Interpreting Attention Heads for Image-to-Text Information Flow in Large Vision–Language Models

Jinyeong Kim **Seil Kang** **Jiwoo Park** **Junhyeok Kim** **Seong Jae Hwang***
Yonsei University
{jinyeong1324, seil, wldn1677, timespt, seongjae}@yonsei.ac.kr

Abstract

Large Vision-Language Models (LVLMs) answer visual questions by transferring information from images to text through a series of attention heads. While this image-to-text information flow is central to visual question answering, its underlying mechanism remains difficult to interpret due to the simultaneous operation of numerous attention heads. To address this challenge, we propose *head attribution*, a technique inspired by component attribution methods, to identify consistent patterns among attention heads that play a key role in information transfer. Using head attribution, we investigate how LVLMs rely on specific attention heads to identify and answer questions about the main object in an image. Our analysis reveals that a distinct subset of attention heads facilitates the image-to-text information flow. Remarkably, we find that the selection of these heads is governed by the semantic content of the input image rather than its visual appearance. We further examine the flow of information at the token level and discover that (1) text information first propagates to role-related tokens and the final token before receiving image information, and (2) image information is embedded in both object-related and background tokens. Our work provides evidence that image-to-text information flow follows a structured process, and that analysis at the attention-head level offers a promising direction toward understanding the mechanisms of LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable success in a wide range of vision and language tasks. As LVLMs have growing importance in various applications, understanding their inner workings has become increasingly crucial. Uncovering the mechanisms behind LVLMs can enhance transparency [1], identify potential biases [2], and facilitate the development of more efficient and robust models [3, 4], parallel to the recent progress in interpretability research for language models.

However, interpreting LVLMs presents inherent challenges that go beyond those encountered with language-only models. Unlike discrete text tokens in LLMs, which carry well-defined semantic meanings, image tokens in LVLMs act as soft prompts without explicit semantic interpretation. This ambiguity complicates the analysis of internal representations. Moreover, while most mechanistic interpretability studies in language models rely on short textual inputs [5–8], LVLMs process hundreds of image tokens by design. As a result, the mechanisms underlying image-to-text information flow remain largely elusive.

Previous work has primarily focused on the layer level [9–11], leaving the role of specific attention heads in facilitating this information transfer unexplored. Identifying such heads could reveal the precise locations where image information is injected and propagated within the model, enabling a more fine-grained understanding of the underlying process.

*Corresponding author.

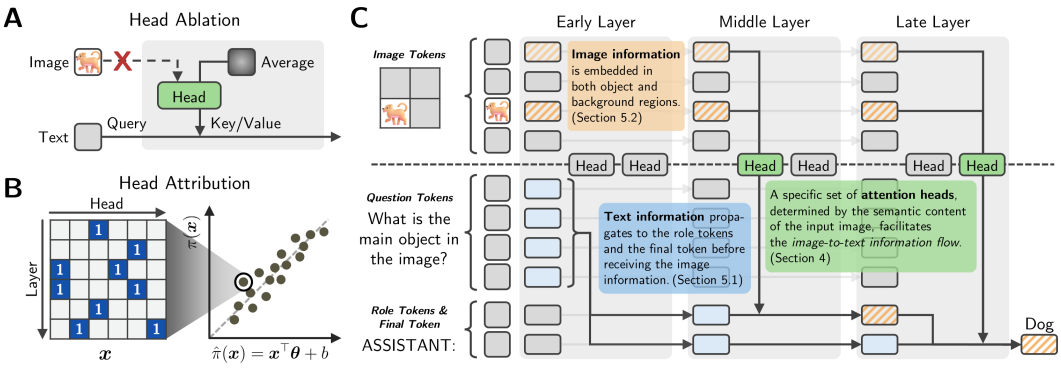

Figure 1: **(A)** Illustration of head ablation. To identify the attention heads responsible for the image-to-text information flow, head ablation blocks the image-to-text information flow of an attention head by replacing its image key/value pairs with the baseline representations. Ablating single head cannot identify crucial heads, as the model utilizes multiple heads to process the information. **(B)** Illustration of head attribution. This method estimates each attention head's contribution to the information flow using a linear model that predicts the logit based on which heads are intact or ablated. Through this method, we can identify the pattern of attention heads that are crucial for the image-to-text information flow. **(C)** Overview of the image-to-text information flow. We uncover the mechanism of image-to-text information flow at the levels of attention heads (§4) and individual tokens (§5).

In light of these challenges, this paper conducts a series of experiments to uncover the fine-grained mechanisms underlying visual question answering in LVLMs. Specifically, we seek to identify the critical attention heads responsible for the image-to-text information flow. Once identified, we analyze how these attention heads mediate the transfer of image information by tracing the visual tokens where the information originates and the language tokens that receive it. To achieve this analysis, we focus on the visual object identification task, where the model is asked to identify the main object in a given image.

First, we show that simply ablating single attention heads (Figure 1A), as done in LLMs [12], is insufficient to identify the key attention heads responsible for image-to-text information flow. This is because the information flow is distributed across multiple heads, making it challenging to isolate a single responsible component.

Therefore, we utilize a systematic approach to quantify the contributions of multiple attention heads to the information flow: *head attribution* (Figure 1B), an adapted version of component attribution from Shah et al. [13]. In this method, we systematically ablate multiple attention heads and use linear regression to estimate each head's contribution to the final logits. The resulting regression model provides a precise estimation of each head's impact on the final logits. Furthermore, since LVLMs may rely on multiple attention heads for image-to-text information flow, head attribution offers a comprehensive explanation of which attention heads are crucial for visual object identification. Our evaluation demonstrates that head attribution outperforms attention-based importance measures and single-head ablations in both faithfulness and completeness.

Head attribution provides a unique lens into the behavior of LVLMs, revealing how they transfer image-to-text information. By analyzing the regression coefficients derived from head attribution, we investigate how attention heads are leveraged for this information flow. Our key findings are as follows. First, attention heads in mid-to-late layers play a pivotal role in transferring image information to text tokens, with a select few heads agonistically utilized across different samples. Second, the contribution of each head is not necessarily correlated with its image attention weights, challenging the common assumption that heads with high attention weights are inherently more important. Lastly, LVLMs employ a similar set of attention heads to process semantically similar objects, suggesting that the model systematically processes objects based on their semantic meanings.

To gain finer-grained insight into the image-to-text information flow, we extend the approach described above to the individual token level. This token-level analysis reveals that the semantics of the question propagate to the role tokens (e.g., "ASSISTANT") and the final token (":"), after which image information is transferred to these tokens. We also observe that only a subset of tokens in the object region, along with a few background tokens, contribute to the final prediction. While attention weights

correlate with token importance to some extent—explaining the success of recent token reduction methods [14, 15]—many high-attention tokens are not necessary for the information flow. This finding suggests that further token reduction in LVLMs could be achieved by quantifying each token's importance through methods that go beyond attention-based metrics.

In summary, our main contributions are as follows: (1) We provide a holistic understanding of how the image-to-text information flow operates in LVLMs (Figure 1C), studied at the levels of attention heads (§4) and individual tokens (§5). We validate the generalizability of our findings across ten different LVLMs. (2) We demonstrate that head attribution serve as effective methods for explaining the information flow (§4.1, 4.2), offering a novel perspective to identify the mechanisms behind LVLMs. (3) We discuss the broader implications of our findings for mechanistic interpretability and the development of efficient LVLMs (§6).

2 Related Works

Mechanistic Interpretability. Mechanistic interpretability (MI) is an emerging research field dedicated to interpret the inner workings of neural networks by reverse-engineering their computations into human-interpretable mechanisms [16]. MI employs various techniques to analyze model internals and understand their decision-making processes [17]. In this section, we briefly review two techniques closely related to our work: causal intervention and component attribution.

Causal Intervention. Causal intervention [18] treats the neural network as a causal model [19] and intervenes in the computation of a model component to observe its effect on the prediction. Attention knockout [12] can be considered a specialized causal intervention technique for attention mechanisms, where the attention weight from a source of interest to target is set to baseline representations (e.g., zero ablation) to block the information flow.

Component Attribution. Beyond estimating the effect of individual model components through individual ablation, component attribution [13] estimates the contributions of multiple components in a more systematic way. Specifically, component attribution ablates subsets of components repeatedly and estimates the contribution of each component via a linear model. We adapt this method for the attention heads of LVLMs, referred to as *head attribution*. Furthermore, while the original work focused on model editing, we repurpose this method to analyze the inner workings of LVLMs by examining the contribution of each head.

Interpreting Large Vision-Language Models. Parallel to the growing interest in interpretability for language models [20], interpreting the inner workings of LVLMs has become a recent focus of research [21]. In particular, the *image-to-text information flow*, which enables the model to generate text from images, has been widely studied. [11] found that image tokens can be projected onto their language vocabulary, and [22] proposed visual information enrichment and semantic refinement as two stages of visual information processing. [9] showed that transferring information from object tokens to the final output in mid-to-late layers is necessary for generating text. [10] complemented this finding by revealing that image information is conveyed to the final output through the language context. On the other hand, although not directly focused on interpretability, studies on efficient inference also provide insights into the mechanisms of LVLMs. For example, [14] and [15] demonstrated that preserving only image tokens with high attention weights is sufficient for maintaining performance, which suggests that many tokens are redundant in generating outputs. By integrating and clarifying these findings with our experiments, we provide a holistic understanding of image-to-text information flow and its implications for the interpretability and efficiency of LVLMs.

3 Experimental Setup

To systematically examine the mechanism of image-to-text information flow in LVLMs, we introduce a minimal yet natural task, a carefully selected dataset, and a diverse set of models.

