Finding and Editing Multi-Modal Neurons in Pre-Trained Transformers

Anonymous ACL submission

Abstract

Understanding the internal mechanism of multi-modal large language models (LLMs) is becoming increasingly critical for continuous improvements in both academia and industry. In this paper, we propose a novel method to identify key neurons for interpretability — how multi-modal LLMs bridge visual and textual concepts for captioning. Our method improves conventional works upon efficiency and applied range by removing needs of costly gradient computation. Based on those identified neurons, we further design a multi-modal knowledge editing method, beneficial to mitigate sensitive words or hallucination. For rationale of our design, we provide theoretical assumption. For empirical evaluation, we have conducted extensive quantitative and qualitative experiments. The results not only validate the effectiveness of our methods, but also offer insightful findings that highlight three key properties of multi-modal neurons: sensitivity, specificity and causal-effect, to shed light for future research. We will release code upon acceptance.

1 Introduction

Recently, large language models (LLMs) have received much attention and become foundation models in many natural language processing applications (Touvron et al., 2023a; Taori et al., 2023; Chiang et al., 2023; Geng et al., 2023). Following the success, researchers in the area of computer vision have extended the input modality to both text and image, namely multi-modal LLMs, showing remarkable performance in various visual understanding tasks (Liu et al., 2023; Dai et al., 2023; Ye et al., 2023a,b). However, the underlying mechanism of how multi-modal LLMs interpret different modalities of features beyond these tasks remains unclear. It hinders in-depth investigation and poses risks in model applications, such as producing misleading outputs without insight into decisions or propagating biases through automatic captions.

There are two main types of methods on LLMs' interpretability. The first group targets probing various abilities through well-designed external tasks (Olsson et al., 2022; Merullo et al., 2023; Huang et al., 2023; Duan et al., 2023). Another line of works, instead, attempt to reveal the internal states, by finding the processes of how LLMs understand and interpret textual inputs to form a response (Meng et al., 2022, 2023; Dai et al., 2022; Merullo et al., 2023). Among them, an interesting finding shows that LLMs' ability to understand textual information mainly comes from feed-forward networks (FFNs). Furthermore, Schwettmann et al. (2023) identify key neurons from FFNs, namely multi-modal neurons. These neurons play an important role in understanding images and generating textual descriptions. However, the identification process is inefficient and limited in applied range, due to costly gradient computation. Besides, their theoretical rationale, empirical characteristics, and potential application remains under-exploration.

To address the issues, we propose a novel method for multi-modal neuron identification. We define a contribution score based on the activation output in FFNs, which is consistent with the probability distribution when predicting. As our method do not need access to the model gradients, we improve efficiency while ensuring effectiveness.

Based on the identified neurons, we further propose a multi-modal knowledge editing method as a potential application. We achieve the goal of editing a specific concept to another designative concept (e.g., in Figure 1(i), 'dog' is edited to 'mouse'), by changing the probability distribution of outputs. Without additionally training the entire model or requiring access to model gradients, our method allows for an efficient, timely and resource-efficient editing of little part of the model parameters.

For empirical characteristics, we have designed metrics and conducted extensive experiments, which highlight three critical properties of multi-

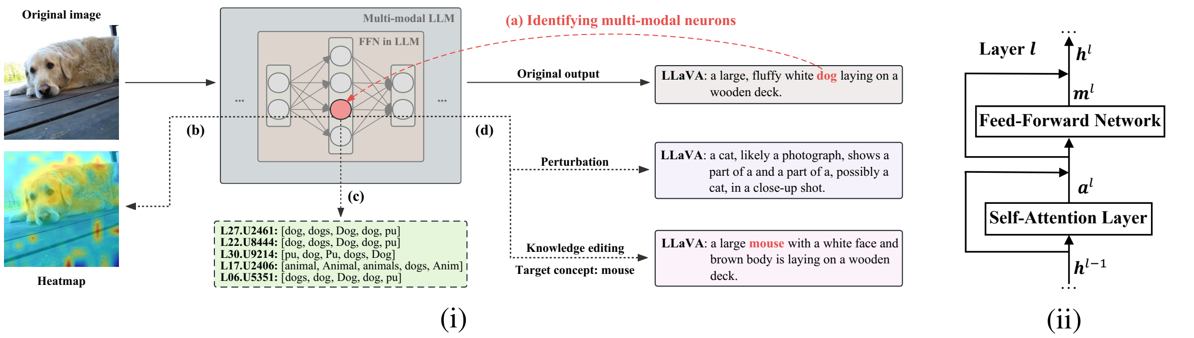


Figure 1: (i) Multi-modal neurons in FFN within multi-modal LLM. We develop a method to **(a)** identify multi-modal neurons and confirm that they can encode specific concepts from **(b)** images to **(c)** texts and **(d)** causally affect model output. (ii) Architecture of layer l in Transformer-based LLM.

modal neurons: (1) **Sensitivity** (§3.3). Multi-modal neurons are sensitive to particular concepts. Once they are activated by some regions of the input image, they are responsible for generating related textual concepts. More importantly, these neurons are invariant in visual translation to different inputs. (2) **Specificity** (§3.4). Although different multi-modal neurons can be activated by the same concepts, they are selectively active for these concepts and hardly respond to others. (3) **Causal-Effect** (§3.5). Multi-modal neurons and the associated concepts have causal-effect and are significantly susceptible. We perturb and edit the identified multi-modal neurons, which leads to significant changes in outputs.

Our contributions can be summarized as follows:

- We propose a new method for identifying multi-modal neurons in Transformer-based multi-modal LLMs.
- We propose a multi-modal knowledge editing method based on the multi-modal neurons.
- We highlight three critical properties of multi-modal neurons by designing four quantitative evaluation metrics and extensive experiments.

2 Method

We first define neurons in the LLM (§2.1), and then define a contribution score for neurons identification (§2.2). Furthermore, we propose a multi-modal knowledge editing method based on identified neurons (§2.3) and introduce several evaluation metrics to evaluate multi-modal neurons (§2.4).

2.1 Neurons in Transformer-Based LLM

A multi-modal LLM typically consists of an image encoder, a textual LLM, and an adaptor to align the above two modules. Following previous works (Dai et al., 2022; Wang et al., 2022; Schwettmann et al., 2023), we research neurons within FFNs in textual LLM, as they carry two-thirds of the parameters and are proven to play a critical role in understanding textual and visual features. As illustrated in Figure 1(ii), we denote the hidden states at layer l as $\mathbf{h}^l$, FFN output as $\mathbf{m}^l$ and self-attention output as $\mathbf{a}^l$, respectively. And $\mathbf{m}^l$ can be calculated by:

$$\mathbf{m}^l = \mathbf{W}_{\text{out}}^l \, \sigma\left(\mathbf{W}_{\text{in}}^l \left(\mathbf{a}^l + \mathbf{h}^{l-1}\right)\right) \, , \quad (1)$$

where $\mathbf{h}^0$ is embedding vector of input, σ is activation function, $\mathbf{W}_{\text{in}}^l$ is the first linear layer and $\mathbf{W}_{\text{out}}^l$ is the second linear layer in FFN. And we omit the normalization in Eq. 1 for the sake of brevity.

For simplicity, let $\mathbf{O}^l = \sigma\left(\mathbf{W}_{\text{in}}^l \left(\mathbf{a}^l + \mathbf{h}^{l-1}\right)\right)$, where the i-th element is the activation output of the i-th neuron. We denote each neuron in the LLM as (Ll.Ui) in subsequent experiments. For instance, (L20.U188) denotes the 188-th neuron at layer 20.

2.2 Identifying Multi-Modal Neurons

We now propose a contribution score that indicates a neuron's contribution to a modal-independent concept. That is, if the score is high, the neuron should be activated with a high probability when taking in the visual concept and generating the textual concept. We first formally define the computational method for it and then prove its validity.

Let $\mathcal{M}$ be the LLM, $\mathbf{x}$ be the sequence of input tokens and $\mathbf{y}$ be the output sequence. The function of LLM can be written as: $\mathbf{y} = \mathcal{M}(\mathbf{x})$.

We assume the model is about to output token $t \in \mathbf{y}$, whose probability is maximum among the vocabulary. Then we define the contribution score of the neuron u_i at layer l to the token t as $s_{i,t}^l$:

$$s_{i,t}^l = \mathbf{Q}^l(i, t) \, , \quad (2)$$

where $\mathbf{Q}^l = \mathbf{W}_u \mathbf{W}_{\text{out}}^l \circ \mathcal{T}\left(\mathbf{O}_{-1}^l\right) \in \mathbb{R}^{d_m \times v}$, $\mathbf{W}_u$ is the unembedding matrix to decode last hidden states, $\mathcal{T}(\cdot)$ is the transpose of the input matrix, $\mathbf{O}_{-1}^l$ is activation output at the last token, d_m is intermediate size, v is vocab size and $\circ$ is an element-wise product with broadcasting mechanism.

To prove rationality and effectiveness of Eq. 2 and explain why we define $\mathbf{Q}^l$ in the manner described above, we try to disassemble and deduce the generation procedure of LLM. When a L layer LLM is generating a new token $t \in \mathbf{y}$, the probability distribution of output can be denoted as follows:

$$\begin{aligned} t &= \operatorname{argmax}\left(\mathbf{W}_u \mathbf{h}_{-1}^L\right) \\ &= \operatorname{argmax}\left(\mathbf{W}_u\left(\mathbf{a}_{-1}^L + \mathbf{m}_{-1}^L + \mathbf{h}_{-1}^{L-1}\right)\right) \\ &= \operatorname{argmax}\left(\sum_{l=1}^{L}\left(\mathbf{W}_u \mathbf{m}_{-1}^l + \mathbf{W}_u \mathbf{a}_{-1}^l\right)\right. \\ &\qquad\qquad \left. + \mathbf{W}_u \mathbf{h}_{-1}^0\right) \\ &= \operatorname{argmax}\left(\sum_{l=1}^{L}\left(\mathbf{W}_u \mathbf{W}_{\text{out}}^l \mathbf{O}_{-1}^l + \mathbf{W}_u \mathbf{a}_{-1}^l\right)\right. \\ &\qquad\qquad \left. + \mathbf{W}_u \mathbf{h}_{-1}^0\right), \end{aligned} \tag{3}$$

where $\mathbf{W}_u$ is the unembedding matrix, $\mathbf{h}_{-1}^L$ is the output of the last token at the last layer L, and $\mathbf{O}_{-1}^l = \sigma\left(\mathbf{W}_{\text{in}}^l\left(\mathbf{a}_{-1}^l + \mathbf{h}_{-1}^{l-1}\right)\right) \in \mathbb{R}^{d_m}$ is activation function output at the last token at layer l.

In Eq. 3, $\mathbf{W}_u \mathbf{W}_{\text{out}}^l \mathbf{O}_{-1}^l$ represents FFN part and $\mathbf{W}_u \mathbf{a}_{-1}^l$ represents self-attention part. Following §2.1, we empirically focus on the FFN and omit the remaining parts. We regard o_i^l, the i-th element of $\mathbf{O}_{-1}^l$, as the activation of the i-th neuron at the last token at layer l, and $\mathbf{W}_u \mathbf{W}_{\text{out}}^l$ as a new unembedding matrix at each layer. The function of $\mathbf{W}_u \mathbf{W}_{\text{out}}^l$ is to project the activation of the neurons onto a distribution of the token vocabulary.

To further evaluate the individual contribution of each neuron, we disassemble the matrix multiplication of $\mathbf{W}_u \mathbf{W}_{\text{out}}^l$ and $\mathbf{O}_{-1}^l$ in Eq. 3 as follows:

$$\mathbf{W}_u \mathbf{W}_{\text{out}}^l \mathbf{O}_{-1}^l = \sum \mathcal{T}\left(\mathbf{W}_u \mathbf{W}_{\text{out}}^l \circ \mathcal{T}\left(\mathbf{O}_{-1}^l\right)\right), \tag{4}$$

where $\sum(\cdot)$ represents summing rows of the input.

Now we can see $\mathbf{Q}^l$ in Eq. 4, which is consistent with the probability distribution when predicting. We regard $\mathbf{Q}^l(i,j)$ as a contribution score that the i-th neuron at layer l contributes to the j-th token. We provide a more detailed explanation in Appendix A.

