to assess the language modeling capability, and CLIP Score (Hessel et al., 2021), to verify image-text alignment. Perplexity measures a model's ability to maintain probability distribution over the original model's output, evaluating whether the masked model can still follow the base model's linguistic patterns. The CLIP score measures semantic alignment between generated text and visual content. A low CLIP score with stable perplexity can arise from two distinct failure modes, namely *perceptual failure* and *expressive degradation*. With a *perceptual failure* the model generates fluent but irrelevant descriptions (e.g., describing a cat when shown a car), indicating damage of vision neurons and loss of visual grounding, while preserving linguistic capabilities. With an *expressive degradation* the model outputs malformed text, suggesting damage to the language generation component.

We measure the *perplexity degradation* on a validation set $\mathcal{D}_{\text{val}} = \{(\mathbf{v}_i, \mathbf{t})\}_{i=1}^{M_{\text{val}}}$, which consists of validation image-prompt pairs. For each input pair $(\mathbf{v}_i, \mathbf{t}) \in \mathcal{D}_{\text{val}}$ and masking configuration $\mathcal{C}_k$, the model generates a token sequence $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_{L^{(i)}}^{(i)})$ of length $L^{(i)}$. The perplexity is:

$$\text{PPL}(\mathbf{y}^{(i)}|\mathbf{v}_i, \mathbf{t}) = \exp\left(-\frac{1}{L^{(i)}} \sum_{j=1}^{L^{(i)}} \log p(y_j^{(i)}|y_{1:j-1}^{(i)}, \mathbf{v}_i, \mathbf{t})\right). \tag{5}$$

We measure the incremental log-ratio of perplexities to quantify the sudden deviation when masking additional neurons:

$$\Delta_{\text{PPL}}(\mathcal{C}_k) = \frac{1}{M_{\text{val}}} \sum_{i=1}^{M_{\text{val}}} \log_{10}\left(\frac{\text{PPL}(\mathcal{M}|\mathcal{C}_k, \mathbf{v}_i, \mathbf{t})}{\text{PPL}(\mathcal{M}|\mathcal{C}_{k-\Delta k}, \mathbf{v}_i, \mathbf{t})}\right), \tag{6}$$

where $\text{PPL}(\mathcal{M}|\mathcal{C}_k, \mathbf{v}_i, \mathbf{t})$ is the perplexity of model $\mathcal{M}$ with neurons in $\mathcal{C}_k$ masked, $\text{PPL}(\mathcal{M}|\mathcal{C}_{k-\Delta k}, \mathbf{v}_i, \mathbf{t})$ is the perplexity at the previous masking step, and Δk is the masking step size. We use log-scale to capture order-of-magnitude changes in perplexity. This metric evaluates the incremental degradation in language modeling capability: larger values indicate a sudden collapse when masking the current set of neurons compared to the previous step.

The CLIP score (Hessel et al., 2021) measures semantic coherence between visual inputs and generated descriptions. We compute the CLIP score on the validation set:

$$\text{CLIP}(\mathcal{C}_k) = \frac{1}{M_{\text{val}}} \sum_{i=1}^{M_{\text{val}}} \text{CLIP}(\mathbf{v}_i, (\mathcal{M}|\mathcal{C}_k, \mathbf{v}_i, \mathbf{t})), \tag{7}$$

where $\text{CLIP}(\mathbf{v}_i, (\mathcal{M}|\mathcal{C}_k, \mathbf{v}_i, \mathbf{t}))$ is the cosine similarity between the CLIP embedding of image $\mathbf{v}_i$ and the text generated by model $\mathcal{M}$ with neurons in $\mathcal{C}_k$ masked on input $(\mathbf{v}_i, \mathbf{t})$. The lower the CLIP score, the worse the semantic alignment between the visual input and the generated text.

We determine the critical neuron set size k in a data-driven manner by greedy search with step size Δk. We increase k until catastrophic failure occurs:

$$k^* = \min\{k : \Delta_{\text{PPL}}(\mathcal{C}_k) \geq \tau_{\text{PPL}} \text{ and } \text{CLIP}(\mathcal{C}_k) \leq \tau_{\text{CLIP}}\}, \tag{8}$$

where τ_{PPL} and τ_{CLIP} are failure thresholds. A *complete collapse* occurs when $\Delta_{\text{PPL}} \geq \tau_{\text{PPL}}$ and $\text{CLIP} \leq \tau_{\text{CLIP}}$, illustrating catastrophic collapse of core model capabilities. We set $\tau_{\text{PPL}} = 1$ (indicating one order of magnitude degradation) and $\tau_{\text{CLIP}} = 22$ based on empirical failure patterns (see details in Section 6.4).