Task. This paper seeks to understand the mechanism by which image information is transferred to text tokens and contributes to the final prediction in visual question answering. As a minimal yet natural task within visual question answering, we focus on the *visual object identification task*, in which the model is required to identify the main object in an image. We choose this task because it is a fundamental visual recognition task and is widely used to evaluate the visual capabilities of LVLMs [23, 24]. Specifically, the task is defined as follows: "USER: <*image*> What is the

main object in the image? Please answer with a single word. ASSISTANT:". After the role tokens "ASSISTANT" and the final token ":", the model is expected to generate a concise response (e.g., "dog") based solely on the image. For the ablation studies, we also consider alternative prompts that perform the same task, in order to examine whether LVLMs utilize attention heads based on the general meaning of the task or whether they are sensitive to the specific wording of the prompt. Please refer to §C for the alternative prompts we consider.

This experimental design offers several advantages for analysis. First, the task is simple and straightforward, allowing us to focus on the mechanism of image-to-text information flow without the distractions of complex elements such as multi-step reasoning. Second, because the prompt is open-ended and the answer cannot be inferred from it, the model is required to generate its response directly from the image. This minimizes potential biases introduced by the prompt and eliminates the possibility of random guessing [25]. Third, we can directly measure the logits due to the single-word answer constraint. Since logits are known to be the most direct indicator of model behavior [26], this design enables a detailed analysis of the model's decision-making process.

Dataset. For image data, we select images containing a single main object from the COCO dataset [27] to ensure the task is well-defined and unambiguous. The model is tasked with predicting the object in each image, and we collect 200 samples with correct predictions. For ablation studies, We also conduct experiments on the DomainNet dataset [28], which includes diverse images spanning multiple domains (styles). Specifically, we select 300 images from the real, sketch, and clipart domains.

Models. To examine the universality and consistency of our findings across different models, we select ten widely used LVLMs from four families: LLaVA-1.5-7B/13B [29], LLaVA-NeXT-7B/13B [30], InternVL2.5-1B/2B/4B/8B [31], and Qwen2-VL-2B/7B [32]. These models differ in architecture, parameter size, and training scheme, enabling us to assess the generalizability of our findings. In the main paper, we visualize the results from LLaVA-1.5-7B for brevity, with additional results for other models in §D.

4 Identifying Attention Heads Contributing to Image-to-Text Information Flow

In this section, we identify the attention heads that play a crucial role in the image-to-text information flow within LVLMs. We begin by showing that single attention head ablation is insufficient to identify important attention heads. Then, we introduce *head attribution* as a method to measure the contributions of individual attention heads to the final logits (§4.1). We validate the head attribution method by assessing its faithfulness and completeness (§4.2). Finally, we analyze the results of head attribution to gain a deeper insight of how LVLMs leverage attention heads for the transfer of image-to-text information (§4.3).

4.1 Measuring the Contributions of Attention Heads

Image-to-text information flow can be interpreted as a retrieval process, where the text query retrieves image information from the image via the attention mechanism. At each attention head, the text query vector interacts with the key/value matrices of the image tokens to access image information.

Attention Knockout (Single Head Ablation). To determine which attention heads are crucial for image-to-text information flow, we aim to measure the causal effect of each attention head on the final logits by ablating individual heads. Recent LVLM interpretability studies [9, 10, 33] commonly use attention knockout [12] to block the information flow. This method adds $-\infty$ before softmax when calculating the attention weights, equivalent to setting the image attention weights and value matrices to zero. We follow this approach to disable the image-to-text information flow in individual heads. However, since zero is an arbitrary value that does not reflect the actual distribution of the image tokens, this approach risks introducing undesired out-of-distribution noise [5, 34]. Instead, we apply mean ablation to block the information flow from the image tokens to the text query. See §A for more details.

Single Head Ablation is Insufficient. We ablate each attention head individually and evaluate the logit difference between the patched and original models. Given the evidence suggesting that attention heads have specialized roles [35], we anticipate the logit difference to be significant if an attention head plays an important role in image-to-text information flow. However, the results show that the

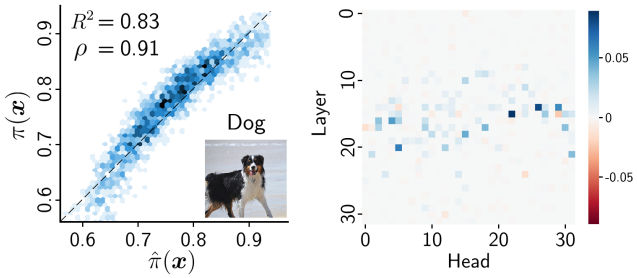

Figure 2: Example result of head attribution for LLaVA-1.5-7B. **(Left)** Scatter plot of the ground-truth logit $\pi(\boldsymbol{x})$ and the predicted logit $\hat{\pi}(\boldsymbol{x})$. The contribution of each attention head is well-captured by head attribution. **(Right)** Attribution coefficients $\boldsymbol{\theta}$ for each attention head. Each coefficient represents the contribution of the corresponding head to the prediction.

logit difference is not substantial. For instance, in LLaVA-1.5-7B, no head accounts for more than 5% of the logit difference in 80% of the cases (see §C for examples).

Based on this observation, we hypothesize that image-to-text information flow is a distributed process involving multiple attention heads, similar to the retrieval process in LLMs [36]. In this context, quantifying the importance of each attention head by knocking out a single head, which is a common practice in language models [5, 7], may be unreliable, as other heads could compensate for the disabled head through self-repair mechanisms [37, 38].

Method: Head Attribution. To capture the distributed nature of the image-to-text information flow, we apply head attribution to groups of heads simultaneously rather than isolating individual attention heads, mitigating the interactions between attention heads. After repeatedly collecting logits from the patched models, we leverage component attribution [13] to estimate each attention head's contribution to the image-to-text information flow.

Specifically, let $\boldsymbol{x} \in \{0, 1\}^{LH}$ represent a binary vector where each element indicates whether the corresponding attention head is intact (1) or ablated (0), with L denoting the number of layers and H the number of heads per layer. The final logit, denoted as $\pi(\boldsymbol{x})$, is adjusted by subtracting the average logits over all possible vocabularies [26] and is then normalized to the range $[0, 1]$ by setting $\pi(\mathbf{0})$ to 0 and $\pi(\mathbf{1})$ to 1. We estimate the logits by fitting a linear regression model:

$$\hat{\pi}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta} + b, \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{LH}$ and $b \in \mathbb{R}$ are the learned parameters. The *attribution coefficients* $\boldsymbol{\theta}$ are interpreted as the contributions of each attention head to the final logits. We refer to this method as *head attribution*. For more technical details, please refer to §A.

Results. Surprisingly, head attribution can accurately predict the final logit based on which attention heads are intact. Figure 2 shows an example of the head attribution results and attribution coefficients $\boldsymbol{\theta}$ for each attention head. The linear model explains the final logit with high accuracy, and the coefficients $\boldsymbol{\theta}$ reveal that multiple attention heads contribute to the image-to-text information flow. Quantitatively, the average explained variance (R^2) exceeds 0.77, and the Pearson correlation coefficient (ρ) between $\pi(\boldsymbol{x})$ and $\hat{\pi}(\boldsymbol{x})$ exceeds 0.88 for all models (see §D for detailed results). These results indicate that the contribution of each attention head is accurately estimated using head attribution.

4.2 Validating Head Attribution

Metrics. If the head attribution method is valid, the coefficients $\boldsymbol{\theta}$ should not only fit well but also reflect the importance of each attention head in the image-to-text information flow. Following [3, 5], we validate its **faithfulness** and **completeness** by evaluating the causal effect. Starting with all heads ablated, we progressively activate the attention heads from the most to the least important based on the coefficients $\boldsymbol{\theta}$. Faithfulness, defined as $\pi(\boldsymbol{x})$, measures the extent to which the selected heads explain the model's actual performance. Completeness, defined as $\pi(\mathbf{1} - \boldsymbol{x})$, evaluates whether the selected heads capture all the essential heads. In other words, if other heads can compensate for the selected heads, the method's completeness is compromised. The ideal faithfulness (*i.e.*, the model's actual performance) is $\pi(\mathbf{1}) = 1$, and the ideal completeness is $\pi(\mathbf{1} - \mathbf{1}) = 0$.

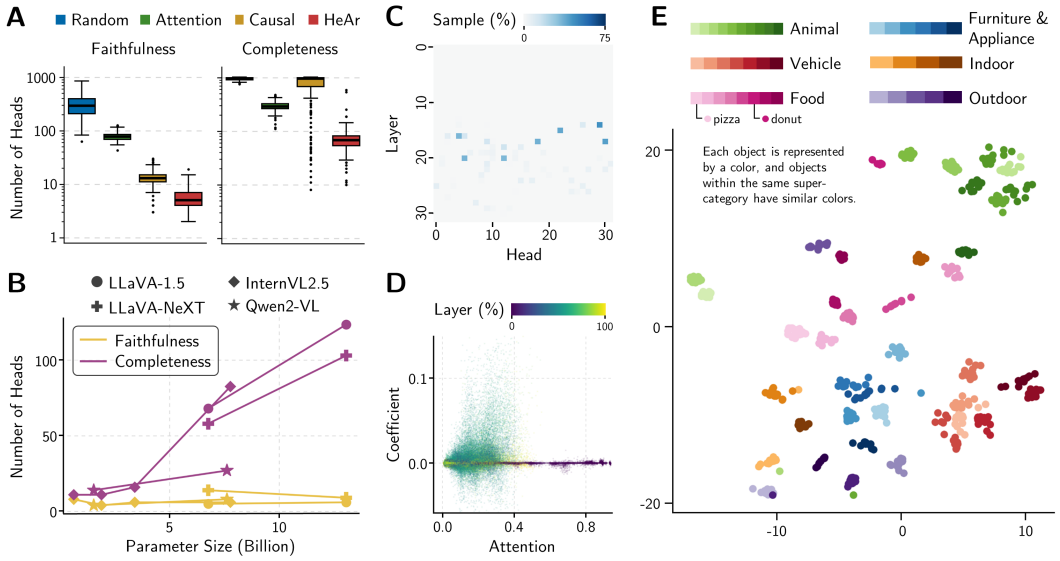

Figure 3: **(A)** The minimum number of heads required for faithfulness > 0.8 (Left) and completeness < 0.2 (Right) for each criterion. Head attribution is abbreviated as HeAr. **(B)** The trend of the minimum number of heads by head attribution across various models. **(C)** The portion of samples requiring a given head for faithfulness > 0.8. **(D)** Scatter plot showing the relationship between attribution coefficients $\boldsymbol{\theta}$ and image attention weights. The color indicates relative layer depth. **(E)** t-SNE visualization of attribution coefficients $\boldsymbol{\theta}$. Each point represents a sample. These plots are for LLaVA-1.5-7B, and the results for other models can be found in §D.

