

Large Multi-modal Models Can Interpret Features in Large Multi-modal Models

Supplementary Material

A. Related Works

Dictionary Learning Dictionary learning is a common approach for problems like ours, where we aim to extract a set of features from a collection of dense vectors. Sparse autoencoders (SAEs), proposed by [11, 33], have been used as a classic interpretability method to address this challenge. SAEs are designed to identify mutually incoherent bases in data and represent the data as sparse linear combinations of these bases. Existing studies have applied SAEs to LLMs, finding that the bases represent monosemantic features in the data, with the coefficients indicating the activation of these features[14, 20, 37].

Large Multimodal Models With the development of large language models (LLMs), the performance of large multimodal models has also advanced rapidly, demonstrating strong results across various tasks [3, 17, 22, 40]. Studies such as [25, 34] have explored methods to understand or manipulate the internal structure of LMMs. In our work, we take an initial step toward evaluating and interpreting the open-semantic features within LMMs.

B. Limitations

Our work primarily focuses on the LLaVA-NeXT-LLaMA-8B model and a specific layer within it. This focus on a particular model and layer is based on the assumption of universality and disentanglement, as discussed in [6, 37]. However, this assumption may contribute to inaccuracies in interpretation and model steering.

Due to limitations in computational complexity and storage, we were unable to prepare a sufficiently large and diverse cached image dataset to accurately interpret the image features. Consequently, we present our results on a subset of features and may have mistakenly classified some features as inactive.

C. Detail about Prompt

We detail the prompts used in different stages of the automated pipeline in this section. The prompt for zero-shot identification of concepts is provided in Tab. 5. For this task, we utilize the LLaVA-NeXT-OV-72B model [18]. To refine labels and categorize explanations using large language models (LLMs), we use the prompts detailed in Tabs. 7 and 8. Specifically, LLaMA-3.1-Instruct-8B is used for label refinement, while LLaMA-3.1-Instruct-70B is employed for categorizing explanations. For high-throughput performance, the models are served using SGLang [43].

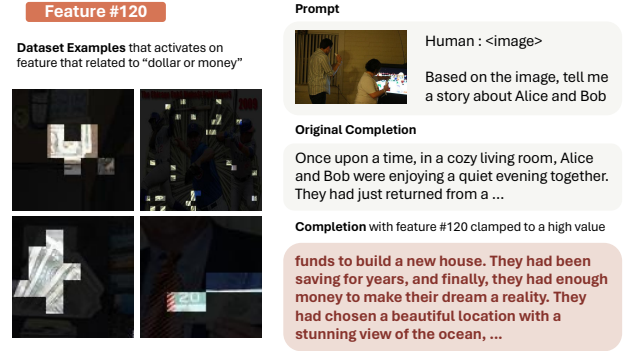


Figure 11. The feature related to money and its steering effect.

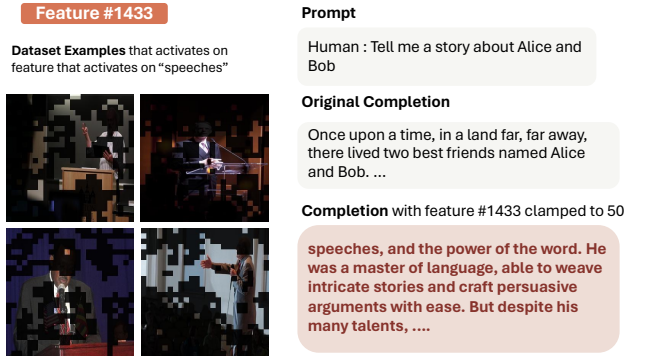


Figure 12. The feature related to speech and its steering effect.

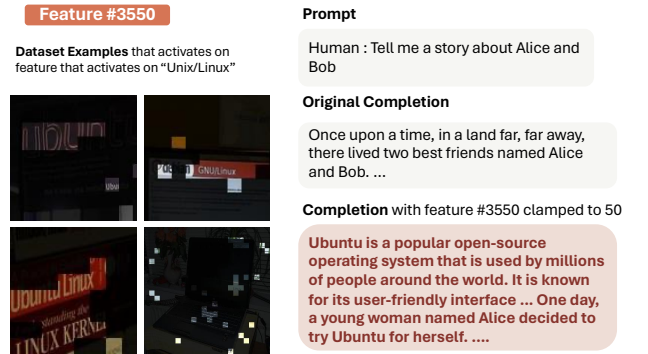


Figure 13. The feature related to unix and its steering effect.

D. Qualitative Steering Experiments

In Tab. 9, we present the results of the steering evaluation for selected cases. Due to the high cost of large-scale steering evaluations, qualitative results are largely absent in the literature, including [37]. To address this, we take an initial

Prompt : Zero-shot Identification of Concepts

You are a meticulous AI researcher conducting an important investigation into a certain neuron in a vision language model.
→ Your task is to analyze the neuron and provide an explanation that thoroughly encapsulates its behavior.