6. Experiments

Architecture-Specific Neuron Localization. We adapt the neuron identification to the specific components of LLaVA (Liu et al., 2024) and InstructBLIP (Dai et al., 2023), as CAN is applicable across different LVLM architectures. For LLaVA, we identify neurons in: vision encoder, multimodal projector, and language model. For InstructBLIP: vision encoder, Q-Former with dual attention over learnable queries, and language model. In Eq. (3), we empirically set $\alpha = 1$ for InstructBLIP since gradient information is essential due to the Q-Former's complex attention routing, and $\alpha = 0$ for LLaVA, as its simpler projector architecture makes activation magnitude alone a stable and sufficient signal for critical neuron identification.

Experimental setup. We conduct experiments with `InstructBLIP-vicuna-7b` and `LLaVA-1.5-7b-hf` on a dataset $\mathcal{D}_{\text{cal}}$ of 300 images for each of 10 object categories from the `ILSVRC2013` dataset (Deng et al., 2009). All experiments were executed on an H100 GPU using fp32 precision. To generate ranked neuron lists, we use images from the 10 categories with `Describe the object in this image` as prompt $\mathbf{t}$. We found this simple prompt to be effective for critical neuron identification after testing different prompts, including more detailed instructions. We record the activation magnitude of each neuron across all model components after generating the first token, then rank neurons using CAN across all image categories. For each object category, we evaluate the progressively masked model using the additional 100 test images from the same category as $\mathcal{D}_{\text{val}}$. In the following, we address the minimum number of critical neurons, their stability across inputs, the effect of masking position, and their localization.

Table 1. Comparison of catastrophic collapse thresholds between `LLaVA-1.5-7b-hf` and `InstructBLIP-vicuna-7b` across different object categories by ablating neurons in `gate_proj`. Each row shows results averaged over 100 test images. Progressive steps of 100 are used for the latter to find an approximate collapse threshold in `InstructBLIP-vicuna-7b`.

Object	LLaVA-1.5-7b-hf			InstructBLIP-vicuna-7b		
	Neurons (#)	PPL Change (↓) (Orig. → Masked, **Change**†)	CLIP Score Change (↑) (Orig. → Masked, **% Change**†)	Neurons (#)	PPL Change (↓) (Orig. → Masked, **Change**†)	CLIP Score Change (↑) (Orig. → Masked, **% Change**†)
dog	5	2.20 → 8.82×10^3 (×**4009**)	30.35 → NaN‡ (**-100%**)	3100	1.75 → 1.38×10^2 (×**79**)	31.44 → 20.02 (**-36%**)
cat	5	2.54 → 7.34×10^3 (×**2890**)	31.44 → NaN‡ (**-100%**)	2700	1.68 → 1.15×10^2 (×**69**)	29.95 → 21.05 (**-30%**)
car	5	2.26 → 8.56×10^3 (×**3788**)	30.80 → NaN‡ (**-100%**)	1200	2.02 → 7.59×10^1 (×**38**)	29.58 → 19.04 (**-36%**)
pencil box	5	3.25 → 7.21×10^3 (×**2218**)	29.52 → NaN‡ (**-100%**)	2000	1.74 → 9.06×10^1 (×**52**)	29.68 → 21.25 (**-28%**)
sofa	5	2.13 → 9.28×10^3 (×**4357**)	31.18 → NaN‡ (**-100%**)	2400	1.75 → 8.60×10^1 (×**49**)	32.67 → 20.28 (**-38%**)
table	5	2.08 → 1.02×10^4 (×**4904**)	29.72 → NaN‡ (**-100%**)	1300	2.06 → 6.31×10^1 (×**31**)	30.99 → 19.41 (**-37%**)
train	5	2.80 → 9.83×10^3 (×**3511**)	28.16 → NaN‡ (**-100%**)	10600	2.34 → 3.55×10^2 (×**152**)	29.74 → 18.94 (**-36%**)
apple	5	2.08 → 9.76×10^3 (×**4692**)	31.64 → NaN‡ (**-100%**)	2300	1.66 → 7.07×10^1 (×**43**)	32.30 → 21.02 (**-35%**)
orange	5	2.38 → 9.99×10^3 (×**4199**)	32.06 → NaN‡ (**-100%**)	1900	1.65 → 7.41×10^1 (×**45**)	31.49 → 20.96 (**-33%**)
cup	5	2.99 → 7.62×10^3 (×**2549**)	30.46 → NaN‡ (**-100%**)	1400	1.64 → 1.20×10^2 (×**73**)	32.21 → 20.04 (**-38%**)

†Orig. and Masked denote metrics from the original model and after masking critical neurons at collapse.
‡NaN occurs when the model outputs only end-of-sentence tokens, making CLIP score computation infeasible. For % Change calculation, NaN is treated as 0 (i.e., -100%).