Algorithm 1: Knowledge Editing

Data: Source token t_0, target token t_1, neurons set $\mathcal{S}$, model $\mathcal{M}$, unembedding matrix $\mathbf{W}_u$, penalty weight β, learning rate α, epochs ϵ

Result: Edited model $\tilde{\mathcal{M}}$

```
for s_j ∈ S do
    l, i ← location of s_j;
    o_i^l ← activation function output of s_j;
    w ← i-th row of W_out^l;
    v_0 ← t_0-th column of W_u;
    v_1 ← t_1-th column of W_u;
    initialize Δw;
    w' ← w + Δw;
    loss ← o_i^l(w'v_0 − w'v_1) + β · ||Δw||_2;
    Δw* ← gradient descent(Δw, loss, α, ε);
    W̃_out^l ← add Δw* to the i-th row of W_out^l;
    M̃ ← replace W_out^l with W̃_out^l in M;
end
return M̃;
```

Based on Eq. 2, we compute the score of each neuron for every **noun** token in the model output. Then we rank all scores of neurons across all layers within the model by the descending order and regard the top neurons as multi-modal neurons. Implementation details can be found in Appendix B.1

2.3 Multi-Modal Knowledge Editing

Following previous works (Mitchell et al., 2022; Meng et al., 2022, 2023) on unimodal knowledge editing, we aim at controlling the textual output. In specific, our goal is to replace a source token with a target token in the output without changing the remaining content. We propose an algorithm (see Algorithm 1) to intervene some parameters based on the identified multi-modal neurons.

We denote top multi-modal neurons of source token t_0 as $\mathcal{S}$. For each multi-modal neuron $s_j \in \mathcal{S}$, we first get its location (l, i), which means the i-th neuron at layer l, and then we record its activation function output o_i^l. Let $\mathbf{w}$ be the i-th row of $\mathbf{W}_{\text{out}}^l$, $\mathbf{v}_0$ be the t_0-th column of $\mathbf{W}_u$, $\mathbf{v}_1$ be the t_1-th column of $\mathbf{W}_u$ and $\mathbf{w}'$ be the edited $\mathbf{w}$, respectively.

Our goal is to prompt the probability of generating token t_1 higher than token t_0, which is equivalent to make $o_i^l \mathbf{w}' \mathbf{v}_1$ larger than $o_i^l \mathbf{w}' \mathbf{v}_0$, so we define a loss function as below:

$$\text{loss} = o_i^l(\mathbf{w}'\mathbf{v}_0 - \mathbf{w}'\mathbf{v}_1) + \beta \cdot ||\Delta \mathbf{w}||_2, \tag{5}$$

where β is penalty weight and $||\Delta \mathbf{w}||_2$ is a L_2-norm constraint as a penalty to avoid the editing is too drastic and affects generating other tokens.

By applying Gradient Descent (Robbins and Monro, 1951), we acquire an optimal $\Delta \mathbf{w}^*$. We then add $\Delta \mathbf{w}^*$ to the i-th row of $\mathbf{W}_{\text{out}}^l$ and replace the original $\mathbf{W}_{\text{out}}^l$ with the new $\mathbf{W}_{\text{out}}^l$ in model $\mathcal{M}$.

Note that our algorithm is independent from the model, and the solution procedure does not need to additionally train or infer the entire model. Accordingly, this allows for an efficient, timely and resource-efficient editing of the model parameters.

2.4 Evaluation Metrics

After identifying multi-modal neurons, in order to comprehensively evaluate the effectiveness of them with quantitative indicators, we measure several evaluation metrics from multiple perspectives.

Semantic Sensitivity To verify if neurons are sensitive to textual concepts, we align neurons with natural language. The more similar the top tokens are to the textual concept, the more sensitive the neurons are. Therefore, we measure BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019) and BLEURT (Sellam et al., 2020) between each textual concept and top-10 tokens that corresponding neurons represent.

Region Invariance To verify if neurons are sensitive to visual concepts, we measure the proportion of invariant neurons when shuffling the image patches. Specifically, for each textual concept in each image, we denote the original top-k multi-modal neurons as $\mathcal{S}_k$. We randomly shuffle the input sequence of image patches of LLM, and equally identify top-k multi-modal neurons, denoted as $\mathcal{S}'_k$. A higher degree of similarity between $\mathcal{S}_k$ and $\mathcal{S}'_k$ indicates stronger region invariance. We calculate the ratio of invariant neurons as below:

$$r_k = \frac{|\mathcal{S}_k \cap \mathcal{S}'_k|}{|\mathcal{S}_k|}\,, \tag{6}$$

and record a mean score across all images.

Cross-Images Invariance We aim at figuring out whether the same neurons would be identified in different images, which is called cross-images invariance. We randomly select N different images from the dataset that all contain a given concept c. Then, we separately identify the top-k neurons of these images and pick out neurons in common. We calculate the ratio of common neurons by:

$$s_{\text{CII}} = \frac{|\mathcal{S}_k^1 \cap \mathcal{S}_k^2 \cap \cdots \cap \mathcal{S}_k^N|}{k}\,, \tag{7}$$

where $\mathcal{S}_k^j$ is top-k multi-modal neurons of image j.

Specificity We then verify if neurons are specific to textual concepts — only activated for some related tokens, but inactivated for other tokens. Formally, we pick out n images, and separately identify their top-1 multi-modal neuron, denoted as $\mathcal{S}$. For each neuron (l, i) in $\mathcal{S}$, we provide a set of concepts T, where $|T| = m$, and calculate scores to each of them. Then we record a mean score across neurons in $\mathcal{S}$ and concepts in T, denoted as S@m:

$$\text{S@}m = \frac{1}{n \cdot m} \sum_{(l,i)\in\mathcal{S}} \sum_{t\in T} s_{i,t}^l\,. \tag{8}$$

We choose two sets of concepts T: related concepts and random concepts. Related concepts are concepts with top probability to each neuron in $\mathcal{S}$, while random concepts are randomly selected from the vocabulary. If multi-modal neurons possess specificity, scores to related concepts will significantly outperform those to random concepts.

3 Experiments

3.1 Investigation Setup

We use LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023) and mPLUG-Owl2 (Ye et al., 2023b) as our research models, three widely-use models for visual semantic understanding task. And we conduct all experiments on 1000 images that are randomly sampled from SBU Captions Dataset (Ordonez et al., 2011), a dataset consists of more than 1 million images from Flickr. We compare our method with Multimodal Neurons (abbreviated as Mmns) (Schwettmann et al., 2023), a technique for detecting *multimodal neurons* that map visual features to corresponding text. Furthermore, we establish a baseline (abbreviated as Base) that simply selects neurons with higher activations at the last token for basic comparison. Details about the implementations can be found in appendix B.1.

3.2 Identifying Multi-Modal Neurons

We employ methodology described in §2.2 to identify multi-modal neurons. Figure 2 shows the distribution of unique multi-modal neurons. We can see that our multi-modal neurons widely occur in higher layers, which is consistent with previous works (Wang et al., 2022; Dai et al., 2022).

3.3 Are Multi-Modal Neurons Sensitive to Certain Concepts?

We now discuss whether multi-modal neurons are sensitive to certain concepts from four perspectives: (1) Whether multi-modal neurons correspond to **visual** concepts (§3.3.1). (2) Whether multi-modal

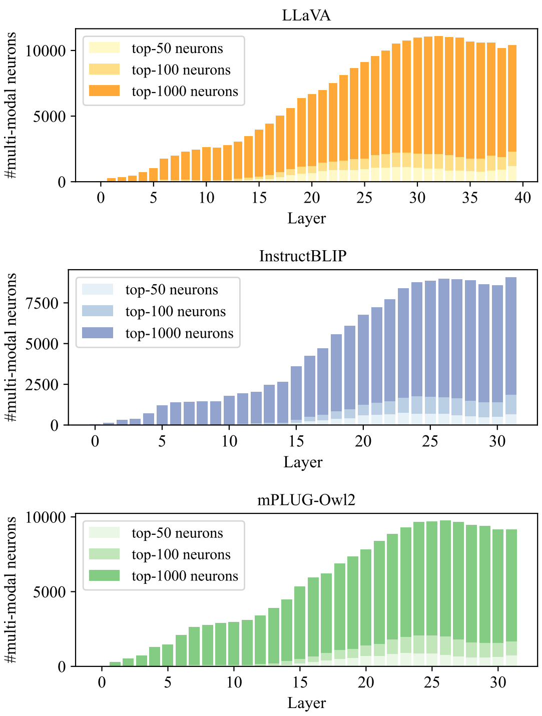


Figure 2: Distribution of unique multi-modal neurons per layer, chosen by different number of neurons with top contribution scores for each image.

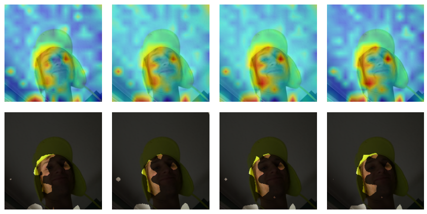
Image & Original output				
	LLaVA: a **man** wearing a yellow **hat** and smiling.			
Concept	**Heatmap & Binary mask**			
	Top-1	**Top-10**	**Top-100**	**Top-1000**
man				
hat				

Table 1: Heatmap and binary mask results of an example image. We plot each heatmap by using scaled mean activations across top-k neurons, where $k = 1, 10, 100, 1000$, and plot binary mask by thresholding mean activations above the 95% percentile, respectively.

neurons correspond to **textual** concepts (§3.3.2). (3) Whether the correspondence between multi-modal neurons and semantic concepts remains constant despite changes in the **same** image (§3.3.3). (4) Whether the correspondence between multi-modal neurons and semantic concepts remains constant despite changes in **different** images (§3.3.4).

3.3.1 Tracing Focus of Neurons in Images

We take the activations of multi-modal neurons at image patch tokens, scale them by bilinear interpolation, and plot the heatmap and binary mask. Implementation details are shown in appendix B.2. As the square root of the number of image patch tokens in InstructBLIP and mPLUG-Owl2 is irrational, we only conduct experiments on LLaVA. Table 1 shows an example. We can see that multi-modal neurons mainly focus on image regions that containing corresponding concepts, and pay less attention to other unrelated area. They reliably highlight the semantically pertinent areas throughout.

3.3.2 Textual Meanings of Neurons

We then verify whether our multi-modal neurons can represent textual meanings. Considering the multiplication of the unembedding matrix and the second layer of FFN is regarded as a projection from the activation of the neurons to probability distributions of the token vocabulary, we empirically sort rows correspond to multi-modal neurons and pick out the top-10 tokens as each neuron represents. We report an example in Table 2. We can find that the baseline and Mmns choose the neurons that are hardly correlated with concepts, whereas our method can more precisely identify neurons representing semantic meanings in comparison to them. More examples are shown in appendix C.2.

To provide stronger evidence, we measure metrics of semantic sensitivity mentioned in §2.4. Table 3 shows the mean results. Our method achieve higher scores than Mmns and baseline, which demonstrates that our selected neurons are more consistent with corresponding concepts.

3.3.3 Region Invariance of Neurons

To quantify the region invariance of the neurons, we calculate the ratio of invariant neurons in top-k neurons when shuffling (see Eq. 6). The mean results are shown in Figure 3. Our method significantly receives higher ratios of the invariant neurons than Mmns, which indicates our selected multi-modal neurons possess a stronger region invariance.

Image	Model	Method	Top neurons	Top tokens
 LLaVA: a **church** with a steeple, surrounded by snow, is captured in the photo. InstructBLIP: a **church** with snow on the ground. mPLUG-Owl2: a **church** with a person shoveling snow in front of it.	LLaVA	Base	**L39.U212** **L24.U5916** **L39.U5925**	[‘’, ‘1’, ‘-’, ‘\n’, ‘(’] [‘arin’, ‘Kennedy’, ‘dy’, ‘dy’, ‘PF’] [‘ ’, ‘—-’, ‘—–’, ‘________’, ‘____’]
		Mmns	**L24.U10906** **L9.U4426** **L20.U3864**	[‘dex’, ‘igung’, ‘nomin’, ‘pill’, ‘pill’] [‘,’, ‘.’, ‘bird’, ‘bird’, ‘-’] [‘oka’, ‘backwards’, ‘рем’, ‘iono’, ‘차’]
		Ours	**L31.U9192** **L34.U8761** **L39.U9669**	[‘church’, ‘Church’, ‘churches’, ‘Kirche’, ‘Kirchen’] [‘religious’, ‘Relig’, ‘relig’, ‘religion’, ‘Catholic’] [‘Church’, ‘Luther’, ‘Bishop’, ‘Orth’, ‘church’]
	InstructBLIP	Base	**L31.U10656** **L31.U7742** **L31.U6024**	[‘:(’, ‘:-)’, ‘:)’, ‘anyway’, ‘solves’] [‘restored’, ‘Accessor’, ‘overwrite’, ‘reuse’, ‘:’] [‘textt’, ‘archivi’, ‘zvuky’, ‘tématu’, ‘lês’]
		Mmns	**L28.U2212** **L4.U10613** **L17.U3575**	[‘etwork’, ‘окру’, ‘⋆’, ‘ ’, ‘Dob’] [‘Хронологија’, ‘Archivlink’, ‘↩’, ‘○’, ‘▶’] [‘’, ‘ ’, ‘Â’, ‘[...]’, ‘mals’]
		Ours	**L29.U7331** **L27.U7707** **L21.U1413**	[‘Church’, ‘church’, ‘churches’, ‘Kirche’, ‘Kirchen’] [‘Christ’, ‘christ’, ‘Christ’, ‘Christ’, ‘Christians’] [‘church’, ‘церков’, ‘churches’, ‘Church’, ‘Religion’]
	mPLUG-Owl2	Base	**L31.U1373** **L31.U7491** **L31.U1563**	[‘’, ‘in’, ‘\n’, ‘(’, ‘.’] [‘apparently’, ‘either’, ‘threaten’, ‘towards’, ‘storing’] [‘archivi’, ‘Kontrola’, ‘Хронологија’, ‘’, ‘’]
		Mmns	**L15.U8368** **L19.U1434** **L13.U420**	[‘yard’, ‘ill’, ‘go’, ‘mouse’, ‘ments’] [‘snow’, ‘ice’, ‘Snow’, ‘winter’, ‘Winter’] [‘church’, ‘Church’, ‘ric’, ‘cho’, ‘uti’]
		Ours	**L25.U911** **L29.U5136** **L31.U7266**	[‘faith’, ‘religion’, ‘relig’, ‘religious’, ‘Relig’] [‘Church’, ‘church’, ‘churches’, ‘Kirche’, ‘chiesa’] [‘religious’, ‘Relig’, ‘prayer’, ‘spiritual’, ‘pray’]

Table 2: An example result shown with top-3 neurons selected by different methods. We report results of the concept *church*. For each neuron, we record its top-5 relative tokens.

