Blavatnik School of Computer Science and AI
Raymond & Beverly Sackler
Faculty of Exact Sciences
Tel Aviv University

Sagol School of Neuroscience
Tel Aviv University

Mental Imagery in Vision-Language Models

Nimrod Lahav

October 2025

Abstract

This report investigates how neurons in vision-language transformer models, such as LLaVA, respond differently to images and text. Using MSCOCO paired data, we extract neuron activations and apply statistical analysis to quantify the preference of modality, providing insight into bias and specialization of artificial neurons. Visualizations of the distributions between layers support the analysis.

1 Introduction

Vision-language models (VLMs) such as LLaVA [5] combine natural language understanding [8] with visual perception [2]. Understanding how their internal neurons respond to different modalities can shed light on whether the model develops representations like *mental imagery* or modality-specific biases. Inspired by neuroscience concepts - mainly the Imagery Debate [3][6] - but also recent interpretation attempts in the field of NLP in Large Language Models, we analyze the activations of LLaVA across thousands of image-text pairs from the MSCOCO dataset [4].

We ask: *Do some neurons show stronger responses to images or to text? How are such preferences distributed across layers?*

2 Method

2.1 Data Extraction

We began by extracting a random* subset of image–caption pairs from the MSCOCO dataset [4]. Each pair consists of a natural image and its corresponding human-written caption, which served as the two modalities of interest (vision and language).

*A list of thousands of images (and matching captions) was created, shuffled once to maintain randomness and reused across activations recording.

2.2 Model Modification

As the LLaVA architecture includes LLaMA [8] and CLIP [7] - as a language model and vision encoder, respectively - the text inputs contain roughly T words, and the image is divided into $T = 576$ patches. It also includes a projection layer that "translates" the visual input into a tensor with the same shape as a prompt encoding. Therefore, both modality inputs are represented as sequences of length T with embedding dimension D, following the patch embedding strategy of Vision Transformers [2].
The canonical LLaVA [5] is a multimodal model, inputting prompts and images. Therefore, a prerequisite of making a valid claim about neuron specialty is removing of confounds - prompt (if an image is processed), or vice versa. Thus, forcing a uni-modal input is a precursor to collection of raw data. In practice, the biggest change was in the processing of input software: Text-only inputs were produced bypassing the vision tower and projector, whereas Image-only intake was achieved by minimizing the prompt in the chat format of the model [5]. Resulting in isolation activation streams.

2.3 Activation Recording

For each input (image-caption pair), we extract activations from every neuron in every layer. This yields a tensor:

$$M \in \mathbb{R}^{T \times L \times N}, \quad M[t, l, n] = \text{activation of neuron } n \text{ at layer } l \text{ for token } t.$$

The results are stored in Table 1.

2.4 Aggregation

Difference in Max Activation
To measure modality sensitivity, we computed the difference in maximum activation values for each neuron, contrasting text vs. image inputs (as seen in Table 2).

For each neuron and input, we compute:

- Max activation: $\max_t M[t, l, n]$
- Mean activation: $\frac{1}{T} \sum_t M[t, l, n]$

We then compute the difference in maximum activations between text and image:

$$\Delta_{l,n,i} = \max_t M_{\text{text},i}[t, l, n] - \max_t M_{\text{image},i}[t, l, n]$$

where i indexes the input.

Neuron-Level Summarization
To summarize modality preference at the neuron level, we averaged these differences across all inputs. For each neuron, we report the mean difference and corresponding t-value (Table 3).

2.5 Statistical Analysis

To summarize modality preference across inputs, we compute the average difference and a one-sample t-test [1]:

$$t_{l,n} = \frac{\bar{\Delta}_{l,n}}{s_{\Delta_{l,n}}/\sqrt{n}}$$

where $\bar{\Delta}_{l,n}$ is the mean difference, $s_{\Delta_{l,n}}$ is the sample standard deviation, and n is the number of inputs.

Table 1: Activation Metrics per Input (example preview)

Input Index	Modality	Layer Index	Neuron Index	Max Activation
1	Text	3	128	0.84
1	Image	3	128	1.12
⋮	⋮	⋮	⋮	⋮
999	Text	7	452	0.91
1000	Image	7	452	1.03

Table 2: Difference in Max Activation by Modality (example preview)

Input Index	Layer Index	Neuron Index	Δ (Text – Image)
1	3	128	-0.28
2	3	256	+0.07
⋮	⋮	⋮	⋮
999	7	452	+0.05
1000	7	452	-0.12

Table 3: Neuron-Level Summarization across Inputs (example preview)