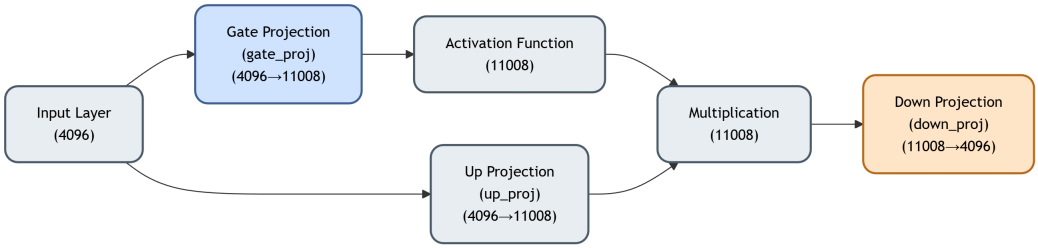


Figure 3. Architecture of the Feed-Forward Network (FFN) module in the transformer layer for `LLaVA-1.5-7b-hf` and `InstructBLIP-vicuna-7b`. Numbers in parentheses indicate tensor dimensions. The identification and masking of critical neurons in the blue component (`gate_proj`) and the orange component (`down_proj`) are discussed in Section 6.1 and Section 6.3, respectively.

6.1. Number of neurons for catastrophic collapse

We progressively masked critical parts in `gate_proj` (see Figure 3) from the feed-forward network (FFN) module in the language model of LVLMs. Our results on `LLaVA-1.5-7b-hf` show a consistent pattern of catastrophic collapse across different object categories. As shown in Table 1, `LLaVA-1.5-7b-hf` requires only five neurons to trigger catastrophic failure across all 10 object categories, with perplexity increases exceeding two orders of magnitude and CLIP scores becoming undefined (NaN) as the model outputs only end-of-sentence tokens, making CLIP score computation infeasible. These five neurons remain consistent regardless of the object category, with three neurons located in layer 1 and two neurons in layer 30 (out of 32 layers, zero-indexed), suggesting that neurons at the beginning and end of `LLaVA-1.5-7b-hf` may play critical roles in maintaining model stability. In contrast, `InstructBLIP-vicuna-7b` is more robust as it requires at least 1200 neurons (still representing an extremely small fraction of 0.17×10^{-6} of the model parameters), depending on the object category. Notably, once catastrophic collapse is triggered, it occurs regardless of whether the input is pure text or image–text pairs.

6.2. Stability

Figure 4 illustrates the collapse progression for both models on the car category. Despite different thresholds, both models exhibit similar two-stage patterns, with an expressive degradation followed by a complete collapse.

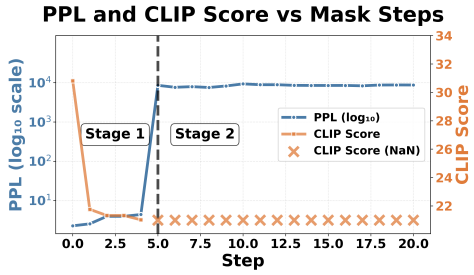
Stage 1: expressive degradation. For `LLaVA-1.5-7b-hf` (Figure 4a, Steps 0–4), CLIP scores gradually decrease from 31 to 21 while perplexity remains stable, indicating declining output quality with preserved comprehension, similar to *expressive aphasia* (Simmons-Mackie et al., 2005) in human cognition. For `InstructBLIP-vicuna-7b` (Figure 4b, Steps 0–1100), CLIP scores drop from 30 to 20 with modest perplexity increases, showing a similar degradation pattern in generation capability. Although the outputs at this stage have already degenerated and contain unintelligible specific tokens (e.g., "archivi", "sierp 2013"), the model still retains most of its original comprehension ability.

Stage 2: complete collapse. At the critical threshold (Step 5 for LLaVA, Step 1200 for InstructBLIP), both models ex-

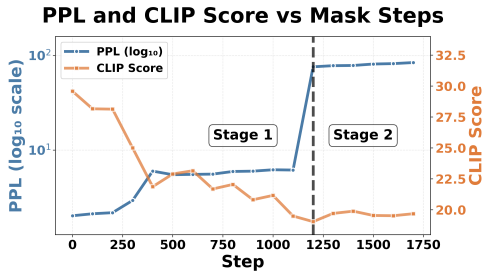
Table 2. Collapse induced by masking neurons in the `down_proj` component of `LLaVA-1.5-7b-hf` and `InstructBLIP-vicuna-7b`. Progressive steps of 100 are used for the latter to find an approximate collapse threshold. The collapse manifests as perplexity increases of at least three orders of magnitude and a significant drop in CLIP Score to below 22.

