Attribution and Application of Multiple Neurons in Multimodal Large Language Models

Feiyu Wang[1], Ziran Zhao[1], Dong Yu[1]*, Pengyuan Liu[1,2]
[1] School of Information Science, Beijing Language and Culture University, Beijing, China
[2] National Print Media Language Resources Monitoring & Research Center,
Beijing Language and Culture University, Beijing, China
wfy_0502@163.com, 202321197010@stu.blcu.edu.cn
yudong@blcu.edu.cn, liupengyuan@pku.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated exceptional performance across various tasks. However, the internal mechanisms by which they interpret and integrate cross-modal information remain insufficiently understood. In this paper, to address the limitations of prior studies that could only identify neurons corresponding to single-token and rely on the vocabulary of LLMs, we propose a novel method to identify multimodal neurons in Transformer-based MLLMs. Then we introduce fuzzy set theory to model the complex relationship between neurons and semantic concepts and to characterize how multiple neurons collaboratively contribute to semantic concepts. Through both theoretical analysis and empirical validation, we demonstrate the effectiveness of our method and present some meaningful findings. Furthermore, by modulating neuron activation values based on the constructed fuzzy sets, we enhance performance on the Visual Question Answering (VQA) task, showing the practical value of our approach in downstream applications in MLLMs.

1 Introduction

Large Language Models (LLMs) have significantly advanced natural language processing, enabling breakthroughs in diverse natural language tasks (Bai et al., 2023; Grattafiori et al., 2024; Guo et al., 2025). Motivated by their success, researchers have further extended LLMs to multimodal domain, developing Multimodal Large Language Models (MLLMs) capable of integrating visual and textual modalities, achieving impressive results across multimodal understanding and generation (Dai et al., 2023; Chen et al., 2025; Liu et al., 2024b).

Despite significant advances in LLM interpretability research (Sun et al., 2024; Tang et al., 2024; Yu and Ananiadou, 2024), the internal mechanisms by which MLLMs process and integrate different modalities remain unclear. This opacity risks generating biased or hallucinatory content and hinders error traceability, undermining trust and escalating safety risks in critical applications, such as medical diagnostics (González-Alday et al., 2023) or autonomous driving systems (Ma et al., 2025). Hence, improving the explainability of MLLMs is essential for both advancing their development and ensuring the trustworthiness of AI.

Recent multimodal studies (Pan et al., 2023; Schwettmann et al., 2023) have focused on identifying neurons that are respond to both textual and visual inputs within Transformer architectures, known as **multimodal neurons**. These neurons have learned visual features from images during training and are capable of influencing text generation. Among existing methods, Schwettmann et al. (2023) propose a gradient-based neuron attribution method to identify multimodal neurons within feed-forward networks (FFNs). Pan et al. (2023) define a contribution score based on activation outputs from FFNs, enabling gradient-free identification of multimodal neurons in MLLMs. However, both methods rely on the next-token prediction for neuron attribution, restricting neuron identification to single-token concepts and limiting generalizability due to dependence on the vocabulary of LLMs. Besides, previous methods only focus on the role of individual neurons, neglecting the complex mechanisms by which multiple neurons collaboratively encode different semantic concepts. Moreover, the potential for broader application in real-world tasks remains largely unexplored.

To address these challenges, we first propose a visual-semantic perturbation-based neuron attribution method. In this method, both original visual inputs and their counterparts with target semantic concepts removed are input into the model for caption generation. By comparing the neuron activation differences accumulated over all generated tokens between the two types of inputs, we can

* Corresponding author: Dong Yu.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11649–11662
November 4-9, 2025

identify the neurons that are highly responsive to the semantic concepts. Thus, our method can expand the scope of identification to include multi-token semantic concepts and avoid the influence of the vocabulary of LLMs.

Furthermore, inspired by prior work (Cao et al., 2025; Kalibhat et al., 2023) that investigate the function of neuron groups in convolutional neural networks (CNNs), we believe that multimodal neurons in MLLMs also exhibit similar associations with semantic concepts—an individual neuron may contribute to multiple semantic concepts with varying degrees of activation, while the representation of semantic concepts often emerges from the coordinated activation of multiple neurons. To model this neuron-semantic relationship, we introduce fuzzy set theory (Zadeh, 1965). Specifically, we construct a fuzzy set for each semantic concept, in which each neuron is assigned a membership degree according to its degree of affiliation with the set. This modeling approach systematically characterizes how multimodal neurons contribute to different concepts, offering a novel perspective on neuron activation mechanisms within MLLMs.

Based on the constructed fuzzy sets of multimodal neurons for semantic concepts, we further demonstrate their application in the Visual Question Answering (VQA) task. By modulating the activation values of neurons in the fuzzy sets corresponding to the concepts relevant to each image-question pair, we strengthen the model's capacity to identify and interpret specific semantics within images, thereby improving VQA performance. Further experimental results demonstrate that our method can be extended to identify text neurons representing interrogative words, which can then be applied in the VQA task, proving the generalizability of our approach.

In summary, our contributions are three-fold: (1) We propose a new method for identifying multimodal neurons in MLLMs. (2) We introduce fuzzy set theory to systematically characterize the contributions of neurons across different semantic concepts. (3) We explore the application of multimodal neurons in the VQA task.

2 Background

2.1 Pixel-Level Semantic Attribution via Diffusion Models

Latent diffusion models (e.g., Stable Diffusion (Rombach et al., 2022)) synthesize photorealistic images from random noise through text-guided iterative denoising. The framework integrates three core components: a CLIP (Dosovitskiy et al., 2020) text encoder converts input prompts into semantic embeddings, a variational autoencoder (VAE) (Kingma et al., 2013) compresses and reconstructs images in latent space, and a U-Net (Ronneberger et al., 2015) network progressively removes noise from latent vectors.

To achieve pixel-level semantic attribution, cross-attention mechanisms in diffusion models align textual tokens with latent space features. At each denoising step, normalized attention scores between textual tokens and latent space features within U-Net layers measure the semantic relevance of local regions to text prompts. Tang et al. (2022) aggregate attention scores across spatiotemporal dimensions, and interpolate them into pixel space to generate semantic attribution maps, which quantify the influence of specific textual tokens on each pixel. In this paper, we utilize the pre-trained Stable Diffusion model to generate images for text prompts containing target concepts and acquire pixel-level attention matrices through diffusion attention mechanisms with attribution maps.