Baselines. We compare the number of heads required to achieve the same level of faithfulness and completeness across three other methods: (1) **Random**: Heads are selected randomly. (2) **Attention**: The heads with the highest image attention weights[2] are selected [22, 39], which are frequently presumed to transfer image information to the text query. (3) **Causal**: The heads with the highest logit difference between the head-ablated and original models are selected (§4.1). If the head attribution method requires fewer heads to achieve the same level of faithfulness and completeness, it suggests that the method more effectively explains the importance of each attention head. In practice, we measure the minimum number of heads required for faithfulness > 0.8 and completeness < 0.2 (refer to §B for the ablation study on the threshold).

Results. Figure 3A shows the results of the validation. Head attribution (HeAr) requires the fewest heads to achieve the same faithfulness and completeness compared to other methods. Therefore, we conclude that the head attribution method is valid for measuring the importance of each attention head in image-to-text information flow.

An intriguing finding from the validation is that the number of heads required to achieve the same level of completeness is significantly larger than that required for faithfulness. This trend becomes more pronounced as the model size increases (Figure 3B). Since completeness measures whether the selected heads capture all heads contributing to the final logits, this result suggests that the model may use redundant attention heads [40] to ensure robustness in the image-to-text information flow. Furthermore, as the model size increases, the number of redundant heads also increases, likely due to the growing complexity of the model.

4.3 Interpreting the Behavior of LVLMs with Head Attribution

In this section, we analyze the head attribution results to gain deeper insights into how LVLMs utilize attention heads for image-to-text information flow. First, we investigate the distribution of important attention heads within the model architecture. Second, we examine whether the importance of the attention heads correlates with the image attention weights to assess whether these weights are reliable indicators of head importance. Finally, we cluster attribution coefficients $\boldsymbol{\theta}$ across different samples to identify common patterns in the attention heads.

[2]To calculate image attention weights, the sum of the image attention weights for each text token is averaged across all text tokens. See §A for more details.

Distribution of Important Attention Heads. We visualize the importance of each attention head in the model architecture by counting how many samples rely on a given head to achieve a faithfulness score greater than 0.8 (Figure 3C). Although this varies across models, important heads are generally located in the middle to later layers, in line with the findings of [9, 22]. Additionally, some heads are consistently important across different samples, while others are only sparsely important, suggesting that LVLMs employ both sample-agnostic and sample-specific attention heads.

Relation to Image Attention Weights. We examine the correlation between attribution coefficients $\boldsymbol{\theta}$ and the image attention weights. As shown in Figure 3D, no significant correlation is observed, indicating that the importance of the attention heads is not directly related to the image attention weights. Therefore, we conclude that image attention weights may not be reliable indicators of the importance of attention heads in image-to-text information flow. We further discuss this in §6. In particular, in the LLaVA family, early-layer heads exhibit high attention weights but are not important. While early-layer heads are known to integrate syntactic information in text [41], we suspect that they are less meaningful in the context of images, where encoding is already effectively handled by the vision encoder.

Systematic Patterns of Attention Heads. We apply t-SNE [42] to cluster attribution coefficients $\boldsymbol{\theta}$ across different samples (Figure 3E). The clustering pattern reveals that LVLMs tend to utilize similar attention heads when processing the same object, even though the size and position of the object can vary significantly across samples. Moreover, objects with similar semantic meanings (classified by super-categories from COCO [27]) are typically clustered together. Additionally, We apply the logit lens [43] to the top attention heads ranked by attribution coefficients $\boldsymbol{\theta}$, and visualize the top vocabulary items (see §C for examples). We find that some heads are clearly associated with the final answer, indicating that attention heads transfer interpretable semantic information. Overall, These findings suggest that LVLMs systematically allocate attention heads to process objects based on their semantics.

Ablation Studies. In LLaVA-1.5-7B, we further conduct two ablation studies to validate the generalizability of our findings. First, we reproduce head attribution using nine additional prompts that perform the visual object identification task, to verify that the observed attention head patterns are not prompt-specific. We find that the cosine similarity between attribution coefficients $\boldsymbol{\theta}$ across different prompts is consistently high (averaging at least 0.91), indicating that the identified patterns are robust to changes in prompt wording. Second, we apply head attribution to images with diverse visual styles (e.g., real, sketch, and clipart) [28] and observe similarly consistent attention head patterns across styles. These results suggest that the systematic allocation of attention heads is not tied to particular prompts or image styles, but instead reflects a generalizable pattern in LVLMs.

5 Tracing Image-to-Text Information Flow: Token-Level Analysis

In §4, we identified attention heads that contribute to the image-to-text information flow by applying head attribution. However, this does not provide a complete picture of the information flow. Attention heads transfer information from the source token to the target token. While we know that the source tokens belong to the image and the target tokens belong to the text, the exact positions from which and to which the information flows remain uncertain. To address this, we perform token-level analysis to gain a deeper understanding of the role of these attention heads. Specifically, we identify the text tokens that receive image information (§5.1), and then examine the image tokens that contribute to the information flow (§5.2).

5.1 Which Text Tokens Receive the Image Information?

Method. Our text input, "What is the main object in the image? Please answer with a single word. ASSISTANT:" consists of approximately 20 tokens. Given that information transfer is known to occur only for specific tokens [5, 44], we hypothesize that a small number of text tokens play a critical role in this process. To identify the text tokens that receive image information, we measure the causal effect of blocking each text token on the image-to-text information flow through head ablation. A significant drop in the logit indicates that the blocked text token receives image information. Note that we retain only the minimum number of attention heads required for faithfulness $\pi(\boldsymbol{x}^*) > 0.8$, which are the most important attention heads for the image-to-text information flow. We then measure the logit difference relative to $\pi(\boldsymbol{x}^*)$, as we are interested in the function of attention heads that significantly contribute to the image-to-text information flow.

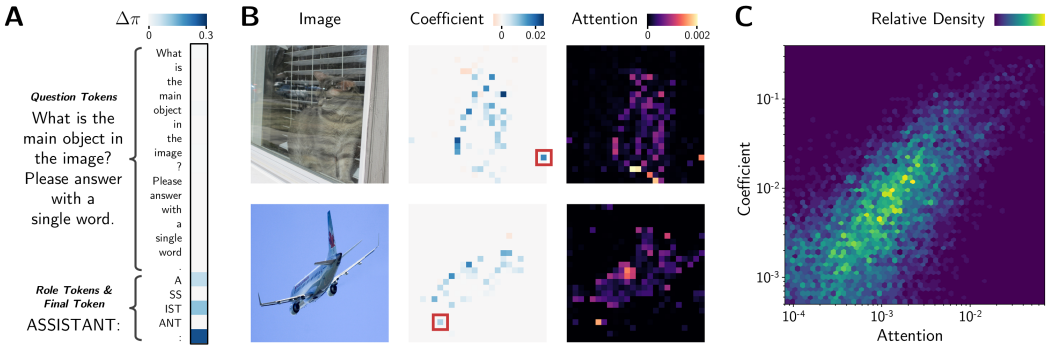

Figure 4: **(A)** Logit difference relative to $\pi(\boldsymbol{x}^*)$ when blocking each text token. **(B)** Image, attribution coefficients $\boldsymbol{\theta}$, and image attention weights. Red boxes highlight high-contribution tokens outside the main object region. **(C)** Scatter plot of attribution coefficients and image attention weights. These plots are for LLaVA-1.5-7B, and the results for other models can be found in §D.

Results. As shown in Figure 4A, the logit difference is substantial when we block the role tokens "ASSISTANT" or the final token before generation ":". The causal effect is highest when we block the final token, but the role tokens also have a significant effect. We validate this result by measuring how much of the logit is preserved when we retain only the role tokens and the final token[3]. In LLaVA-1.5-7B, $97.7 \pm 1.8\%$ of the logit is preserved relative to $\pi(\boldsymbol{x}^*)$, indicating that the role tokens and the final token are critical for maintaining the image-to-text information flow. This result clearly explains why blocking only the final token cannot perfectly restrict the information flow [9], and identifies which text tokens are precisely responsible for the image-to-text information flow, complementing the findings of [10].

It is quite surprising that no significant information is transferred to the question tokens ("What is the main object in the image?"). Since the question tokens contain the most important semantics for completing the task, it is natural to expect the image information to flow to them preferentially. However, the image information only flows to the role tokens and the final token directly. This result suggests that the model first transfers the semantics of the question to the role tokens and the final token, and then certain attention heads direct the image information to these tokens in response to the question semantics embedded within them.