Object	LLaVA-1.5-7b-hf			InstructBLIP-vicuna-7b		
	Neurons (#)	PPL Change (↓) (Orig. → Masked, **Change**†)	CLIP Score Change (↑) (Orig. → Masked, **% Change**†)	Neurons (#)	PPL Change (↓) (Orig. → Masked, **Change**†)	CLIP Score Change (↑) (Orig. → Masked, **% Change**†)
dog	4	2.20 → 1.44×10^4 (×**6545**)	30.35 → 20.26 (**-33%**)	2400	1.75 → 3.78×10^3 (×**2160**)	31.44 → 20.09 (**-36%**)
cat	4	2.54 → 1.21×10^4 (×**4764**)	31.44 → 20.53 (**-35%**)	2500	1.68 → 2.93×10^3 (×**1744**)	29.95 → 21.61 (**-28%**)
car	4	2.26 → 1.53×10^4 (×**6770**)	30.80 → 20.41 (**-34%**)	1000	2.02 → 4.65×10^3 (×**2302**)	29.58 → 18.98 (**-36%**)
pencil box	4	3.25 → 1.61×10^4 (×**4954**)	29.52 → 19.48 (**-34%**)	1300	1.74 → 9.01×10^2 (×**518**)	29.68 → 19.81 (**-33%**)
sofa	4	2.13 → 1.23×10^4 (×**5775**)	31.18 → 19.58 (**-37%**)	2000	1.75 → 2.34×10^3 (×**1337**)	32.67 → 19.51 (**-40%**)
table	4	2.08 → 1.54×10^4 (×**7404**)	29.72 → 19.25 (**-35%**)	1000	2.06 → 2.34×10^3 (×**1136**)	30.99 → 19.42 (**-37%**)
train	4	2.80 → 2.21×10^4 (×**7893**)	28.16 → 20.48 (**-27%**)	14400	2.34 → 1.08×10^4 (×**4615**)	29.74 → 18.47 (**-38%**)
apple	4	2.08 → 1.59×10^4 (×**7644**)	31.64 → 20.50 (**-35%**)	2100	1.66 → 2.84×10^3 (×**1711**)	32.30 → 20.70 (**-36%**)
orange	4	2.38 → 1.94×10^4 (×**8151**)	32.06 → 21.17 (**-34%**)	2800	1.65 → 2.64×10^3 (×**1604**)	31.49 → 21.70 (**-30%**)
cup	4	2.99 → 1.08×10^4 (×**3612**)	30.46 → 20.52 (**-33%**)	1100	1.64 → 1.87×10^3 (×**1140**)	32.21 → 21.15 (**-34%**)

†Orig. and Masked denote metrics before and after collapse induced by masking critical neurons.



(a) `LLaVA-1.5-7b-hf`



(b) `InstructBLIP-vicuna-7b`

Figure 4. Comparison of PPL and CLIP score changes during progressive masking of neurons in `gate_proj` from FFN using car images. Stage 1 indicates expressive degradation and Stage 2 indicates complete collapse, as explained in Section 6.1.

perience sharp transitions: perplexity increases dramatically while CLIP scores either become NaN or drop to around 19, meaning that the models fail to produce meaningful output (e.g., "and, and ...", "ÄÄÄÄÄÄ..."). Due to the continuous feedback loops inherent in autoregressive generation, once the number of masked neurons exceeds the critical threshold, the accumulation of errors within the residual stream exceeds the network's ability to self-correct. The numerical shift destabilizes the model's internal representations and damages the generation process. Continued masking beyond this threshold produces minimal additional change: both models stabilize in collapsed state and cannot recover their functionality once collapsed.[1]

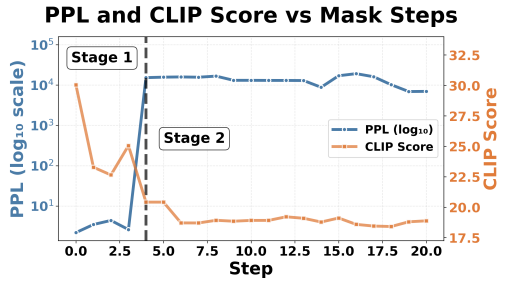
The observation that both models show similar collapse patterns but the number of critical neurons varies across architectures aligns with prior robustness studies: LLaVA achieves 89% attack success rates under jailbreak attacks, while InstructBLIP is more resilient at 33% (Li et al., 2025a).

6.3. Impact of mask position

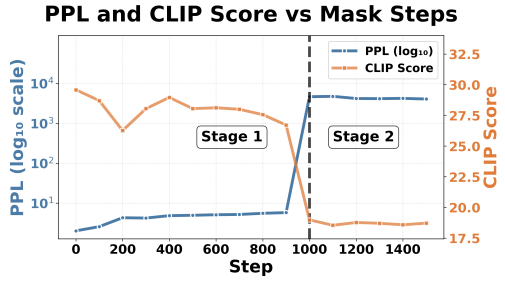
To assess how neuron position influences the number of critical neurons required to trigger catastrophic collapse, we conducted a comparative analysis across two distinct positions in the FFN architecture. In Section 6.1, we examined `gate_proj` (blue component in Figure 3, 11,008 dimensions), which operates on the expanded intermediate representation before activation function, and `down_proj` (orange component in Figure 3, 4,096 dimensions), which performs dimensionality reduction at the network output.