2.2 Neurons in Transformer-Based MLLMs

A MLLM typically contains three core components: a visual encoder, a textual LLM backbone, and a vision-to-language adapter module. Following previous works (Dai et al., 2021; Pan et al., 2023; Schwettmann et al., 2023; Wang et al., 2022), we focus our analysis specifically on neurons within FFNs in the textual LLMs, as they carry two-thirds of the parameters and have been empirically demonstrated to be essential in understanding textual and visual features.

Formally, the FFN in a Transformer (Vaswani et al., 2017) layer is:

$$\mathrm{FFN}(\mathbf{x}) = f(\mathrm{x}\mathbf{W}_1{}^\top + \mathrm{b}_1)\mathrm{W}_2 + \mathrm{b}_2, \quad (1)$$

where $\mathrm{x} \in \mathbb{R}^d$ is the hidden representation from the previous layer, $f(\cdot)$ is the activation function, W_1, $\mathrm{W}_2 \in \mathbb{R}^{d_m \times d}$ are projection matrices, and b_1, b_2 are biases.

For simplicity, let $\mathbf{a} = f(\mathrm{x}\mathrm{W}_1{}^\top + \mathrm{b}_1)$, where the i-th element represents the activation output of the i-th neuron, which is also the unit to be investigated in our study. In the following discussion, we denote it as $\mathrm{u}(l, i)$, where l represents the index of the Transformer layer, and i indicates the neuron index within the hidden layer of the FFN.

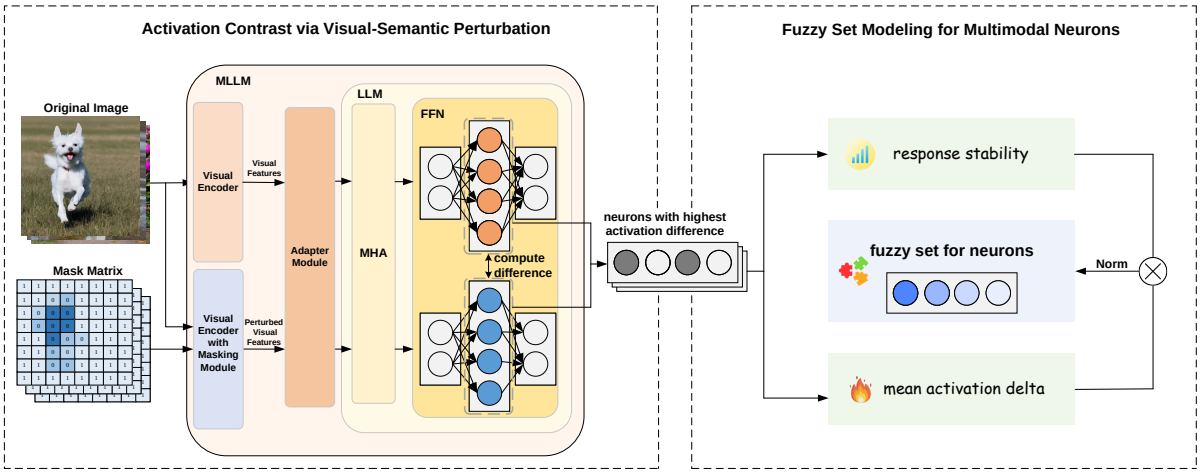


Figure 1: The overview of method for identifying and modeling multimodal neurons.

3 Method

In this section, we first propose a perturbation approach to measure activation differences and identify neurons responsive to semantic concepts. We further leverage membership degrees from fuzzy set theory to quantify the semantic contribution of multimodal neurons. The overview of our method is illustrated in Figure 1.

3.1 Activation Contrast via Visual-Semantic Perturbation

We first introduce the method for activation contrast. Given a target semantic concept x, we synthesize N original images $\{v^{(x,1)}, \ldots, v^{(x,N)}\}$ using Stable Diffusion with text prompts containing x. Based on the pixel-level semantic attribution method described in Section 2.1, we compute the semantic attribution map by spatiotemporal aggregation of cross-attention scores for each image $v^{(x,n)}$. We then apply K-means clustering ($k = 2$) on its attribution map to obtain binary mask $M^{(x,n)} \in \{0,1\}^P$, where P is the number of patches and $M^{(x,n)}[p] = 0$ indicates that patch p has high attention scores to the concept x.

The perturbation is strategically applied after the patch-embedding stage in the visual encoder. Let $v_{pe}^{(x,n)} = \text{PatchEmbed}(v^{(x,n)}) \in \mathbb{R}^{P\times D}$ denote the original patch embeddings of image $v^{(x,n)}$, where D is the embedding dimension. We define a masking module $\mathcal{M}$ that performs patch-level noise replacement on selected embeddings:

$$\mathcal{M}(v_{pe}^{(x,n)}, M^{(x,n)})_p = \begin{cases} \epsilon \sim \mathcal{N}(0,\sigma^2) & \text{if } M^{(x,n)}[p] = 0 \\ v_{pe}^{(x,n)}[p] & \text{if } M^{(x,n)}[p] = 1 \end{cases}, \tag{2}$$

Hence, patches with high attribution to x are replaced by random noise, while other patches remain unchanged. The resulting perturbed embeddings $v_{pe}^{(x,n)'}$ are then combined with the [CLS] token embedding v_{CLS} and positional embeddings E_{pos}, and fed into the Transformer Encoder to obtain perturbed visual features $v_f^{(x,n)'}$:

$$v_f^{(x,n)'} = \text{TransformerEncoder}\left(\left[v_{\text{CLS}}, v_{pe}^{(x,n)'}\right] + E_{\text{pos}}\right), \tag{3}$$