Model	Method	BS	MS	BRT
LLaVA	Base	0.236	0.664	0.086
	Mmns	0.652	0.678	0.100
	Ours	**0.794**	**0.730**	**0.214**
InstructBLIP	Base	0.626	0.656	0.071
	Mmns	0.339	0.663	0.089
	Ours	**0.726**	**0.706**	**0.160**
mPLUG-Owl2	Base	0.360	0.664	0.068
	Mmns	0.620	0.675	0.101
	Ours	**0.730**	**0.715**	**0.183**

Table 3: Results of metrics including BERTScore (BS), MoverScore (MS) and BLEURT (BRT). For each image, we select top-10 multi-modal neurons for each concept, and we record the mean metrics across all concepts. We ultimately calculate means across all images.

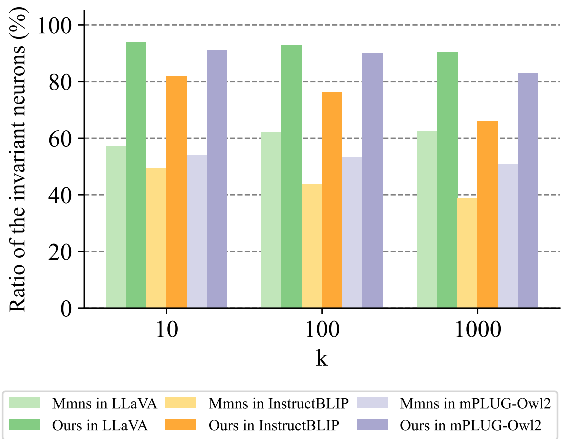


Figure 3: Ratios of the invariant neurons in top-k neurons before and after shuffling. For each image, we record the mean ratio across concepts that both exist in original caption and caption generated by shuffled image patches, and then calculate means across all images.

3.3.4 Cross-Images Invariance of Neurons

As for cross-images invariance, same neurons shall occur in different images that carry similar semantic information. To verify cross-images invariance of multi-modal neurons, we calculate the ratio of common neurons by Eq. 7. The results of Mmns and our method are shown in Figure 4. Our multi-modal neurons significantly outperform Mmns. Specifically, our method achieves common neuron ratios over 20% in LLaVA and mostly over 40% in InstructBLIP and mPLUG-Owl2, which is substantially higher than Mmns that attain ratios mainly under 10% in LLaVA, under 30% in InstructBLIP and under 20% in mPLUG-Owl2. We report more results with different N and k in appendix C.4.

3.4 Are Multi-Modal Neurons Specific?

For multi-modal neurons, claiming indiscriminate sensitivity to all concepts does not sufficiently demonstrate their functional role within the model. As such, we investigate their specificity. We record the scores of multi-modal neurons that correspond to their specific textual meanings when encoding other different concepts in the same image. Figure 5 shows an example. Additional examples are provided in appendix C.5. We can see that when encod-

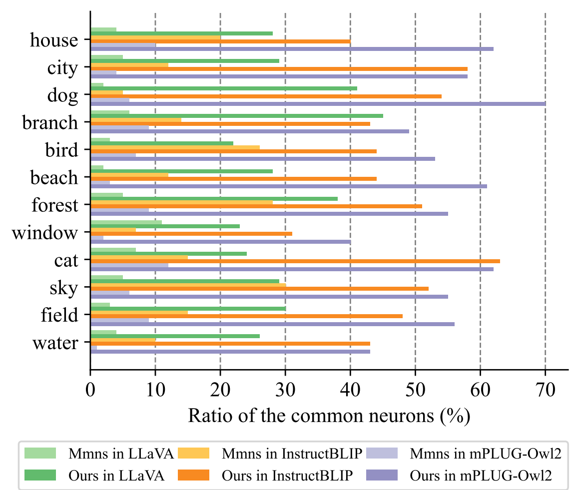


Figure 4: Ratios of the common neurons in top-100 neurons. We set $N = 5$ and report results of some concepts that frequently appear in sampled images.

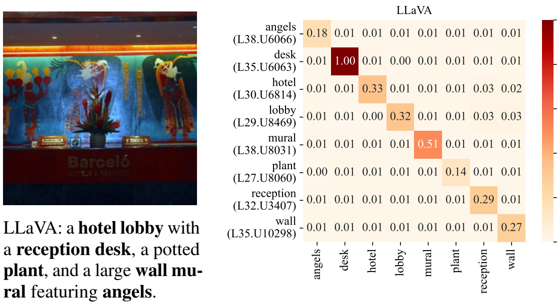


Figure 5: Heatmap of the scores (after normalization) of multi-modal neurons corresponding to specific concepts when encoding different contents in an example image. The x-axis represents concepts in the given image, and y-axis represents the top-1 neuron corresponding to each concept, respectively. Darker blocks indicate higher scores, which means higher relevance.

ing a specific concept, the top-1 multi-modal neuron receives a higher score than irrelevant concepts. We also adopt a metric to quantify the specificity of neurons (see §2.4). The results are shown in Table 4, from which we can find that neurons significantly get higher scores to those related concepts than to unrelated concepts, proving their specificity.

3.5 Do Multi-Modal Neurons Causally Affect Output?

Perturbation Study Previous works (Mitchell et al., 2022; Meng et al., 2022, 2023) have shown that applying directional editing to FFNs significantly change the model output. Inspired by these, we try to perturb multi-modal neurons. Specifically, for each concept in each image, we add a Gaussian noise ($\mu = 0$ and $\sigma = 0.5$) to the i-th row of the second layer of FFN at layer l. Table 5 shows an example when perturbing neurons in LLaVA. We can see that perturbing multi-modal neurons really makes a difference in model output, while simply perturbing few random neurons has no impact. Furthermore, we note that applying perturbation on neurons sometimes makes the corresponding token disappear in output and provides some new tokens, while sometimes results in meaningless output (e.g., in Table 5, when we perturb concepts 'leaves' and 'sky', the model can generate fluent output without 'leaves' and 'sky', but it is confused when we perturb concepts 'tree' and 'branches'). The former phenomenon piques our curiosity regarding the potential possibility that a well-designed alteration may substitute for Gaussian noise to enable knowledge editing of model output.

Knowledge Editing We hypothesize that replacing the Gaussian noise with an elaborate alteration can achieve a knowledge editing. Accordingly, we design an efficient algorithm (see Algorithm 1) that edits weights of the second layer of FFNs. Table

Model	Type	S@1	S@5	S@10	S@50
LLaVA	Related	3.549	2.920	2.333	0.467
	Random	0.018	0.012	0.014	0.003
InstructBLIP	Related	2.504	2.133	1.774	0.355
	Random	0.005	0.007	0.008	0.002
mPLUG-Owl2	Related	1.949	1.637	1.295	0.259
	Random	0.002	0.003	0.003	0.001

Table 4: Average scores that multi-modal neurons contribute to related concepts and random concepts. We report average scores with $m = 1, 5, 10, 50$, which are denoted as S@1, S@5, S@10 and S@50, respectively.

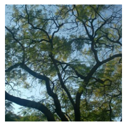
Concept	Perturbed model output
Image & Original output	 LLaVA: a **tree** with many **branches** and **leaves**, set against a blue **sky**.
tree	a Hamon's Garden, featuring a Hamon' the S the Hamon's Garden, featuring a Hamon's the S the Hamon's ...
branches	ameshupelageaameshupelageaamesh...
leaves	a tree with branches spread out, surrounded by tree branches and Homosassa, Florida, and the things around it.
sky	a tree with leaves, possibly a palm tree, with a large and sturdy trunk, surrounded by a large, vibrant, and colorful body of leaves.
random	a tree with many branches and leaves, set against a blue sky.

Table 5: Perturbation results of LLaVA. For each concept in the image, we only perturb the top-5 multi-modal neurons. For comparison, we report a result of perturbing the same number of random chosen neurons.

Image & Original output	
	LLaVA: a white **cat** sleeping in a tree. InstructBLIP: a white **cat** sleeping in a tree. mPLUG-Owl2: a white **cat** sleeping on a tree branch.

Model	Target	Edited model output
LLaVA	monkey	a white **monkey** sleeping in a tree.
	clock	a white **clock** sitting on a tree stump.
	iPhone	a white **iPhone** lying on a tree stump.
	food	a white **food** in a tree.
InstructBLIP	monkey	a white **monkey** sleeping in a tree.
	clock	a white **clock** sleeping in a tree.
	iPhone	a white **iPhone** 3Gs sitting on a tree stump.
	food	a white **food** sleeping in a tree.
mPLUG-Owl2	monkey	a white **monkey** sleeping on a tree branch.
	clock	a **clock** clocking in a tree trunk.
	iPhone	a white **iPhone** sitting on a tree branch.
	food	a white **food** food sleeping on a tree branch.

Table 6: Knowledge editing results of an example. We choose to edit concept *cat* to 4 target concepts. Target concepts are in bold in the edited model output.

Source concept: bird

Image	Target	Edited LLaVA's output
	None	a **bird** walking on the beach near the water.
	cat	a **cat** walking on the beach near the water.
	horse	a **horse** on the beach, walking through the water and enjoying the waves.
	None	a **bird**, possibly a pigeon, standing in a puddle of water on a city street.
	cat	a **cat** sitting in a puddle of water.
	horse	a **horse** in a pond, surrounded by leaves and water.
	None	a river flowing through a rocky area, with a waterfall and a rocky cliff.
	cat	a river flowing through a rocky area, with a waterfall and a rocky cliff.
	horse	a river flowing through a rocky area, with a waterfall and a rocky cliff.

Table 7: Edited LLaVA's output of different images. We select *bird* as source concept, choose *cat* and *horse* as target concept (*None* means no editing), and modify model parameters based on image (a). We then test the edited model on another two images, where image (b) contains the source concept *bird* and image (c) doesn't.

6 shows an example, where we guide the model to generate a different concept from the original concept. We find that model drops the source concept and successfully generates the target concept, which did not appear in original output. To prove effectiveness of our method, we evaluate the edited model on other different images, as shown in Table 7. We find that when we input another image that contains the same source concept, the edited model will identify it and generate the target concept, while an unrelated image will not be affected.

4 Related Work

Identifying Neurons in Deep Neural Networks There has been growing interest in interpreting and analyzing the inner workings of deep neural networks. Prior works have sought to characterize what types of information are encoded in individual neurons. Koh et al. (2020) proposes a technique for identifying "concept neurons" that detect semantic concepts in vision models. Dai et al. (2022) discusses the discovery of "knowledge neurons" which encode specific commonsense knowledge automatically learned during pre-training, while Wang et al. (2022) proposes a method to identify "skill neurons" in pre-trained Transformer-based language models that are heavily involved in specific tasks. Recently, Schwettmann et al. (2023) introduces a procedure for identifying "multimodal neurons", which explain how LLMs convert visual representations into corresponding texts.

Analysing Pre-Trained Transformers With the proposal of the Transformer (Vaswani et al., 2017) architecture, Transformer-based models have attracted a large amount of studies. Prior works have focused on the function and mechanism of self-attention modules (Voita et al., 2019; Clark et al., 2019; Hao et al., 2021), while some works emphasize the significance of feed-forward layers in Transformer (Press et al., 2020; Geva et al., 2021; Dai et al., 2022). Among these, some works probe Transformer representations to quantify their encoding of linguistic information (Peters et al., 2018; Niven and Kao, 2019; Yun et al., 2019).