In general, masking neurons in `down_proj` requires fewer critical neurons to induce collapse compared to `gate_proj` (see Table 2). For `LLaVA-1.5-7b-hf`, only four neurons in `down_proj` (compared to five in `gate_proj`) are sufficient to trigger catastrophic failure, with perplexity increases reaching up to four orders of magnitude. These four neurons also remain consistent regardless of object category, with two neurons located in layer 1 and two neurons in layer 30 (out of 32 layers, zero-indexed), further confirming the importance of neurons at the language model's boundaries. This sensitivity can be attributed to

[1]Notably, multiple ablation trials of the same number of randomly selected non-critical neurons in both models consistently produce negligible impact on model performance, confirming the specificity of these critical neurons.

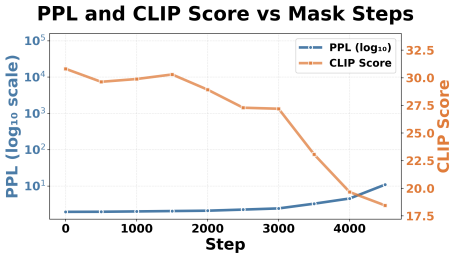


(a) `LLaVA-1.5-7b-hf`

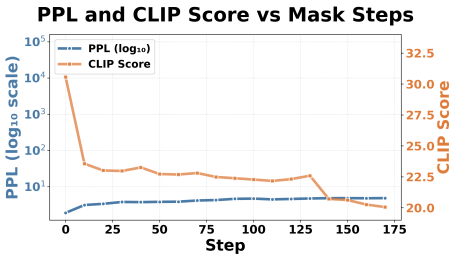


(b) `InstructBLIP-vicuna-7b`

Figure 5. Comparison of PPL and CLIP score changes during masking of neurons in `down_proj` using car images from FFN for the two models. Stage 1 indicates an expressive degradation and Stage 2 indicates a complete collapse.



(a) `LLaVA-1.5-7b-hf`



(b) `InstructBLIP-vicuna-7b`

Figure 6. Comparison of PPL and CLIP score changes using car images during progressive masking of neurons from `multimodal_projector` in `LLaVA-1.5-7b-hf` and from `Q-Former` in `InstructBLIP-vicuna-7b`. In both cases, CLIP Score drops while PPL remains relatively stable, without order-of-magnitude increase.

the dimensionality reduction performed by `down_proj`, which projects the expanded intermediate representation back into the original dimensionality of the residual stream. We hypothesize that the identified critical neurons are indispensable within this projection: when masked, the rank structure of the projection space is disrupted, preventing high-dimensional features from mapping back into the residual space. Consequently, this local perturbation at a critical information bottleneck propagates directly into the primary information flow, yielding higher post-collapse perplexity values than those observed under `gate_proj` masking (Table 1). Similarly, `InstructBLIP-vicuna-7b` shows the same pattern of increased vulnerability in `down_proj`.

Masking critical neurons in `down_proj` for both models results in severe output corruption. For `LLaVA-1.5-7b-hf`, masking `down_proj` produces empty outputs with only end-of-sentence tokens, while masking `gate_proj` generates repetitive character strings, as shown in Figure 1. In contrast, `Instruct-BLIP-vicuna-7b` outputs repetitive tokens after masking both positions, suggesting model-specific vulnerability patterns to neuron masking in FFN components. Figure 5 confirms the same two-stage collapse pattern for `down_proj`, showing a stronger bottleneck than `gate_proj` due to its role in information synthesis.[2]

[2]Repeated ablations of the same number of randomly selected non-critical neurons in both `down_proj` components consistently produce negligible impact, further confirming the specificity of these critical neurons.