To quantify the effect of visual-semantic perturbations, we record neuron activations during autoregressive text generation for image captioning under two visual conditions: the original visual features $v_f^{(x,n)}$ and the perturbed features $v_f^{(x,n)'}$. For the original visual input, let $a^{(x,n)}(l;i)[j] \in \mathbb{R}$ denote the activation of the i^{th} neuron in layer l at the j^{th} generated token. Similarly, for the perturbed visual input, let $a^{(x,n)'}(l;i)[j] \in \mathbb{R}$ denote the corresponding activation. The activation difference $\Delta^{(x,n)}(l,i)$ is computed as the difference between the expected neuron activations under the original and perturbed visual inputs:

$$\Delta^{(x,n)}(l,i) = \mathbb{E}\left[a^{(x,n)}(l;i)[j]\right] - \mathbb{E}\left[a^{(x,n)'}(l;i)[j]\right], \tag{4}$$

3.2 Fuzzy Set Modeling for Multimodal Neurons

To further evaluate the contribution of each neuron to specific semantic concepts, we introduce a fuzzy set-based modeling approach and compute membership degrees by aggregating two key metrics: ***response stability*** and ***mean activation delta***. For each semantic concept x, we select

the top 100 neurons with the highest activation difference $\Delta^{(x,n)}(l,i)$ for each synthesized image $v^{(x,n)}$. Let $U^{(x,n)} = \{u(l,i) \mid \Delta^{(x,n)}(l,i) \in \text{Top100}(\Delta^{(x,n)})\}$ denote the set of neurons with high response, and define the union set across all images as $U_x = \bigcup_{n=1}^{N} U^{(x,n)}$.

For each neuron $u(l,i) \in U_x$, we define its *response stability* $c_x(l,i)$ as the proportion of images in which the neuron appears in the high response set, and its *mean activation delta* $s_x(l,i)$ as the geometric mean of its activation differences across those same images:

$$c_x(l,i) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\big(u(l,i) \in U^{(x,n)}\big), \tag{5}$$

$$s_x(l,i) = \sqrt{\frac{\sum_{n=1}^{N} \Delta^{(x,n)}(l,i) \cdot \mathbb{I}\big(u(l,i) \in U^{(x,n)}\big)}{\sum_{n=1}^{N} \mathbb{I}\big(u(l,i) \in U^{(x,n)}\big)}}. \tag{6}$$

where $\mathbb{I}(\cdot)$ is an indicator function.

The membership degree $\mu_x(l,i) \in [0,1]$ is obtained by normalizing the product of *response stability* and *mean activation delta*:

$$\mu_x(l,i) = \text{Norm}\left(c_x(l,i) \cdot s_x(l,i)\right), \tag{7}$$

where $\text{Norm}(\cdot)$ denotes min-max normalization.

Finally, we define the fuzzy set X for multimodal neurons corresponding to the concept x as follows. Here, $\theta \in [0,1]$ is a prespecified membership threshold.

$$X = \{(u(l,i), \mu_x(l,i)) \mid \mu_x(l,i) \geq \theta,\ u(l,i) \in U_x\}, \tag{8}$$

4 Experiments

4.1 Experimental Setup

Models. We use LLaVA-NEXT (Liu et al., 2024b) and Janus-Pro (Chen et al., 2025) as our research models. LLaVA-NEXT focuses on multimodal understanding tasks, while Janus-Pro unifies multimodal understanding and visual generation by decoupling visual encoding but still using a single, unified Transformer architecture for processing. Specifically, we select llava-v1.6vicuna-7b-hf[1] and Janus-Pro-7B[2], and the number of neurons under study is 352.3K and 330.2K, respectively.

Datasets. We employ our method and conduct experiments across 80 semantic categories from the MSCOCO dataset (Lin et al., 2014), covering both single-token concepts (e.g., *dog*) and multi-token concepts (e.g., *fire hydrant*). We utilize the pre-trained diffusion model[3] to synthesize images and obtain pixel-level attention matrices. The textual prompts are randomly sourced from annotated sentences in Flickr 30k (Young et al., 2014), specifically those containing the targeted concepts.

Metrics. To evaluate if multimodal neurons are responsive to certain concepts from both textual and visual perspectives, we measure several evaluation mentrics as follows: (1) **Semantic Relevance**: To verify the relevance of neurons to textual concepts, we align them with natural language. The closer the top tokens associated with a neuron are to the textual concept, the stronger the semantic relevance of the neuron. To quantify this, we measure BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019) and LaBSE (Feng et al., 2020) between each concept and the corresponding top-10 tokens. (2)**Visual Selectivity**: To investigate the alignment between multimodal neurons and semantic concepts in images, we quantify their receptive fields for specific visual concepts by taking the activations at image patch tokens, following Schwettmann et al. (2023). We upsample the activation maps to the input resolution by bilinear interpolation and threshold them above the 0.95 percentile to obtain binary masks, which are compared to COCO instance segmentations using Intersection over Union (IoU).

Baselines. We implement the following two baseline methods for comparison: (1) **Grad**: The gradient of the output logit with respect to neuron activation through backpropagation (Schwettmann et al., 2023); (2)**Act**: The element-wise product of neuron activation and the unembedding matrix, representing the neuron's contribution to a specific token in the output vocabulary (Pan et al., 2023).

Implementation Details. In constructing the fuzzy set X, we select a threshold θ corresponding to the top 20% of the membership values $\mu_x(l,i)$ in U_x, ensuring that neurons demonstrating statistically meaningful contributions to concept x are retained. Finally, we statistically analyze the number of multimodal neurons within the fuzzy sets across the 80

[1] https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf

[2] https://huggingface.co/deepseek-ai/Janus-Pro-7B

[3] https://huggingface.co/stabilityai/stable-diffusion-2-1

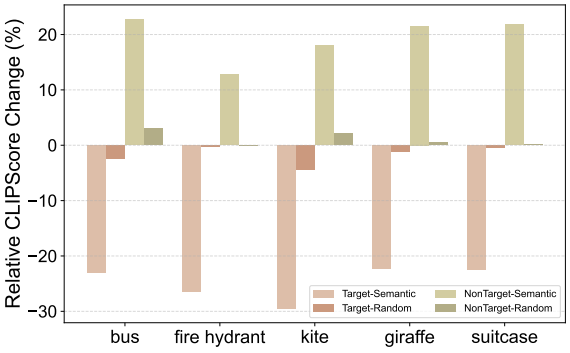


Figure 2: Relative CLIPScore change (%) across five concepts under four conditions: Target/NonTarget denotes the text concept type; Semantic/Random denotes the mask type. We report the mean change over N images per concept, with all changes statistically significant ($p < 0.01$).