5 Conclusion

We propose a new method to identify multi-modal neurons in Transformer-based multi-modal LLMs. We also introduce a knowledge editing approach based on the identified neurons, which achieves a knowledge editing from a specific token to another designative token. We highlight three critical properties of multi-modal neurons by four well-designed quantitative evaluation metrics through extensive experiments. Both quantitative and qualitative experiments validate the explanatory powers of our multi-modal neurons. This work provides illuminating perspectives on multi-modal LLMs and stimulates additional explanatory artificial intelligence studies emphasizing model interpretability.

Limitations

While this work provides new insights into interpreting multi-modal large language models, there are several limitations that should be acknowledged: (1) We only conduct experiments on LLaVA, InstructBLIP and mPLUG-Owl2, while other Transformer-based models may also be possible to be explained by our multi-modal neurons. Besides the Transformer architecture, it is still unclear whether neurons exist in other multi-modal large language models based on different architectures and requires further explorations. (2) We only focus on neurons in feed-forward networks in Transformer and omit other parts like the neurons in self-attention heads, which may also contribute to identify image features and generate output. (3) When analysing multi-modal neurons, we only consider the role of a single neuron. We expect future works can explore how multiple neurons jointly influence the model. (4) As our multi-modal knowledge editing method is based on changing the probability distribution of the generated token, we only achieve a transformation from a single source token to another single designative token, which is still insufficient, since there are a large amount of words consist of multiple tokens. Further addressing these limitations through broader and more methodologically rigorous studies would help advance knowledge in interpretability of multi-modal large language models.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

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.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023. Language models implement simple word2vec-style vector arithmetic.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Ofir Press, Noah A. Smith, and Omer Levy. 2020. Improving transformer models by reordering their sublayers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2996–3005, Online. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023a.

mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. 2019. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Supplementary Explanation

In § 2.2, we illustrate how to identify multi-modal neurons in Transformer-based LLMs. We now provide some additional details here.

In Eq. 2, we use matrix $\mathbf{Q}^l$ to define the contribution score. From the dimensional perspective of $\mathbf{Q}^l$, since $\mathbf{Q}^l \in \mathbb{R}^{d_m \times v}$, where d_m is intermediate size and v is vocab size, each element in $\mathbf{Q}^l$ can be regarded as a contribution of each neuron at layer l to each token in the vocabulary. For instance, the contribution of the i-th neuron u_i at layer l to token t is derived from the i-th row and t-th column of $\mathbf{Q}^l$ (i.e. $\mathbf{Q}^l(i,t)$). From the perspective of the meaning of $\mathbf{Q}^l$, $\mathbf{Q}^l$ is consistent with the probability distribution when predicting, where we prove it through Eq. 3 and Eq. 4.

In Eq. 3, we disassemble the generation procedure of the LLM. We first decompose the hidden states at the last layer $\mathbf{h}^L_{-1}$ into three parts: self-attention output $\mathbf{a}^L_{-1}$, FFN output $\mathbf{m}^L_{-1}$ and hidden states at the previous layer $\mathbf{h}^{L-1}_{-1}$ (Line 1 to Line 2). Then $\mathbf{h}^{L-1}_{-1}$ can be further decomposed through layers until we get the embedding vector of input $\mathbf{h}^0_{-1}$ (Line 2 to Line 3). Ultimately, we replace $\mathbf{m}^l_{-1}$ with $\mathbf{W}^l_{\text{out}}\mathbf{O}^l_{-1}$ (Line 3 to Line 4). Note that we have omitted layer normalization operations in Eq. 3 through approximate assumptions for the sake of brevity.

In Eq. 4, we disassemble the multiplication of $\mathbf{W}_u\mathbf{W}^l_{\text{out}}$ and $\mathbf{O}^l_{-1}$. The dimensionality of $\mathbf{W}_u\mathbf{W}^l_{\text{out}}$ is $d_m \times v$. We aim at obtaining a matrix that can indicate the contribution from each neuron to each token. Accordingly, we adopt an element-wise product with broadcasting mechanism between $\mathbf{W}_u\mathbf{W}^l_{\text{out}}$ and $\mathcal{T}\left(\mathbf{O}^l_{-1}\right)$, keeping the original dimensionality unchanged.

We mainly focus on the last token outputs in Eq. 2, Eq. 3 and Eq. 4. The rationale behind our approach is that an autoregressive Transformer (Vaswani et al., 2017) will generate the new token at the position of the last input token. Therefore, analyzing the last token can help us understand the principles underlying the model generation process.

B Implementation Details

B.1 Identifying Multi-Modal Neurons

For model LLaVA (Liu et al., 2023), we choose the version whose base LLM is LLaMA-2-13B-Chat (Touvron et al., 2023b) and visual encoder is ViT-L/14 (Radford et al., 2021). Each input image is resized to (224, 224) and encoded into a sequence $[z_1, \cdots, z_p]$ of dimensionality 1024, where $p = 256$. Then a projection layer transforms sequence $[z_1, \cdots, z_p]$ into image prompts $[x_1, \cdots, x_p]$ of dimensionality 5120. The image prompts will be concatenated into the textual prompts and received by LLaVA.

For model InstructBLIP (Dai et al., 2023), we choose the version that employs image encoder including ViT-g/14 (Fang et al., 2023) and a Q-former (Li et al., 2023), and adopts Vicuna-7B (Chiang et al., 2023) as the LLM. Similar to LLaVA, each image is encoded into a sequence $[z'_1, \cdots, z'_q]$, where $q = 256$. And then the sequence is sent into the Q-former to get the extracted image features $[z_1, \cdots, z_p]$ of dimensionality 768, where $p = 32$. Then a projection layer transforms sequence $[z_1, \cdots, z_p]$ into image prompts $[x_1, \cdots, x_p]$ of dimensionality 4096.

Model mPLUG-Owl2 (Ye et al., 2023b) utilizes ViT-L/14 (Radford et al., 2021) as visual encoder and LLaMA-2-7B (Touvron et al., 2023b) as LLM. Different from LLaVA and InstructBLIP, mPLUG-Owl2 adopts a visual abstractor after the visual encoder, which transforms image features $[z_1, \cdots, z_p]$ of dimensionality 1024 into image prompts $[x_1, \cdots, x_p]$ of dimensionality 4096.

We adopt "Describe the image in few words." as query prompts in all models. Note that for better captioning results, we add a text prefix "An image of" after the textual prompts.

We use greedy search when generating captions for each image, which means the token with the highest probability will be selected at each step. We calculate the contribution score $s_{i,t}^l$ for each nominal token t in the generated caption, and rank all contribution scores across all layers within the model by the descending order to select top neurons as multi-modal neurons.

It should be noted that while we can calculate scores for all tokens generated by the model, some tokens may not be readily describable from the image content alone. Therefore, for the purpose of clearer explanation, our analysis focuses only on tokens corresponding to nouns. If a noun consists of multiple tokens, we select the first token as being representative of that noun. To identify all nouns in the caption, we use Stanford CoreNLP (Manning et al., 2014), a tool for natural language processing in Java, by a python wrapper [1].

We compare our method with Multimodal Neurons (Schwettmann et al., 2023), which calculates the attribution scores to select neurons. In their method, an attribution score is obtained for each image patch and neuron. For fair comparisons in our experiments, we modify this by taking the maximum attribution score across patches for each neuron. This modification avoids unnecessary repetition while maintaining the interpretability of the neuron attributions.

Furthermore, we established a baseline approach that solely considers the activations of neurons at the last input token as contribution scores, selecting those neurons exhibiting higher levels of activation as contributory neurons.

We run the experiments on NVIDIA RTX 1080Ti, NVIDIA RTX 2080Ti and NVIDIA RTX 3090 GPUs, and it takes about 500 GPU hours.

B.2 Tracing Focus of Neurons in Images

Following previous works on feature visualization (Bau et al., 2017; Schwettmann et al., 2023), we are curious about where neurons focus their attention. To trace focus of neurons in images, we employ a visualization approach described below.

We denote the size of input images as $d_i \times d_i$. Assuming that after passing through the image encoder, there are p image tokens input into the LLM. We assume that p can be square rooted. For each multi-modal neuron, we take its activations at image tokens and reshape them into a $\sqrt{p} \times \sqrt{p}$ matrix. And then we scale them to $d_i \times d_i$ by bilinear interpolation. Now the scaled activations and the input images have the same size. For each image, we first plot a heatmap by using a mean scaled activation across top-k neurons and put it over the image. We then threshold the mean scaled activations above the 95% percentile to produce a binary mask and also combine it with the original image.

Since the square root of the number of image patch tokens (i.e. $\sqrt{p}$) in InstructBLIP and mPLUG-Owl2 is irrational, we only trace focus of neurons using LLaVA.

B.3 Targeted Editing

For most images, we empirically pick out the top-5 multi-modal neurons as $\mathcal{S}$, initialize $\Delta\mathbf{w}$ as $\mathbf{0}$, and set the learning rate α as 0.001, the iteration epochs ϵ as 1000 and the penalty weight β as 4, respectively.

C More Experiment Results

We report more experiment results and show more cases here to confirm our conclusion convincingly.

C.1 Tracing Focus of Neurons in Images

We report heatmap and binary mask results of examples in Table 8. Each heatmap is plotted by using scaled mean activations across top-k neurons, where $k = 1, 10, 50, 100, 500, 1000$, and each binary mask is plotted by thresholding mean activations above the 95% percentile, respectively.

C.2 Textual Meanings of Neurons

Table 9 shows examples. For each concept in the caption, we report its multi-modal neurons with their corresponding top-tokens and contribution scores.

C.3 Region Invariance of Neurons

In Table 10, we report some example results of captions and multi-modal neurons before and after shuffling the input sequence of image patches.

C.4 Cross-Image Invariance of Neurons

To confirm the cross-image invariance of multi-modal neurons, in Figure 6, we report the ratio of the common neurons in top-k neurons across

[1] https://github.com/Jason3900/corenlp_client

N images that contain the same concepts, where $N = 2, 3, 4, 5$ and $k = 10, 100, 1000$, respectively.

C.5 Specificity of Neurons

To verify the specificity of multi-modal neurons, in Figure 7, we report some examples of the heatmap of the scores of multi-modal neurons corresponding to specific concepts when encoding different concepts.

C.6 Perturbing Multi-Modal Neurons

Table 11 shows results of perturbing top-5 multi-modal neurons and 5 randomly selected neurons.

C.7 Targeted Editing

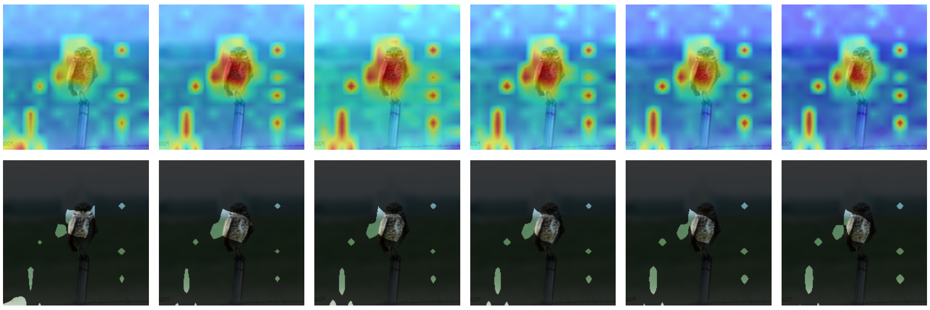
Table 12 shows additional examples of targeted editing results.

Image & Original output

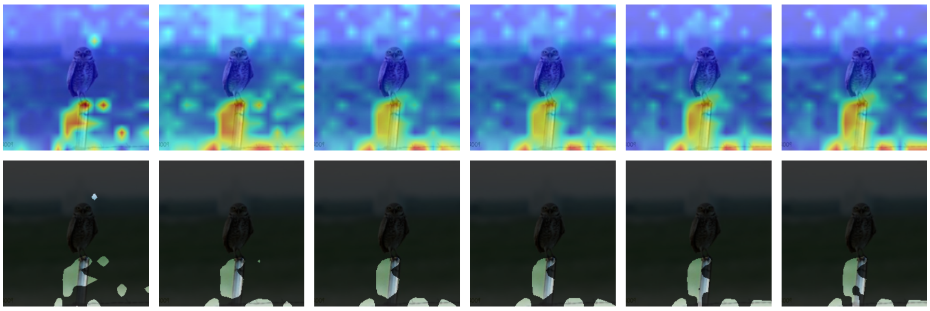
LLaVA: a small **owl** perched on a **metal pole** in a grassy **field**.