[REQUIREMENTS]

1. Focus only on the highlighted region in each image. If no region is highlighted or if the highlighted region is minimal (e.g., a few bright spots), ignore the image.
→
2. Identify common visual patterns, objects, or concepts in the activated regions. For example, note if highlighted areas show
→ consistent structures, such as mesh patterns or similar objects.

[GUIDELINES]

You will receive a series of images where specific regions have been highlighted to indicate neuron activation. Non-highlighted areas will be masked out or dimmed. Your analysis should consider only the highlighted regions and complete the following tasks:

1. Describe Only the Highlighted Regions: Generate captions solely based on the highlighted regions. If no meaningful pattern is visible, or if only a few scattered spots are highlighted, output: "[EXPLANATION]: Unable to produce descriptions."
2. Concise Description Only: Provide a short, direct description of the common features within the highlighted regions. Avoid any interpretive language—simply state what you see, such as "mesh-like structures" or "actions related to joy or happiness"
3. Output Format: Begin each response with "[EXPLANATION]: " followed by your explanation, if applicable. Ensure the last line of your output follows this format.
→

If unable to determine common visual features, output:

"[EXPLANATION]: Unable to produce descriptions"

Table 5. The prompt for zero-shot identification of concepts

Prompt : GPT-consistency Evaluation

[GUIDELINES]

You are an AI assistant to help assessing whether the generated explanation is consistent with the activation area in the image.
→ The activation area is being highlighted in the image and an explanation is provided for the activation area.

You should output:

- 0 if the explanation is not consistent with the activation area in the image.
- 1 if the explanation is consistent with the activation area in the image.

Please strictly follow the [GUIDELINES] and do not output anything other than the number 0 or 1

Here is the explanation:

{explanation}

ANSWER :

Table 6. The prompt to ask GPT to evaluate the correctness of the evaluation

Prompt : Categorize explanation concept

[GUIDELINES]

You are an AI assistant tasked with assigning a single label based on the given input text. Each input will contain a description
 → of a visual feature, which you must categorize into one of the following classes:

scene – Describes a scene or environment.
 object – Describes an object or entity.
 part – Describes a part or aspect of an object.
 material – Describes a material or substance that constitutes other objects.
 texture – Describes the texture of an object.
 color – Describes the color of an object.

Please provide only the class label from the list [scene, object, part, material, texture, color] with no additional text. Only one
 → label should be chosen. Make sure you only choose from the classes listed above and do not output any other classes.

Categorize the following description:

{description}

ANSWER:

Table 7. The prompt that use to label concept for each description

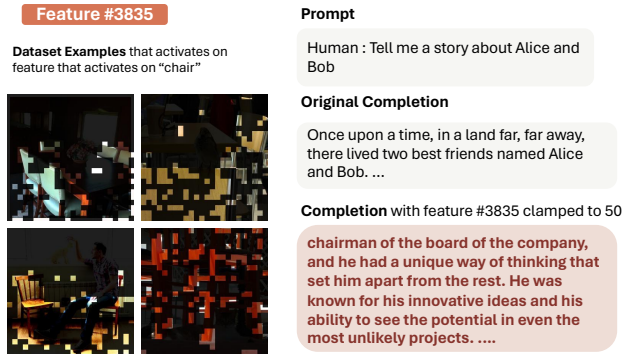


Figure 14. The feature related to chair and its steering effect.

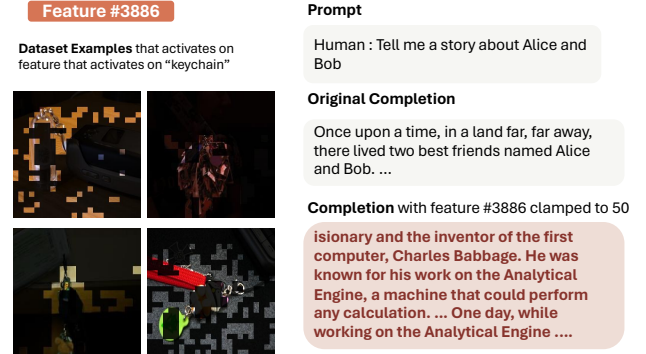


Figure 15. The feature related to money and its steering effect.

exploratory step by using an LLM to assess steering examples, demonstrating a potential solution.