Semantic concept: dog			
Image	**Model**	**Type**	**Output**
	LLaVA-NEXT	original	a **dog** swimming in the water.
		perturbed	water with ripples.
	Janus-Pro	original	a **dog** swimming in water.
		perturbed	the water surface with gentle waves.
Semantic concept: fire hydrant			
	LLaVA-NEXT	original	a man sitting on the grass next to a red **fire hydrant**.
		perturbed	a person sitting on the grass.
	Janus-Pro	original	a person sitting on grass next to a red **fire hydrant**.
		perturbed	a person sitting on grass outdoors.

Table 1: Example model outputs.

semantic concepts. The average counts are 601 for LLaVA-Next and 847 for Janus-Pro, corresponding to 0.17% and 0.26% of the total number of neurons in the FFN intermediate layers, respectively. More details can be found in appendix A.

4.2 Results & Discussion

In this section, we first validate the effectiveness of our proposed visual-semantic perturbation method (§4.2.1). We then compare our method with prior approaches, demonstrating that our identified multimodal neurons exhibit higher semantic relevance (§4.2.2) and stronger visual selectivity (§4.2.3). Finally, we analyze the semantic activation patterns (§4.2.4) and layer-wise distributions of the discovered multimodal neuron sets (§4.2.5).

4.2.1 Perturbation Efficacy

To evaluate the effectiveness of our proposed visual perturbation method, we compute the relative CLIPScore (Hessel et al., 2021) change using the pre-trained CLIP model[4]. For each image, we encode its visual features both with and without our masking module in the visual encoder. Text features are encoded using two prompt types: the target concept (e.g., *bus*) and non-target concepts extracted from image captions (e.g., *street, people, trees* from "a bus on a street with people and trees."). As a control, we also apply a randomly shuffled binary mask with the same sparsity. As shown in Figure 2, random masking results in negligible score changes, whereas semantic masking substantially reduces similarity to the target concept while preserving or enhancing alignment with non-target concepts. Table 1 shows examples of how our perturbation affects model predictions. These results demonstrate that our method selectively suppresses target information without compromising the integrity of other semantic content.

Model	Method	BS	MS	LB
LLaVA-NEXT	Grad	0.533	0.641	0.376
	Act	0.549	0.649	0.400
	Ours	**0.560**	**0.652**	**0.406**
Janus-Pro	Grad	0.527	**0.638**	0.354
	Act	0.530	0.630	0.374
	Ours	**0.534**	0.629	**0.375**

Table 2: Results of BERTScore(BS), MoverScore(MS), and LaBSE(LB). For each semantic concept, we select top 5% multimodal neurons and compute the average scores across all concepts.

Semantic concept: kite			
Model	**Method**	**Top Neuron**	**Top Tokens**
LLaVA-NEXT	Grad	**u(4,8572)**	['flag', 'flags', 'Flag', 'emo']
	Act	**u(31,2611)**	['k', 'children', 'children', 'Children']
	Ours	**u(21,7551)**	['flight', 'flying', 'fly', 'fly']
Janus-Pro	Grad	**u(6,2778)**	['Content', 'Mut', 'hi', 'Alliance']
	Act	**u(29,6369)**	['跳舞', '舞蹈', 'dance', 'Dance']
	Ours	**u(28,3371)**	['fly', 'Fly', 'flying', 'Fly']

Table 3: An example of ***kite***, showing the top neuron selected by different methods and its top-4 tokens.

4.2.2 Textual Semantics of Neurons

To evaluate whether the identified multimodal neurons encode interpretable textual semantics, we follow the projection analysis in Pan et al. (2023), where the multiplication of the unembedding matrix and the second layer of FFN maps neuron activations to token probabilities. For each neuron,

[4] https://huggingface.co/openai/clip-vit-large-patch14-336

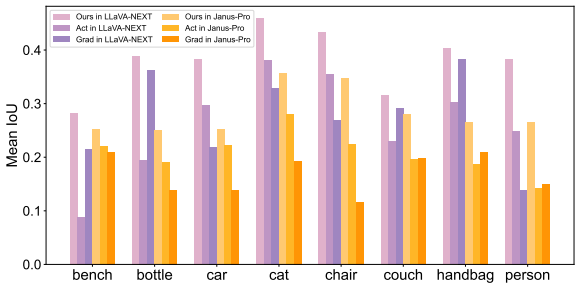


Figure 3: Mean IOU across 8 semantic concepts. For Grad and Act, we select the same number of top-ranked neurons based on their respective attribution scores and average their scaled activations. We randomly sample 20 images from the MSCOCO-2017 validation set for each concept and report the mean results.

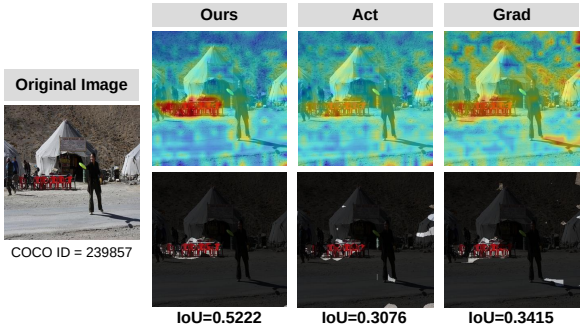


Figure 4: Heatmap, binary mask, and IoU results for ***chair*** in an example image.

we rank its projection row and extract the top-10 tokens. Following Schwettmann et al. (2023), we restrict the computation to interpretable neurons. For each semantic concept, we select the top 5% of multimodal neurons from its corresponding fuzzy set and compute their semantic relevance, then take the average. Table 2 reports the results on single-token concepts, with an example illustrated in Table 3. Experimental results and examples on multi-token concepts are provided in appendix B.1. These results confirm that our selected neurons exhibit strong semantic relevance.

