Do VLMs Have Bad Eyes? Diagnosing Compositional Failures via Mechanistic Interpretability

Ashwath Vaithinathan Aravindan, Abha Jha, Mihir Kulkarni
University of Southern California
{vaithina, abhajha, mkulkarn}@usc.edu

Abstract

Vision-Language Models (VLMs) have shown remarkable performance in integrating visual and textual information for tasks such as image captioning and visual question answering. However, these models struggle with compositional generalization and object binding, which limit their ability to handle novel combinations of objects and their attributes. Our work explores the root causes of these failures using mechanistic interpretability techniques. We show evidence that individual neurons in the MLP layers of CLIP's vision encoder represent multiple features, and this "superposition" directly hinders its compositional feature representation which consequently affects compositional reasoning and object binding capabilities. We hope this study will serve as an initial step toward uncovering the mechanistic roots of compositional failures in VLMs. The code and supporting results can be found here.

1. Introduction

Vision-Language Models (VLMs) have demonstrated impressive capabilities in tasks like image captioning, visual question answering, and zero-shot classification by integrating visual and textual information. However, they often struggle with *compositionality*—the ability to understand and reason about novel combinations of familiar concepts [15]. In this context, compositionality refers to recognizing combinations of known objects and attributes.

A related issue is the *binding problem* [1], where the model fails to associate specific attributes (e.g., color, shape, position) with the correct object. Robust binding is essential for fine-grained and coherent multimodal understanding.

While techniques have been proposed to encourage Convolutional Neural Networks (CNNs) to develop compositional representations [19], VLMs face additional complexity due to their need to handle both visual and textual data simultaneously.

We hypothesize that compositional failures in Vision-Language Models (VLMs) stem from **superposition** [18], where multiple concepts are entangled within shared representational subspaces. This entanglement may lead to the incorrect binding of objects, attributes, and relationships within multimodal tasks. To investigate this, we employ mechanistic interpretability techniques to analyze internal representations and diagnose failure points. Our central research question is: *Are entangled or misaligned representations within the CLIP vision encoder the root cause for failures in compositional reasoning?*

We perform several experiments with CLIP's vision encoder and note our observations about feature entanglement at the neuron level. We use Grad-CAM to analyze the spatial attention patterns of CLIP in response to compositional text prompts. These visualizations revealed consistent attribute–object binding failures, where the model incorrectly attended to multiple unrelated regions or failed to isolate the correct object-attribute pair.

To further investigate the underlying cause of these failures, we conduct a neuron-level analysis of CLIP's MLP activations. Using a synthetic dataset of simple shapes, colors, and positions, we identified "feature neurons", neurons that respond selectively to specific visual attributes, and used Shannon entropy to quantify their selectivity.

The results of our experiments strongly suggest that individual neurons encode multiple visual concepts, proving the presence of superposition. We also show that superposition of features in neurons is related to how separable the features are in the output embedding space.

2. Related Work

Compositionality and Binding in VLMs. Previous studies [1, 20] have shown that VLMs struggle with associating attributes like color, shape, and count to specific objects, often leading to compositional failures. While these works document the failures, our work goes further by investigating internal representations—specifically superposition and attention misalignment—as potential causes.

CLIP Interpretability and Concept Alignment. Recent work [5, 10] quantifies how CLIP-like models align with human-understandable concepts using attention and text decomposition. In contrast, our work is focused on uncovering why these models fail at binding—by tracing failure cases to specific neurons and spatial attention behaviors. We use techniques from tools like Prisma [4, 7] for analyzing the CLIP models.

Superposition. Elhage et al.[3] explores polysemanticity of neurons using small ReLu networks and by limiting the data to sparse input features. Olah et al.[13] expands superposition to vision tasks and proposes that superposition exists across circuits within the network. Our work focuses on neuron-level superposition and also studies how it affects Object binding and compositionality problems for vision and multimodal tasks.

3. Methodology

3.1. Datasets

Following are the datasets we use throughout our experiments:

1. CIFAR-10 dataset that contains 60000 images that belong to 10 different classes. [8]
2. A toy Shapes dataset we created, containing 500 images with simple shapes of different colors.

3.2. Model

We use the pretrained `CLIP-ViT-L/14` [14] model and its associated pre-processor due to the wide use of CLIP and its variants as the vision encoder in most VLMs. We pick the simplest variant to allow easier examination of internal representation. The vision encoder component takes a single RGB image of size 224x224 and outputs an embedding of length 768.

