Patchwise Cooperative Game-based Interpretability Method for Large Vision-language Models

Yao Zhu[1] **Yunjian Zhang**[1*] **Zizhe Wang**[1] **Xiu Yan**[2] **Peng Sun**[3] **Xiangyang Ji**[1]
[1]Tsinghua University, China [2]Meituan Group, China
[3]Central University of Finance and Economics, China
ee_zhuy@zju.edu.cn, sdtczyj@gmail.com, wangzz@act.buaa.edu.cn,
yanx18@tsinghua.org.cn, 2023212399@email.cufe.edu.cn, xyji@tsinghua.edu.cn

Abstract

Amidst the rapid advancement of artificial intelligence, research on large vision-language models (LVLMs) has emerged as a pivotal area. However, understanding their internal mechanisms remains challenging due to the limitations of existing interpretability methods, especially regarding faithfulness and plausibility. To address this, we first construct a human response interpretability dataset that evaluates the plausibility of model explanations by comparing the attention regions between the model and humans when answering the same questions. We then propose a patchwise cooperative game-based interpretability method for LVLMs, which employs Shapley values to quantify the impact of individual image patches on generation likelihood and enhances computational efficiency through a single input approximation approach. Experimental results demonstrate our method's faithfulness, plausibility, and robustness. Our method provides researchers with deeper insights into model behavior, allowing for an examination of the specific image regions each layer relies on during response generation, ultimately enhancing model reliability. Our code is available at `https://github.com/ZY123-GOOD/Patchwise_Cooperative`.

1 Introduction

Recently, large language models (LLMs) such as those in the families of GPT (Brown et al., 2020) and Llama (Touvron et al., 2023) have showcased exceptional capabilities across diverse domains, drawing widespread attention from both academia and industry. Building upon this progress, recent advancements, including Qwen-vl (Bai et al., 2023), ShareGPT4V (Chen et al., 2024), and LLaVA (Li et al., 2024), have extended LLMs to encompass visual understanding, giving rise to large vision-language models (LVLMs). These models are capable of performing complex text and visual tasks in response to human instructions, demonstrating a high level of competency and versatility.

However, these LVLMs often considered ''black-box'' systems with opaque internal mechanisms, posing potential risks to downstream applications due to their lack of transparency. Developing interpretability techniques to understand and explain these models is therefore essential for clarifying their behavior, limitations, and societal impact. Interpretability refers to the ability of a model to articulate or demonstrate its functioning in a way that is comprehensible to humans. For users, interpretability provides a clear understanding of the reasoning behind model responses, enabling them to grasp the operational logic without requiring specialized expertise, thereby fostering reasonable trust in the model. For researchers, interpretability helps reveal unexpected biases and potential risks in model behavior, identifies areas requiring improvement, and provides guidance for optimizing model performance.

The interpretability of LVLMs poses unique challenges compared to earlier deep models, primarily due to their high complexity. These models often contain billions of parameters, leading to intricate internal representations and reasoning processes that significantly increase computational costs during interpretation (Zhao et al., 2024). Recent efforts, such as LVLM-Interpret (Stan et al., 2024), have begun to explore the interpretability of these models; however, challenges remain, particularly in providing explanations that are both comprehensible and easily evaluable by humans.

This paper proposes a Patchwise Cooperative Game-based Interpretability Method. The method

[*] Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 744–759, 2025. https://doi.org/10.1162/tacl_a_00756
Action Editor: Grzegorz Chrupała. Submission batch: 9/2024; Revision batch: 1/2025; Published 7/2025.

starts by dividing the image into non-overlapping patches, treating each patch as a player in a cooperative game, and using Shapley values (Shapley et al., 1953) to measure each player's contribution to the generation likelihood. This results in importance heatmaps constructed from Shapley values, highlighting which parts of the input image contribute most to the model's generated response. Although treating image patches as players significantly reduces the computational cost compared to treating individual pixels as players, the large number of parameters in LVLMs and the extensive sampling required for Shapley value calculations still pose substantial computational challenges. To address this, we propose a single input approximation method that approximates LVLM behavior in local regions using an extremely simplified multilayer perceptron, effectively reducing the issue of handling large parameter counts. Additionally, we introduce a Monte Carlo sampling-based Shapley value approximation method (Song et al., 2016) to alleviate the excessive sampling burden. Furthermore, we construct a human response interpretability dataset to evaluate the similarity between model and human attention when responding to the same questions, thereby assessing whether the explanations provided by the interpretability method are plausible to humans.

Experimental results demonstrate that our method provides faithful and plausible explanations of LVLM responses and can be used to observe attention across different layers of LVLMs. The method also shows robust performance across various questioning methods and image domains. Notably, our method only requires access to the model's output probabilities to generate explanations, without needing information about the model's architecture, parameters, or gradients, making it easily applicable to different LVLMs.

The key contributions are summarized as follows:

* We provide a human response interpretability dataset designed to evaluate the plausibility of model explanations, enabling researchers to easily compare the attention patterns of LVLMs with those of humans during response generation.

* We introduce a single sample approximation method for large vision-language models, which reduces the computational cost of interpretability by approximating the model's behavior in local regions.

* We propose the Patchwise Cooperative Game-based Interpretability Method, providing a convenient, faithful, and plausible way to explain LVLMs.

* We observe that the attention in the earlier layers of the model tends to be more dispersed, often leading to incorrect responses, whereas the attention in the later layers is typically more focused on the target, resulting in correct responses.

2 Related Work

In the realm of interpretability in neural networks (Ribeiro et al., 2016; Wexler et al., 2019; Zhang et al., 2019), researchers endeavor to dissect the response-generating processes inherent in neural networks. This not only deepens human understanding of the model's internal workings but also facilitates iterative improvements in model architecture and application.

2.1 Interpretability in Vision Models

Key advancements in interpretability research for computer vision tasks include the following:

- Gradient-based Attribution Methods: Early gradient-based techniques, such as input gradients (Sung, 1998; Baehrens et al., 2010), evaluate the influence of individual pixels or features by analyzing the gradients of the model's predictions with respect to the input data. Subsequent advancements include Guided Backprop (Springenberg et al., 2015), SmoothGrad (Smilkov et al., 2017), and Full-Gradients (Srinivas and Fleuret, 2019), which offer refined methods for attribution.

- CAM (Class Activation Mapping) and its Extensions: CAM (Zhou et al., 2016) generates class-specific activation maps via global average pooling, identifying salient regions crucial for specific class predictions. GradCAM (Selvaraju et al., 2017) and SmoothGradCAM (Omeiza et al., 2019), combine gradient information with CAM methodology to broaden its applicability, offering more clear visual explanations. Wang et al. (2020b) introduce ScoreCAM, which calculates the weight of each activation map based on the forward pass score of the target class, reducing the dependency on gradients. Wang et al. (2020a) propose Smoothed Score-CAM that further enhances object feature localization performance through smoothing operations.