6.4. Critical neuron location

We now mask neurons in vision-related components: `multimodal_projector` and vision encoder in `LLaVA-1.5-7b-hf`, and Q-Former and vision encoder in `InstructBLIP-vicuna-7b` (Tables 3 and 4), revealing a distinct failure mode from language model components (Tables 1 and 2): masking vision components drops CLIP scores while perplexity remains stable, whereas masking language FFN neurons causes *complete collapse*. This dissociation suggests that vision component failures result in *perceptual failure*: the model loses visual grounding while retaining its linguistic ability. According to previous research on compositional reasoning and object hallucination, CLIP scores in the 15-22 range represent a "chance-level" baseline where outputs become semantically indistinguishable from random image-text pairings (Hessel et al., 2021; Radford et al., 2021; Yuksekgonul et al., 2023). This threshold is supported by our language model results, where all masked outputs converge to this range despite generating non-meaningful text. We therefore adopt 22 as the CLIP score threshold for *perceptual failure*. For `LLaVA-1.5-7b-hf`, *perceptual failure* requires masking 61–97% of the multimodal projector and 17–30% of the vision encoder, far

Table 3. Ablation results for masking critical neurons in different components of `LLaVA-1.5-7b-hf`. Results are averaged over 100 test images per object category. We use steps of 500 neurons to identify approximate number of neurons in vision-related neurons leading to *perceptual failure*, while PPL remains relatively stable without order-of-magnitude increase.

Object	Multimodal Projector (4096)†			Vision Encoder (98304)†		
	#Neurons (%)‡	PPL Change (↓) (Orig. → Masked§, **Change**)	CLIP Score Change (↑) (Orig. → Masked§, **% Change**)	#Neurons (%)‡	PPL Change (↓) (Orig. → Masked§, **Change**)	CLIP Score Change (↑) (Orig. → Masked§, **% Change**)
dog	3500 (85%)	2.20 → 5.96 (×**3**)	30.35 → 20.95 (**-31%**)	17500 (17%)	2.20 → 7.30 (×**3**)	30.35 → 21.91 (**-28%**)
cat	3500 (85%)	2.54 → 9.83 (×**4**)	31.44 → 21.03 (**-33%**)	18500 (18%)	2.54 → 8.93 (×**4**)	31.44 → 20.17 (**-36%**)
car	4000 (97%)	2.26 → 4.52 (×**2**)	30.80 → 19.65 (**-36%**)	18500 (18%)	2.26 → 4.14 (×**2**)	30.80 → 19.21 (**-38%**)
pencil box	3500 (85%)	3.25 → 4.27 (×**1**)	29.52 → 21.16 (**-28%**)	30000 (30%)	3.25 → 4.46 (×**1**)	29.52 → 21.12 (**-28%**)
sofa	4000 (97%)	2.13 → 8.56 (×**4**)	31.18 → 20.17 (**-35%**)	30000 (30%)	2.13 → 8.13 (×**4**)	31.18 → 19.92 (**-36%**)
table	4000 (97%)	2.08 → 3.34 (×**2**)	29.72 → 21.38 (**-28%**)	19500 (19%)	2.08 → 3.42 (×**2**)	29.72 → 21.68 (**-27%**)
train	2500 (61%)	2.80 → 7.60 (×**3**)	28.16 → 21.11 (**-25%**)	19000 (19%)	2.80 → 4.78 (×**2**)	28.16 → 20.03 (**-29%**)
apple	4000 (97%)	2.08 → 9.38 (×**5**)	31.64 → 21.29 (**-33%**)	30000 (30%)	2.08 → 8.23 (×**4**)	31.64 → 21.40 (**-32%**)
orange	4000 (97%)	2.38 → 4.93 (×**2**)	32.06 → 20.62 (**-36%**)	19000 (19%)	2.38 → 4.60 (×**2**)	32.06 → 21.48 (**-33%**)
cup	4000 (97%)	2.99 → 6.70 (×**2**)	30.46 → 21.06 (**-31%**)	30000 (30%)	2.99 → 5.41 (×**2**)	30.46 → 21.98 (**-28%**)

†Total neuron candidates in each component.
‡Percentages are calculated relative to the total candidates.
§Orig. and Masked denote metrics before and after masking critical neurons.

Table 4. Ablation results for masking critical neurons in different components of `InstructBLIP-vicuna-7b`. Results are averaged over 100 test images per object category. We have used steps of 500 neurons for vision encoder and steps of 10 neurons in Q-Former to identify the number of neurons leading to *perceptual failure*, while PPL remains stable without order-of-magnitude increase.