3.3. Visual Grounding with CLIP and Gradient-Based Localization

Although CLIP [14] has a strong performance on tasks like image retrieval and zero-shot classification, it often struggles to bind attributes to objects (e.g., confusing 'a red square' with 'a red circle') or correctly interpret spatial and relational signals. One possible reason for these failures is misalignment in the model's attention mechanisms. That is, even if CLIP correctly identifies relevant features, it may not bind them to the correct object instance. To investigate this, we use gradient-based attribution—specifically Grad-CAM [16]—to visualize which parts of an image CLIP attends to when making a prediction for a given text prompt.

3.3.1. Activation and Gradient Extraction

To understand which visual features influence CLIP's predictions, we register hooks on the final MLP layer of the vision encoder to capture both activations and gradients. During the forward pass, we record the activation maps for the image. Then, we perform a backward pass on the probability corresponding to a specific text prompt, which gives us the gradient of the prediction with respect to the image features.

3.3.2. Grad-CAM Computation

We compute a Grad-CAM heatmap for each prompt using the following steps:

1. A forward pass computes the similarity score S_t between the image and a text prompt t.
2. A backward pass computes the gradient of S_t with respect to the activation map $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$ from a chosen layer in the vision transformer:

$$\frac{\partial S_t}{\partial \mathbf{A}} \in \mathbb{R}^{H\times W\times C}$$

3. We average the gradient across spatial locations to compute importance weights for each channel:

$$\alpha_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial S_t}{\partial \mathbf{A}_{ij}^{c}}$$

4. These weights are used to compute a weighted combination of the channels in the activation map:

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_c \alpha_c \mathbf{A}^c\right)$$

This procedure reveals the spatial regions that contribute the most to the CLIP decision for a given text prompt. By analyzing these attention patterns across different prompts and images, we gain insight into how CLIP binds - or fails to bind - semantic attributes to objects in its visual representation.

3.4. Analyzing CLIP Activations

To better understand the behavior of the CLIP vision encoder and to search for evidence of superposition, we focus our analysis on the activations of the MLP layers. MLPs are a natural target for such investigation, as they constitute the majority of a transformer's parameters. Prior work has shown that MLPs in large language models function as key–value stores [6, 12], and are responsible for encoding factual and compositional knowledge. We hypothesize that MLPs in the CLIP vision encoder may play a similarly important role in storing visual concepts and their associations.

3.4.1. Toy Shapes Dataset

To facilitate this analysis, we construct a toy dataset of 500 synthetic images. Each image contains 1 to 5 objects, with features randomly sampled from a controlled set of attributes: 5 shapes (circle, triangle, square, pentagon, hexagon), 6 colors (red, green, blue, pink, black,

Algorithm 1 Identifying Feature-Selective Neurons in CLIP Vision Encoder

for each neuron n in all MLP layers **do**
  Rank all images by activation of neuron n
  Select top-k images with highest activation
  for each feature f_i in the dataset **do**
    Count occurrences o_i of f_i across the top-k images
  end for
  for each feature f_i in the dataset **do**
    Compute feature affinity value $a_i = \frac{o_i}{\sum_{j=1}^{N} o_j}$
  end for
  Compute Shannon entropy
end for
Sort neurons by ascending entropy
return Neurons with lowest entropy as highly selective neurons

yellow), and 5 spatial locations (top-left, top-right, bottom-left, bottom-right, center). A few example images from this dataset are shown in Figure 2.

The low visual and semantic complexity of this dataset enables more direct correlation between object-level features and neuron activations in CLIP. This setting allows us to isolate and interpret the behavior of individual neurons, making it easier to identify feature selectivity and potential cases of superposition.

3.4.2. Methodology

For our analysis, we record the activation of every neuron in the final linear layer of each MLP block within the CLIP vision encoder. The CLIP-ViT-L/14 model [14] consists of 24 transformer blocks, and each block contains an MLP with two layers with 1024 neurons each. For this study, we consider only the neurons on the output layer. This results in a total of 24,576 MLP neurons whose activations are monitored across the dataset.

We perform Algorithm 1 to get highly selective neurons. The Shannon Entropy [17] metric helps us identify *interesting* neurons that respond selectively to a small subset of features while remaining inactive for others. Entropy effectively quantifies the degree of bias a neuron exhibits toward particular features. A lower entropy indicates a peaked distribution implying strong preference for a few features while higher entropy suggests a uniform distribution, meaning the neuron responds similarly across diverse features and is therefore less informative for our analysis.