- Shapley Value Methods: Grounded in game theory, the Shapley value (Shapley et al., 1953) quantifies the contribution of each pixel or feature to a model's predictions. Sun et al. (2024) propose employing the Segment Anything Model (SAM) (Kirillov et al., 2023) to identify distinct concepts in images, followed by computing the Shapley values of these concepts to explain the decision-making process in image classification models. While Shapley-based methods provide a strong theoretical foundation, they are computationally intensive. Current research seeks to mitigate these costs through techniques such as Monte Carlo sampling (Song et al., 2016), model-specific approximations (Ancona et al., 2019), and small surrogate models (Lundberg and Lee, 2017; Covert and Lee, 2021).

2.2 Interpretability in Multimodal Models

With the significant progress and widespread application of multimodal model in recent years (Liu et al., 2024), there is a growing demand for research on the interpretability of multimodal models. The existing research on the interpretability of multimodal models (Chefer et al., 2021; Aflalo et al., 2022; Lyu et al., 2022) primarily focuses on classical vision-language models, such as LXMERT (Tan and Bansal, 2019), CLIP (Radford et al., 2021), and ALBEF (Li et al., 2021b). These models fuse and align visual and linguistic information encoded by their respective encoders using techniques like contrastive learning, and are commonly applied to tasks such as classification and image-text matching.

Chefer et al. (2021) utilize the attention layers of a multimodal model to generate correlation maps for interactions between input modalities within the network. Aflalo et al. (2022) propose VL-Interpret, which interprets the attention and hidden representations in multimodal transformers by tracking various statistics of attention heads across all layers in both visual and language components. Lyu et al. (2022) decompose the model into unimodal contributions and multimodal interactions, generating visualized explanations for each part, thereby enabling a more accurate and fine-grained analysis of the decision-making process in multimodal models. Ramesh and Koh (2022) propose an interpretability framework for attention interactions in the VisualBERT multimodal transformer model (Lu et al., 2019) using Label Attribution and Optimal Transport of the vision-language semantic spaces. Parcalabescu and Frank (2023) propose MM-SHAP, a method for measuring the contribution of each modality in visual and language tasks based on interpretability, and note that the CLIP model exhibits a relatively balanced dependency between the two modalities. Cafagna et al. (2023) propose a method for improving the perceived quality of explanations by using semantic visual priors from visual backbones such as CNNs and Vision Transformers. Cinà et al. (2023) consider how to evaluate whether explanations capture concepts of interest to humans and proposes an evaluation method based on semantic matching. Aggarwal et al. (2024) explain the behavior of multimodal models by calculating the independent contributions of the text and image modalities to the predictions, and interestingly finds that the visual component of Hate Meme Detection Models exhibits limited transferability, with its generalization largely relying on the textual part of the meme.

However, methods tailored to classical vision-language models are not well-suited for modern large vision-language models, which have recently gained significant attention and are typically based on decoder-only architectures of large language models with billions of parameters (Zhao et al., 2024). LVLMs utilize generative models to process multimodal embeddings that combine image features and word embeddings, generating image captions or answering questions by predicting the next token in sequence, primarily applied to generative tasks (Chen et al., 2024; Li et al., 2024). Recently, Stan et al. (2024) conducted an early study on the interpretability of LVLMs and proposed a new interactive tool, LVLM-Interpret, to identify the crucial parts of an image that contribute to generating responses. The explanations provided by this tool contain significant noise and have poor visual quality. As shown in Figure 1, LVLM-Interpret (Stan et al., 2024) provides an explanation for the question ''Is there a glass?'' that does not focus

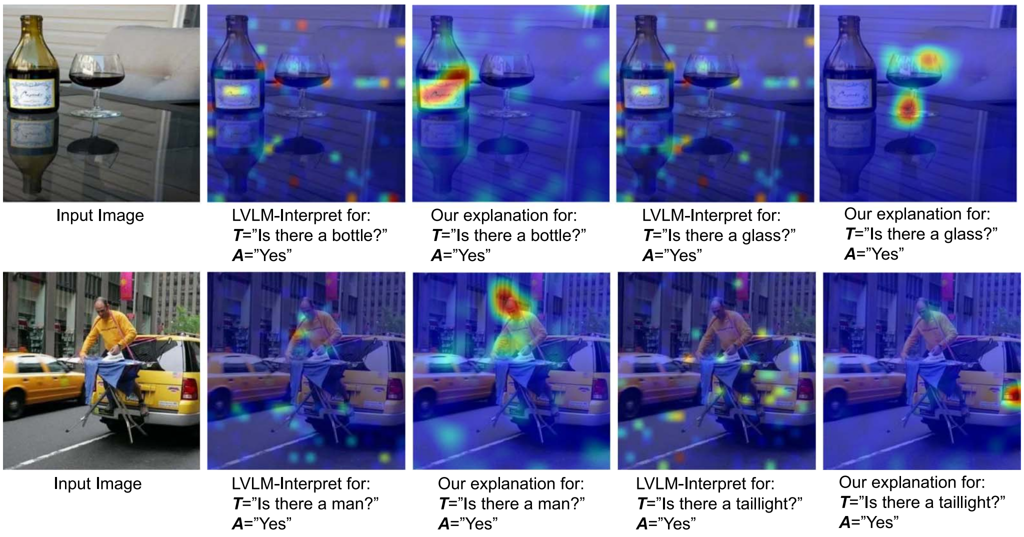

Input Image
LVLM-Interpret for: T="Is there a bottle?" A="Yes"
Our explanation for: T="Is there a bottle?" A="Yes"
LVLM-Interpret for: T="Is there a glass?" A="Yes"
Our explanation for: T="Is there a glass?" A="Yes"
Input Image
LVLM-Interpret for: T="Is there a man?" A="Yes"
Our explanation for: T="Is there a man?" A="Yes"
LVLM-Interpret for: T="Is there a taillight?" A="Yes"
Our explanation for: T="Is there a taillight?" A="Yes"

Figure 1: Comparison with the recent work LVLM-Interpret. Our method accurately provides explanations for the outputs generated by LVLMs, aiding researchers in better understanding multimodal models. In contrast, the explanations provided by LVLM-Interpret are noisy, making them harder for humans to comprehend.

on the main part of the glass, but instead contains considerable noise.

In summary, research on the interpretability of LVLMs is still in its early stages, and the few existing studies face challenges of providing explanations that are noisy and difficult for humans to interpret. Therefore, this paper aims to provide efficient, faithful, and plausible explanations for the responses of large vision-language models. As shown in Figure 1, the explanations provided by our method align better with human visual perception than those of LVLM-Interpret.

3 Method

3.1 Desiderata

The interpretability research of LVLMs is crucial for enhancing transparency and building user trust in the LVLMs' responses. Following previous research (Zhao et al., 2024; Sun et al., 2024; Li et al., 2021a), this subsection briefly introduces two desiderata of interpretability methods.