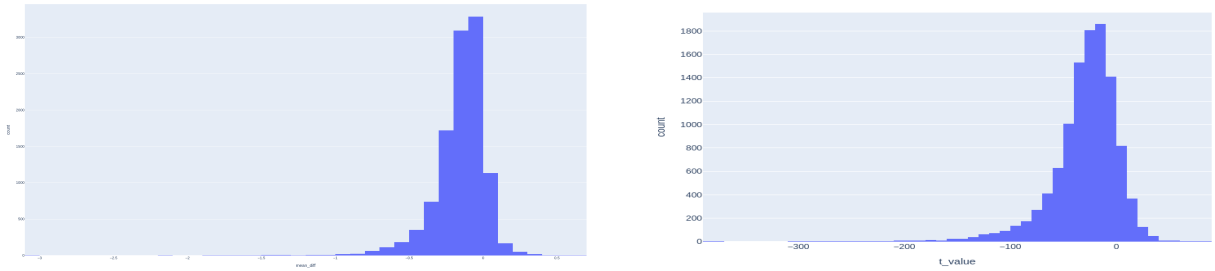
Layer Index	Neuron Index	t-value	Mean Difference
3	128	2.15	-0.14
3	256	-1.82	+0.06
⋮	⋮	⋮	⋮
7	452	3.27	-0.11

Note: The tables above illustrate the structure of the data produced at each stage of our analysis. The values shown here are only randomly generated examples, included for illustrative purposes, and do not represent the actual results collected during our experiments.

2.6 Visualization

Finally, to visualize the overall distribution of modality differences, we generated histograms of the above metrics for each layer using Plotly's histogram plotting functionality. This

provided insight into whether entire layers exhibit systematic biases toward one modality. For each layer, we visualize the distribution of mean differences and t-values using histograms as shown in Figure 1, the distributions of t-values and mean differences reveal distinct modality preferences. These reveal whether neurons in early vs. late layers show systematic modality biases.



(a) Mean difference distribution.

(b) t-value distribution.

Figure 1: t-value vs. mean difference distributions across neurons.

Afterwards, a final graph plotting mean difference across layers (Figure 2) and enabling a more generalized point of view of the immense data processed in this project.

3 Results and Discussion

3.1 Layer-wise Modality Differences

Figure 2 presents the mean modality difference across the 32 transformer layers of the LLaVA model. Each data point represents the average difference in neuron activation between text-only and multimodal (image–text) inputs for a given layer, with the shaded blue region denoting the 95% confidence interval (CI). The red dashed line indicates a linear regression fit to the trend.

Across the network depth, the mean difference exhibits a nearly monotonic decline, indicating a progressive suppression of text-only activations relative to multimodal activations. Early layers (layers 0–5) show near-zero or slightly positive differences, suggesting that textual and multimodal representations remain similar in the lower feature hierarchy. However, beginning around layer 8, a clear downward trend emerges, reaching strongly negative values by the deepest layers (25–31). The regression fit yields an R^2 of approximately 0.97, demonstrating a highly linear relationship between layer depth and mean modality difference.

3.2 Confidence and Robustness of the Trend

The narrow confidence intervals across all layers indicate low variance among neurons and high statistical reliability of the observed pattern. This stability suggests that the trend is not driven by outlier neurons but is instead a consistent property across the neuron population. In other words, the "forgetting" of text-only representations is a systematic effect of the model architecture and fine-tuning procedure rather than random variability.

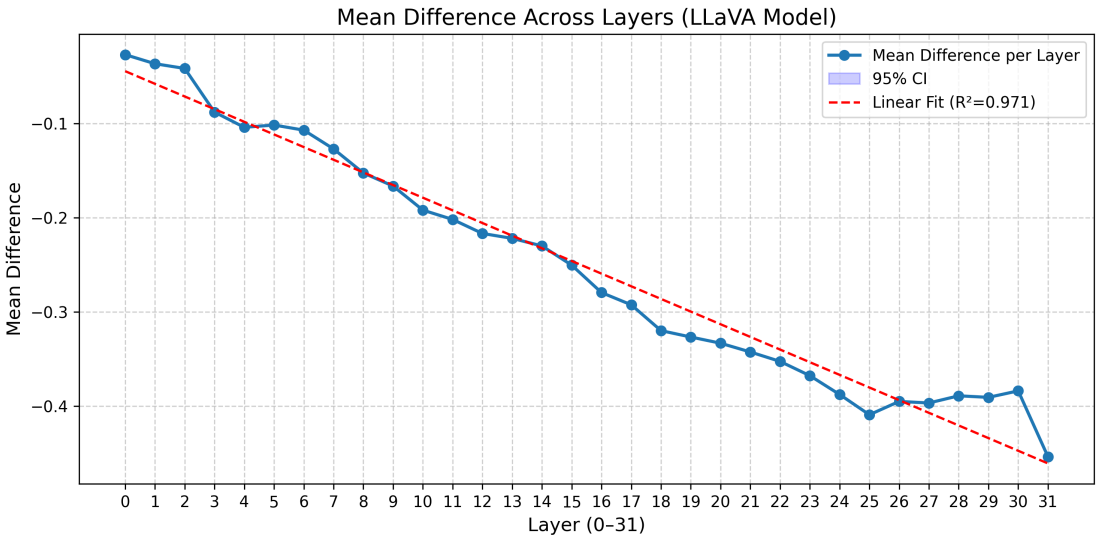


Figure 2: Mean difference between text-only and multimodal activations across transformer layers. Shaded region denotes the 95% confidence interval; the red dashed line represents a linear regression fit.