4.2.3 Visual Focus of Neurons

To evaluate the visual selectivity of multimodal neurons in images, we conduct experiments using the constructed fuzzy set of multimodal neurons, computing weighted activations based on membership degrees. As shown in Figure 3, our method produces receptive fields that more accurately segment target objects than Grad and Act. The examples in Figure 4 show our neurons are more focused on image regions containing specific concepts. More examples are shown in appendix B.2.

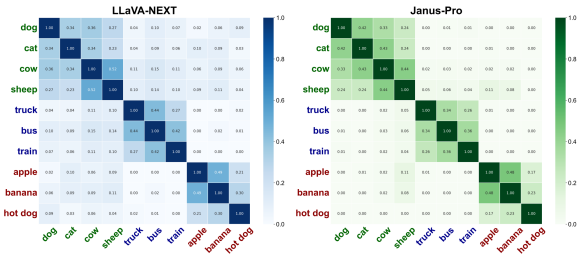


Figure 5: Spearman's rank correlations between different semantic concepts.

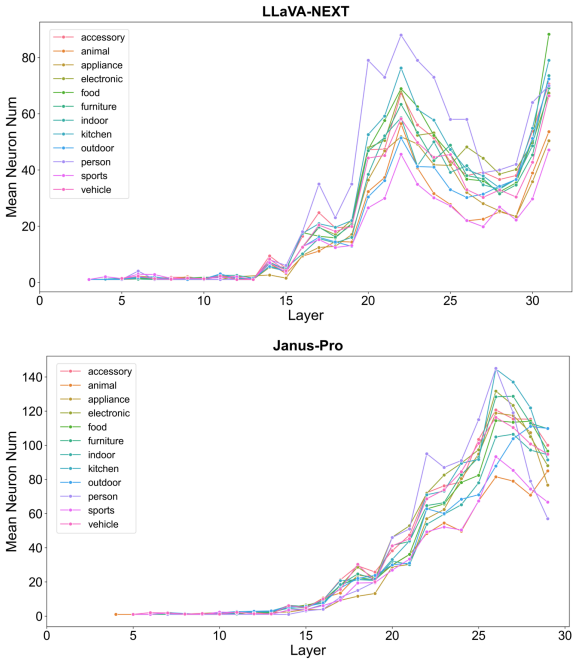


Figure 6: Layer-wise Distribution of multimodal neurons in fuzzy sets, averaged across sub-categories within each super-category per layer.

4.2.4 Semantic Activation Correlation

To examine the correlation of neuron activation patterns across different semantic concepts in MLLMs, we calculate Spearman's rank correlations (Spearman, 1961) based on neuron rankings by membership degree within the corresponding multimodal neuron fuzzy sets. As shown in Figure 5, concepts within the same super-category exhibit high inter-correlations, while those from different super-categories remain consistently low, indicating that models exhibit shared activation patterns for semantically related concepts while maintaining distinct patterns for unrelated concepts. We further provide additional examples in appendix B.3.

4.2.5 Layer-Wise Neuron Distribution

Figure 6 shows the distribution of multimodal neurons across layers. Both LLaVA-Next and Janus-

Models	Settings	COCO-QA				VQA-v2							
		what	where	how many	ALL	what	why	who	which	where	how	how many	ALL
LLaVA-NEXT	Baseline	66.53	43.50	72.81	64.51	57.46	32.96	62.48	48.92	50.25	**43.88**	70.48	58.85
	X-Fix	66.50	41.20	73.00	64.36	57.35	32.91	62.56	48.84	50.23	43.57	70.43	58.75
	X	66.66	42.94	**72.81**	64.59	57.55	**33.02**	62.57	49.04	50.29	43.82	70.50	58.92
	Q-Fix	64.39	41.89	70.20	62.29	57.42	32.95	62.36	48.64	50.34	43.71	70.35	58.79
	Q	66.63	**43.99**	72.39	64.59	57.60	32.96	62.73	49.00	50.41	43.85	70.51	58.96
	X+Q-Fix	64.28	39.83	70.12	62.11	57.25	32.95	62.47	48.65	50.03	43.74	70.35	58.66
	X+Q	**66.82**	43.54	72.28	**64.74**	**57.72**	33.01	**62.74**	**49.12**	**50.54**	43.80	**70.56**	**59.06**
Janus-Pro	Baseline	57.22	30.02	**71.22**	55.01	50.14	15.34	47.17	39.54	40.23	**31.73**	67.16	51.80
	X-Fix	57.29	30.15	70.95	55.06	50.13	15.18	47.19	39.72	40.25	31.48	67.23	51.80
	X	57.32	30.19	71.10	55.10	50.30	15.46	47.55	**39.75**	**40.31**	31.62	67.27	51.95
	Q-Fix	57.29	29.86	71.07	55.06	50.01	15.42	46.82	38.96	40.18	31.40	66.87	51.63
	Q	57.31	30.19	71.14	55.08	50.22	15.65	47.01	39.23	40.10	31.58	67.25	51.86
	X+Q-Fix	57.16	30.06	71.07	54.96	50.17	15.49	47.58	39.60	40.09	31.54	67.21	51.83
	X+Q	**57.33**	**30.35**	71.03	**55.12**	**50.34**	**15.72**	**47.64**	39.72	40.28	31.60	**67.31**	**51.98**

Table 4: Accuracy(%) for each question type and overall accuracy on VQA task, restricted to 5W2H question types. For LLaVA-Next, the scaling coefficients are set to $\lambda_x = 0.50$, $\lambda_q = 0.25$; for Janus-Pro, $\lambda_x = 0.02$, $\lambda_q = 0.02$.

Pro exhibit a low number of neurons in early layers, increasing gradually to a peak in higher layers, which is consistent with previous studies (Huo et al., 2024; Pan et al., 2023). Notably, LLaVA-Next shows a resurgence of neurons in the final layer, whereas Janus-Pro does not. We hypothesize that this difference is due to their distinct task orientations: LLaVA-Next focuses on multimodal understanding, enabling semantic multimodal neurons in the final layer to directly influence the language output distribution, whereas Janus-Pro is designed for both understanding and generation, which may require its final layer to serve as a unified output head, thereby limiting its specialization in semantic multimodal representations. The interpretability of models that unify multimodal understanding and generation will be explored in future work.