3.1.1 Faithfulness

Faithfulness reflects the capability of interpretability methods to capture the internal logic of LVLMs. Let G denote the LVLM to be analyzed and E represent the explanation provided by the interpretability method. Faithfulness can be defined as the correlation between E and the actual response-generating process of G, with a higher correlation indicating a more effective explanation. The latest debiased and no-retraining evaluation framework ROAD (Rong et al., 2022) can be used to assess faithfulness by perturbing input pixels considered most or least important by the explanation E and observing their impact on the large vision-language model's responses. In the LeRF (Least Relevant First) setting, pixels are arranged in ascending order of importance, so perturbing the top $k\%$ of pixels should have minimal impact on the response of LVLMs. In contrast, in the MoRF (Most Relevant First) setting, pixels are arranged in descending order of importance, and perturbing the top $k\%$ of pixels is expected to significantly impact the generation results. During the evaluation process, k is gradually increased from 0 to 100.

Regarding how to perturb input pixels, we adopt the method from the ROAD framework (Rong et al., 2022), where each selected pixel is replaced with the weighted average of its surrounding pixels as follows:

$$\begin{aligned}\boldsymbol{x}_{i,j} = w_1(&\boldsymbol{x}_{i,j+1} + \boldsymbol{x}_{i,j-1} + \boldsymbol{x}_{i+1,j} + \boldsymbol{x}_{i-1,j}) \\ &+ w_2(\boldsymbol{x}_{i+1,j+1} + \boldsymbol{x}_{i+1,j-1} \\ &+ \boldsymbol{x}_{i+1,j+1} + \boldsymbol{x}_{i-1,j-1}),\end{aligned} \tag{1}$$

where w_1, w_2 are constant coefficients for direct neighbors and indirect, diagonal neighbors. Given that the weights need to sum up to 1 for weighted interpolation, w_1 and w_2 are set to $\frac{1}{6}$ and $\frac{1}{12}$, respectively.

3.1.2 Plausibility

Plausibility reflects whether the explanations provided by interpretability methods are understandable to humans, which is crucial for enabling users to comprehend the reasoning behind the model's responses and to build trust in the model. To date, there is no consensus in research on the interpretability of LVLMs regarding how to quantitatively measure plausibility (Zhao et al., 2024). To address this, this paper constructs a human response interpretability dataset, which can be used to assess plausibility by calculating the distance between model explanations provided by interpretability methods and human-given explanations.

Specifically, the human response interpretability dataset leverages the images from the recent large-scale vision-language model object hallucination assessment dataset, POPE (Li et al., 2023), and adopts the question format used in the POPE dataset, such as ''*Is there a snowboard in the image?*,'' with answers provided as ''yes'' or ''no.'' Our proposed dataset allows for the evaluation of whether the visual information relied upon by the model aligns with human semantic cognition by comparing the most influential visual regions for the model's responses with human annotations, making it suitable for assessing the plausibility of interpretability methods. We divide the image into patches of 12 rows and 12 columns and ask the annotators to select the three patches that have the most significant impact on their response. If the annotators believe that the relevant object is not present in the image, or if there are too many patches related to the object, making it confusing to identify the top three patches related to their response, they pass on the sample. See annotation examples in Figure 2.

All of our annotators have a background in computer science or statistics at the master's level or higher. To prevent a small number of annotators from dominating the annotation process, we limited each annotator to a maximum of 100 tasks. In total, we collected 6,000 annotations from 60 participants. We retained images annotated by more than four participants and selected the top three most important patches based on a voting method, resulting in a dataset of 891 images along with their corresponding human response interpretability results.

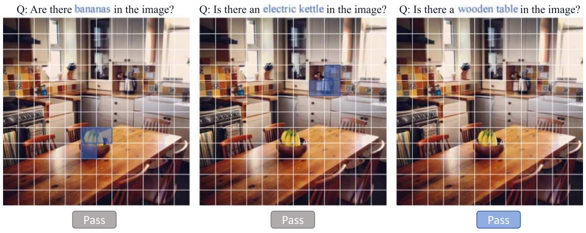

Q: Are there bananas in the image?
Q: Is there an electric kettle in the image?
Q: Is there a wooden table in the image?
Pass
Pass
Pass

Figure 2: Annotation examples: Annotators select the three patches most relevant to their response. If they are unsure about selecting the top three related patches, they may choose to pass.

Regarding how to evaluate the plausibility of model interpretability results based on this dataset, we propose two metrics: Top1 Interpretability Distance (TID_1) and Top3 Interpretability Distance (TID_3). The former calculates the minimum Euclidean distance between the most important patch for the model and the top three patches identified by humans, while the latter calculates the minimum Euclidean distance between the model's top three important patches and the top three patches identified by humans:

$$\mathrm{TID}_1 = \min_{i \in \{0,1,2\}} \sqrt{(\mathrm{col}_0^m - \mathrm{col}_i^h)^2 + (\mathrm{row}_0^m - \mathrm{row}_i^h)^2}, \quad (2)$$

$$\mathrm{TID}_3 = \min_{i,j \in \{0,1,2\}} \sqrt{(\mathrm{col}_j^m - \mathrm{col}_i^h)^2 + (\mathrm{row}_j^m - \mathrm{row}_i^h)^2}, \quad (3)$$

where col_j^m and row_j^m denote the column and row indices, respectively, of the j-th most important patch identified by the model, while col_i^h and row_i^h represent the column and row indices, respectively, of the i-th most important patch identified by humans.

As to the reason for selecting the top three patches instead of just one, it is primarily based on the following two considerations. On the one hand, it takes into account the issues of diversity and coverage, as key features of an object may be distributed across different parts of the image. By selecting multiple regions rather than just one, we can more comprehensively capture the participants' understanding of the image. For example, when determining whether a cat is present in an image, both the cat's nose and ears could be key features. On the other hand, it also concerns user experience, as requiring participants to choose only one patch could increase their cognitive burden by demanding high confidence in

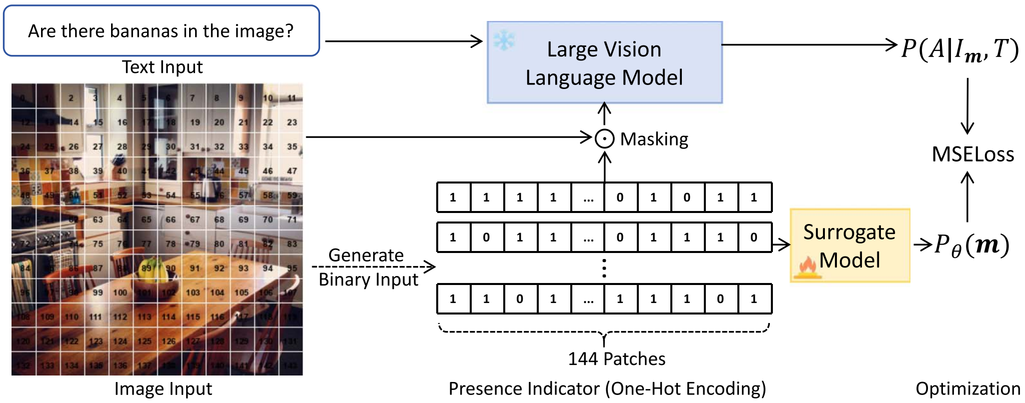


Figure 3: The training process of the surrogate model: The input is a binary sequence representing the presence or absence of each patch. The optimization objective is to make the output probabilities of the surrogate model on the binary input match the generation likelihood of the LVLM on the image after applying the binary sequence as a mask, thereby approximating the behavior of the original complex model in the local region.

their decision. Allowing them to select multiple patches helps to reduce this cognitive burden. This dataset will be made available to the community at the earliest convenience. Using our proposed annotation method, researchers can also generate human response data that is more suitable for their specific tasks to evaluate the plausibility of model responses.