The Shannon entropy of the feature-occurrence distribution for a neuron is computed as:

$$\text{Entropy} = \sum_{i=1}^{N} -a_i \log a_i$$

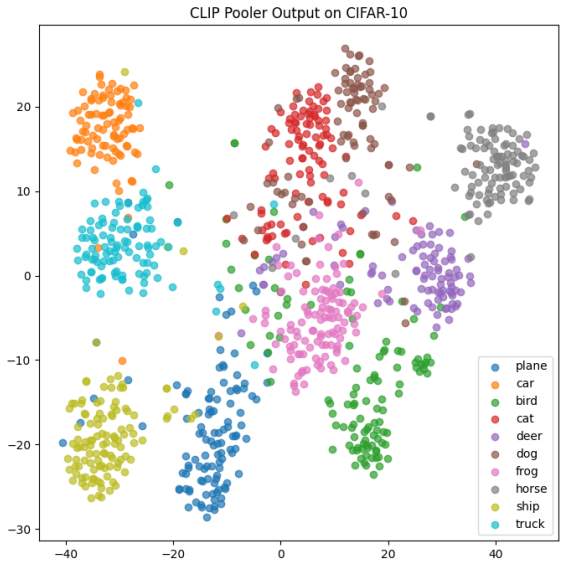


Figure 1. t-SNE plot of CLIP pooler outputs on CIFAR-10. Each point represents a single image embedding colored by its ground-truth class label.

$$\text{where } a_i = \frac{o_i}{\sum_{j=1}^{N} o_j}$$

Here, a_i denotes feature affinity of the neuron to a feature i and it is computed from o_i which denotes the average number of occurrences of feature i among the neuron's top-k activating images, and N is the total number of features. In our experimental setting, $k = 30$ and $N = 16$, corresponding to 5 shapes, 6 colors, and 5 spatial positions.

Neurons with the lowest entropy values are prioritized for further analysis, as they are more likely to encode disentangled, semantically meaningful features.

4. Experiments

4.1. CLIP on CIFAR

To understand how CLIP's vision encoder represents image categories, we extract the pooler outputs (final [CLS] embeddings) for a subset of CIFAR-10 images and projected them into 2D space using t-SNE, a non-linear dimensionality reduction technique, to visualize the embedding space. This visualization is shown in Figure 1.

The figure reveals several key observations:

1. Class-specific clusters are clearly separable for most categories, indicating that CLIP's frozen vision encoder retains discriminative features even when trained without supervision on CIFAR-10.
2. Some overlap is observed among semantically similar categories (e.g., dog, cat, horse, deer), hinting at potential superposition in the latent space, where related con-

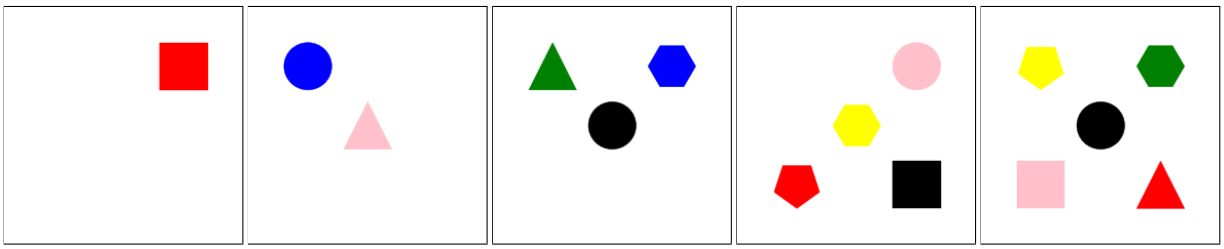


Figure 2. Sample images from the toy dataset used to study MLP activations in CLIP vision encoder

cepts are partially entangled.

These insights validate that the vision encoder itself maintains a strong semantic prior. The proximity between related clusters strongly hints at embedding-level superposition.

We found that because of the rich features of the CIFAR classes, the CLIP neuron activations were harder to analyze. Therefore, we created a simpler dataset, described in 3.4, consisting of shapes and colors with minimum feature overlap.

4.2. Analysis of Compositional Failures via Grad-CAM

To better understand the cause of compositionality failures in CLIP, we begin with a qualitative analysis using Grad-CAM. Grad-CAM is employed here due to its simplicity, transparency and strong grounding in prior interpretability work [20]. Specifically, we ask: when CLIP fails to bind attributes (e.g., color, shape, or spatial relations) to objects, can this be observed directly in its visual attention? This section investigates whether the model's failure modes are reflected in the spatial localization of its attention, as captured by gradient-based attribution.

We evaluate CLIP's ability to correctly associate multimodal concepts by comparing Grad-CAM visualizations for a set of text prompts applied to the same input image. Table 1 presents these comparisons. Below, we analyze the results case-by-case.

4.2.1. Geometric Shapes Image

Prompt: ***"a green circle"***
This case demonstrates successful grounding: CLIP correctly localizes the green circle, showing that when the prompt exactly matches the image content, the model is capable of precise visual grounding.
Prompt: ***"a green square"***
There is no green square in the image; however, CLIP's attention is split between the red square and the green circle. This illustrates a *binding problem*—the model detects color and shape as separate, unbound features. Rather than suppressing attention when the full conjunction is missing, CLIP incorrectly activates over both partial matches.