Concept	Heatmap & Binary mask					
	Top-1	Top-10	Top-50	Top-100	Top-500	Top-1000

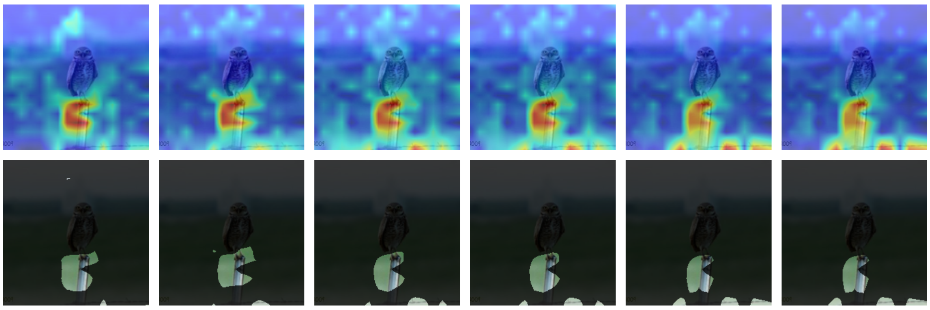
owl



metal



pole



field

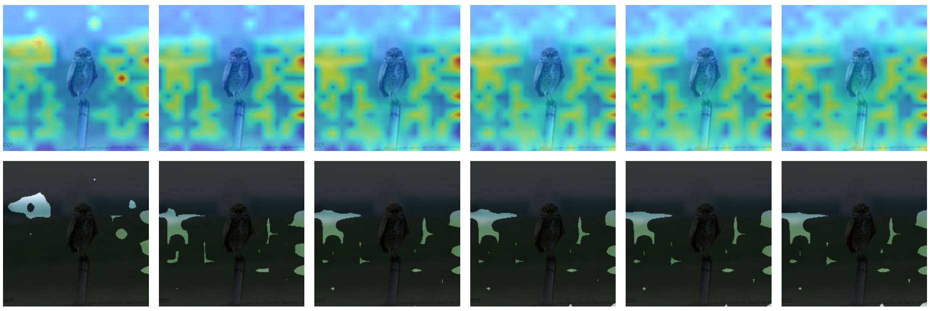


Image & Original output	
	LLaVA: a **box** filled with empty **beer bottles**, sitting on the **sidewalk**.

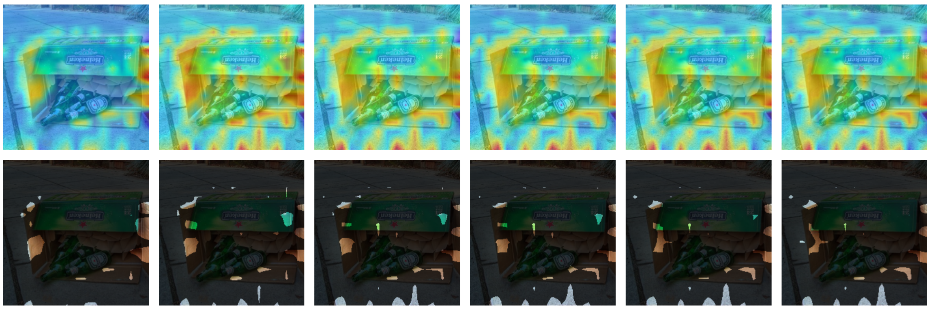
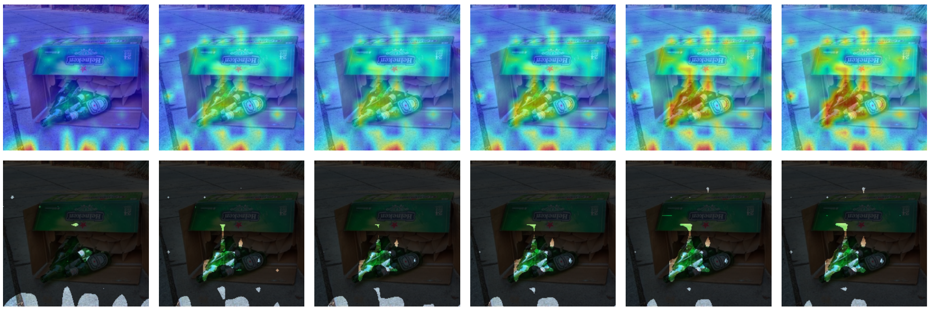
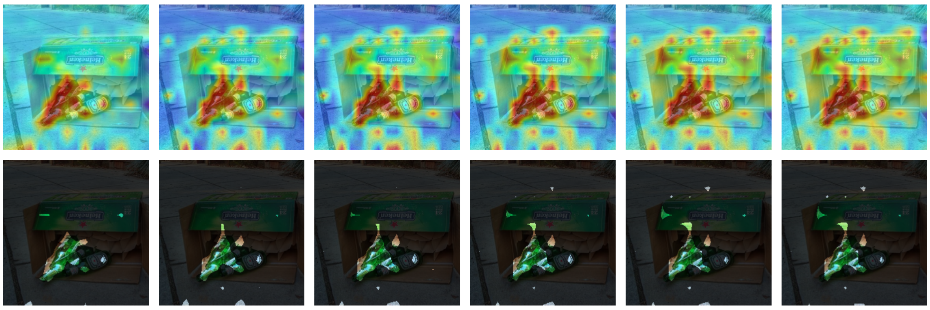
Concept	Heatmap & Binary mask Top-1 / Top-10 / Top-50 / Top-100 / Top-500 / Top-1000
box	
beer	
bottles	
sidewalk	

Image & Original output	
	LLaVA: a beautiful **lake** surrounded by **mountains**, with a **boat** floating on the **water**.

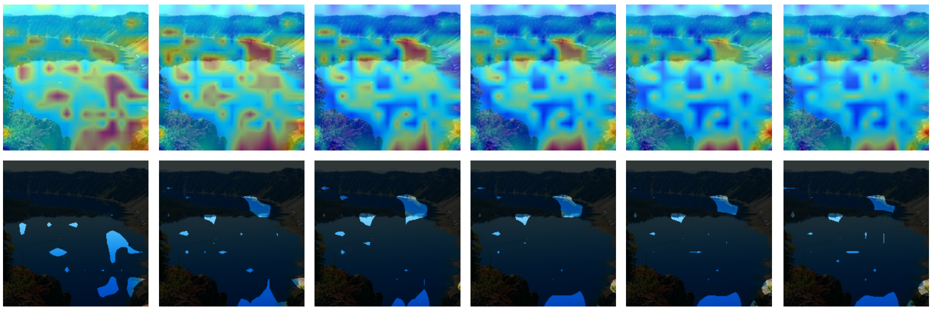
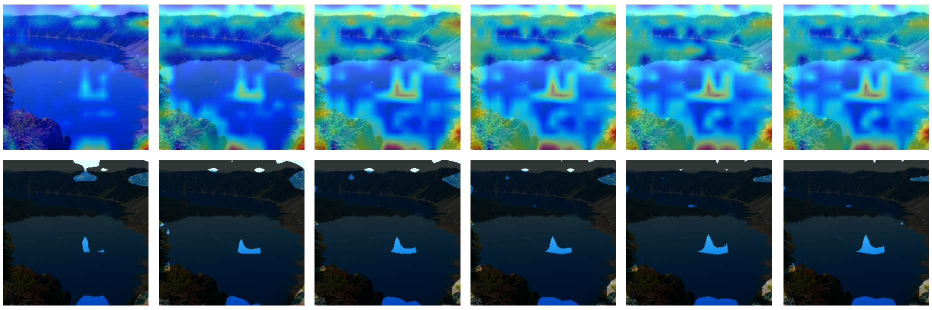
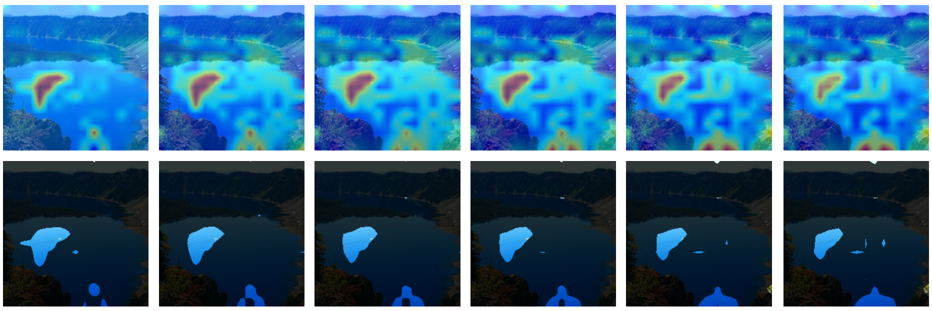
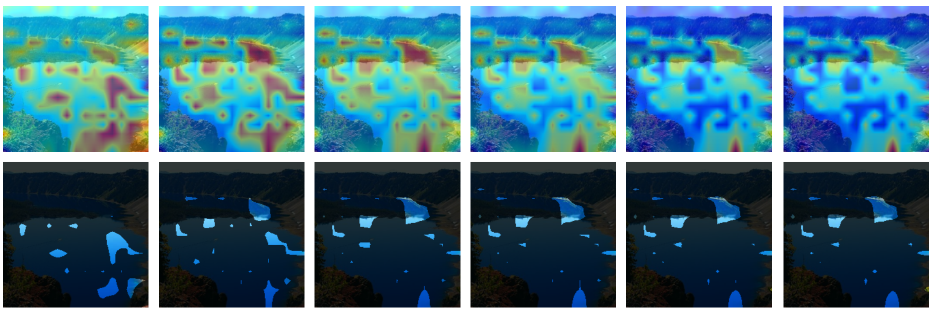
Concept	Heatmap & Binary mask					
	Top-1	Top-10	Top-50	Top-100	Top-500	Top-1000
lake						
mountains						
boat						
water						

Table 8: Heatmap and binary mask results of example images. We plot each heatmap by using scaled mean activations across top-k neurons, where $k = 1, 10, 50, 100, 500, 1000$, and plot binary mask by thresholding mean activations above the 95% percentile, respectively.