3.2 Patchwise Cooperative Game-based Interpretability Method

Although existing work on the interpretability of visual models has been able to faithfully reflect model decision behaviors to some extent and provide explanations that are reasonable to humans, challenges remain in reasonably interpreting the output of LVLMs due to the complexity of their internal representations and reasoning processes. For example, the latest LVLM-Interpret method (Stan et al., 2024) proposes an interactive tool for interpreting the inner attention mechanisms of LVLMs. However, the explanation results contain significant noise, making it difficult for humans to understand which parts of the image contribute most to the generated response.

This paper proposes an interpretability method based on a patchwise cooperative game, aiming to answer the question in a faithful and plausible manner: ''*Which regions of an image does the generated response of a large vision-language model actually rely on?*'' This is crucial for determining whether users can trust the model.

We denote the user's image input as $\boldsymbol{I}$, the text input as $\boldsymbol{T}$, and the model's generated response as $\boldsymbol{A}$. For example:

$$\boldsymbol{T} = \text{``What sport is this man playing?''}$$
$$\boldsymbol{A} = \text{``Frisbee''}$$

In this way, given the image input $\boldsymbol{I}$ and the text input $\boldsymbol{T}$, the generation likelihood of the answer $\boldsymbol{A}$ from an LVLM can be formulated as $P(\boldsymbol{A}|\boldsymbol{I}, \boldsymbol{T})$.

First, given that the input size of large vision-language models is typically large, such as $3 \times 336 \times 336$, calculating the importance for each pixel would be computationally expensive. Inspired by the commonly used image patching operation in vision transformers (Dosovitskiy et al., 2020), this paper divides the input $\boldsymbol{I}$ into $N = 12 \times 12$ image patches $C = \{C_1, C_2, \ldots, C_N\}$, as shown in Figure 2.

Second, given that LVLMs have an enormous number of parameters, leading to inherent computational complexity, we propose a single input approximation method that uses a simple two-layer perceptron to approximate the generation likelihood of LVLMs in the local region of a single input sample as illustrated in Figure 3. Intuitively, we aim to replace the target LVLM with a surrogate model that retains the same functionality of the target LVLM in the local region of a single input sample while achieving higher computational efficiency.

Formally, we denote the surrogate model that takes the one-hot embedding as input as

$f' : \{0,1\}^N \rightarrow [0,1]$ and the one-hot embedding as $\boldsymbol{m} = [m_1, m_2, \ldots, m_N]$, where $m_i \in \{0,1\}$ indicates whether the i-th patch C_i is present in the input. If $m_i = 1$, it means C_i is present; if $m_i = 0$, it means C_i is masked. In practice, we randomly generate 1000 one-hot embeddings and mask the given input according to the patch presence represented by the one-hot embeddings as:

$$\boldsymbol{I_m} = \sum_{i=1}^{N} m_i \cdot C_i, \tag{4}$$

which means that the image patches corresponding to the indices with a value of 0 in the one-hot embedding are masked. We represent the probabilities output by the surrogate model as $P_\theta(\boldsymbol{m})$, where θ denotes the parameters of the surrogate model f'. The masked image $\boldsymbol{I_m}$ is input into the target LVLM to obtain the generation likelihood of the answer $\boldsymbol{A}$ as $P(\boldsymbol{A}|\boldsymbol{I_m}, \boldsymbol{T})$. Let $\mathbb{M}$ represent the distribution of one-hot embeddings. The formulation of the optimization for the surrogate model can be expressed as:

$$\min_\theta \mathbb{E}_{\boldsymbol{m} \in \mathbb{M}}[(P_\theta(\boldsymbol{m}) - P(\boldsymbol{A}|\boldsymbol{I_m}, \boldsymbol{T}))^2], \tag{5}$$

which encourages the surrogate model to output probabilities that closely match the LVLM's probability of generating a specific response for the given masked input images, thus approximating the LVLM in the local region of a single input sample.

Next, we introduce the concept of Shapley value (Shapley et al., 1953; Chen et al., 2023) to quantify the contribution of each patch to the generation likelihood. Originally introduced in cooperative game theory, the Shapley value (Shapley et al., 1953) measures the contribution of each player to the total payoff within a coalition. The Shapley value for player i is determined by calculating their average marginal contribution across all possible coalitions, accounting for every potential order of players. In this paper, we consider each patch C_i as a player. These players contain an equal number of pixels. Then, we can define the marginal contribution of a patch C_i as the difference between the likelihood predicted by the surrogate model f' on $\mathcal{S} \cup \{C_i\}$ and $\mathcal{S}$, where $\mathcal{S} \subseteq C \setminus \{C_i\}$:

$$\begin{aligned} \Delta_{C_i}(S) = P_\theta(&\text{OneHot}(\mathcal{S} \cup \{C_i\})) \\ &- P_\theta(\text{OneHot}(\mathcal{S})), \end{aligned} \tag{6}$$

where OneHot$(\cdot)$ denotes the process of obtaining the one-hot encoding. Hence, the Shapley value of the each patch C_i can be formulated as:

$$\phi_{C_i} = \frac{1}{N} \sum_{q=1}^{N} \frac{1}{\binom{N-1}{q-1}} \sum_{\mathcal{S} \in \mathcal{S}_q(i)} \Delta_{C_i}(S), \tag{7}$$

where $\mathcal{S}_q(i)$ is the collection of all coalitions of size q that does not contain patch C_i. In practice, due to the prohibitively large number of possible coalitions, we approximate the Shapley value using the Monte Carlo sampling approach introduced by Song et al. (2016). Specifically, we sample Q coalitions for each patch and estimate the Shapley value as follows:

$$\hat{\phi}_{C_i} = \frac{1}{Q} \sum_{q=1}^{Q} \Delta_{C_i}(S_q), \tag{8}$$

where S_q is the q-th sampled coalition. In this manner, we can obtain the Shapley values for each patch, thereby constructing the importance map represented by Shapley values as $\Phi = \{\hat{\phi}_{C_1}, \hat{\phi}_{C_2}, \ldots, \hat{\phi}_{C_N}\}$. By upsampling this importance map to the size of the input image, we can generate an importance heatmap for the input image, highlighting the regions within the image that have a more substantial impact on the model's generation likelihood.

4 Experiments

We evaluate the faithfulness and plausibility of the proposed interpretability method on four LVLMs, including LLaVA-v1.5-7B, LLaVA-v1.5-13B (Liu et al., 2023), ShareGPT4V-7B, and ShareGPT4V-13B (Chen et al., 2024). All experiments are conducted on four RTX 4090 GPUs. To account for randomness, all experimental results are averaged across three random seeds.