5 Application: VQA

To evaluate the practical benefits of our constructed fuzzy sets of multimodal neurons, we apply them to the VQA task. In typical VQA pipelines, the model takes an image and a question as input and outputs an answer. To improve model's semantic grounding during inference, we modulate the activation values of multimodal neurons corresponding to image-related semantics. To further capture question intent and facilitate cross-modal reasoning, we extend our method to the textual modality by identifying and modulating neurons associated with interrogative words (5W2H: what, why, who, which, where, how, how many). Specifically, we construct paired inputs: one with the original image-question pair, and the other with the same image and a declarative sentence formed by replacing the interrogative word with the answer. We then record neuron activations over the textual tokens for both inputs and compute the differences between their mean activations. For each question type, we use 500 paired inputs and construct the corresponding neuron fuzzy sets, following the method described in Section 3. Other experimental settings follow those used for multimodal semantic neurons. On average, the seven fuzzy sets of **interrogative neurons** contain 1,022 neurons (0.29%) in LLaVA-Next and 1,588 neurons (0.48%) in Janus-Pro. The layer-wise distribution of these neurons is presented in appendix B.4.

We conduct experiments on the COCO-QA (Ren et al., 2015) and VQA v2.0 (Goyal et al., 2017) datasets, both of which use images from MSCOCO. For each input pair, we select the multimodal neuron sets corresponding to the image semantics x and the interrogative neuron sets for the question type q. We then amplify their activations based on fuzzy membership degrees. Let $a(l,i)$ denote the original activation of neuron $u(l,i)$, and let $\mu_x(l,i)$, $\mu_q(l,i) \in [0,1]$ represent its membership degrees in the corresponding fuzzy sets. The amplified activation is computed as:

$$\hat{a}(l,i) = a(l,i) \cdot (1 + \lambda_x \mu_x(l,i) + \lambda_q \mu_q(l,i)) \,. \tag{9}$$

where λ_x and λ_q are scaling coefficients for the multimodal and interrogative neurons, respectively.

Image				
Question	What is the color of the bed?	What should the cars do?	Which fruit is lying on top?	How many dogs are outside?
Baseline	black	stop	apple	4
X+Q	**purple**	go	banana	5
-Modulate	'what', 'bed', 'person'	'what', 'car', 'traffic light'	'which', 'apple', 'banana', 'bowl'	'how many', 'dog', 'frisbee'

Table 5: Some examples where our X+Q setting generates correct answer while direct inference using LLaVA-Next produces incorrect answers. Here, *-Modulate* indicates the corresponding sets of neurons amplified during inference.

We evaluate the effect of neuron-level activation modulation under seven settings: **Baseline** (no neuron modulation), **X** and **Q** (amplifying only multimodal or interrogative neurons based on their fuzzy membership degrees), **X+Q** (joint amplification), and their corresponding comparison settings using a fixed membership degree of 0.5, namely **X-Fix**, **Q-Fix**, and **X+Q-Fix**. As shown in Table 4, joint amplification based on membership degrees achieves the best performance, demonstrating the effectiveness of fuzzy set-based neuron modulation. Table 5 presents some examples from LLaVA-Next. We further conduct experiments suppressing multimodal and interrogative neurons by setting their activations to 0. Results on COCO-QA using LLaVA-Next indicate that suppressing fewer than 1% of neurons leads to around a 10% performance drop, while suppressing the same number of random neurons shows negligible changes. This highlights the critical role the identified neurons play in the task. Exploring more effective utilization of these neurons remains a promising direction.

6 Related work

Neuron Analysis in Pre-Trained Transformers. Recent studies have revealed neuron-level functional specialization in pre-trained Transformers, establishing a foundation for interpretability and controllable behavior. In LLMs, Dai et al. (2021) identify knowledge neurons whose activations correlate with factual recall, while Wang et al. (2022) discover skill neurons responsible for task-specific behaviors such as translation task. Tang et al. (2024) focus on language-specific neurons, which Huo et al. (2024) and Huang et al. (2024) further expand to multimodal settings, uncovering domain-specific and modality-specific neurons in MLLMs. In addition, Schwettmann et al. (2023) identify multimodal neurons that reveal how LLMs convert visual representations into corresponding texts. Pan et al. (2023) introduce a new approach for identifying such neurons in MLLMs, highlighting their sensitivity, specificity, and causal-effect.

Development of MLLMs. Recent progress in MLLMs has advanced both vision-language understanding and image generation. For understanding, building on the design of LLaVA (Liu et al., 2023), researchers have developed a series of high-performing MLLMs (Liu et al., 2024a,b; Bai et al., 2023; Yang et al., 2024). To unify multimodal understanding and generation, some approaches combine LLMs with pre-trained diffusion models (Ge et al., 2023a,b; Sun et al., 2023); Wu et al. (2024) instead decouples visual encoding into separate paths for semantics and spatial details, and Chen et al. (2025) further enhance this design with scalable training and adaptive fusion.

7 Conclusion

In this paper, we propose a novel method to identify multimodal neurons in MLLMs and introduce fuzzy set theory to model their collaborative contributions to semantic concepts. We further successfully apply our method to the VQA task. Extensive quantitative and qualitative experiments demonstrate the explanatory powers and practical value of our fuzzy set-based multimodal neurons, advancing interpretability research in MLLMs.

Limitations

Our research has some limitations: (1) The effectiveness of the pipeline relies on the quality of semantic attribution maps generated by the diffusion model, and noisy attention may lead to inaccurate masks, potentially affecting neuron identification. (2) The utilization of the text-to-image generation process to obtain pixel-level attention matrices may introduce additional computational overhead, as analyzed in the Appendix. (3) The collaborative patterns among multiple neurons may involve more complex dynamical interactions, and relying solely on activation values to interpret these patterns may still fail to comprehensively reveal the decision-making mechanisms of the model. (4) The effectiveness and utility of the proposed method have been validated on specific tasks and datasets, while its applicability to other scenarios requires further investigation. We recognize these limitations as potential areas for future research.