3.3 Implications for Multimodal Representation and Forgetting

The observed decline in mean difference provides strong evidence of **text-only forgetting** in multimodal transformer models. Early layers maintain linguistic representations similar to those in the original pretrained LLaMA backbone, preserving much of the model's text memory. As depth increases, multimodal fine-tuning reshapes activations, aligning them with vision-conditioned features. This alignment repurposes neurons previously tuned to linguistic features, leading to a measurable weakening of purely textual activations.

In effect, the model's internal representation transitions from a predominantly linguistic space to a joint multimodal embedding space. This pattern mirrors the phenomenon of *representational drift*, where neural representations shift toward task-specific features following new training objectives. The nearly linear decline across depth suggests that this drift occurs gradually, with mid-to-late layers acting as a transition zone where the model balances textual grounding with multimodal alignment.

3.4 Relation to Multimodal Memory

From a cognitive perspective, these results parallel mechanisms of interference in human memory: as new modalities or associations are learned, earlier single-modality pathways become attenuated. In LLaVA, this manifests as the loss of text-specific activation strength in favor of shared multimodal features. The stability of the CIs further supports the interpretation that this forgetting is structured, not stochastic.

3.5 Summary

Overall, the results demonstrate that LLaVA's multimodal fine-tuning introduces a layer-dependent reorganization of internal representations. The consistent negative slope in Figure 2 quantitatively captures how the model's "memory" of its text-only training is gradually overwritten as visual features become dominant. This supports recent findings that multimodal alignment, while enabling cross-modal reasoning, can come at the cost of eroding unimodal linguistic precision.

4 Conclusion

This study quantified and visualized layer-wise modality differences in the LLaVA multimodal model to investigate how text representations evolve after multimodal fine-tuning. The analysis revealed a clear and consistent negative trend in mean modality difference across layers, indicating that text-only activations are progressively weakened as depth increases. This provides quantitative evidence of **text-only forgetting**, a phenomenon in which multimodal adaptation repurposes linguistic neurons to serve image–text alignment rather than pure language processing.

From a representational perspective, early transformer layers preserve the linguistic memory inherited from the pretrained LLM, while deeper layers transition toward a joint embedding space optimized for multimodal reasoning. This transition reflects a structured form of representational drift, where the network's capacity for textual specificity diminishes in favor of cross-modal integration.

5 General Mathematical Foundations

Most computations in this model rely on matrix multiplication, as most distinct values are one of many in a given tensor.

- Artificial Neuron: A single neuron computes a weighted sum of inputs followed by a non-linear activation:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

But in our context, a "neuron" is in fact the model's transformer neuron activation. At layer l, the hidden state (or activation) for token t is computed as:

$$h_t^{(l)} = \mathrm{LN}\Big(h_t^{(l-1)} + \mathrm{MHA}\Big(h_t^{(l-1)}, H^{(l-1)}, H^{(l-1)}\Big)\Big)$$

followed by a feed-forward update:

$$h_t^{(l)} = \mathrm{LN}\Big(h_t^{(l)} + f\Big(W_2\sigma(W_1 h_t^{(l)} + b_1) + b_2\Big)\Big)$$

Here, σ denotes the sigmoid gating used in SwiGLU activations, while f is an element-wise nonlinearity such as GELU or SiLU. Together, they define a gated activation

mechanism:

$$\text{SwiGLU}(x) = f(W_1 x + b_1) \odot \sigma(W_1' x + b_1')$$

which has been shown to improve performance in modern transformer models such as LLaMA [8].

- Multi-Layer Perceptron (MLP): MLPs enrich per-token representations via stacked layers:

$$h = f(Wx + b)$$

- Scaled Dot-Product Attention: Attention contextualizes token embeddings [9]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

- Multi-Head Attention: Multiple heads capture diverse relationships [9]:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O$$

- t-Statistic: We rely on the one-sample t-test:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\mu = 0$ under the null hypothesis (no modality preference).

References

[1] David Diez and Christopher D Barr. Openintro statistics fourth edition. 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 6 2021.

[3] Stephen Michael Kosslyn. Information representation in visual images. *Cognitive Psychology*, 7:341–370, 1975.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. 2 2015.

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 12 2023.

[6] Zenon W. Pylyshyn. The imagery debate: Analogue media versus tacit knowledge. *Psychological Review*, 88:16–45, 1 1981.

[7] 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. 2 2021.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2 2023.

[9] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Technical report, 2023.