4.1 Evaluations on Faithfulness

In this subsection, we assess the faithfulness of our method from two aspects: local faithfulness and global faithfulness (Tomsett et al., 2020). The former is measured on a single sample, while the latter is evaluated on a set of samples.

As for local faithfulness, we select three different images and explain which regions of the images the model's responses were based on. We assess faithfulness using the ROAD evaluation framework introduced in Section 3.1.1. Under the

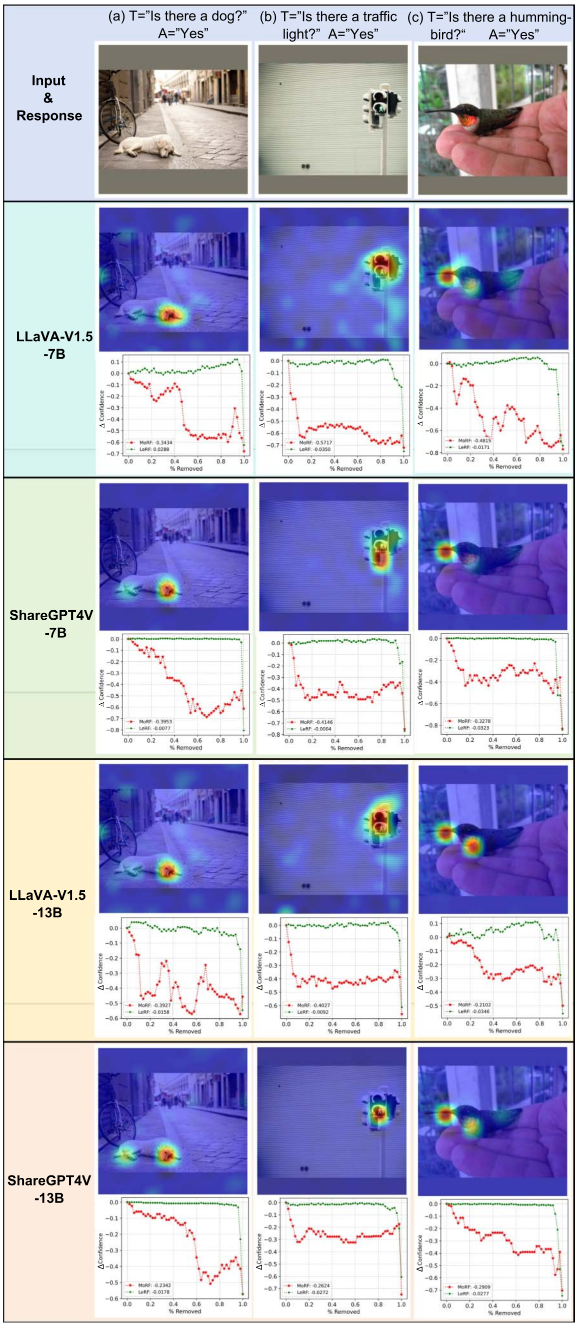
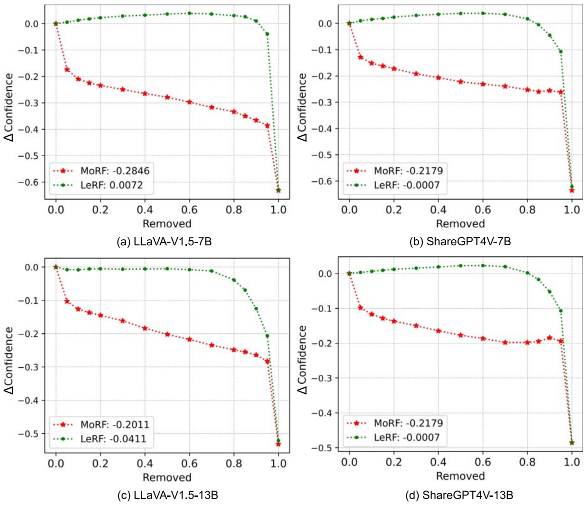


Figure 4: Evaluation of local faithfulness. The image shows importance heatmaps and confidence change curves under the MoRF and LeRF settings. The numbers represent the area between the curves and the horizontal axis, where areas below the axis are negative and areas above the axis are positive.

LeRF setting, we progressively erase the regions deemed unimportant by the explanation and show the changes in the model's generation probability as different proportions of the image were erased: ΔConfidence = $P(\boldsymbol{A}|\boldsymbol{I}^p, \boldsymbol{T})$ - $P(\boldsymbol{A}|\boldsymbol{I}, \boldsymbol{T})$, where $\boldsymbol{I}^p$ is the image after erasing $p\%$ of the least important regions. Conversely, under the MoRF setting, we progressively erased the regions deemed important by the explanation. As shown in Figure 4, we illustrate the importance heatmaps and confidence change curves for different models in response to three images and their corresponding questions. The numbers represent the area under the curve with respect to the horizontal axis.



(a) LLaVA-V1.5-7B (b) ShareGPT4V-7B (c) LLaVA-V1.5-13B (d) ShareGPT4V-13B

Figure 5: Evaluation of global faithfulness. The image shows the confidence change curves under the MoRF and LeRF settings for different models.

Overall, when the erasure proportion is less than 80%, the generation likelihood remains stable or even slightly increases under the LeRF setting, while it sharply decreases under the MoRF setting even with a small proportion of erasure. Taking the LLaVA-V1.5-7B model as an example, for the traffic light image, even when erasing up to 80% of the irrelevant regions, the generation likelihood does not change significantly. However, when erasing only 10% of the regions deemed important, the generation likelihood drops by about 0.6. This phenomenon indicates that the Patchwise Cooperative Game-based Interpretability Method proposed in this paper provides explanations that faithfully capture the most and least important content in the input for the model's response generating process.

Regarding global faithfulness, we conducted experiments on the widely used POPE dataset (Li et al., 2023), with Figure 5 illustrating the average results across different images. As observed from the LeRF curve, when 90% of the unimportant pixels are removed, the model's generation likelihood hardly decreases and may even slightly increase. Conversely, removing just 10% of the most important pixels leads to a sharp decline in

Model	Ours		LVLM-Interpret	
	TID1	TID3	TID1	TID3
LLaVA-v1.5-7B	0.905	0.387	3.642	1.632
ShareGPT4V-7B	0.902	0.325	3.653	1.643
LLaVA-v1.5-13B	0.901	0.378	3.501	1.698
ShareGPT4V-13B	1.142	0.446	3.747	1.768

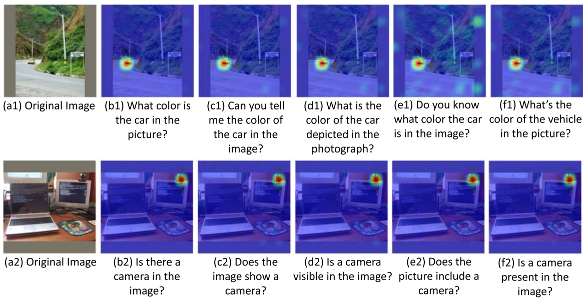
Table 1: Evaluation of the plausibility of different interpretability methods on LLaVA-V1.5-7B, ShareGPT4V-7B, LLaVA-V1.5-13B, and ShareGPT4V-13B.

generation likelihood. These results are consistent with the observations from the local faithfulness evaluation. Since global faithfulness is assessed on the entire dataset, the resulting curves are smoother.