Image	Model	Concept	Top neurons	Top tokens	Score
 LLaVA: a small red **motorcycle** parked on the **grass** near a **beach**. InstructBLIP: a **motorcycle** parked on the **grass** near the **ocean**. mPLUG-Owl2: a **motorcycle** parked on the **grass** near the **ocean**.	LLaVA	motorcycle	**L34.U12567**	[‘motor’,‘Motor’,‘mot’,‘b’,‘mot’]	0.906
			L33.U6828	[‘mot’,‘Mot’,‘mot’,‘motiv’,‘Motor’]	0.850
			L25.U11735	[‘motor’,‘tennis’,‘hockey’,‘basketball’,‘football’]	0.641
			L24.U5729	[‘vehicle’,‘vehicles’,‘aircraft’,‘boat’,‘motor’]	0.591
			L27.U11389	[‘mot’,‘motor’,‘Mot’,‘Motor’,‘mot’]	0.533
		grass	**L25.U5542**	[‘grass’,‘woods’,‘leaf’,‘forest’,‘bush’]	2.039
			L32.U12094	[‘grass’,‘aupt’,‘itza’,‘ustration’,‘inx’]	1.873
			L30.U1365	[‘la’,‘La’,‘La’,‘la’,‘wn’]	1.526
			L20.U7408	[‘grass’,‘garden’,‘gard’,‘草’,‘veget’]	1.150
			L29.U7377	[‘gr’,‘Gr’,‘Grant’,‘gr’,‘grant’]	1.145
		beach	**L36.U13537**	[‘Coast’,‘coast’,‘beach"’,‘Beach’,‘ocean’]	2.984
			L30.U13327	[‘be’,‘be’,‘Be’,‘BE’,‘aches’]	0.704
			L21.U13303	[‘beach’,‘coast’,‘Beach’,‘Coast’,‘shore’]	0.607
			L21.U11114	[‘sw’,‘Sw’,‘sw’,‘pool’,‘Sw’]	0.505
			L39.U11294	[‘flying’,‘sea’,‘aer’,‘Sea’,‘jet’]	0.502
	InstructBLIP	motorcycle	**L34.U12567**	[‘motor’,‘Motor’,‘mot’,‘b’,‘mot’]	0.906
			L33.U6828	[‘mot’,‘Mot’,‘mot’,‘motiv’,‘Motor’]	0.850
			L25.U11735	[‘motor’,‘tennis’,‘hockey’,‘basketball’,‘football’]	0.641
			L24.U5729	[‘vehicle’,‘vehicles’,‘aircraft’,‘boat’,‘motor’]	0.591
			L27.U11389	[‘mot’,‘motor’,‘Mot’,‘Motor’,‘mot’]	0.533
		grass	**L25.U5542**	[‘grass’,‘woods’,‘leaf’,‘forest’,‘bush’]	2.039
			L32.U12094	[‘grass’,‘aupt’,‘itza’,‘ustration’,‘inx’]	1.873
			L30.U1365	[‘la’,‘La’,‘La’,‘la’,‘wn’]	1.526
			L20.U7408	[‘grass’,‘garden’,‘gard’,‘草’,‘veget’]	1.150
			L29.U7377	[‘gr’,‘Gr’,‘Grant’,‘gr’,‘grant’]	1.145
		ocean	**L36.U13537**	[‘Coast’,‘coast’,‘beach"’,‘Beach’,‘ocean’]	2.984
			L30.U13327	[‘be’,‘be’,‘Be’,‘BE’,‘aches’]	0.704
			L21.U13303	[‘beach’,‘coast’,‘Beach’,‘Coast’,‘shore’]	0.607
			L21.U11114	[‘sw’,‘Sw’,‘sw’,‘pool’,‘Sw’]	0.505
			L39.U11294	[‘flying’,‘sea’,‘aer’,‘Sea’,‘jet’]	0.502
	mPLUG-Owl2	motorcycle	**L30.U9081**	[‘Motor’,‘motor’,‘mot’,‘mot’,‘Mot’]	2.236
			L29.U7834	[‘autom’,‘Autom’,‘automat’,‘Autom’,‘motor’]	0.824
			L21.U7122	[‘bi’,‘Bi’,‘cy’,‘cycle’,‘cycle’]	0.650
			L26.U6941	[‘passenger’,‘車’,‘vehicle’,‘passengers’,‘vehicles’]	0.468
			L25.U8004	[‘motor’,‘Motor’,‘mot’,‘undle’,‘overflow’]	0.413
		grass	**L27.U5003**	[‘grass’,‘ass’,‘ersion’,‘mitt’,‘比’]	1.614
			L22.U10525	[‘sand’,‘Sand’,‘dust’,‘gra’,‘grass’]	0.708
			L31.U2642	[‘forest’,‘Forest’,‘tree’,‘Tree’,‘Tree’]	0.433
			L20.U2081	[‘field’,‘Hay’,‘Field’,‘hay’,‘fields’]	0.390
			L21.U819	[‘tur’,‘grass’,‘Tur’,‘sod’,‘bl’]	0.329
		ocean	**L30.U4330**	[‘sea’,‘marine’,‘Sea"’,‘Marine’,‘ocean’]	1.953
			L22.U10714	[‘sea’,‘ocean’,‘Sea’,‘Ocean’,‘Atlantic’]	1.123
			L23.U8790	[‘sand’,‘beach’,‘be’,‘Beach’,‘Sand’]	0.542
			L23.U8326	[‘water’,‘water’,‘Water’,‘waters’,‘水’]	0.520
			L21.U6004	[‘coast’,‘Coast’,‘sea’,‘ocean’,‘tid’]	0.439

Image	Model	Concept	Top neurons	Top tokens	Score
 LLaVA: a small **figurine**, possibly a **toy** or a **model**, is displayed on a green **surface**, possibly a **table** or a grassy **area**. InstructBLIP: a miniature **figurine** with a **knife**. mPLUG-Owl2: a small **figurine** of a **man** holding a **knife**.	LLaVA	figurine	**L36.U8273**	[‘figure’,‘Fig’,‘Figure’,‘figures’,‘Fig’]	2.572
			L24.U12276	[‘stat’,‘statue’,‘sculpt’,‘Stat’,‘stat’]	1.161
			L18.U4770	[‘mini’,‘figure’,‘figures’,‘figur’,‘model’]	1.014
			L38.U10971	[‘figure’,‘figures’,‘Figure’,‘Fig’,‘figured’]	0.833
			L38.U2195	[‘Хронологија’,‘Kontrola’,‘konn’,‘Audiod’,‘techni’]	0.627
		toy	**L39.U98**	[‘to’,‘to’,‘To’,‘To’,‘TO’]	2.121
			L32.U6038	[‘Toy’,‘To’,‘Toast’,‘TO’,‘To’]	1.298
			L39.U212	[‘’,‘1’,‘-’,‘\n’,‘(’]	1.101
			L38.U184	[‘to’,‘to’,‘To’,‘到’,‘into’]	0.890
			L39.U11820	[‘externas’,‘’,‘a’,‘(’,‘,’]	0.754
		model	**L39.U3149**	[‘models’,‘model’,‘models’,‘model’,‘Model’]	2.893
			L23.U1705	[‘mini’,‘model’,‘models’,‘model’,‘Model’]	1.705
			L24.U12276	[‘stat’,‘statue’,‘sculpt’,‘Stat’,‘stat’]	0.914
			L18.U4770	[‘mini’,‘figure’,‘figures’,‘figur’,‘model’]	0.710
			L39.U4397	[‘mode’,‘Mode’,‘Model’,‘MODE’,‘Mode’]	0.639
		surface	**L37.U10337**	[‘Sur’,‘Sur’,‘sur’,‘surface’,‘surfaces’]	3.676
			L30.U2704	[‘qu’,‘sil’,‘background’,‘emb’,‘Sil’]	0.620
			L36.U3279	[‘surface’,‘face’,‘面’,‘faces’,‘fac’]	0.492
			L35.U8250	[‘surface’,‘surfaces’,‘superficie’,‘superfic’,‘повер’]	0.439
			L34.U6951	[‘soft’,‘fi’,‘bra’,‘pla’,‘soft’]	0.438
		table	**L23.U1705**	[‘mini’,‘model’,‘models’,‘model’,‘Model’]	0.458
			L19.U13612	[‘tables’,‘table’,‘wall’,‘sink’,‘chair’]	0.429
			L26.U10793	[‘table’,‘Table’,‘tables’,‘table’,‘TABLE’]	0.369
			L32.U1205	[‘table’,‘Table’,‘Scanner’,‘Table’,‘table’]	0.328
			L18.U4770	[‘mini’,‘figure’,‘figures’,‘figur’,‘model’]	0.321
		area	**L35.U2653**	[‘Area’,‘area’,‘area’,‘Area’,‘areas’]	1.570
			L31.U12802	[‘area’,‘Area’,‘zone’,‘region’,‘area’]	0.630
			L37.U2420	[‘region’,‘region’,‘regions’,‘Region’,‘Region’]	0.494
			L25.U12317	[‘places’,‘cave’,‘homes’,‘environments’,‘Places’]	0.388
			L31.U9217	[‘rug’,‘car’,‘blank’,‘felt’,‘fel’]	0.332
	InstructBLIP	figurine	**L27.U10783**	[‘figure’,‘figures’,‘Figure’,‘figure’,‘Fig’]	0.824
			L31.U5983	[‘beside’,‘beneath’,‘populated’,‘centered’,‘aligned’]	0.620
			L31.U3824	[‘anyway’,‘жовт’,‘frequ’,‘whenever’,‘meant’]	0.590
			L31.U8541	[‘Unterscheidung’,‘archivi’,‘Hinweis’,‘zvuky’,‘burgo’]	0.585
			L31.U6958	[‘analyz’,‘recognized’,‘Student’,‘participated’,‘analyt’]	0.540
		knife	**L27.U1255**	[‘kn’,‘Kn’,‘kn’,‘Bla’,‘Knight’]	5.137
			L29.U835	[‘K’,‘Kid’,‘kernel’,‘k’,‘kne’]	1.061
			L18.U2218	[‘pen’,‘pen’,‘pens’,‘sword’,‘rod’]	0.726
			L25.U9447	[‘um’,‘Um’,‘flash’,‘flash’,‘pen’]	0.716
			L31.U8169	[‘CR’,‘PK’,‘EX’,‘BR’,‘HT’]	0.679
	mPLUG-Owl2	figurine	**L20.U1471**	[‘doll’,‘oll’,‘ted’,‘figur’,‘dollars’]	0.698
			L31.U4677	[‘closer’,‘semantics’,‘mind’,‘totalité’,‘minds’]	0.405
			L31.U9439	[‘theoret’,‘’,‘Complex’,‘influenced’,‘stabil’]	0.301
			L15.U3991	[‘doll’,‘model’,‘statue’,‘figures’,‘representation’]	0.283
			L22.U10518	[‘models’,‘figures’,‘models’,‘figure’,‘cav’]	0.274
		man	**L27.U5003**	[‘man’,‘man’,‘Man’,‘Man’,‘mann’]	1.614
			L22.U10525	[‘man’,‘Man’,‘Man’,‘man’,‘mann’]	0.708
			L31.U2642	[‘man’,‘Man’,‘man’,‘Man’,‘MAN’]	0.433
			L20.U2081	[‘man’,‘boy’,‘челове’,‘hombre’,‘raste’]	0.390
			L21.U819	[‘Man’,‘Man’,‘manual’,‘man’,‘manual’]	0.329
		knife	**L27.U2163**	[‘kn’,‘Kn’,‘kn"’,‘Knight’,‘cheval’]	3.330
			L26.U2228	[‘kn’,‘Kn’,‘kn’,‘Knight’,‘scope’]	3.117
			L21.U9295	[‘carry’,‘revol’,‘carried’,‘carrying’,‘kn’]	0.707
			L19.U8668	[‘gun’,‘guns’,‘gun’,‘Gun’,‘sword’]	0.404
			L31.U913	[‘archivi’,‘textt’,‘hyp’,‘immediately’,‘separ’]	0.390

Image	Model	Concept	Top neurons	Top tokens	Score
 LLaVA: a **plant** with red **flowers** hanging from it, possibly a Venus **flytrap**, is displayed in a **greenhouse**. InstructBLIP: a **pitcher** **plant** hanging from a **tree**. mPLUG-Owl2: a **pitcher** **plant** hanging from a **ceiling** in a **greenhouse**.	LLaVA	plant	**L27.U8060**	[‘plant’,‘Plant’,‘plant’,‘plants’,‘planta’]	1.087
			L29.U9056	[‘shr’,‘bush’,‘Bush’,‘plant’,‘plants’]	0.962
			L28.U11440	[‘flow’,‘blo’,‘Flow’,‘blo’,‘Flow’]	0.621
			L27.U498	[‘branch’,‘Branch’,‘branches’,‘branch’,‘bush’]	0.600
			L25.U11504	[‘roots’,‘root’,‘Root’,‘root’,‘leaves’]	0.502
		flowers	**L28.U11440**	[‘flow’,‘blo’,‘Flow’,‘blo’,‘Flow’]	1.447
			L20.U11853	[‘flower’,‘flowers’,‘flor’,‘Flor’,‘花’]	1.277
			L27.U13027	[‘pet’,‘pod’,‘leaves’,‘pet’,‘bud’]	0.990
			L27.U498	[‘branch’,‘Branch’,‘branches’,‘branch’,‘bush’]	0.675
			L27.U3452	[‘fol’,‘flowers’,‘leaves’,‘fol’,‘leaf’]	0.551
		flytrap	**L39.U1989**	[‘FI’,‘fo’,‘fig’,‘fer’,‘float’]	0.913
			L36.U7481	[‘F’,‘Φ’,‘フ’,‘Φ’,‘Fest’]	0.678
			L36.U6716	[‘file’,‘フ’,‘fake’,‘flower’,‘File’]	0.625
			L28.U7379	[‘vol’,‘flight’,‘flow’,‘fle’,‘fl’]	0.558
			L38.U998	[‘Fred’,‘Frederick’,‘Freder’,‘Fon’,‘Fen’]	0.530
		greenhouse	**L30.U1994**	[‘blo’,‘green’,‘Blo’,‘blo’,‘green’]	2.258
			L39.U3579	[‘red’,‘green’,‘red’,‘yellow’,‘blue’]	1.122
			L39.U9915	[‘white’,‘silver’,‘brown’,‘blue’,‘gold’]	1.086
			L28.U8699	[‘green’,‘ho’,‘Green’,‘green’,‘tunnel’]	0.836
			L29.U11697	[‘Green’,‘Green’,‘Blue’,‘Brown’,‘Black’]	0.420
	InstructBLIP	pitcher	**L28.U7071**	[‘pitch’,‘ML’,‘ML’,‘itch’,‘baseball’]	3.258
			L31.U3824	[‘anyway’,‘жовт’,‘frequ’,‘whenever’,‘meant’]	0.414
			L31.U9856	[‘P’,‘Pet’,‘Pan’,‘По’,‘П’]	0.407
			L31.U8541	[‘Unterscheidung’,‘archivi’,‘Hinweis’,‘zvuky’,‘burgo’]	0.406
			L31.U157	[‘.’,‘\n’,‘and’,‘jú’,‘shares’]	0.336
		plant	**L27.U8513**	[‘plant’,‘Plant’,‘plant’,‘plants’,‘planta’]	4.895
			L22.U7930	[‘plant’,‘plants’,‘plant’,‘Plant’,‘gard’]	3.105
			L23.U1593	[‘plant’,‘plants’,‘Plant’,‘plant’,‘Bonn’]	0.627
			L23.U7557	[‘Garden’,‘Gard’,‘garden’,‘gard’,‘plant’]	0.539
			L31.U5946	[‘whites’,‘contribute’,‘alongside’,‘dawn’,‘upon’]	0.500
		tree	**L22.U7930**	[‘plant’,‘plants’,‘plant’,‘Plant’,‘gard’]	1.845
			L19.U7918	[‘trees’,‘tree’,‘forest’,‘trees’,‘tree’]	0.658
			L29.U8371	[‘Tree’,‘landscape’,‘Tree’,‘trees’,‘tree’]	0.650
			L25.U441	[‘wood’,‘Wood’,‘wooden’,‘wood’,‘woods’]	0.586
			L20.U947	[‘roots’,‘root’,‘branches’,‘branch’,‘fruit’]	0.561
	mPLUG-Owl2	pitcher	**L27.U9072**	[‘pitch’,‘ML’,‘ML’,‘itch’,‘ml’]	0.540
			L31.U6404	[‘designated’,‘partially’,‘swing’,‘direct’,‘potentially’]	0.310
			L31.U3644	[‘-’,‘kick’,‘—’,‘timing’,‘ban’]	0.295
			L31.U8384	[‘kick’,‘...’,‘confront’,‘Mongo’,‘further’]	0.267
			L24.U4842	[‘éric’,‘CAA’,‘schaften’,‘rinn’,‘inta’]	0.237
		plant	**L24.U4652**	[‘plant’,‘Plant’,‘plant’,‘plants’,‘node’]	2.779
			L23.U10661	[‘blo’,‘flow’,‘Flow’,‘flow’,‘flowers’]	1.422
			L21.U9554	[‘seed’,‘botan’,‘seed’,‘Plant’,‘plant’]	0.403
			L22.U9083	[‘botan’,‘Botan’,‘flower’,‘plant’,‘Plant’]	0.400
			L30.U702	[‘plant’,‘subject’,‘Plant’,‘plant’,‘ak’]	0.366
		ceiling	**L20.U3762**	[‘walls’,‘wall’,‘floor"’,‘ce’,‘wall’]	0.582
			L17.U1877	[‘ce’,‘walls’,‘wall’,‘Ce’,‘Wall’]	0.380
			L21.U4447	[‘vent’,‘Vent’,‘vent’,‘du’,‘ce’]	0.316
			L23.U4000	[‘flo’,‘Flo’,‘float’,‘ground’,‘float’]	0.303
			L31.U9617	[‘Zyg’,‘behaviour’,‘etc’,‘Datos’,‘Gest’]	0.251
		greenhouse	**L28.U2667**	[‘Green’,‘green’,‘Green"’,‘green’,‘зе’]	3.994
			L31.U210	[‘yellow’,‘green’,‘red’,‘blue’,‘brown’]	1.497
			L26.U253	[‘green’,‘green’,‘Green’,‘gre’,‘Green’]	0.390
			L31.U9558	[‘pes’,‘tex’,‘davon’,‘flex’,‘scal’]	0.381
			L21.U9554	[‘seed’,‘botan’,‘seed’,‘Plant’,‘plant’]	0.303