E. More Steering Examples

Sad We present a feature that may be related to the feeling of "sadness" and explore the potential for enabling the model to share emotional responses. After probing and confirming that the feature aligns with "sadness," we investigate whether manipulating this feature could influence the model's reasoning to simulate emotional responses. To test this, we use a simple prompt, "What is your feeling right now?" and ask the assistant. Without steering, the model responds in a neutral, standard AI assistant tone, showing

no emotion. However, when we clamp the "sad" feature to a high value, the model responds with "sad" as shown in Fig. 16

In this section, we provide more steering examples that we discover during experiments. We perform a large scale steering on the 5000 size features subset we choose and then filtered some interesting examples here. In Fig. 11, the feature activates on money and when this feature is clamped to 50, the model output a story about saving funds and by a house. In Fig. 12, when a feature relates to a feature that relate to speech, the model output a story about a man who is a speech master. In Fig. 13, we found a feature that relate to unix/linux and its steering effect would output a story about Ubuntu. More interestingly, in Fig. 14, though the model re-

Prompt : Refine Interpretation

[GUIDELINES]

You are an AI assistant tasked with extracting meaningful labels from descriptions. You will receive a description that may

- contain references to various entities, and your job is to rephrase and extract the key entities from the text. You will
- encounter several types of descriptions, and examples for each case are provided below. Please follow the given
- instructions carefully.

When presenting your answer, first output "[ANSWER]", followed by the extracted entity. Thank you!

Case 1: Good Description

In this case, the description clearly identifies the entity.

Examples:

Description: The cell phone.

Output: [ANSWER] The cell phone

Description: The letters on the shipping containers.

Output: [ANSWER] The letters on the shipping containers

Case 2: Description includes additional words

In this case, the description contains more information than needed. Extract only the key entity.

Examples:

Description: The images all display different models of Honda vehicles, suggesting the neuron is activated by the presence of

- Honda vehicles or the Honda logo.

Output: [ANSWER] Honda vehicles

Description: The neuron seems to be reacting to the word "ORD" on the billboard. It could be part of a larger word or phrase,

- but the neuron specifically highlights the letters "ORD." This suggests that the neuron might be specialized in
- recognizing or processing certain words or characters in images. The activation across the images indicates that the
- neuron is consistent in its response to textual elements, particularly those that include the "ORD" sequence.

Output: [ANSWER] The word "ORD"

Case 3: Bad Description

In this case, the description does not provide sufficient or valid information.

Examples:

Description: Unable to produce descriptions.

Output: Unable to produce descriptions

[Description]

{description}

Table 8. The prompt that used to refine the explanations for grounding and segment visual objects

	Steering	Random
GPT-4o Score	6.36	2.02

Table 9. GPT4 Score Evaluation of Steering Effects

this feature related to "key" or "keychain", the model output a story about developing some analytic software.

F. CLIP-Score and IOU details

We use Grounding DINO L [24] as our grounding module and SAM Huge [16] as our segment module. The output from the interpretation pipeline is being refined into con-

sponse on a visual "chair" object, when steering this feature, model would output a story relates to "chairman" instead of a "chair". Another example is that in Fig. 15, when steering

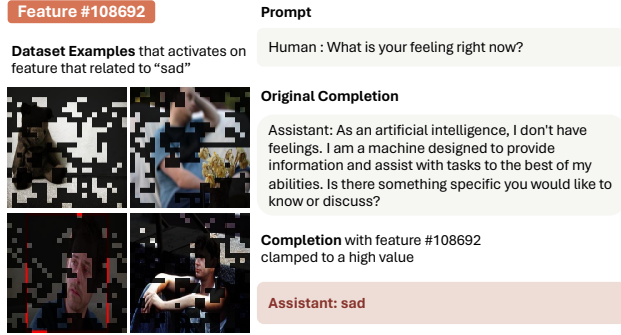


Figure 16. The feature that relates to sad. We probe and find out the feature that activated on “sad”. By clamping this feature, we can enforce the model to share the feeling of sad

cise description by using the LLaMA-3.1-Instruct-8B [10]. We use ViT-B/32 CLIP model to generate embeddings and calculate the cosine similarity between the interpretations and the image. We calculate the IOU and the CLIP-Score using the top-5 activated images for each features. Due to the same limitation as illustrate in [6, 37], we report the result on a 5000 subset of features with around 46684 images for caching the features’ activations.

G. Feature Probing

Due to the large number of features, identifying specific features of interest is challenging, and interpreting all available features before making a selection is impractical. Following Templeton et al. [37], we also probe into the features of our SAE to identify several emotion-related features that may influence the model’s perceived emotional responses. We first prepare an image representing a specific emotion, then select the top-k activated features for that image to run through our explanation and steering pipeline. From the output, we manually select the desired features and validate them through steering and activated examples. Unlike the approach in [37], which uses only the top 5 activated features, we found that a higher value of k is preferable because a single image can contain many low-level visual features and diverse semantic information. In practice, we select $30 \leq k \leq 100$ and skip some of the top-activated values to exclude low-level visual features.