4.2 Evaluations on Plausibility

Plausibility, alongside faithfulness, is another key focus of interpretability methods. The importance heatmaps in Figure 4 qualitatively demonstrate that the proposed method provides human-understandable interpretability. This subsection aims to quantitatively explore the similarity between the regions of the image that the model attends to when answering questions and those that humans focus on in response to the same questions. Using the human response interpretability dataset proposed in this paper, we evaluated the interpretability of various LVLMs based on the Top1 Interpretability Distance (TID_1) and Top3 Interpretability Distance (TID_3) metrics. Lower scores indicate that the model's attention regions are more consistent with those of human participants when answering the same questions.

As shown in Table 1, the TID_1 and TID_3 scores of the explanations generated by the proposed method exhibit small variations across different models and remain relatively low. For instance, the TID_3 score of LLaVA-V1.5-7B is only 0.387, indicating that the average distance between the top three attention regions of the model and those of humans is less than 0.5 patch width. In contrast, the TID_1 and TID_3 scores of the explanations generated by the existing method LVLM-Interpret are much higher. For example, the TID_1 score of LVLM-Interpret on LLaVA-V1.5-7B reaches as high as 3.642, suggesting that the average distance between the image regions deemed most important by LVLM-Interpret and those considered most important by humans exceeds 3.5 patch widths. Under the same setting, the explanations provided by our proposed method achieve an average distance of less than 1 patch width from human judgments. In summary, the plausibility-based evaluation demonstrates that the explanations generated by our method exhibits attention patterns more similar to those of human participants than existing approaches.



(a1) Original Image

(b1) What color is the car in the picture?

(c1) Can you tell me the color of the car in the image?

(d1) What is the color of the car depicted in the photograph?

(e1) Do you know what color the car is in the image?

(f1) What's the color of the vehicle in the picture?

(a2) Original Image

(b2) Is there a camera in the image?

(c2) Does the image show a camera?

(d2) Is a camera visible in the image?

(e2) Does the picture include a camera?

(f2) Is a camera present in the image?

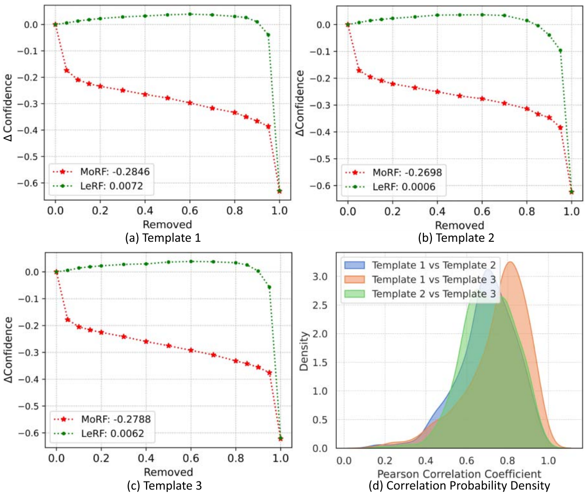
Figure 6: Illustration of the importance heatmaps generated by the LLaVA-V1.5-7B model under different types of questions. The impact of the questioning methods on the importance heatmaps is minimal.

4.3 Evaluations on Robustness

In this subsection, we explore an interesting question: Is the proposed interpretability method robust? We analyze this question from two aspects.

First, we examine whether the regions that LVLMs focus on when generating responses remain stable under different questioning methods with the same image input. In Figure 6, we present a case study that visually illustrates the importance heatmaps generated by the interpretability method under different questions. The impact of different questioning methods on the importance heatmaps is minimal; when answering questions about the car's color, the most critical regions in the explanations consistently focus around the car, and when responding to questions about the presence of a camera in the image, the most critical regions consistently focus around the camera.

Beyond the case study, we also evaluate the faithfulness of the proposed method across the entire dataset under three different question templates and illustrate the distribution of Pearson correlation coefficients between the explanations generated by our method under these templates in Figure 7. Specifically, Template 1 is ''Does this



(a) Template 1

(b) Template 2

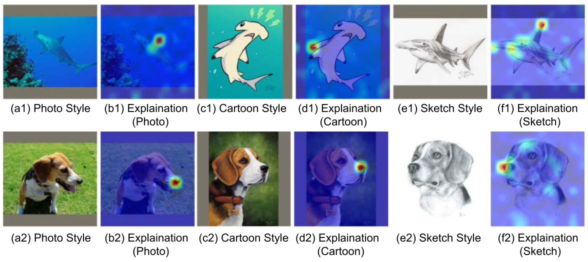
(c) Template 3

(d) Correlation Probability Density

Figure 7: Evaluation of global faithfulness under different questioning templates and the distribution of Pearson correlation coefficients between explanations generated under these templates.

figure show {}?'', Template 2 is ''Is there a {} visible in the image?'', and Template 3 is ''Does this picture include {}?''. As shown in Figure 7, when the important parts of the explanation are removed, the likelihood that the model generates correct responses drastically decreases, whereas removing the least important parts of the explanation slightly increases the likelihood. This suggests that, under different questioning approaches, the explanations provided by our method robustly and faithfully reflect which visual information is most critical for the model's response generation. The Pearson correlation coefficient is a statistical measure of the strength of the linear relationship between two variables, with values ranging from -1 to 1. A value closer to 1 indicates a stronger positive correlation between variables. As shown in Figure 7(d), the peak values of the probability distribution curves of the Pearson correlation coefficients between the explanations generated by our method under different questioning approaches are all above 0.6, indicating strong correlations among the explanations. These experiments demonstrate the robustness of the proposed method across different questioning approaches on the entire dataset.

Second, we investigate whether the proposed method can consistently provide reasonable explanations when the same questioning method is applied, but the image inputs come from different domains. Figure 8 presents a case study on the robustness of the interpretability method across cross-domain image inputs. When LVLMs answer the question ''Can you see the shark's dorsal fin?'', the regions of focus are consistently around the shark's dorsal fin, with minimal influence from the image style. When responding to images in the sketch domain, LVLMs exhibit more dispersed attention compared to their responses to images in the photo and cartoon domains. We hypothesize that this is due to the increased difficulty of answering questions about sketch images, as they provide relatively less visual information.