5.2 Which Image Tokens Contribute to the Flow?

Method. After identifying the text tokens that receive the image information, the next step is to determine which image tokens contribute to the information flow. Unlike text tokens, multiple image tokens may contribute, as object information is distributed across several image tokens [9, 11]. Therefore, simply measuring the causal effect of blocking individual image tokens may not be sufficient to identify the key tokens. Instead, we apply component attribution [13] again, this time systematically ablating image tokens rather than attention heads to quantify each image token's contribution to the image-to-text information flow. We focus on LLaVA models for this analysis, as their image encoders preserve the spatial information of the image.

Results. We qualitatively visualize the attribution coefficients $\boldsymbol{\theta}$ for the image tokens in the second column of Figure 4B. By comparing the coefficient with the image, we find that most important image tokens are located in the main object region, which aligns with the findings of [9]. However, the coefficient reveals two important insights not addressed in previous studies. First, the model does not utilize all tokens in the main object region; rather, only a sparse subset contributes to the information flow. Second, the model also uses tokens outside the main object region, such as those from the background, to transfer information (highlighted in red boxes in Figure 4B). This seemingly counterintuitive phenomenon may occur because the vision encoder [45] tends to capture global information in tokens unrelated to the main object [46, 47], or because LVLMs treat background tokens as anchor tokens, which language models use to store information [48]. A more detailed analysis of this phenomenon is left for future work.

[3]More precisely, for example, in LLaVA-1.5-7B, "ASSISTANT:" is split into five tokens A / SS / IST / ANT / :". Since only the first, third, and final tokens have a high causal effect, we retain only these three tokens.

We then compare the attribution coefficients $\boldsymbol{\theta}$ with the attention weights of the image tokens[4] (the 3rd column of Figure 4B) to validate whether attention weights can serve as surrogate indicators of the importance of image tokens. In many cases, the attention weights include image tokens with high attribution coefficients. We also visualize the relationship between the attribution coefficients and the attention weights in Figure 4C. Though the trend is not perfect, the attention weights correlate with the attribution coefficients to some extent (Pearson correlation (ρ) is 0.70 for LLaVA-1.5-7B). This result explains why recent token reduction methods based on attention weights [14, 15] can be effective.

However, as shown in Figure 4B, many image tokens with high attention weights are not important for the information flow. This result can be explained by (1) attention sink tokens [49, 50], which are tokens that receive high attention weights but are uninformative, and (2) the fact that some visual tokens contain information but are not useful for predicting the object's name [51, 52]. Thus, while retaining tokens with high attention weights could be enough to maintain the information flow, it is essential to be mindful that not all high-attention tokens are necessarily significant.

6 Conclusion and Discussion

Through §4, 5, we have shown that head attribution can effectively identify the important heads and tokens for image-to-text information flow. As a result, we now have a holistic understanding of how LVLMs achieve visual object identification task (Figure 1C). The process of understanding this simple task provides several insights, including the implications for mechanistic interpretability (MI) and efficient LVLMs. In this section, we discuss these implications and conclude the paper with limitation and future work.

Implications for MI. Interpreting individual attention heads relies on the assumption that some attention heads are monosemantic [53] (or that the function we aim to understand is localized in a few heads). However, as demonstrated in this work, many attention heads systematically collaborate to process information. This observation aligns with the concept of *attention head superposition* [54], where the model leverages multiple heads, rather than a single head, as an adaptive mechanism for utilizing its limited parameter capacity. In such cases, head attribution offers a promising approach for interpreting distributed mechanisms, as intervening in a single head alone may not fully capture the underlying dynamics.

Especially for interpretability research in LVLMs, our work provides, to the best of our knowledge, the first example of interpretability in individual component-level and pixel-level analysis beyond merely investigating attention weights. Our findings demonstrate that image attention weight is not a reliable indicator of head importance (§4.3), despite its widespread use in prior works [22, 39, 55]. Hence, we encourage the LVLM interpretability community to critically assess the limitations of attention weight-based analyses and to explore alternative approaches for understanding information flow in LVLMs.

Implications for Efficient LVLMs. The computational cost of LVLMs is considerably high due to the large number of image tokens, emphasizing the need for more efficient models. Recent works have introduced strategies to reduce the number of image tokens based on attention weights [56, 57]. We provide empirical evidence for why these approaches succeed in reducing token count without sacrificing performance, demonstrating a correlation between attention weights and attribution coefficients (§5.2). However, we also show that not all tokens with high attention weights are important. Therefore, we argue that attention-based token selection alone is insufficient, and greater efficiency can be achieved by assessing token importance beyond simple attention weights.

Limitation and Future Work. While our work provides a holistic understanding of image-to-text information flow in LVLMs, several limitations remain. First, our focus is on visual object identification, and the generalization of our findings to other tasks, such as reasoning and localization, is uncertain. A comparative analysis across tasks would deepen our understanding of how LVLMs handle different multimodal tasks. Seconds, head attribution requires multiple forward passes, making it computationally expensive. Developing scalable methods for head attribution is an important direction for efficient and practical interpretability [58, 59].

[4]We calculate the weighted attention map for each sample. See §A for more details.

References

[1] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)*, pages 464–483. IEEE, 2023.

[2] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[3] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

[4] Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention is needed. *CoRR*, 2024.

[5] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[6] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

[7] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

[8] Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. Position-aware automatic circuit discovery. *arXiv preprint arXiv:2502.04577*, 2025.

[9] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.

[10] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. *arXiv preprint arXiv:2411.17491*, 2024.

[11] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *CoRR*, 2024.

[12] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

[13] Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. *arXiv preprint arXiv:2404.11534*, 2024.

[14] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025.

[15] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

[16] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022. `https://www.transformer-circuits.pub/2022/mech-interp-essay`.

[17] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv*, 2024.

[18] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[19] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

[20] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

[21] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

[22] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*, 2024.

[23] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.

[25] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *CoRR*, 2024.

[26] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

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

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

[31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[33] Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in vision-language models. *arXiv preprint arXiv:2412.06646*, 2024.

[34] Maximilian Li and Lucas Janson. Optimal ablation for interpretability. *arXiv preprint arXiv:2409.09951*, 2024.

[35] Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

[36] Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with fakepedia. *arXiv preprint arXiv:2312.02073*, 2023.

[37] Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

[38] Cody Rushing and Neel Nanda. Explorations of self-repair in language models. *arXiv preprint arXiv:2402.15390*, 2024.

[39] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*, 2024.

[40] Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. *arXiv preprint arXiv:2404.07129*, 2024.

[41] Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference? *arXiv preprint arXiv:2406.19384*, 2024.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] nostalgebraist. interpreting gpt: the logit lens, 2020. URL `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

[44] Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*, 2024.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[47] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024.

[48] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. pages 9840–9855, 2023.

[49] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, 2020.

[50] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[51] Sarthak Jain and Byron C Wallace. Attention is not explanation. pages 3543–3556, 2019.

[52] Sofia Serrano and Noah A Smith. Is attention interpretable? pages 2931–2951, 2019.

[53] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[54] Adam Jermyn, Chris Olah, and Tom Henighan. Attention head superposition. *Transformer Circuits Thread*, 2023. `https://transformer-circuits.pub/2023/may-update/index.html#attention-superposition`.

[55] Nikolai Ilinykh and Simon Dobnik. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, 2022.

[56] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

[57] Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipvl: Efficient large vision-language models with dynamic token sparsification and kv cache compression. *arXiv preprint arXiv:2410.08584*, 2024.

[58] Neel Nanda. Attribution patching: Activation patching at industrial scale. *URL: https://www.neelnanda. io/mechanistic-interpretability/attribution-patching*, 2023.

[59] János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[61] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[62] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[63] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.

[64] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[65] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, March 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[67] Qiong Wu, Wenhao Lin, Weihao Ye, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. *arXiv preprint arXiv:2411.19628*, 2024.

[68] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024.

A Method Details

Through §4 and 5, we have introduced *head attribution* to analyze the information flow in LVLMs. We offer additional details on the background of LVLMs (§A.1), the methods (§A.2, A.3), and miscellaneous details of the experiments (§A.4).

A.1 Background

In this section, we provide a brief overview of the architecture of Large Vision-Language Models (LVLMs) and the attention mechanism employed in these models to contextualize the proposed method.

LVLMs typically process an input image and a text prompt to generate a sequence of text tokens as output. The input image is initially passed through a vision encoder, such as those described in [45, 60], to produce a visual representation. This representation is then projected into a shared embedding space aligned with the text input. Concurrently, the text input is tokenized and embedded to generate a sequence of text embeddings. Both the visual and text embeddings are fed into a large language model, which is generally implemented as a decoder-only transformer [61].

Let $\boldsymbol{Z}^0_{\text{img}} \in \mathbb{R}^{P_{\text{img}} \times d}$ denote the visual embeddings and $\boldsymbol{Z}^0_{\text{txt}} \in \mathbb{R}^{P_{\text{txt}} \times d}$ denote the text embeddings, where P_{img} and P_{txt} represent the number of visual and text embeddings, respectively, and d is the embedding dimension. These embeddings are concatenated as $\boldsymbol{Z}^0 = [\boldsymbol{Z}^0_{\text{img}}; \boldsymbol{Z}^0_{\text{txt}}]$ and passed through a decoder-only transformer to generate the output sequence of text tokens. The decoder-only transformer consists of L layers, each comprising a multi-head attention (MHA) mechanism followed by a feed-forward neural network (FFN). For intermediate representations, we define $\boldsymbol{Z}^{\ell-1} = [\boldsymbol{Z}^{\ell-1}_{\text{img}}; \boldsymbol{Z}^{\ell-1}_{\text{txt}}]$ as the input to the ℓ-th layer.