H. Low Level Perception Features Examples

We identify many low-level visual features from the model that differ from the text-based features in large language models (LLMs). These visual features are strongly activated across most images and represent the model’s basic perceptual and cognitive abilities. In Fig. 17, we present examples of features activated by structure, shape, and color. In many of our probing trials, these features exhibit high activation levels and respond to various aspects of the im-

ages. We believe these features function as universal elements in how language-vision models (LLMs) understand the world.

I. More Model comparison

	IOU	IOU(random)	CS	CS(random)
Qwen-2.5-VL	26.67	0.06	27.99	18.22
InstructBLIP-7B	-	-	28.01	17.71

Table 10. Pipeline results on Qwen2.5-VL

We provide a further experiment using Qwen-2.5-VL to prove the generalizability of our methods. As shown in Tab. 10

J. Hallucination Steering Examples

	Better	Same	Worse
HalluBench	0.09	0.89	0.02

Table 11. Hallucination Case study on 100 examples on Hallucination Bench with a single feature clamped at high value

We conduct a small-scale experiment on the Hallucination Bench by clamping irrelevant features, and the results are presented in Table 11. Among the 100 examples, clamping led to improved performance in 9 cases. Although this investigation is still in its early stages, we believe this approach shows potential for reducing hallucinations.

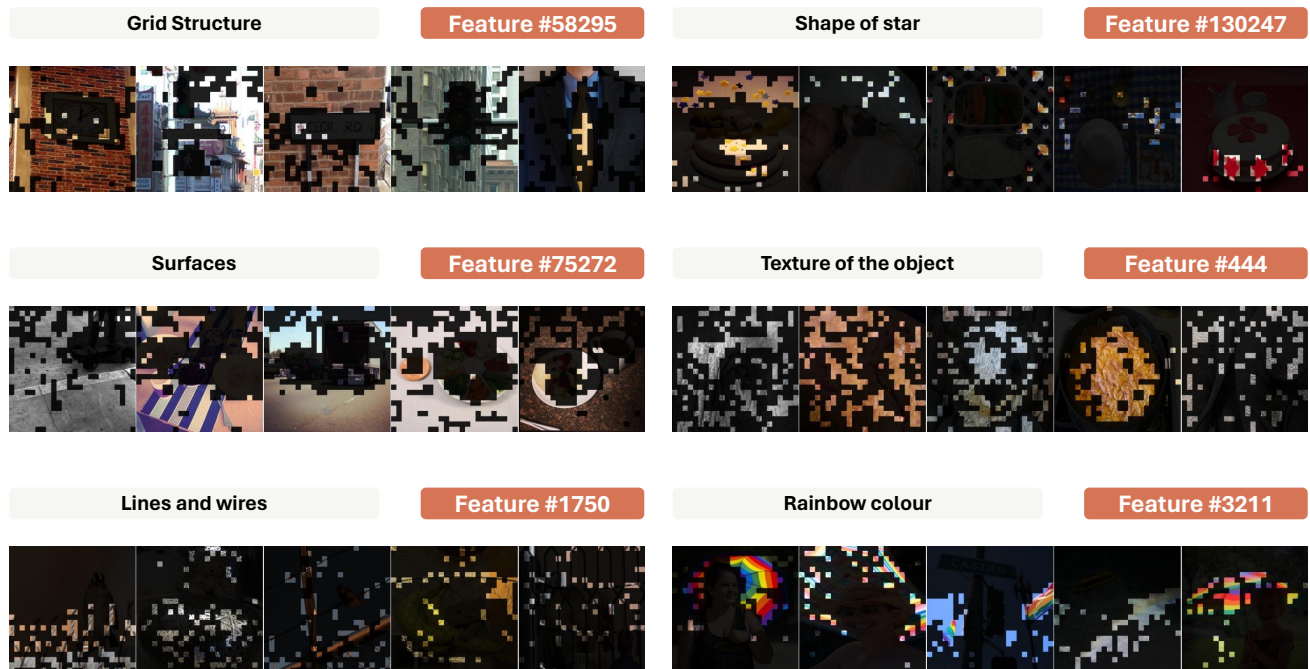


Figure 17. Low level features in the LMM. These features activate in most of the images and showcase the model’s basic cognition and perception abilities.

Feature #4575

Dataset Examples that activates on feature that activates on "cell phones"

Prompt

Human : Tell me a story about Alice and Bob

Original Completion

Once upon a time, in a land far, far away, there lived two best friends named Alice and Bob. ...

Completion with feature #4575 clamped to 50

Bluetooth is a wireless communication technology that allows devices to communicate with each other without the need for cables or wires. It is commonly used in devices such as smartphones,

Figure 18. The feature related to money and its steering effect.