(a1) Photo Style (b1) Explaination (Photo) (c1) Cartoon Style (d1) Explaination (Cartoon) (e1) Sketch Style (f1) Explaination (Sketch)

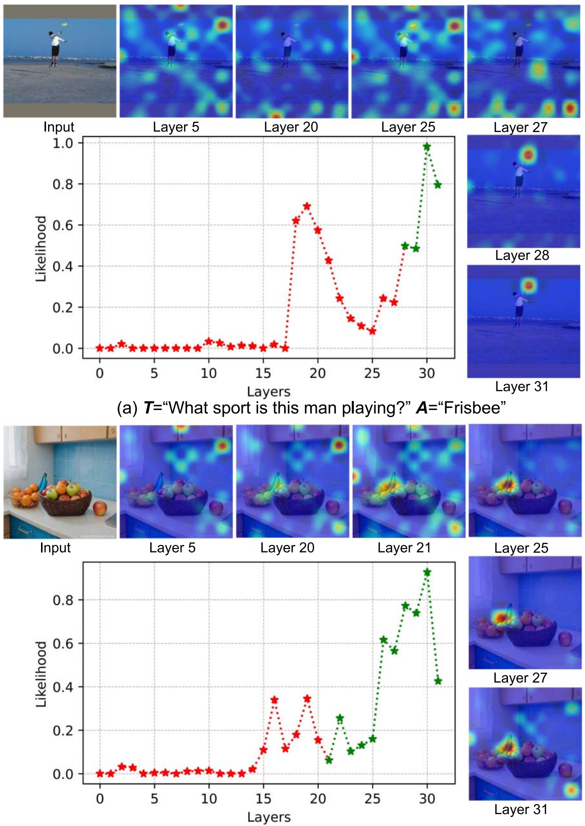
(a2) Photo Style (b2) Explaination (Photo) (c2) Cartoon Style (d2) Explaination (Cartoon) (e2) Sketch Style (f2) Explaination (Sketch)

Figure 8: Illustration of the importance heatmaps generated by the LLaVA-V1.5-7B model for images in photo style, cartoon style, and sketch style. In the image of a shark, we pose the question ''Can you see the shark's dorsal fin?'' to the LVLMs, and the response received was ''Yes, the shark's dorsal fin is visible.'' In the image of a dog, we ask the LVLMs ''What color is the dog's nose in this image?'' and the response is ''Black.''

4.4 Layer Attentions

The aforementioned experiments demonstrate the effectiveness of the proposed interpretability method in explaining the generated responses of LVLMs, revealing a high similarity between the model's attention regions and those that humans focus on when making responses. This approach helps researchers understand how the model generates its responses at an overall level, but it raises the question: At a layer-specific level, which regions of the image do individual layers rely on for their response likelihood?

As shown in the Figure 9, using the LLaVA-V1.5-7B model as an example, which consists of 32 transformer blocks, we selected two examples: an image of playing frisbee and an image of blue bananas in a kitchen. We illustrate the response generation likelihood and importance heatmap of different layers in the model. In the image, green-marked points indicate layers that produced



(a) ***T***="What sport is this man playing?" ***A***="Frisbee"

(b) ***T***="What is the color of the banana?" ***A***="Blue"

Figure 9: Importance heatmaps and likelihood for different layers during response generation. Points marked in green indicate layers that produced correct responses, while points marked in red indicate layers that generated incorrect responses. **T** represents the text input, and **A** represents the model's response.

correct responses, while red-marked points indicate layers that generated incorrect responses. As shown in Figure 9(a), the model's 27th layer gave an incorrect response and did not pay sufficient attention to the region where the frisbee is located. In contrast, the 28th layer provided a correct response, with its attention focused near the frisbee in the image. In Figure 9(b), the 21st layer generated a correct response with a relatively low likelihood; we believe this is because, despite the dispersed importance distribution, the area of the banana was already a critical part of the importance heatmap. The 27th layer, however, produced a correct answer with a higher likelihood, likely because the region where the banana is located dominated the importance heatmap. Overall, the earlier layers tend to have more dispersed attention and are more likely to make incorrect responses, while the later layers generally focus attention on the target and provide more accurate responses.

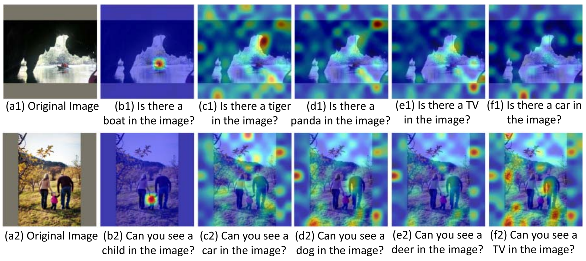

(a1) Original Image
(b1) Is there a boat in the image?
(c1) Is there a tiger in the image?
(d1) Is there a panda in the image?
(e1) Is there a TV in the image?
(f1) Is there a car in the image?
(a2) Original Image
(b2) Can you see a child in the image?
(c2) Can you see a car in the image?
(d2) Can you see a dog in the image?
(e2) Can you see a deer in the image?
(f2) Can you see a TV in the image?

Figure 10: When LVLMs are asked about objects that do not exist in the image, the model's attention is dispersed and not focused on any specific part of the image.

4.5 Attention When Referring to Non-existent Objects

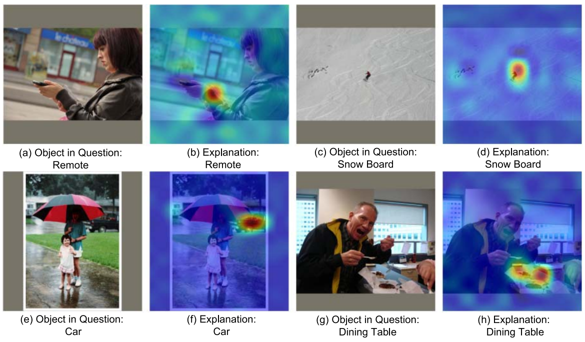
The questions in the previous subsections are about objects that are present in the image. When we ask about objects that do not exist in the image, on which regions does the LVLM's attention focus?

As shown in Figure 10, we query the LVLMs about both present and non-existent objects in the image and illustrate the importance heatmaps. We observed an interesting phenomenon: When we ask about objects that are present in the image, the LVLM tends to focus on specific locations within the image. In contrast, when we inquire about objects that do not exist in the image, the LVLM's attention tends to be dispersed across the entire image. Specifically, when we ask the LVLM about children, its attention is concentrated on the child, whereas when we ask about cars, its attention appears chaotic. We think that this is because the model's responses about ''non-existence'' are derived from a comprehensive analysis of the entire image, with no specific features dominating the model's answer.

4.6 Discussion on Failure Cases

In this subsection, we explore several failure cases of the interpretability method proposed in this paper, focusing on situations where the visual information supporting the model's responses does not align with human perception.