4.2.2. Dog And Pot Image (A Dog Behind a Flower Pot)

Prompt: ***"a dog"***
The model successfully identifies the presence of a dog and focuses its attention primarily on the animal. Notably, the Grad-CAM heatmap reveals concentrated activation around semantically salient regions such as the snout, paws, and face. This focused attention on characteristic parts suggests that CLIP's visual encoder attends to discriminative features when grounding basic object categories. It also indicates that, in the absence of compositional cues or relational modifiers, the model is capable of forming relatively clean object-level representations.
Prompt: ***"a dog behind a pot"*** This spatially compositional prompt results in scattered attention across the image, including the dog, the background, and pot regions. CLIP appears unable to interpret and localize relational or positional language in a grounded way, indicating a limitation in *compositional grounding*.

These results demonstrate that CLIP's internal representations, while effective at coarse image-text alignment, lack robust feature binding. The model tends to process object categories and attributes (e.g., shape, color, spatial relation) independently, often resulting in incorrect or ambiguous localization.

4.3. MLP Analysis — Feature Neurons

Through the neuron activation analysis described in Section 3.4, we identified several neurons with strong preferences for specific visual features. Figure 3 shows the distribution of the entropy values of all MLP neurons in CLIP-ViT-L/14 computed over the Toy Shapes dataset. A subset of low-entropy neurons—referred to as **feature neurons**—are presented in Tables 2, 3, 4 along with their entropy value & percentile, top activating features, average activation when the feature is present, and images from the dataset that produce the highest activations in that neuron.

As shown in the tables, the low-entropy neurons exhibit clear selectivity and consistent activation patterns in response to specific features. Their top-activating images predominantly contain a single, distinct object characterized by the target feature, providing strong evidence for their specialization. In contrast, the high-entropy neurons display no

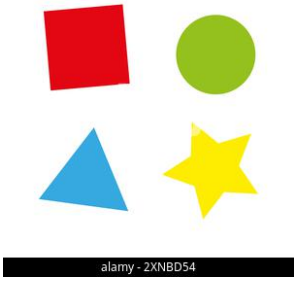
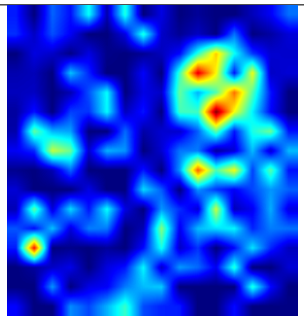
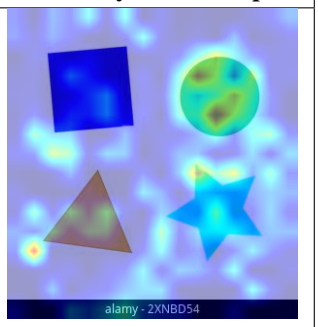
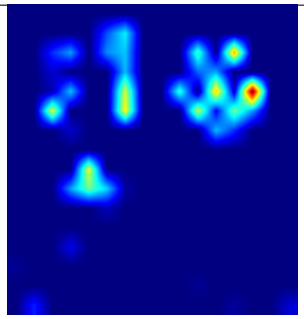
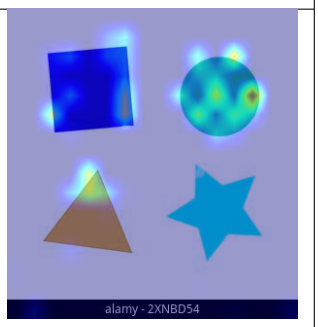
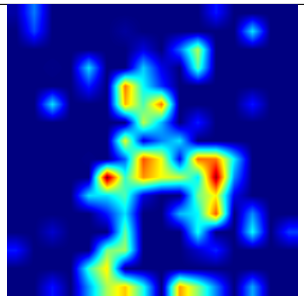
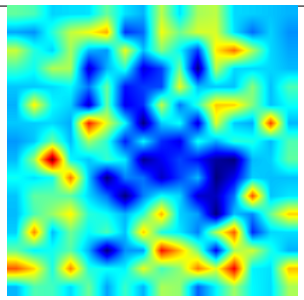
Original Image	Text Prompt	Grad-CAM Heatmap	Overlayed Heatmap
	*a green circle*		
	a green square		
	*a dog*		
	dog behind pot		

Table 1. Grad-CAM results for different text prompts. Each row shows how CLIP's attention shifts for various descriptions of the same image. Incorrect or partial attention localization reveals binding failures (e.g., attending to both green circle and red square for "green square" prompt).

clear or consistent visual pattern across their top activations. This does not imply that these neurons are meaningless or uninformative; rather, it is likely that their preferred features are underrepresented or absent in our toy dataset.