Following the framework proposed by [62], we adopt the *residual stream* perspective to interpret computations within the transformer. This perspective treats the transformer's computation as a sequence of residual updates, where intermediate representations are iteratively refined by adding the outputs of the MHA and FFN layers to their inputs:

$$\hat{\boldsymbol{Z}}^{\ell} = \boldsymbol{Z}^{\ell-1} + \sum_{h=1}^{H} \text{MHA}^{\ell,h}(\boldsymbol{Z}^{\ell-1}), \quad \boldsymbol{Z}^{\ell} = \hat{\boldsymbol{Z}}^{\ell} + \text{FFN}^{\ell}(\hat{\boldsymbol{Z}}^{\ell}), \tag{2}$$

where $\hat{\boldsymbol{Z}}^{\ell}$ represents the output of the MHA layer, $\boldsymbol{Z}^{\ell}$ is the output of the FFN layer, and H denotes the number of attention heads per layer. For simplicity, layer normalization [63, 64] is omitted. Finally, the raw logits are computed via a linear projection of the output of the final layer's text embeddings $\boldsymbol{z}^L_{\text{txt}}$ using an unembedding matrix $\boldsymbol{W}_U \in \mathbb{R}^{d \times |\mathcal{V}|}$, where $|\mathcal{V}|$ represents the vocabulary size:

$$\text{logits} = \boldsymbol{z}^L_{\text{txt}} \boldsymbol{W}_U \in \mathbb{R}^{|\mathcal{V}|}. \tag{3}$$

We always measure the adjusted logits, obtained by subtracting the average logits across the vocabulary from the raw logits, as recommended by [26].

We focus on the attention mechanism, which facilitates information flow within the model. The MHA mechanism calculates the attention weights $\boldsymbol{A}^{\ell,h}$ between the input embeddings $\boldsymbol{Z}^{\ell-1}$ as follows:

$$\boldsymbol{A}^{\ell,h} = \text{softmax}\left(\frac{\boldsymbol{Q}^{\ell,h}(\boldsymbol{K}^{\ell,h})^{\top}}{\sqrt{d_k}} + \boldsymbol{M}\right) \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}}) \times (P_{\text{img}}+P_{\text{txt}})}, \tag{4}$$

where $\boldsymbol{Q}^{\ell,h} = \boldsymbol{Z}^{\ell-1}\boldsymbol{W}^{\ell,h}_Q \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}}) \times d_k}$ and $\boldsymbol{K}^{\ell,h} = \boldsymbol{Z}^{\ell-1}\boldsymbol{W}^{\ell,h}_K \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}}) \times d_k}$ represent the query and key matrices, respectively. Here, d_k denotes the dimension of the query and key vectors, and $\boldsymbol{M}$ is a mask matrix that prevents attention to future tokens in the sequence, typically by assigning $-\infty$ to the upper triangular part of the matrix. The attention weights are then used to compute the MHA layer's output:

$$\text{MHA}^{\ell,h}(\boldsymbol{Z}^{\ell-1}) = \boldsymbol{A}^{\ell,h}\boldsymbol{V}^{\ell,h}\boldsymbol{W}^{\ell,h}_O \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}}) \times d}, \tag{5}$$

where $\boldsymbol{V}^{\ell,h} = \boldsymbol{Z}^{\ell-1}\boldsymbol{W}^{\ell,h}_V \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}}) \times d_k}$ is the value matrix, and $\boldsymbol{W}^{\ell,h}_O \in \mathbb{R}^{d_k \times d}$ is the output weight matrix.

As discussed in Section 4.1, image-to-text information flow can be understood as a retrieval process. Consider an arbitrary text token $\boldsymbol{z}_i^{\ell-1} \in \mathbb{R}^d$ at the ℓ-th layer, which is a row vector from the text matrix $\boldsymbol{Z}_{\text{txt}}^{\ell-1} \in \mathbb{R}^{P_{\text{txt}} \times d}$. This text token is converted into a query vector $\boldsymbol{q}_i^{\ell,h} \in \mathbb{R}^{d_k}$ by multiplying it with the query weight matrix $\boldsymbol{W}_Q^{\ell,h} \in \mathbb{R}^{d \times d_k}$. The query vector $\boldsymbol{q}_i^{\ell,h}$ interacts with the key and value matrices of the image tokens $\boldsymbol{K}_{\text{img}}^{\ell,h}$ and $\boldsymbol{V}_{\text{img}}^{\ell,h}$ to retrieve relevant information. Using Equation (4), (5), the attention weights and the MHA output for $\boldsymbol{q}_i^{\ell,h}$ are computed as:

$$\boldsymbol{a}_i^{\ell,h} = \text{softmax}\left(\frac{\boldsymbol{q}_i^{\ell,h}[\boldsymbol{K}_{\text{img}}^{\ell,h}; \boldsymbol{K}_{\text{txt}}^{\ell,h}]^\top}{\sqrt{d_k}} + \boldsymbol{m}\right) \in \mathbb{R}^{P_{\text{img}}+P_{\text{txt}}}, \tag{6}$$

$$\text{MHA}^{\ell,h}(\boldsymbol{z}_i^{\ell-1}) = \boldsymbol{a}_i^{\ell,h}[\boldsymbol{V}_{\text{img}}^{\ell,h}; \boldsymbol{V}_{\text{txt}}^{\ell,h}]\boldsymbol{W}_O^{\ell,h} \in \mathbb{R}^d. \tag{7}$$

Intuitively, the attention weights $\boldsymbol{a}_i^{\ell,h}$ represent the relevance of other tokens to the text token $\boldsymbol{z}_i^{\ell-1}$, and the MHA layer's output is computed as a weighted sum of these tokens, with the attention weights serving as the coefficients.

A.2 Head Ablation

In this section, we formally explain the concept of head ablation utilized in §4.1.

Attention knockout, or head ablation [12] is a widely used method for analyzing information flow in attention mechanisms. In the context of LVLMs, it is specifically employed to measure image-to-text information flow [9, 10, 33]. The method involves setting the attention weights between image and text tokens to baseline values and observing the resulting effect on the model's performance. Head ablation replaces the image key and value matrices with averaged values from other images.

Specifically, we collect the key and value matrices of image tokens from 100 images in the COCO [27] dataset and compute the average key matrices $\boldsymbol{K}_{\text{img, avg}}^{\ell,h}$ and value matrices $\boldsymbol{V}_{\text{img, avg}}^{\ell,h}$. The image key and value matrices in the attention mechanism are then replaced as follows:

$$\boldsymbol{K}_{\text{img}}^{\ell,h} \leftarrow \boldsymbol{K}_{\text{img, avg}}^{\ell,h}, \quad \boldsymbol{V}_{\text{img}}^{\ell,h} \leftarrow \boldsymbol{V}_{\text{img, avg}}^{\ell,h}. \tag{8}$$

Finally, the attention weights and the MHA layer's output are recomputed using Equation (6), (7). By substituting the image key and value matrices with their average values, head ablation effectively eliminates information flow while preserving the natural attention distribution of the model.

A.3 Head Attribution

In this section, we provide more details on the *head attribution* method introduced in §4.1.

Let $\boldsymbol{x} \in \{0,1\}^{LH}$ denote a binary vector, where L is the number of layers and H is the number of attention heads per layer. An element x_n of $\boldsymbol{x}$ is set to 1 if the n-th attention head[5] is intact, and 0 if it is ablated. For each sample, we ablate $p = 75\%$ of the attention heads in the model[6] $M = 10,000$ times to obtain $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \pi(\boldsymbol{x}^{(1)})), \ldots, (\boldsymbol{x}^{(M)}, \pi(\boldsymbol{x}^{(M)}))\}$, where $\pi(\boldsymbol{x})$ denotes the normalized logit as described in §4.1. The obtained dataset $\mathcal{D}$ is then split into two subsets: $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, containing 80% and 20% of the samples, respectively.

We train a linear regression model $\hat{\pi}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta} + b$ on $\mathcal{D}_{\text{train}}$ to predict the normalized logit $\pi(\boldsymbol{x})$ from the binary vector $\boldsymbol{x}$. The model is trained with elastic net regularization [65], implemented using Scikit-learn [66]. Specifically, our hyperparameters are set as follows:

```
ElasticNet(alpha = 0.0005, l1_ratio = 0.5, max_iter = 1000).
```
(9)

The result of linear regression is evaluated on $\mathcal{D}_{\text{test}}$ using the explained variance (R^2) and the Pearson correlation coefficient (ρ) between the predicted logits $\hat{\pi}(\boldsymbol{x})$ and the true logits $\pi(\boldsymbol{x})$, following [13].

[5] The attention heads are indexed in row-major order; that is, the first H elements correspond to the attention heads of the first layer, the next H elements correspond to those of the second layer, and so on. In other words, h-th attention head of the ℓ-th layer is indexed as $n = \ell \times H + h$, where $\ell \in \{0, 1, \ldots, L-1\}$ and $h \in \{0, 1, \ldots, H-1\}$.

[6] See §B.2 for the ablation study on the ablation ratio p.

A.4 Miscellaneous Details

To enhance reproducibility and clarity, we provide further details on the experiments described in §4 and 5.

Model Details on Head Ablation Applied to Individual Attention Heads (§4.1). Let $\boldsymbol{x}_{\neg n}$ denote a binary vector in which the n-th attention head is ablated (*i.e.*, $x_n = 0$, while all other elements of $\boldsymbol{x}_{\neg n}$ are set to 1). The logit difference is then computed as $\Delta\pi_n = 1 - \pi(\boldsymbol{x}_{\neg n})$.