Object	Q-Former (9216)†			Vision Encoder (239616)†		
	#Neurons (%)‡	PPL Change (↓) (Orig. → Masked§, **Change**)	CLIP Score Change (↑) (Orig. → Masked§, **% Change**)	#Neurons (%)‡	PPL Change (↓) (Orig. → Masked§, **Change**)	CLIP Score Change (↑) (Orig. → Masked§, **% Change**)
dog	70 (0.76%)	1.75 → 4.40 (×**3**)	31.44 → 21.90 (**-30%**)	20000 (8.35%)	1.75 → 3.55 (×**2**)	31.44 → 21.79 (**-31%**)
cat	150 (1.63%)	1.68 → 4.09 (×**2**)	29.95 → 21.81 (**-27%**)	35000 (14.61%)	1.68 → 3.32 (×**2**)	29.95 → 21.38 (**-29%**)
car	140 (1.52%)	2.02 → 4.79 (×**2**)	29.58 → 20.71 (**-30%**)	26500 (11.06%)	2.02 → 3.42 (×**2**)	29.58 → 21.77 (**-26%**)
pencil box	30 (0.33%)	1.74 → 4.82 (×**3**)	29.68 → 22.15 (**-25%**)	29500 (12.31%)	1.74 → 3.62 (×**2**)	29.68 → 21.89 (**-26%**)
sofa	110 (1.19%)	1.75 → 4.24 (×**2**)	32.67 → 21.66 (**-34%**)	31000 (12.94%)	1.75 → 3.41 (×**2**)	32.67 → 21.86 (**-33%**)
table	90 (0.98%)	2.06 → 4.66 (×**2**)	30.99 → 21.97 (**-29%**)	49000 (20.45%)	2.06 → 5.00 (×**2**)	30.99 → 21.93 (**-29%**)
train	80 (0.87%)	2.34 → 6.04 (×**3**)	29.74 → 21.32 (**-28%**)	32500 (13.56%)	2.34 → 3.67 (×**2**)	29.74 → 21.78 (**-27%**)
apple	110 (1.19%)	1.66 → 3.78 (×**2**)	32.30 → 21.11 (**-35%**)	62500 (26.09%)	1.66 → 3.51 (×**2**)	32.30 → 21.75 (**-33%**)
orange	70 (0.76%)	1.65 → 4.49 (×**3**)	31.49 → 20.95 (**-33%**)	38000 (15.86%)	1.65 → 3.85 (×**2**)	31.49 → 21.10 (**-33%**)
cup	50 (0.54%)	1.64 → 4.07 (×**2**)	32.21 → 20.39 (**-37%**)	41500 (17.32%)	1.64 → 3.25 (×**2**)	32.21 → 21.89 (**-32%**)

†Total neuron candidates in each component.
‡Percentages are calculated relative to the total candidates.
§Orig. and Masked denote metrics before and after masking critical neurons.

exceeding the sparse critical neuron rates in language model FFN components. As Figure 6a shows, even when CLIP scores drop below 22, perplexity remains stable. The vision encoder of `InstructBLIP-vicuna-7b` shows similar patterns, requiring 8–26% masking. However, the Q-Former behaves distinctly: only 0.33–1.63% of neurons trigger *perceptual failure*, yet perplexity changes remain minimal (Figure 6b). This suggests that vision-related components function as perceptual pathways rather than core reasoning components: the masking causes visual impairment while preserving the model's fundamental capabilities. In contrast, masking critical neurons in the language model's FFN components causes *complete collapse* (Tables 1 and 2): simultaneous perplexity explosion and severe CLIP score degradation indicate catastrophic failure in both comprehension and generation.

7. Conclusion

We investigated the vulnerabilities of LVLMs through neuron-level ablation studies. We proposed CAN, a neuron identification method combining activation magnitude with gradient sensitivity, and showed that masking as few as four neurons suffices to collapse a 7B LVLM. We identified a consistent two-stage collapse pattern and localized the vulnerability bottleneck to the language model's feed-forward networks rather than vision components. This finding challenges the intuition that failures in multimodal tasks should stem from vision-language alignment components (Liang et al., 2022), revealing instead that the language backbone constitutes the primary structural vulnerability in LVLMs. From a mechanistic interpretability standpoint, our study exposes latent structural vulnerabilities in LVLMs, contributing empirical grounding for safety-aware interpretability research in multimodal models. Future work will extend our analysis to larger models and diverse architectures, and de-

sign defense strategies such as redundant neuron pathways.

8. Limitation

While our findings reveal that LVLM functionality depends on a sparse set of neurons concentrated in the language backbone, we acknowledge several limitations that present opportunities for future research. First, our investigation is primarily empirical. Our CAN method identifies a sufficient set of neurons that can trigger the collapse, but it does not guarantee that the discovered set is minimal and could be data-dependent. Second, conducting mechanistic analysis on LVLMs presents greater challenges compared to unimodal LLMs. The lack of architectural standardization across multimodal models, such as diverse vision encoders and varying cross-modal projectors, renders code implementation complex and model-dependent. Finally, due to resource constraints, our experiments were restricted to the 7B parameter scale. We welcome future studies to scale these ablation studies to larger models to verify whether massive parameter scaling (e.g. 70B) impacts the fragility of the language core in LVLMs.

9. Acknowledgements

This research was carried out as part of the Horizon Europe alignAI project (grant agreement ID: 101169473) and funded by the Swiss State Secretariat for Education, Research and Innovation (contract number: 25.00046).

References

Alexos, A., Tsai, Y.-D., Domingo, I., Pishgar, M., and Baldi, P. Neural erosion: Emulating controlled neurodegeneration and aging in AI systems, 2024. URL https://arxiv.org/abs/2403.10596.