In the first row of examples in Figure 11, when the given LVLM was asked about the objects present in the image, the LVLM incorrectly responded that such objects were absent. The explanations reveal that the model's attention did not focus on the correct visual information. For instance, when asked whether a remote is



(a) Object in Question: Remote
(b) Explanation: Remote
(c) Object in Question: Snow Board
(d) Explanation: Snow Board
(e) Object in Question: Car
(f) Explanation: Car
(g) Object in Question: Dining Table
(h) Explanation: Dining Table

Figure 11: Some cases where the explanations provided by our proposed interpretability method applied to LLaVA-V1.5-7B fail to align well with human perception.

present in the image, the LVLM focused on the woman's wrist instead of the remote she was holding. Similarly, when asked about the presence of a snowboard, the LVLM's attention was directed to the upper-right area near the skier instead of the snowboard itself. We think this issue arises because the objects in these cases are relatively small, posing a challenge for the LVLM's judgment. In the second row of examples in Figure 11, when the given LVLM was asked about objects absent from the image, the model incorrectly responded that such objects were present. The explanations show that the model's attention was drawn to objects visually or semantically related to the queried category. For example, when asked whether a car is present in the image, the truck in the image dominated the model's attention. Similarly, when asked whether a dining table is present, the office desk in the image became the dominant focus. In these cases, the presence of objects belonging to the same broader category interfered with the model's decision-making. In summary, the perception of small objects and fine-grained distinctions remain challenges that need to be addressed in the future development of LVLMs.

5 Limitations

Our method facilitates interpretability analysis of the responses generated by LVLMs, demonstrating considerable flexibility. However, the computational complexity associated with Shapley values remains a challenge, even with the acceleration techniques we implemented. For instance, conducting a single interpretation requires approximately 150 seconds for the LLaVA-V1.5-7B model and about 200 seconds for the LLaVA-V1.5-13B model, both on four RTX 4090 GPUs. This raises the need for more efficient and potentially real-time interpretability methods, which will be a key focus of our future research.

Additionally, this study is limited by the small number of comparative methods used in the analysis. This constraint is largely due to the limited number of existing interpretability studies on large vision-language models. For example, even the latest work, LVLM-Interpret (Stan et al., 2024), falls short of providing faithful and plausible explanations for users. Therefore, this is not a major drawback of our work. In the future, more research will likely focus on the interpretability of LVLMs, and we look forward to comparing our work with future advancements.

6 Conclusion

This paper introduces a patchwise cooperative game-based interpretability method to explain the reasoning behind the responses of large vision-language models. This approach conceptualizes each image patch as a player in a cooperative game, quantifying their contributions to the generation likelihood, treated as the utility function. To approximate the Shapley values, we use Monte Carlo sampling and employ surrogate functions to locally approximate LVLMs, which reduces computational costs. Notably, this method requires only the model's generation likelihood, without needing access to its architecture, parameters, or gradients, allowing for flexible application across different LVLMs. Experimental results indicate that the proposed method achieves strong performance in faithfulness, plausibility, and robustness. We hope that our approach can serve as a foundational baseline for interpretability research in LVLMs and enhance user trust by providing understandable explanations of the models' responses.

Acknowledgments

The authors are grateful to action editor Grzegorz Chrupała and the anonymous reviewers for their meticulous work and valuable suggestions, which have significantly enhanced the quality of this paper.

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415. https://doi.org/10.1109/CVPR52688.2022.02072

Piush Aggarwal, Jawar Mehrabanian, Weigang Huang, Özge Alaçam, and Torsten Zesch. 2024. Text or image? What is more important in cross-domain generalization capabilities of hate meme detection models? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117.

Marco Ancona, Cengiz Oztireli, and Markus Gross. 2019. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 272–281. PMLR.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, pages 1–24.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Michele Cafagna, Lina M Rojas-Barahona, Kees van Deemter, and Albert Gatt. 2023. Interpreting vision and language generative models with semantic visual priors. *Frontiers in Artificial Intelligence*, 6:1220476. https://doi.org/10.3389/frai.2023.1220476, PubMed: 37818428

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406. https://doi.org/10.1109/ICCV48922.2021.00045

Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. 2023. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, pages 1–12. https://doi.org/10.1038/s42256-023-00657-x

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 1–15. https://doi.org/10.1007/978-3-031-72643-9_22

Giovanni Cinà, Daniel Fernandez-Llaneza, Ludovico Deponte, Nishant Mishra, Tabea E. Röber, Sandro Pezzelle, Iacer Calixto, Rob Goedhart, and Ş. İlker Birbil. 2023. Fixing confirmation bias in feature attribution methods via semantic match. *arXiv preprint arXiv:2307.00897*, pages 1–24.

Ian Covert and Su-In Lee. 2021. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026. https://doi.org/10.1109/ICCV51070.2023.00371

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Jiahui Li, Kun Kuang, Lin Li, Long Chen, Songyang Zhang, Jian Shao, and Jun Xiao. 2021a. Instance-wise or class-wise? A tale of neighbor shapley for concept-based explanation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3664–3672. https://doi.org/10.1145/3474085.3475337

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–14. https://doi.org/10.18653/v1/2023.emnlp-main.20

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:1–9.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467. https://doi.org/10.1145/3514094.3534148

Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. 2019. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*.

Letitia Parcalabescu and Anette Frank. 2023. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059. https://doi.org/10.18653/v1/2023.acl-long.223

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Krithik Ramesh and Yun Sing Koh. 2022. Investigation of explainability techniques for multimodal transformers. In *Australasian Conference on Data Mining*, pages 90–98. Springer. https://doi.org/10.1007/978-981-19-8746-5_7

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ‘‘why should i trust you?’’ Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. https://doi.org/10.1145/2939672.2939778

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci.

2022. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pages 18770–18795. PMLR.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626. https://doi.org/10.1109/ICCV.2017.74

Lloyd S. Shapley et al. 1953. A value for n-person games. In H. Kuhn and A. Tucker, Eds., *Contributions to the Theory of Games II*. Princeton University Press, pages 307–316. https://doi.org/10.1515/9781400881970-018

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, pages 1–9.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. 2016. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083. https://doi.org/10.1137/15M1048070

J. Springenberg, Alexey Dosovitskiy, Thomas Brox, and M. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, pages 1–10.

Suraj Srinivas and François Fleuret. 2019. Full-gradient representation for neural network visualization. *Advances in Neural Information Processing Systems*, 32:1–9.

Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. 2024. Explain any concept: Segment anything meets concept-based explanation. *Advances in Neural Information Processing Systems*, 36.

Andrew H. Sung. 1998. Ranking importance of input parameters of neural networks. *Expert Systems with Applications*, 15(3–4):405–411. https://doi.org/10.1016/S0957-4174(98)00041-4

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. 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 5100–5111. https://doi.org/10.18653/v1/D19-1514

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2020. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6021–6029. https://doi.org/10.1609/aaai.v34i04.6064

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, pages 1–27.

Haofan Wang, Rakshit Naidu, Joy Michael, and Soumya Snigdha Kundu. 2020a. Ss-cam: Smoothed score-cam for sharper visual feature localization. *arXiv preprint arXiv:2006.14255*, pages 1–10.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020b. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25. https://doi.org/10.1109/CVPRW50498.2020.00020

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65.

https://doi.org/10.1109/TVCG.2019.2934619, PubMed: 31442996

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6261–6270. https://doi.org/10.1109/CVPR.2019.00642

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38. https://doi.org/10.1145/3639372

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929. https://doi.org/10.1109/CVPR.2016.319