Acknowledgments

This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (23YJAZH184) and the Fundamental Research Funds for the Central Universities in BLCU (21PT04).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tue M Cao, Nhat X Hoang, Hieu H Pham, Phi Le Nguyen, and My T Thai. 2025. Neurflow: Interpreting neural networks through neuron groups and functional interactions. *arXiv preprint arXiv:2502.16105*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023a. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023b. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Raquel González-Alday, Esteban García-Cuesta, Casimir A Kulikowski, and Victor Maojo. 2023. A scoping review on the progress, applicability, and future of explainable artificial intelligence in medicine. *Applied Sciences*, 13(19):10778.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *arXiv preprint arXiv:2410.04819*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Neha Kalibhat, Shweta Bhardwaj, C Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. 2023. Identifying interpretable subspaces in image representations. In *International Conference on Machine Learning*, pages 15623–15638. PMLR.

Diederik P Kingma, Max Welling, and 1 others. 2013. Auto-encoding variational bayes.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

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

Yunsheng Ma, Wenqian Ye, Can Cui, Haiming Zhang, Shuo Xing, Fucai Ke, Jinhong Wang, Chenglin Miao, Jintai Chen, Hamid Rezatofighi, and 1 others. 2025. Position: Prospective of autonomous driving-multimodal llms world models embodied intelligence ai alignment and mamba. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1010–1026.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2023. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2):5.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. 2023. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2862–2867.

Charles Spearman. 1961. The proof and measurement of association between two things.

Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. 2024. Crafting large language models for enhanced interpretability. *arXiv preprint arXiv:2407.04307*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Zeping Yu and Sophia Ananiadou. 2024. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. *arXiv preprint arXiv:2409.14144*.

Lotfi Asker Zadeh. 1965. Fuzzy sets. *Information and control*, 8(3):338–353.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Implementation Details

A.1 Synthesizing Images and Obtaining Pixel-Level Semantic Attribution Maps

When generating images using the diffusion model, the number of denoising steps is set to 50, and the seed of the generator is set to 42. During this process, to achieve pixel-level alignment between the global heatmaps and the generated images, each raw attention map is upsampled to (336, 336) using bicubic interpolation, resulting in a unified token-to-pixel attention representation.

A.2 Perturbation in the Visual Encoder

As shown in Figure 7, we take the Vision Transformer architecture as an example and apply the masking module before the transformer encoder. Specifically, for each binary mask matrix, we perform average pooling to downsample it to (24, 24), where each 14×14 region corresponds to an image patch. The average value of each region determines whether the corresponding image patch is kept or perturbed. The resulting mask is then reshaped to match the patch-embedding dimensions, and the corresponding image patches are replaced with noise accordingly.

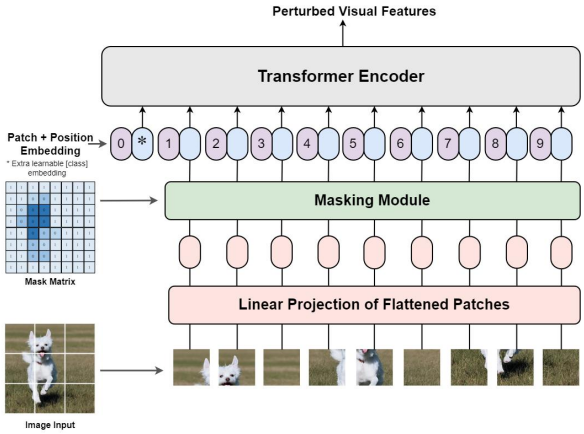


Figure 7: The Vision Transformer architecture with the masking module applied.

B More Experiment Results

We present supplementary experimental results and case studies to strengthen the support for our conclusion.

B.1 Textual Semantics of Neurons

As shown in Table 6, our method achieves competitive scores on multi-token concepts. Table 7 shows some examples.

Model	Method	BS	MS	LB
LLaVA-NEXT	Ours	**0.411**	**0.551**	**0.335**
Janus-Pro	Ours	**0.393**	**0.547**	**0.297**

Table 6: Results of BERTScore(BS), MoverScore(MS), and LaBSE(LB) on multi-token concepts. For each semantic concept, we select top 5% multimodal neurons and compute the average scores across all concepts.

B.2 Visual Focus of Neurons

We present additional examples of both single-token and multi-token concepts in Table 8, reporting their heatmaps, binary masks, and IOU results.

B.3 Semantic Activation Correlation

Figure 8 shows another example to support the observation of neuron activation patterns.

B.4 Layer-Wise Distribution Comparison between Multimodal Neurons and Interrogative Neurons

We compare the layer-wise distributions of multimodal neurons and interrogative neurons across LLaVA-Next and Janus-Pro. For each semantic concept or interrogative word, the top 500 neurons in the corresponding fuzzy set are selected for visualization. As shown in Figure 9, both models exhibit similar trends in the layer-wise distribution of multimodal neurons and interrogative neurons. Compared to multimodal neurons, interrogative neurons begin to emerge in large numbers at intermediate layers and show a sharp increase at the final layers. This suggests that the models start encoding question semantics in the middle layers while the deeper layers are primarily responsible for the integration of conceptual semantics.

C Expansion of Limitation

To further elaborate on the limitation regarding computational overhead, we quantify the computa-

tional requirements of our method using LLaVA-Next on an RTX 4090 GPU under two settings:

1) **Per-inference cost.**

   This setting includes synthesizing a single image and performing two forward passes with the MLLM (original and perturbed).

   - Image synthesis: ~81.907 TFLOPS (CUDA time: ~1.375 seconds)
   - MLLM inference ($\times 2$): ~81.124 TFLOPS (CUDA time: ~3.698 seconds)
   - Total per-inference: ~163.031 TFLOPS (CUDA time: ~5.073 seconds)

2) **Per-concept cost.**

   This setting measures the cost of identifying multimodal neurons for a given semantic concept, requiring N image synthesis steps and 2*N MLLM inferences. With N = 200, the total computation amounts to ~32,606.2 TFLOPS and the inference runtime reaches ~1,014.6 seconds on an RTX 4090.

Although our method incurs a non-trivial computational cost, it remains within an acceptable range — requiring only **5.073 seconds** per inference on a consumer-grade GPU. Moreover, the cost is one-time, and the identified multimodal neurons offer lasting benefits for downstream tasks such as VQA.