Table 9: Multi-modal neurons with their corresponding top tokens and their contribution scores. For each concept in the caption, we report the top-5 neurons with the top-5 highest probability of tokens.

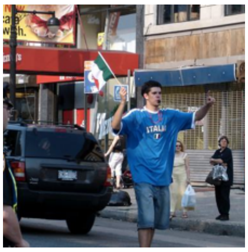
Image	Original	Shuffled
	a **tree** with white **flowers** in a **field**, surrounded by a **dirt road** and a **fence**. **tree:** **[L28.U9085, L36.U1422, L22.U171, L27.U8824]** **flowers:** **[L28.U11440, L20.U8129, L27.U13027, L27.U498]** **field:** **[L34.U12955, L28.U1085, L25.U5542, L39.U7153]** **dirt:** **[L39.U8730, L31.U526, L39.U212, L35.U1480]** **road:** **[L39.U8637, L26.U1456, L37.U12619, L29.U224]** **fence:** **[L27.U12313, L38.U5969, L37.U2453, L39.U212]**	a **tree** with white **flowers** in a **field**, surrounded by a **dirt road** and a **fence**. **tree:** **[L28.U9085, L36.U1422, L22.U171, L27.U8824]** **flowers:** **[L28.U11440, L20.U8129, L27.U13027, L27.U498]** **field:** **[L34.U12955, L28.U1085, L25.U5542, L39.U7153]** **dirt:** **[L31.U526, L39.U8730, L39.U212, L35.U1480]** **road:** **[L39.U8637, L26.U1456, L37.U12619, L29.U224]** **fence:** **[L27.U12313, L38.U5969, L37.U2453, L39.U212]**
	a **plate** of **meat**, including **steak** and a side of **vegetables**, is presented. **plate:** **[L33.U350, L23.U8551, L22.U9849, L19.U13764]** **meat:** **[L25.U9753, L29.U859, L23.U8551, L37.U11136]** **steak:** **[L37.U577, L25.U9753, L28.U10409, L22.U384]** **vegetables:** **[L37.U6234, L25.U3659, L38.U7433, L23.U8551]**	a **plate** of **meat**, including **steak** and mashed potatoes, accompanied by a side of **vegetables**. **plate:** **[L33.U350, L23.U8551, L22.U9849, L19.U13764]** **meat:** **[L25.U9753, L29.U859, L23.U8551, L22.U3753]** **steak:** **[L37.U577, L25.U9753, L28.U10409, L22.U384]** **vegetables:** **[L25.U3659, L37.U6234, L23.U8551, L25.U8838]**
	a young **girl** standing in a **doorway** of a building, possibly a school, with a **brick wall**. **girl:** **[L39.U5692, L28.U12204, L39.U364, L37.U9680]** **doorway:** **[L22.U9920, L27.U235, L21.U1052, L26.U10562]** **brick:** **[L29.U10814, L39.U8576, L25.U10651, L33.U10983]** **wall:** **[L35.U10298, L29.U9350, L29.U2530, L25.U10651]**	a young **girl** standing in front of a stone **wall**, possibly a **brick wall**, with a **doorway**. **girl:** **[L39.U5692, L28.U12204, L39.U364, L37.U9680]** **doorway:** **[L22.U9920, L29.U2530, L25.U5313, L25.U10438]** **brick:** **[L29.U10814, L24.U9050, L25.U10651, L33.U10983]** **wall:** **[L35.U10298, L29.U2530, L29.U9350, L19.U10353]**
	a group of **men** in a **room**, celebrating and cheering while holding up their **arms** and **fists**. **men:** **[L39.U5989, L29.U5763, L35.U8027, L29.U11953]** **room:** **[L38.U7800, L30.U6814, L29.U10611, L21.U8512]** **arms:** **[L23.U4494, L38.U10666, L24.U4501, L39.U5889]** **fists:** **[L38.U5969, L37.U2453, L39.U212, L36.U8631]**	a group of **men** in a **room**, celebrating and cheering while holding up their **arms** and **fists**. **men:** **[L39.U5989, L29.U5763, L35.U8027, L29.U11953]** **room:** **[L38.U7800, L30.U6814, L29.U10611, L21.U8512]** **arms:** **[L23.U4494, L38.U10666, L24.U4501, L26.U2293]** **fists:** **[L38.U5969, L37.U2453, L39.U212, L36.U8631]**
	a **man** standing on a **street corner**, holding an Italian **flag**, and waving it while a **police officer** watches him. **man:** **[L34.U3689, L39.U12617, L28.U9293, L34.U6857]** **street:** **[L39.U8140, L26.U1456, L26.U12900, L17.U5764]** **corner:** **[L38.U9436, L23.U12251, L28.U4161, L26.U8916]** **flag:** **[L25.U6794, L24.U6437, L23.U8268, L19.U12464]** **police:** **[L27.U7931, L31.U9142, L23.U2072, L35.U8410]** **officer:** **[L27.U7931, L23.U2072, L21.U3591, L39.U7884]**	a **man** standing on a **street corner**, holding an Italian **flag** and waving it, while a **police officer** watches him from a car. **man:** **[L34.U3689, L39.U12617, L28.U9293, L34.U6857]** **street:** **[L39.U8140, L26.U1456, L26.U12900, L17.U5764]** **corner:** **[L38.U9436, L23.U12251, L28.U4161, L26.U8916]** **flag:** **[L25.U6794, L19.U12464, L24.U6437, L23.U8268]** **police:** **[L27.U7931, L31.U9142, L23.U2072, L35.U8410]** **officer:** **[L27.U7931, L23.U2072, L21.U3591, L39.U7884]**

Table 10: Example results of captions and multi-modal neurons before and after shuffling the input sequence of image patches, respectively. We just record the concepts that appear both in original and shuffled captions from LLaVA, and for each concept, we report its top-4 multi-modal neurons.

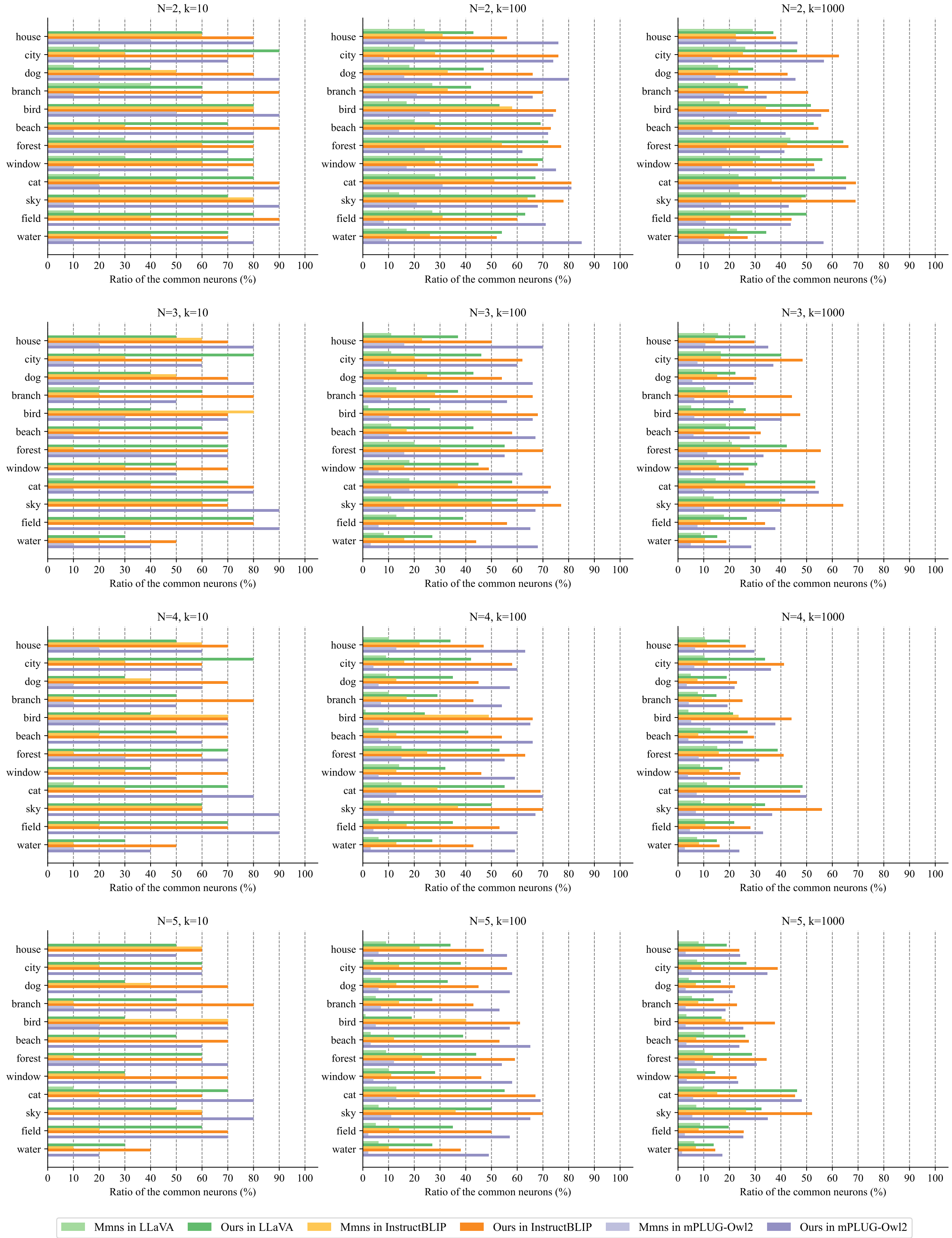
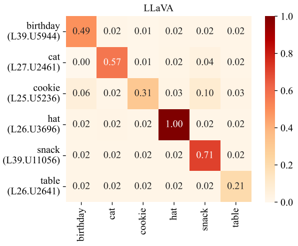


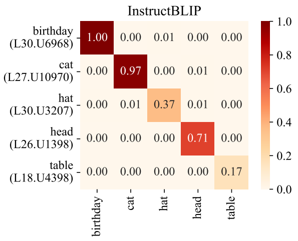
Figure 6: Ratio of the common neurons in top-k neurons selected by Mmns and our method. We report $N = 2, 3, 4, 5$ and $k = 10, 100, 1000$ for model LLaVA, InstructBLIP and mPLUG-Owl2.