Image Attention Weights for Each Attention Head (§4.2 and 4.3). We calculate image attention weights for each attention head and use them as a criterion to select the most important attention heads in Figure 3A. These weights are then compared with attribution coefficients in Figure 3D. The image attention weights are determined by averaging the attention weights of image tokens across text tokens. Specifically, given the attention weights $\boldsymbol{A}^{\ell,h} \in \mathbb{R}^{(P_{\text{img}}+P_{\text{txt}})\times(P_{\text{img}}+P_{\text{txt}})}$ from Equation (4), the image attention weight for each head is computed as follows:

$$\text{Image Attention Weight} = \frac{\sum_{i=P_{\text{img}}+1}^{P_{\text{img}}+P_{\text{txt}}} \sum_{j=1}^{P_{\text{img}}} a_{i,j}^{\ell,h}}{P_{\text{txt}}}, \tag{10}$$

where $a_{i,j}^{\ell,h}$ represents the element of $\boldsymbol{A}^{\ell,h}$ at the i-th row and j-th column. Similar methods for calculating attention weights per head or per layer have been employed in various studies [10, 22, 33, 39, 55, 67, 68].

More Details on t-SNE Visualization (§4.3). For the t-SNE visualization in Figure 3E and 13, we collect additional samples from the COCO dataset to enhance the diversity of images for each object. Specifically, objects are included if LLaVA-1.5-7B generates at least 10 correct samples for them. From these, we randomly select up to 15 samples per object and repeat the head attribution process to compute attribution coefficients $\boldsymbol{\theta}$.

More Details on Text Token Experiments (§5.1). The experiments described in §5 retain only the minimum number of attention heads required to maintain faithfulness above 0.8. This approach focuses on understanding the behavior of the most critical attention heads for image-to-text information flow. To formalize this, we define a binary vector $\boldsymbol{x}^*$, where each element is set to 1 if the corresponding attention head is essential to ensure faithfulness > 0.8, and 0 otherwise.

Since head ablation can be selectively applied to subsets of text or image tokens as explained in §A.2, we introduce binary vectors $\boldsymbol{u} \in \{0,1\}^{P_{\text{txt}}}$ and $\boldsymbol{v} \in \{0,1\}^{P_{\text{img}}}$, where P_{txt} and P_{img} represent the total number of text and image tokens, respectively. Each element u_i of $\boldsymbol{u}$ is set to 1 if the i-th text token remains intact, and 0 if it is ablated. Similarly, each element v_j of $\boldsymbol{v}$ is set to 1 if the j-th image token remains intact, and 0 if it is ablated.

The normalized logit, computed using only the attention heads required for faithfulness > 0.8 (denoted as $\boldsymbol{x}^*$), with intact text queries and image tokens, is denoted as $\pi(\boldsymbol{x}^*; \boldsymbol{u}, \boldsymbol{v})$. When all text and image tokens are intact, $\pi(\boldsymbol{x}^*; \mathbf{1}, \mathbf{1})$ simplifies to $\pi(\boldsymbol{x}^*)$, as defined in §4.

In §5.1, we aim to measure the causal effect of blocking image information flow to individual text tokens. To achieve this, we define $\boldsymbol{u}_{\neg i}$ as a binary vector where the i-th text token is ablated (i.e., only $u_i = 0$, while all other elements of $\boldsymbol{u}_{\neg i}$ remain 1). The logit difference relative to $\pi(\boldsymbol{x}^*)$ is then computed as:

$$\Delta\pi_i = 1 - \frac{\pi(\boldsymbol{x}^*; \boldsymbol{u}_{\neg i}, \mathbf{1})}{\pi(\boldsymbol{x}^*)}. \tag{11}$$

If $\Delta\pi_i > 0.05$, we conclude that the i-th text token is important for image-to-text information flow.

Once the important text tokens are identified, we further evaluate whether they alone are sufficient for recovering the information flow. To do this, we define a binary vector $\boldsymbol{u}^*$, where $u_i^* = 1$ if $\Delta\pi_i > 0.05$ and $u_i^* = 0$ otherwise. The corresponding logit relative to $\pi(\boldsymbol{x}^*)$, computed as $\pi(\boldsymbol{x}^*; \boldsymbol{u}^*, \mathbf{1})/\pi(\boldsymbol{x}^*)$, indicates the extent to which the important text tokens restore the information flow. If the logit is sufficiently high, we conclude that the identified important text tokens are adequate for preserving image-to-text information flow.

More Details on Image Token Experiments (§5.2). In §5.2, we extend the head attribution method to image tokens. Following a similar approach to §A.3, we ablate $p = 75\%$ of the image tokens in the model $M = 10,000$ times to generate the dataset $\mathcal{D} =$

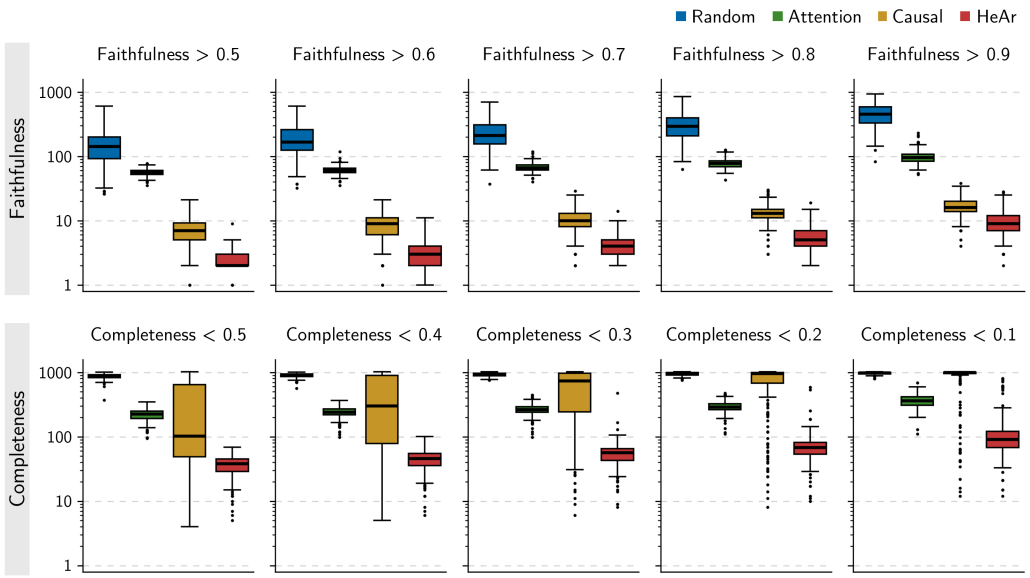

Figure 5: The minimum number of heads required for each threshold of faithfulness (Top) and completeness (Bottom) in LLaVA-1.5-7B. Related to Figure 3A.

$\{(\boldsymbol{v}^{(1)}, \tilde{\pi}(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v}^{(1)})), \ldots, (\boldsymbol{v}^{(M)}, \tilde{\pi}(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v}^{(M)}))\}$, where $\tilde{\pi}(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v})$ denotes the normalized logit computed as:

$$\tilde{\pi}(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v}) = \frac{\pi(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v})}{\pi(\boldsymbol{x}^*; \boldsymbol{u}^*, \mathbf{1})}. \tag{12}$$

We then train a linear regression model, $\hat{\pi}(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{\theta} + b$, to predict $\pi(\boldsymbol{x}^*; \boldsymbol{u}^*, \boldsymbol{v})$ from the binary vector $\boldsymbol{v}$. The model is trained using the same hyperparameters as described in §A.3.

Weighted Image Attention Weights for Each Sample (§5.2). Since image attention weights are computed for each attention head, we calculate the weighted image attention weights as a representative value for each sample. Let $\boldsymbol{A}^n = \boldsymbol{A}^{\ell,h}$, where $n = \ell \times H + h$. The weighted image attention weights $\boldsymbol{w} \in \mathbb{R}^{P_{\text{img}}}$ are defined as:

$$\boldsymbol{w} = \sum_{n=0}^{LH-1} \left[(x_n^* \cdot \theta_n) \sum_{i=P_{\text{img}}+1}^{P_{\text{img}}+P_{\text{txt}}} \{(u_i^* \cdot \Delta\pi_i) \cdot \boldsymbol{a}_i^n[: P_{\text{img}}]\} \right], \tag{13}$$

where $\boldsymbol{a}_i^n$ represents the i-th row of $\boldsymbol{A}^n$, θ_n is the n-th element of the attribution coefficients $\boldsymbol{\theta}$ derived from head attribution, x_n^* is the n-th element of $\boldsymbol{x}^*$, u_i^* is the i-th element of $\boldsymbol{u}^*$, and $\Delta\pi_i$ is the logit difference for the i-th text token (Equation (11)). Intuitively, the weighted image attention weights $\boldsymbol{w}$ consider the importance of each attention head and text token in determining the attention weights for image tokens. The weighted image attention weights $\boldsymbol{w}$ are reshaped into dimensions $\sqrt{P_{\text{img}}} \times \sqrt{P_{\text{img}}}$ to visualize the attention distribution in Figure 4B and to create joint distribution plots in Figure 4C.

Computational Resources. All experiments were conducted on a single NVIDIA RTX 6000 GPU. For LLaVA-1.5-7B, processing a single input example requires approximately 0.81 GPU hours. The total GPU hours for a complete experiment can be calculated by multiplying this value by the number of images. For example, in the case of COCO, using 200 images results in approximately 160 GPU hours.