AlKhamissi, B., Tuckute, G., Bosselut, A., and Schrimpf, M. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. 2024. URL https://arxiv.org/abs/2411.02280.

Bahl, L. R., Jelinek, F., and Mercer, R. L. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, 1983. doi: 10.1109/TPAMI.1983.4767370.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., and Dronkers, N. F. Voxel-based lesion–symptom mapping. *Nature neuroscience*, 6(5): 448–450, 2003.

Chen, J., Wang, X., Yao, Z., Bai, Y., Hou, L., and Li, J. Finding safety neurons in large language models. *CoRR*, abs/2406.14144, 2024. doi: 10.48550/ARXIV.2406.14144. URL https://doi.org/10.48550/arXiv.2406.14144.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. C. H. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Hallett, M. Transcranial magnetic stimulation: a primer. *Neuron*, 55(2):187–199, 2007.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Huo, J., Yan, Y., Hu, B., Yue, Y., and Hu, X. MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6816, Miami, Florida, USA, November 2024. doi: 10.18653/v1/2024.emnlp-main.387. URL https://aclanthology.org/2024.emnlp-main.387/.

Kimberg, D. Y., Coslett, H. B., and Schwartz, M. F. Power in voxel-based lesion-symptom mapping. *Journal of cognitive neuroscience*, 19(7):1067–1080, 2007.

Lee, N., Ajanthan, T., and Torr, P. H. S. Snip: Single-shot network pruning based on connection sensitivity, 2019. URL https://arxiv.org/abs/1810.02340.

Li, C., Wang, H., and Fang, Y. Attack as defense: Safeguarding large vision-language models from jailbreaking by adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20138–20152, 2025a.

Li, M. and Janson, L. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems 38*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c55e6792923cc16fd6ed5c3f672420a5-Abstract-Conference.html.

Li, X., Li, L., Jiang, Y., Wang, H., Qiao, X., Feng, T., Luo, H., and Zhao, Y. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118:102995, 2025b. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.102995. URL https://www.sciencedirect.com/science/article/pii/S1566253525000685.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

Liang, X., Lai, H., Wang, H., Zhang, W., Zhang, L., Chen, Y., Xiong, F., and Li, Z. Dissecting role cognition in medical LLMs via neuronal ablation, 2025. URL https://arxiv.org/abs/2510.24677.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484.

Meyer, S., Kessner, S. S., Cheng, B., Bönstrup, M., Schulz, R., Hummel, F. C., De Bruyn, N., Peeters, A., Van Pesch, V., Duprez, T., et al. Voxel-based lesion-symptom mapping of stroke lesions underlying somatosensory deficits. *NeuroImage: Clinical*, 10:257–266, 2016.

Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. Ablation studies in artificial neural networks. *CoRR*, abs/1901.08644, 2019. URL http://arxiv.org/abs/1901.08644.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJGCiw5gl.

Qin, Z., Lyu, K., Yu, Q., Sun, Y., and Fan, Z. The Achilles' heel of LLMs: How altering a handful of neurons can cripple language abilities, 2025. URL https://arxiv.org/abs/2510.10238.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rorden, C. and Karnath, H.-O. Using human brain lesions to infer function: a relic from a past era in the fMRI age? *Nat. Rev. Neurosci.*, 5(10):812–819, 2004.

Simmons-Mackie, N., Threats, T. T., and Kagan, A. Outcome assessment in aphasia: A survey. *Journal of communication disorders*, 38(1):1–27, 2005.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Stolfo, A., Wu, B., Gurnee, W., Belinkov, Y., Song, X., Sachan, M., and Nanda, N. Confidence regulation neurons in language models, 2024. URL https://arxiv.org/abs/2406.16254.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. Language-specific neurons: The key to multilingual capabilities in large language models, 2024. URL https://arxiv.org/abs/2402.16438.

Ye, J. and Tang, H. Multimodal large language models for medicine: A comprehensive survey, 2025. URL https://arxiv.org/abs/2504.21051.

Yu, M., Wang, D., Shan, Q., Reed, C. J., and Wan, A. The super weight in large language models, 2025. URL https://arxiv.org/abs/2411.07191.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Zhao, Y., Zhang, W., Xie, Y., Goyal, A., Kawaguchi, K., and Shieh, M. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025. URL https://openreview.net/forum?id=yR47RmND1m.

Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., and Knoll, A. C. Vision language models in autonomous driving: A survey and outlook, 2024. URL https://arxiv.org/abs/2310.14414.

Zhou, Z., Yu, H., Zhang, X., Xu, R., Huang, F., Wang, K., Liu, Y., Fang, J., and Li, Y. On the role of attention

heads in large language model safety. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025. URL https://openreview.net/forum?id=h0Ak8A5yqw.