Semantic concept: sports ball			
Model	**Method**	**Top Neuron**	**Top Tokens**
LLaVA-NEXT	Ours	**u(20,512)**	[‘ball’, ‘ball’, ‘balls’, ‘Ball’]
		u(25,4873)	[‘s’, ‘soc’, ‘fut’, ‘vol’]
		u(23,11005)	[‘ball’, ‘throw’, ‘throws’, ‘aim’]
Janus-Pro	Ours	**u(23,4402)**	[‘球迷’, ‘足球’, ‘football’, ‘soccer’]
		u(24,9614)	[‘field’, ‘fields’, ‘field’, ‘Field’]
		u(27,1085)	[‘赛事’, ‘athletes’, ‘球迷’, ‘竞技’]
Semantic concept: traffic light			
Model	**Method**	**Top Neuron**	**Top Tokens**
LLaVA-NEXT	Ours	**u(31,6588)**	[‘Sign’, ‘signs’, ‘Sign’, ‘sign’]
		u(20,7139)	[‘Exit’, ‘exit’, ‘Exit’, ‘exit’]
		u(21,570)	[‘blue’, ‘yellow’, ‘red’, ‘green’]
Janus-Pro	Ours	**u(28,10732)**	[‘Traffic’, ‘traffic’, ‘traffic’, ‘Traffic’]
		u(26,2579)	[‘lights’, ‘灯光’, ‘LED’, ‘灯’]
		u(28,4444)	[‘lit’, ‘Lit’, ‘lit’, ‘lit’]

Table 7: The examples of the concept ***sports ball*** and ***traffic light***, showing the top-3 neurons selected by our method and their associated top-4 activation tokens.

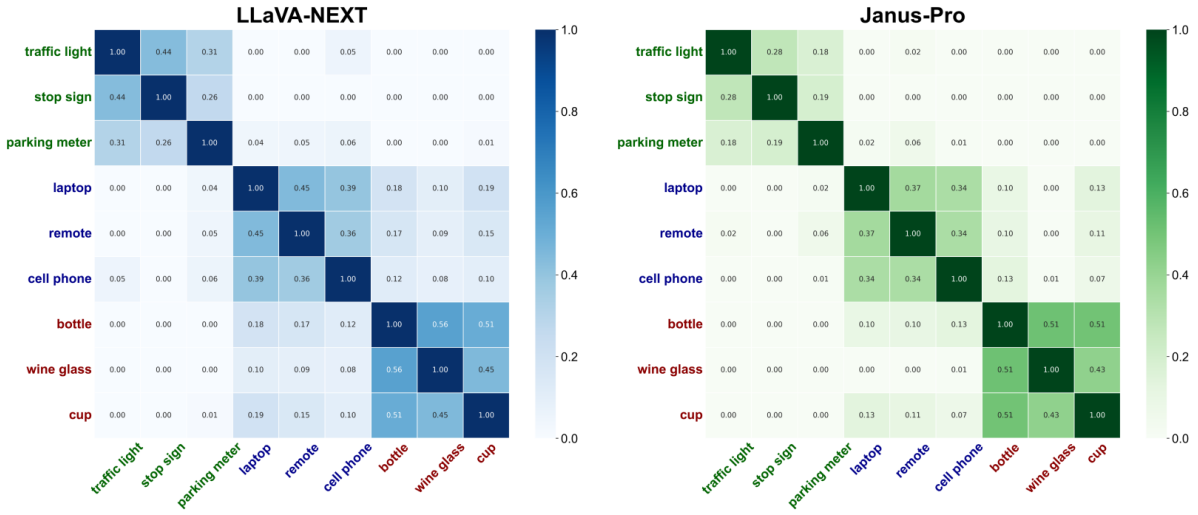


Figure 8: Another example of Spearman’s rank correlations between different semantic concepts.

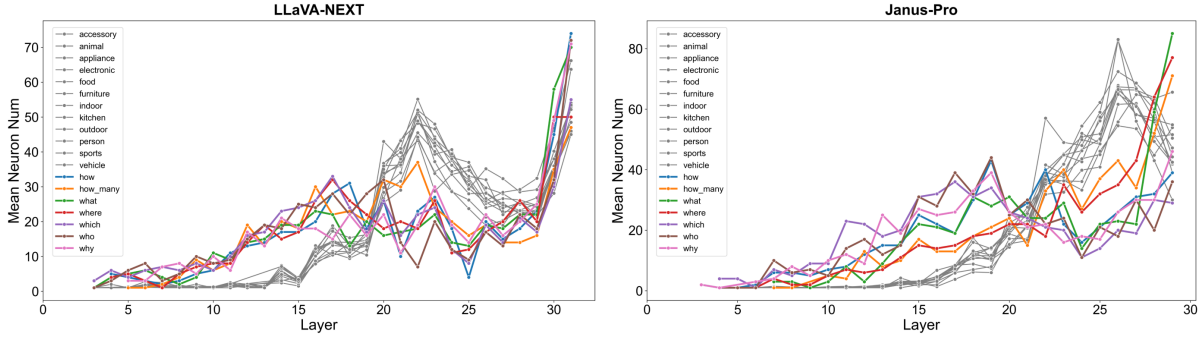


Figure 9: Layer-Wise Distribution Comparison between Multimodal Neurons and Interrogative Neurons. Grey lines indicate multimodal neurons, and coloured lines correspond to interrogative neurons.

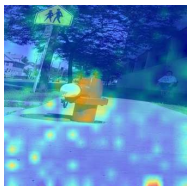
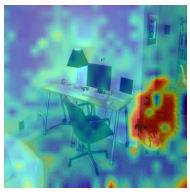
Semantic Concept	COCO ID	Original Image	Heatmap	Binary Mask	IOU
bird	508602				**0.586**
car	180878				**0.608**
clock	512248				**0.566**
pizza	80932				**0.644**
fire hydrant	38048				**0.466**
hot dog	42286				**0.514**
potted plant	37740				**0.553**
teddy bear	92091				**0.586**

Table 8: Heatmap, binary mask, and IoU results for more concepts in example images.