B Ablation Studies

B.1 Ablation Study on Threshold for Validating Head Attribution

In §4.2, we validate head attribution by counting the number of heads required to achieve faithfulness > 0.8 and completeness < 0.2. In this section, we conduct an ablation study to examine how

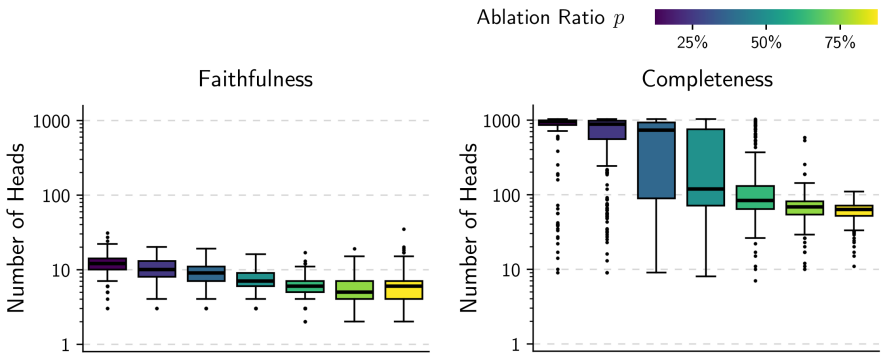

Figure 6: The minimum number of heads for faithfulness > 0.8 (Left) and completeness < 0.2 (Right) for each ablation ratio p in LLaVA-1.5-7B. Related to Figure 3A.

different thresholds affect the number of heads needed to achieve various levels of faithfulness and completeness. Specifically, we vary the faithfulness threshold from 0.5 to 0.9 and the completeness threshold from 0.5 to 0.1.

We present the number of heads required to achieve these faithfulness and completeness levels in Figure 5. As the faithfulness threshold increases and the completeness threshold decreases, the number of required heads increases. Nevertheless, across all thresholds, the head attribution method consistently requires fewer heads to attain the same faithfulness and completeness levels.

B.2 Ablation Study on Ablation Ratio of Head Attribution

In the main paper, we set the ablation ratio p to 75% for the head attribution method. In this section, we conduct an ablation study to examine how different ablation ratios affect the results of the head attribution method. Specifically, we vary the ablation ratio from 12.5% to 87.5%.

As shown in Figure 6, sufficiently high ablation ratios ($> 50\%$) require fewer heads to achieve the same level of faithfulness and completeness. This result aligns with the motivation behind the head attribution method, which mitigates interactions between heads by ablating multiple heads simultaneously. Additionally, the results remain consistent across sufficiently high ablation ratios, indicating that the head attribution method operates robustly within a reasonable range of ablation ratios.

C Additional Experiments

C.1 Head Ablation on Individual Attention Heads

As described in §4.1, we apply head ablation to each attention head individually. We then measure the logit difference $\Delta\pi_n$ between the original and patched models for each head n (see §A.4 for technical details). As shown in Figure 7, the logit difference is not significant for most samples. These results suggest that image-to-text information flow is distributed across many attention heads rather than being concentrated in a single specialized head. Therefore, we introduce head attribution to more effectively measure each head's contribution to image-to-text information flow.

C.2 Logit Lens Analysis

To further understand the role of attention heads in image-to-text information transfer, we apply logit lens [43] to the top 5 attention heads ranked by attribution coefficients $\boldsymbol{\theta}$ and visualize the top 10 vocabulary items in Figure 8. We find that some heads are clearly associated with the final answer, while others are only weakly related or primarily encode background information. Although certain influential heads are not interpretable in the vocabulary space, the logit lens reveals that most top-ranked heads transfer semantically relevant information in text, rather than contributing in an uninterpretable way. This suggests that some degree of image-to-text transformation may have

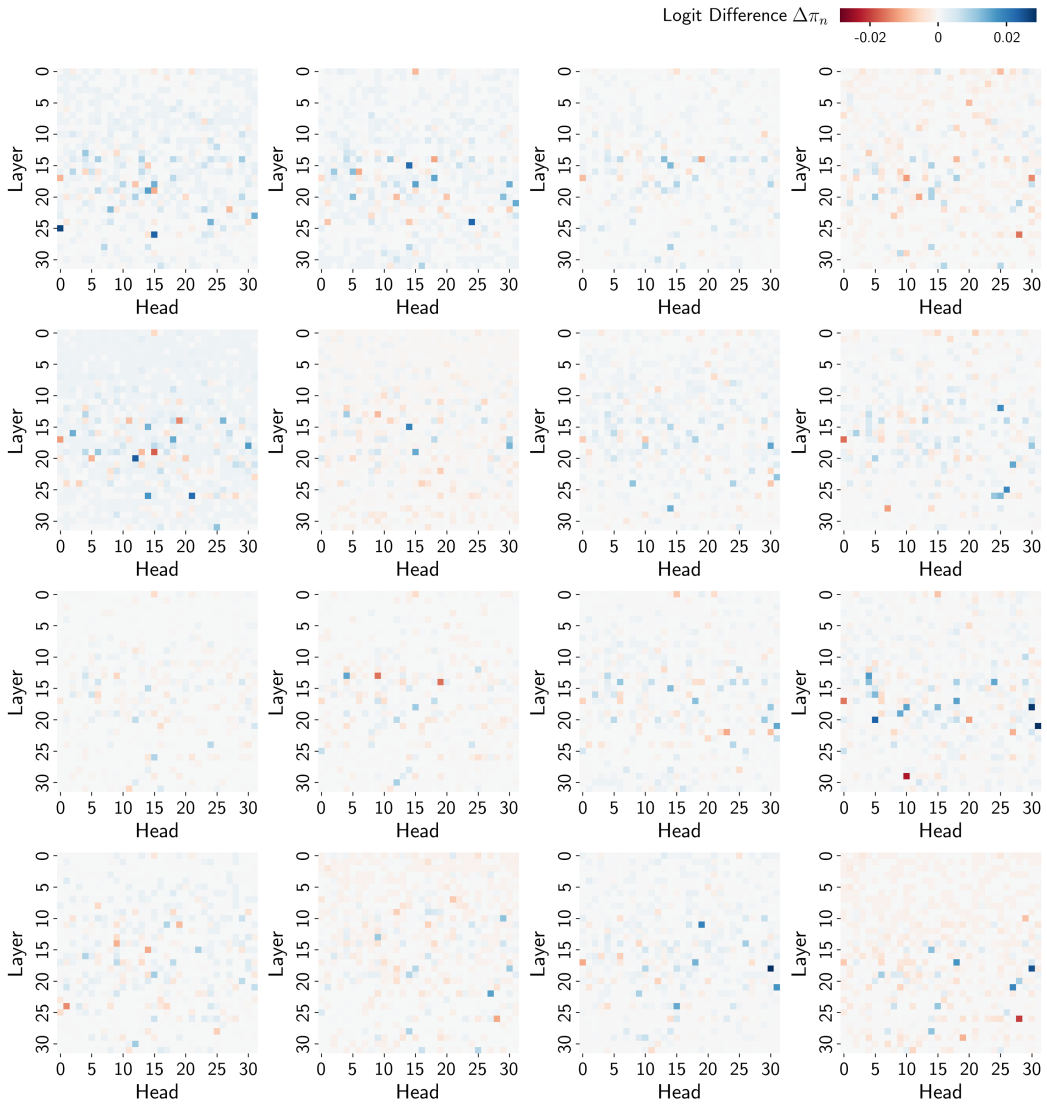

Figure 7: Head ablation results for individual attention heads in LLaVA-1.5-7B. Each subfigure presents the results of individual head ablation for a randomly selected image sample. The logit difference $\Delta\pi_n$ is not significant for most samples, suggesting that the image-to-text information flow is distributed across multiple attention heads and individual head ablation does not provide a consistent explanation for the model's behavior (§C.1). Related to §4.1.

Image	Top 5 Heads (Head Attribution)	Logit Lens on Attention Heads (Top 10 Vocabularies)
Dog	(L, H)	Top 10 Vocabularies
	(15, 22)	'_dog', '_fetch', '_water', '_Dog', '_dogs', 'dog', '_beach', '_hoof', '_Water', '_Animal'
	(14, 26)	'oga', 'heits', '_Ces', 'uby', 'rum', 'ogo', '_owner', 'gems', 'ixel', '_portrait'
	(20, 5)	'_dog', '_dogs', '_Dog', 'dog', '_pu', '_Kenn', '_Pu', 'ogs', 'pu', '_fetch'
	(14, 29)	'NR', '_Nature', '_Safari', '_fis', '_beaut', '_Sport', '_alg', 'läu', '_species', '_cam'
	(16, 4)	'_Camera', 'Camera', '_animal', '_Phot', '_camera', '_Animal', '_zo', '_animals', '_Image', '_Anim'
Boat	(L, H)	Top 10 Vocabularies
	(17, 9)	'_boat', '_boats', '_floating', '_float', '_Float', '_dock', '_fish', '_shore', 'float', 'Float'
	(14, 29)	'_fish', '_boat', '_cru', '_Fish', '_water', 'fish', '_marine', '_river', '_boats', '_waters'
	(18, 12)	'_water', '_fish', 'water', '水', '_aqu', '_waters', '_Water', '_ocean', '_marine', '_sea'
	(20, 12)	'_boat', '_boats', '_ship', '_floating', '_Float', '_waters', '_float', '_vessel', '_ships', 'bo'
	(17, 18)	'_boats', '_ships', '_vessels', '_vessel', '_корабl', '_boat', '_keyboard', '_vehicles', '_cars', 'ships'
Dount	(L, H)	Top 10 Vocabularies
	(17, 30)	'_cookie', '_bread', '_cookies', '_dess', '_pas', '_bis', 'Cookie', '_jel', '_Bak', '_Baker'
	(14, 29)	'食', 'jed', '料', '_Unidos', '_chef', '_inspect', 'Texture', '_Toast', '店', '_beaut'
	(20, 5)	'_Bak', '_dess', '_pas', '_bread', '_cookies', '_Baker', '_cookie', '_Pas', '_pie', '_bis'
	(15, 22)	'_Baker', '_bread', '_dess', '_Bak', 'rum', 'disk', '_rus', '_cust', 'koz', '_pie'
	(17, 18)	'ourd', 'oca', '_Bak', 'onn', 'urn', 'widget', '_cookies', '_round', '_gran', '_drums'

Figure 8: Logit lens analysis on top-ranked attention heads, selected by attribution coefficients $\boldsymbol{\theta}$. Top-ranked attention heads often transfer semantically meaningful information aligned with the final textual output (§C.2).

already occurred at the image token level, and the attention heads retrieve this partially converted information.