4.4. MLP Analysis — Evidence of Superposition

Many of the feature neurons identified in our analysis exhibit signs of *superposition* [3], that is, they respond to multiple distinct features simultaneously. For example, the 563^{rd} neuron in the 16^{th} transformer block (highlighted first in Table 2) shows high activation for both circular and square shapes. To verify this, we analyzed the patch-wise activation maps of this neuron across several test images.

Figure 4 presents a set of images alongside the corresponding activation maps of this neuron. In the first exam-

ple, we observe that the neuron's activation is strongest (denoted by yellow and red regions) in areas containing its preferred features namely circles, squares, and the colors pink and red. In contrast, regions containing unrelated features such as triangles or blue objects elicit weak responses (blue regions in the heatmap). This suggests the neuron does not indiscriminately fire, but instead responds selectively to a combination of semantically meaningful features.

To further support this observation, we examine activation maps on single-object images. In the second and third rows of Figure 4, the neuron activates strongly only in the presence of one or more of its preferred features. In the final row, where the image contains none of the relevant features, the activation is low throughout the entire spatial map indicated by uniformly blue regions.

These results provide strong evidence that the neuron simultaneously encodes multiple visual concepts, validating the presence of superposition. Understanding such behavior is crucial for downstream compositional reasoning, as neurons that entangle multiple features may introduce ambiguity when precise attribute–object bindings are required.

As a side note, one reason for the activations in Figure 4 are not very spatially focused on the locations of the objects in the image, is that image tokens pass through many layers of self-attention, which mixes and redistributes information across tokens. As a result, neurons in the later layers may respond to features that are no longer in the same location as they were in the original image. Another possible reason, one that complements the first, is that ViTs often use background tokens as places to gather and process information. Darcet et al.[2] found that ViTs tend to select less important tokens, often from background regions, and use them as pooling spots to combine information from other tokens.

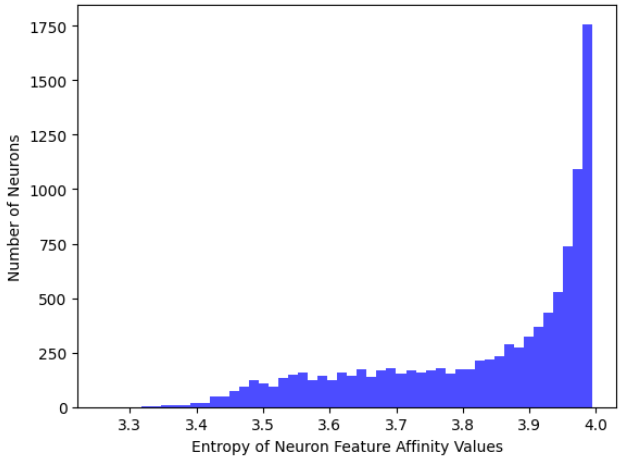


Figure 3. Distribution of the entropy of neuron feature-affinity values in CLIP-ViT-L/14, evaluated on the Toy Shapes dataset.

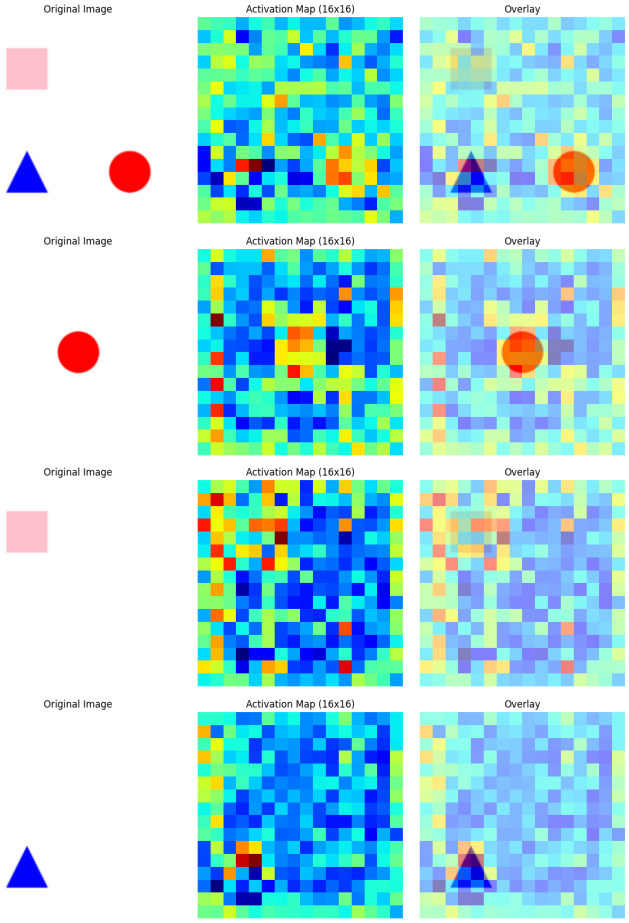


Figure 4. Images from the toy dataset and the patch-wise activation map of 563rd neuron in layer 16 showing affinity of this neuron to square and circle shapes (first three rows). The neuron shows largely low activations on the bottom-most image containing only a triangle, which is not a target feature for this neuron.