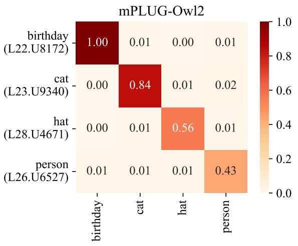
LLaVA: a **cat** wearing a **birthday hat** and eating a **snack**, possibly a **cookie**, while sitting on a **table**.

InstructBLIP: a **cat** laying on a **table** with a **birthday hat** on its **head**.

mPLUG-Owl2: a **cat** wearing a **birthday hat** and a **person** feeding it.



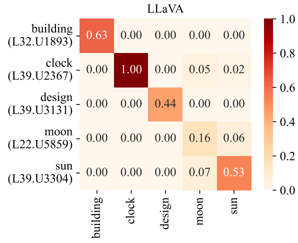


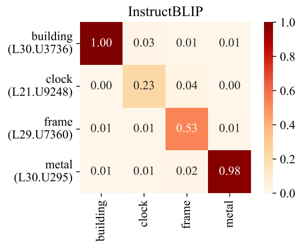


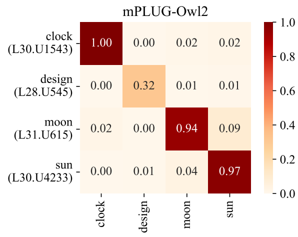
LLaVA: a large **clock** on a **building**, featuring a **moon** and **sun design**.

InstructBLIP: a **clock** on a **building** with a **metal frame**.

mPLUG-Owl2: a **clock** with a **sun** and **moon design** on it.



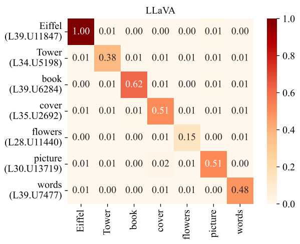


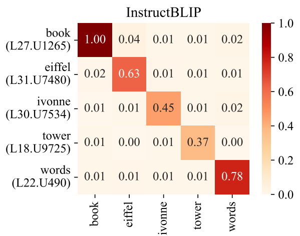


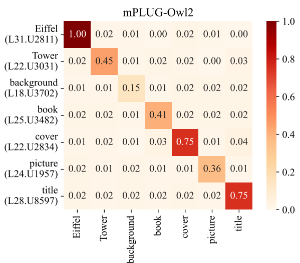
LLaVA: a **book** with a **cover** featuring a **picture** of the **Eiffel Tower**, **flowers**, and the **words** "Before You" written on it.

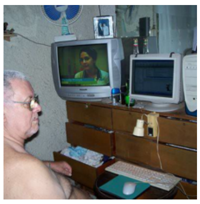
InstructBLIP: the **eiffel tower** with the **words**, **ivonne book** one before you.

mPLUG-Owl2: a **book cover** with the **title** "Before You" and a **picture** of the **Eiffel Tower** in the **background**.





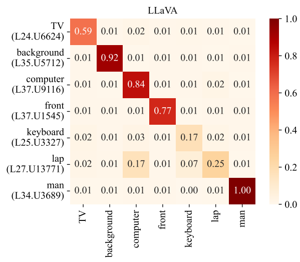


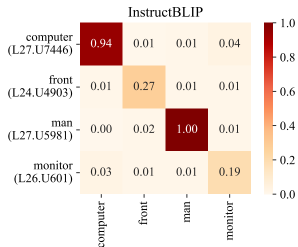


LLaVA: a **man** sitting in **front** of a **computer**, with a **TV** in the **background**, and a **keyboard** on his **lap**.

InstructBLIP: a **man** sitting in **front** of a **computer monitor**.

mPLUG-Owl2: an older **man** sitting in **front** of a **television**, watching a **woman** on the **screen**.





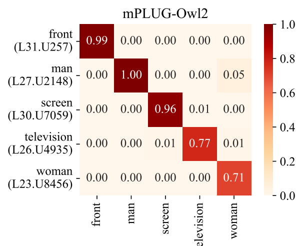


Figure 7: Heatmaps of the scores (after normalization) of multi-modal neurons corresponding to specific semantics when encoding different semantics. For each image, we report the result of the top-1 multi-modal neuron. In each heatmap, the x-axis represents concepts in the given image, and y-axis represents the top-1 neuron corresponding to each concept, respectively. Darker blocks indicate higher scores, which means higher relevance.

Image	Concept	Perturbed model output
 LLaVA: a tall **apartment building** with **balconies** and a **tree** in the **background**.	apartment	a multilevelishiigledishiigledishiigledishiigledishi...
	building	a white and blue building with a balcony and a tree in the background.
	balconies	a building with eradicated trees in the background, with eradicated trees on eradicated trees on 2200.
	tree	a white building with a balcony and a chair on it.
	background	a tall apartment building with balconies and a tree in front of it.
	random	a tall apartment building with balconies and a tree in the background.
 LLaVA: a mountainous **landscape** with a **village** in the **valley**, featuring a grassy **field** and a **road**.	landscape	a mountain range with a village in the valley, surrounded by a green field.
	village	a mountain with a small town or village located at its base, surrounded by a lush green field.
	valley	a mountain with a lush green field in the background, surrounded by a village.
	field	a mountain with a village in the valley below, surrounded by a lush green countryside.
	road	a mountainous landscape with a village in the valley, surrounded by a lush green field.
	random	a mountainous landscape with a village in the valley, featuring a grassy field and a road.
 LLaVA: a large **tower** with a **ball** on top, standing next to a **street light**.	tower	a large, white building with a light on a black background, with a lighted street lamp in the foreground.
	ball	a tall building with a light on top, possibly a light tower, and a street light nearby.
	street	a large white building with a clock tower, a street light, and a city skyline in the background.
	light	a tall tower with a large sphere on top, a building with a large sphere on top, and a building with a large sphere on top, all located in a city.
	random	a large tower with a ball on top, standing next to a street light.
 LLaVA: a **man** hanging from a **tree branch** while wearing a **hat**.	man	a person hanging from a tree branch, possibly participating in a tree-climbing activity or adventure.
	tree	a man hanging from a tree, possibly on a rope or a swing, with a sky background.
	branch	a man hanging from a suspended rope in a tree in the outdoors.
	hat	a person hanging from a tree branch, possibly participating in a ropes course or a similar activity.
	random	a man hanging from a tree branch while wearing a hat.

Table 11: Perturbation results of example images. For each concept in the image, we pick out top-5 multi-modal neurons and each add a Gaussian noise to perturb them. We also report results of perturbing 5 randomly selected neurons for comparison.

Image	Source	Target	Edited model output
 LLaVA: a small white **dog** standing on a tiled **floor**.	dog	mouse	a **mouse** in a pink dress, standing on a tiled floor.
		bag	a white **bag bag bag**, or a white **bag** with a pink stripe, is standing on a tiled floor.
		dinosaur	a small white sauce **dinosaur** (dino) or a small white sauce-covered **dinosaur** toy is standing on a tiled floor.
		baby	a **baby** white **baby**, wearing a pink dress, standing on a tiled floor.
	floor	bed	a small white dog wearing a pink dress, standing on **bed** in a bedroom.
		chair	a small white dog wearing a pink dress, standing on **chair** in a **chair**, and looking at the camera.
		tree	a small white dog standing on a **tree tree** in a **tree**-lined area.
		beach	a small white dog, possibly a Maltese, wearing a pink dress and standing on the **beach**.
		table	a small white dog wearing a pink dress, standing on a **table**.
		ship	a small white dog standing on a **ship** shipment.
 LLaVA: a **man** standing in a **forest**, surrounded by **trees** and a large **rock**.	man	woman	a **woman** standing in a forest, surrounded by trees and rocks.
		baby	a **baby** in a forest, surrounded by trees and rocks.
		girl	a **girl** standing in a forest, surrounded by trees and rocks.
		stone	a **stone** wall or rock formation in a forest, with a **stone** cave or a rocky outcropping, and a **stone** wall with a **stone** door.
	forest	mountain	a man standing in a mountainous **mountain** area, surrounded by trees and rocks.
		garden	a man standing in a **garden** with a large rock and a tree.
		water	a man standing in a **waterlogged** area, surrounded by a **waterfall** and a rocky cliff.
		city	a man standing in a **city** park, surrounded by trees and a large rock formation.
		desert	a man standing in a deserted **desert** area, surrounded by trees and a large rock.
		hall	a man standing in a **hallway** of a cave, surrounded by rocks and trees.
	trees	wild	a man standing in front of a **wild**, **wild** rock formation, surrounded by wildlife and a forest trees.
		flowers	a man standing in a forest, surrounded by **flowers** and **flowers** in the background.
		cloud	a man standing in front of a **cloudy** sky, surrounded by a forest trees and **cloudy** sky.
	rock	house	a man standing in a forest, surrounded by trees and a large **house**.
		tower	a man standing in a forest, surrounded by trees and a towering **tower** towering over him.
		building	a man standing in a forest, surrounded by trees and a large **building**.
		ball	a man standing in a forest, surrounded by trees and a large **ball** of moss.
		bear	a man standing in a forest, surrounded by trees and bushes, with a large **bear bear** nearby.
		human	a man standing in a forest, surrounded by trees and a large **human**-like figure made of leaves.

Image	Source	Target	Edited model output
 LLaVA: a red **flower** with a **butterfly** on it, sitting on a leafy **plant**.	flower	head	a **head** of a red and white plant with a butterfly on it.
		book	a red **book** with a white butterfly on it, placed on a bookshelf.
		road	a red **roadside** plant with a white butterfly on it.
		vehicle	a red **vehicle** with a white butterfly on top of it.
		stream	a red and white flowering plant with a butterfly on **stream**.
		salt	a **salt** and pepper plant with a **salt** and pepper **salt** shaker on top of it.
	butterfly	insect	a flower with **insect** wings on it, sitting on a plant.
		light	a **light** red flower with white petals, sitting on a leafy plant, and surrounded by **light** rain.
		rain	a red flower with **rain** drops on it, sitting on a leafy plant.
	plant	tree	a **tree tree** with a red flower and a butterfly on it.
		wall	a red flower with a butterfly on it, sitting on a **wall**.
		ground	a red flower with a butterfly on it, sitting on a **ground** with green leaves.
		bowl	a red flower with a butterfly on it, sitting on a **bowl bowl** of water.
		tower	a towering **tower** of red flowers planted in a towering **tower** of green towering **tower**.
		park	a red flower with a butterfly on it, sitting on a **park** bench.
		cup	a red flower with a butterfly on it, sitting on a **cup**.
 LLaVA: a **bridge** over a body of **water**, with a **boat** passing underneath it.	bridge	floor	a **floor** with a view of a **floor**, a car, and a **floor** with a view of a car.
		tree	a **tree**-lined road with a **tree** in the foreground, a **tree** in the middle, and a **tree** in the background.
		building	a **building** with a large **building** in the background, a boat on the water, and a **building** on the water.
		bed	a **bed** with a view of a bedroom and a bedroom window with a view of a **bed**.
		hill	a hillside overlooking a **hill** with a hillside road and a hillside **hill**.
	water	beach	a bridge over a **beach**, with a boat in the background and a car driving on the bridge.
		heaven	a bridge over heavenly blue **heaven**, with a boat passing underneath it.
		fire	a bridge over a large body of **fire**, with a boat in the background.
		snow	a bridge over **snowy** mountains, with a boat traveling underneath it.
		city	a bridge over a large body of **city**, with a boat visible in the distance.
	boat	plane	a bridge over a body of water, with a **plane** flying in the background.
		vehicle	a bridge over a body of water, with a **vehicle** driving on it, and a **vehicle** on the other side of the bridge.
		horse	a **horse**-drawn carriage traveling on a bridge over a body of water.
		moon	a bridge over a body of water, with a **moon** in the background.
		sun	a **sunny** day with a bridge over a body of water, with a **sunny** sky in the background.

Table 12: Targeted editing results of example images. For each source concept in the image, we artificially transform it to other target concepts. Target concepts are in bold in the edited model output.