4.5. Effects of Superposition

4.5.1. Motivation

Building on the evidence of superposition, we hypothesize that such neuron-level superposition erodes the model's ability to form clean, compositional object–attribute bindings. In other words, if the same low-entropy neuron fires for both "red" and "square", the model may confuse a red circle for a green square.

4.5.2. Method

Taken from the 16 features in our shapes dataset, we create pairs of features. For every ordered feature pair $\langle f_1, f_2 \rangle$ we:

1. **Quantify superposition** using for the 1000 lowest-entropy neurons. The measure for Superposition or **S**,

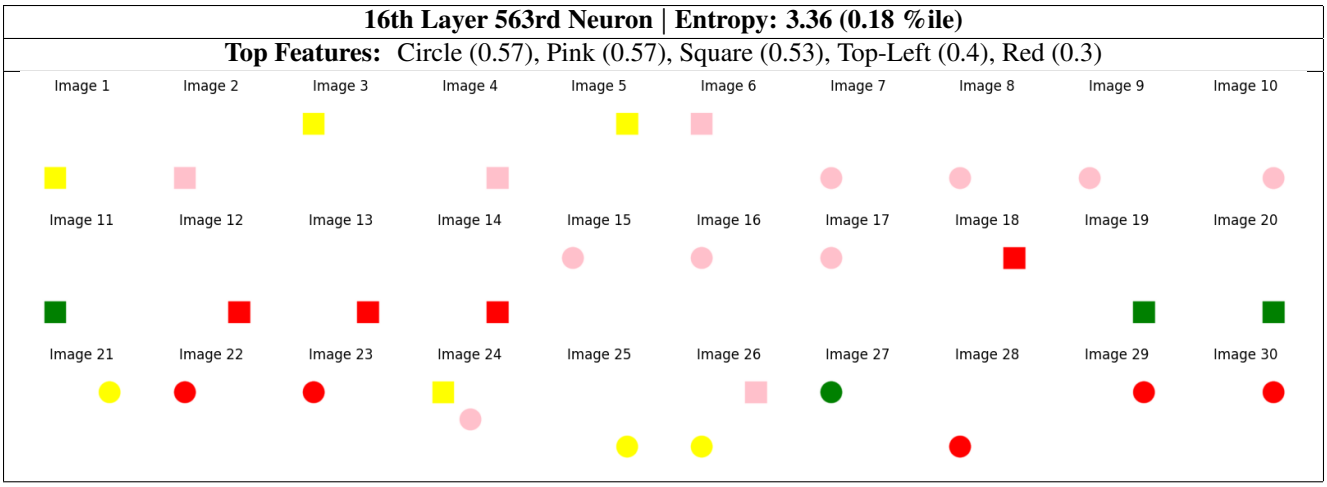


Table 2. Top features and Top activating images for 563rd Neuron in the 16th Layer. This neuron activates the most when handling images containing square and circle shapes and the color pink.

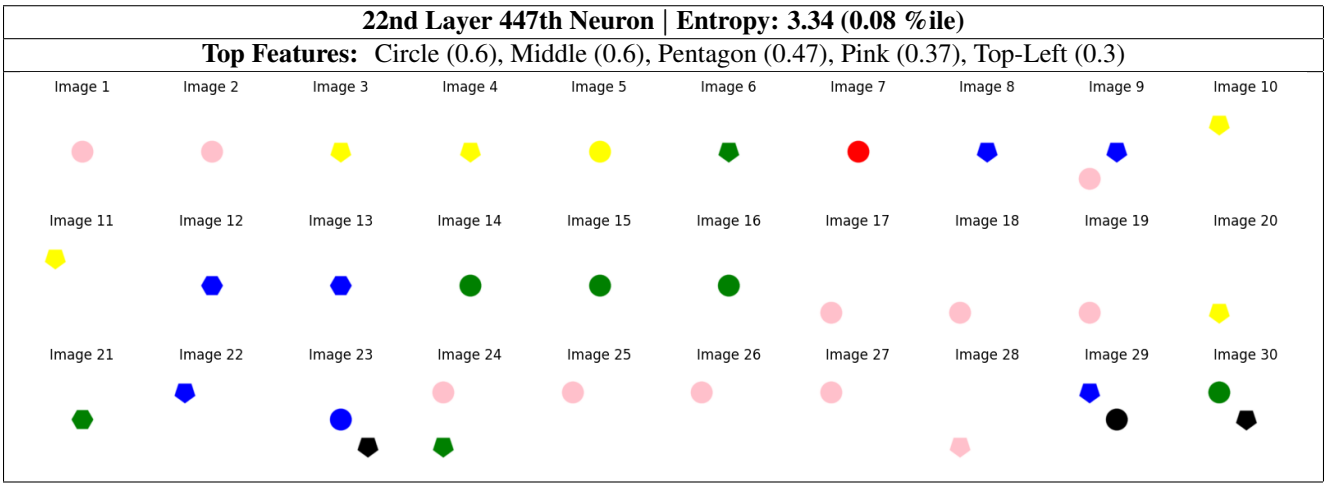


Table 3. Top features and Top activating images for 447th Neuron in the 22nd Layer. This neuron activates the most when handling images containing the circle and pentagon shapes and the middle position.

measured between features f_1 and f_2 is

$$\mathbf{S}(f_1, f_2) = \sum_{i=1}^{n} \frac{a_{f_1}^{(i)} + a_{f_2}^{(i)}}{\sum_{j=1}^{16} a_{f_j}^{(i)}}$$

where $a_{f_j}^{(i)}$ is the affinity of feature j to neuron i.

2. **Quantify compositional separability** of the corresponding image embeddings. The metrics used for this are
   - Cluster-center distance $\mathbf{D}(f_1, f_2)$: the Euclidean distance between the two single-feature centroids (larger is better).
   - Misclassification rate $\mathbf{M}(f_1, f_2)$: fraction of embeddings whose nearest centroid is of the *wrong* feature (smaller is better).

4.5.3. Results

Scatter plots of Measure of Superposition or S versus D and M are shown in Figures 5, 6 respectively.

The key observations from these plots are -

1. *Inverse relation with distance.* Feature pairs that are superimposed more (high S) exhibit **smaller cluster gaps** (low D) in embedding space. This indicates that superposition pulls concept representations closer together, reducing their geometric separability.
2. *Direct relation with error.* The same high-S pairs suffer **higher misclassification rates** M, showing that entangled neurons translate into increased attribute–object binding mistakes.

The experiment demonstrates a connection, albeit a weak correlation that warrants further investigation, be-

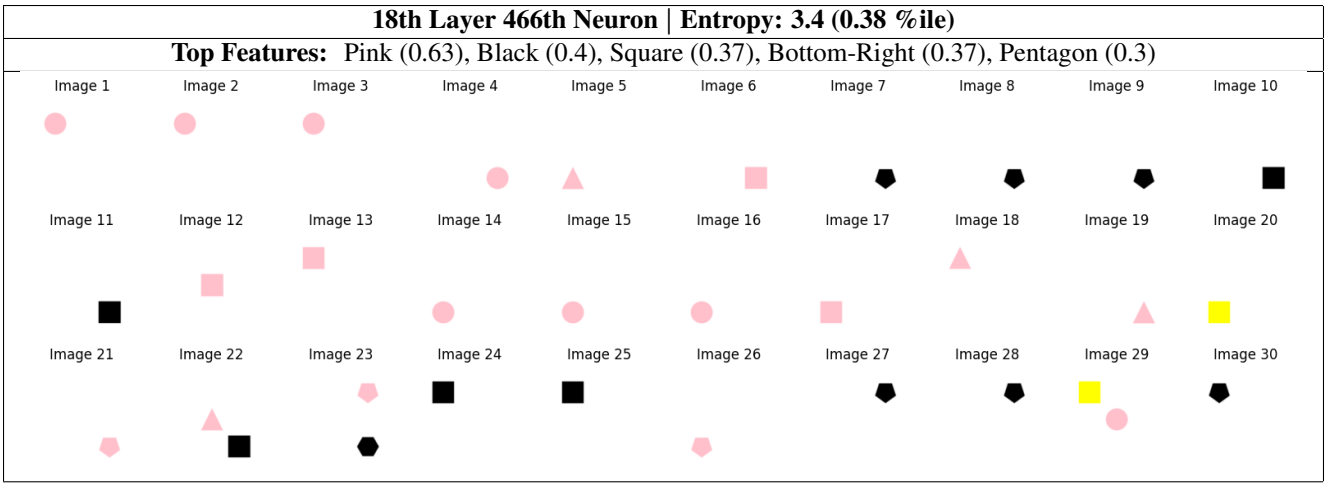


Table 4. Top features and Top activating images for 466th Neuron in the 18th Layer. This neuron activates the most when handling images containing pink and black shapes.

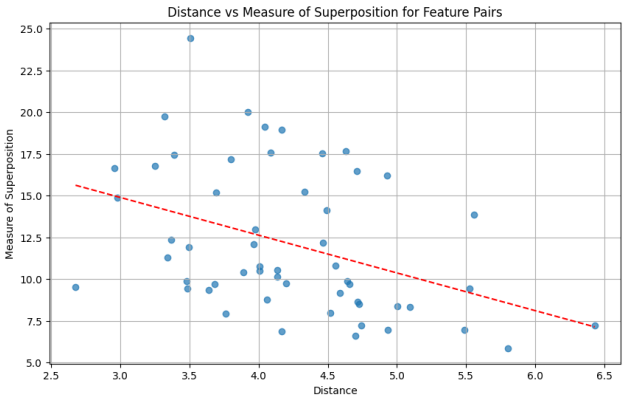


Figure 5. Distance (**D**) vs Measure of Superposition (**S**) for Feature Pairs

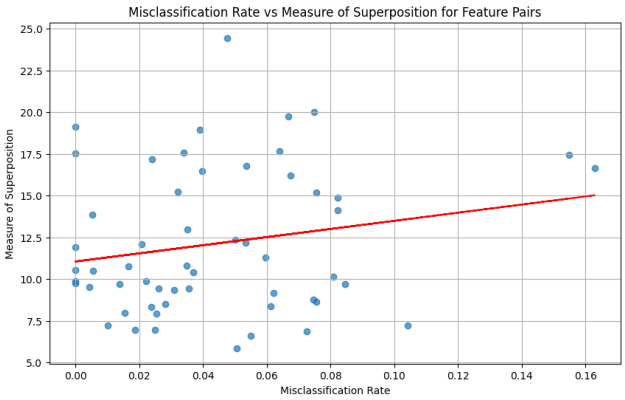


Figure 6. Misclassification Rate (**M**) vs Measure of Superposition (**S**) for Feature Pairs

tween superposition inside CLIP's MLP layers and its external difficulties with compositional reasoning.

5. Limitations and Future Work

This study is limited to the CLIP-ViT-L/14 model, whereas most modern vision-language models (VLMs) adopt variants such as SigLIP[11] and alternative architectures like SAM [9]. Future work could strengthen our claims by establishing a more direct causal link between weak compositional feature representations and failures in compositional reasoning. Additionally, extending our analysis to more complex, feature-annotated datasets would facilitate a deeper investigation into the relationship between neuronal feature superposition and compositional reasoning limitations.

6. Conclusion

Our study uncovers a mechanistic connection between internal feature representation and CLIP's image embeddings.

1. **Neuron-level superposition.** Feature entanglement is present not only at the embedding-level but all the way down to individual neurons, many of which may encode multiple, semantically unrelated attributes.
2. **Impact on compositionality.** The stronger this superposition, the weaker CLIP's ability to bind objects and attributes: high entanglement predicts smaller embedding-space separation and higher misclassification rates on compositional tasks.

These findings establish superposition as a key bottleneck for object–attribute compositionality and motivate future work on disentangling neuron activations to improve CLIP's feature representation.

References

[1] Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37: 113436–113460, 2024. 1

[2] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 6

[3] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Bowman, Jacob Chen, Nicholas Joseph, Chris Olah, and Dario Amodei. Toy models of superposition. *Transformer Circuits Thread*, 2022. 2, 5

[4] Neel Elhage, Amjad Nanda, and Charlie Joseph. Laying the foundations for vision and multimodal mechanistic interpretability, 2023. Accessed: March 12, 2025. 2

[5] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition, 2024. 2

[6] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020. 2

[7] Sonia Joseph. Vit prisma: A mechanistic interpretability library for vision transformers. https://github.com/soniajoseph/vit-prisma, 2023. 2

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[9] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 8

[10] Avinash Madasu, Yossi Gandelsman, Vasudev Lal, and Phillip Howard. Quantifying and enabling the interpretability of clip-like models, 2024. 2

[11] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. 8

[12] 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. 2

[13] 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. 2

[14] OpenAI. Clip vit-l/14 - hugging face. https://huggingface.co/openai/clip-vit-large-patch14, 2021. Accessed: 2025-04-11. 2, 3

[15] Jacob Russin, Sam Whitman McGrath, Danielle J Williams, and Lotem Elber-Dorozko. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *arXiv preprint arXiv:2405.15164*, 2024. 1

[16] 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. *International Journal of Computer Vision*, 128(2):336–359, 2019. 2

[17] Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. 3

[18] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. 1

[19] Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5058–5067, 2017. 1

[20] Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14151, 2024. 1, 4