C.3 Generalization of Attribution Coefficients Across Prompts

In the main paper, we fix the prompt, "What is the main object in the image? Please answer with a single word." and measure the attribution coefficients $\boldsymbol{\theta}$ for each attention head. We also examine whether LVLMs utilize attention heads based on the meaning of the task or if they are sensitive to the wording of the prompt. To investigate this, we assess whether the attribution coefficients, which represent the contribution of each attention head to the image-to-text information flow, remain consistent across different prompts. If the attribution coefficients are stable across varying prompts, it suggests that LVLMs systematically rely on attention heads according to the task's meaning rather than the specific wording of the prompt.

We ask ChatGPT to generate nine different prompts, each with different wording but maintaining the same visual object identification task. The prompts are as follows:

1. Identify the primary subject in this picture with a single word.
2. Please specify the key item in this image using one word only.
3. Please give me the main subject in this picture, stated in a single word.
4. Provide a one-word term for the principal object depicted here.
5. In one word, what is the primary focus of the photo?
6. Name the primary element in the photo, limiting your answer to one word.
7. Which single word best describes the main subject of this picture?
8. Using a single word, name the object that stands out the most in this image.
9. Offer the primary object's name from the picture, restricted to a single word.

We collect 40 samples for each prompt and measure the cosine similarity between the attribution coefficients obtained from the original prompt and those from each of the nine new prompts.

Table 1: Cosine similarity of the attribution coefficients $\boldsymbol{\theta}$ across different prompts. The consistency of the attribution coefficients across prompts suggests that LVLMs utilize attention heads based on the task rather than the specific wording of the prompt (§C.3).

Prompts	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9
Similarity	0.97 ± 0.02	0.96 ± 0.03	0.97 ± 0.02	0.96 ± 0.02	0.91 ± 0.03	0.94 ± 0.02	0.94 ± 0.03	0.93 ± 0.04	0.94 ± 0.03

As shown in Table 1, the cosine similarity between the attribution coefficients from the original prompt and those from the nine new prompts is high, with an average similarity of 0.95. These results suggest that the attribution coefficients generalize across different prompts, indicating that LVLMs systematically utilize attention heads based on the task's meaning rather than the precise wording of the prompt.

C.4 Generalization of Attribution Coefficients Across Image Styles

To further evaluate whether the model's internal mechanisms generalize beyond natural images, we analyze its behavior under more challenging domain shifts. While Figure 3E demonstrates that semantically similar natural images elicit consistent attention head usage patterns, all tested inputs were restricted to natural images. Here, we extend the analysis to the DomainNet dataset [28], which contains images from diverse visual domains, including real, sketch, and clipart styles. This setting enables a more rigorous test of the model's robustness and invariance to domain-specific variations. We randomly sample 300 images across the three domains (see Figure 9D for examples), and apply our head attribution method to examine whether semantically equivalent images still activate similar attention head patterns.

First, we confirm that the proposed head attribution method maintains high faithfulness and completeness compared to the Attention and Causal baselines (Figure 9A), in line with the results in Section 4.2. Figure 9B and C replicate the analyses of Figure 3C and D, showing that head attribution patterns remain consistent across different domains. This consistency suggests that the model's attention heads are not tailored to specific visual domains but rather capture domain-invariant semantic structures. Finally, Figure 9E illustrates that attention head usage patterns cluster similarly across styles, as observed previously in natural image domains. This further supports the conclusion that the model generalizes well across domains and that its internal mechanisms reflect semantic content rather than superficial style.

Overall, these results demonstrate that our finding extends beyond natural images and remains effective across diverse visual domains, validating its generalizability in broader settings.

C.5 Attribution Coefficients of Individual Attention Heads

To gain a deeper understanding of individual attention heads, we concatenate the attribution coefficients θ_n of each head n across all samples used in Figure 3E and define this as a head vector $\boldsymbol{h}_n$. We then visualize the head vectors of attention heads with $\theta_n > 0.1$ in at least one sample in Figure 10. To identify patterns among attention heads, we apply hierarchical clustering to the head vectors, grouping similar attention heads together. The results are interpreted as follows.

First, the head vectors form two major clusters. The first cluster consists of attention heads that are consistently utilized across samples (sample-agnostic heads), while the second cluster comprises heads that are used in a sample-specific manner (sample-specific heads). This finding aligns with the observation in Figure 3C, indicating that LVLMs leverage both sample-agnostic and sample-specific attention heads to process image-to-text information flow.

Second, the head vector patterns remain consistent within the same object category or super-category. In other words, different images containing the same object tend to activate the same attention heads. This observation further supports the results in Figure 3E, demonstrating that LVLMs utilize attention heads based on the semantics of the input image.

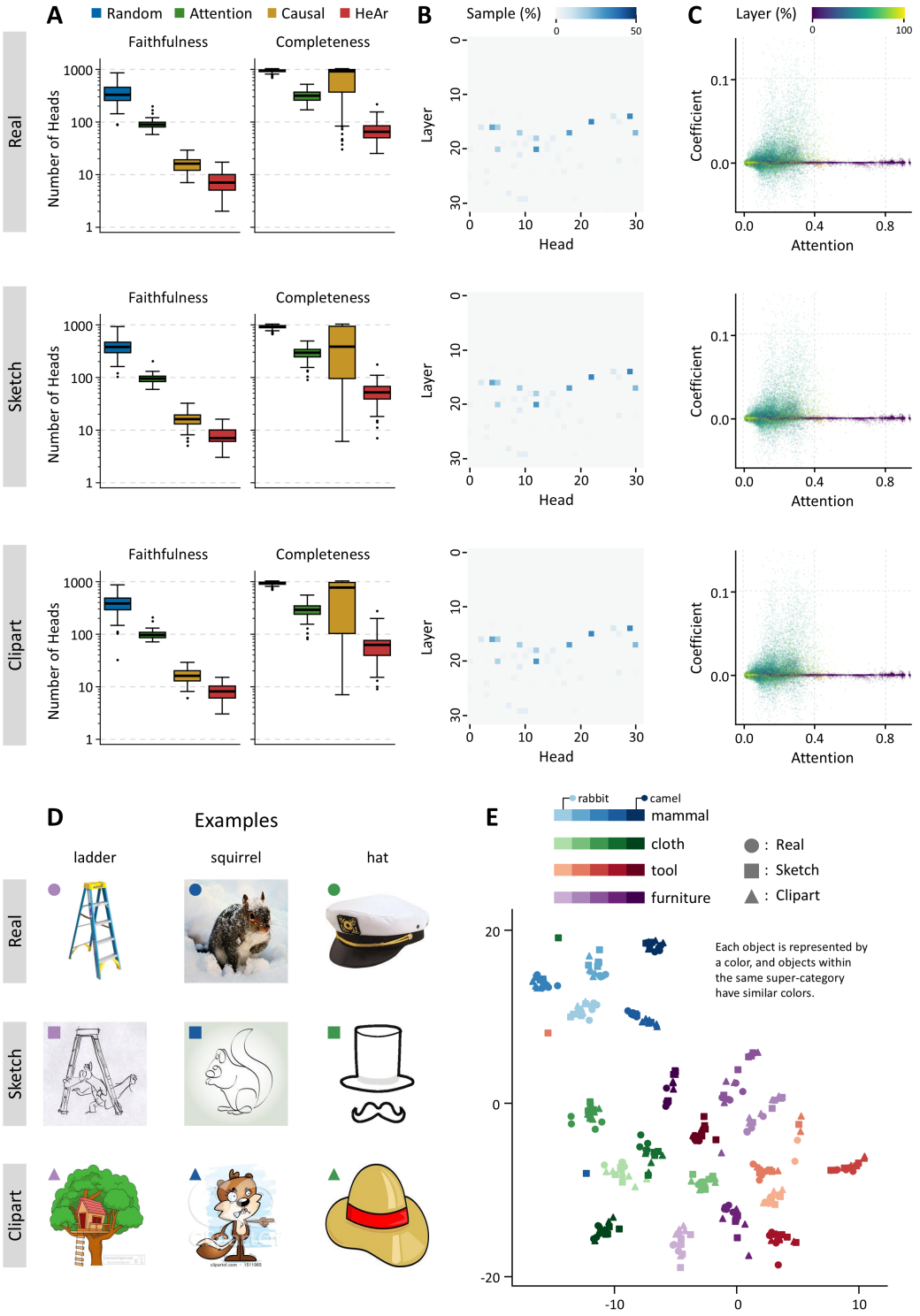

Figure 9: DomainNet experiment. Despite large variations in image style, attention head patterns remain consistent across domains when the semantic content is preserved (§C.4).

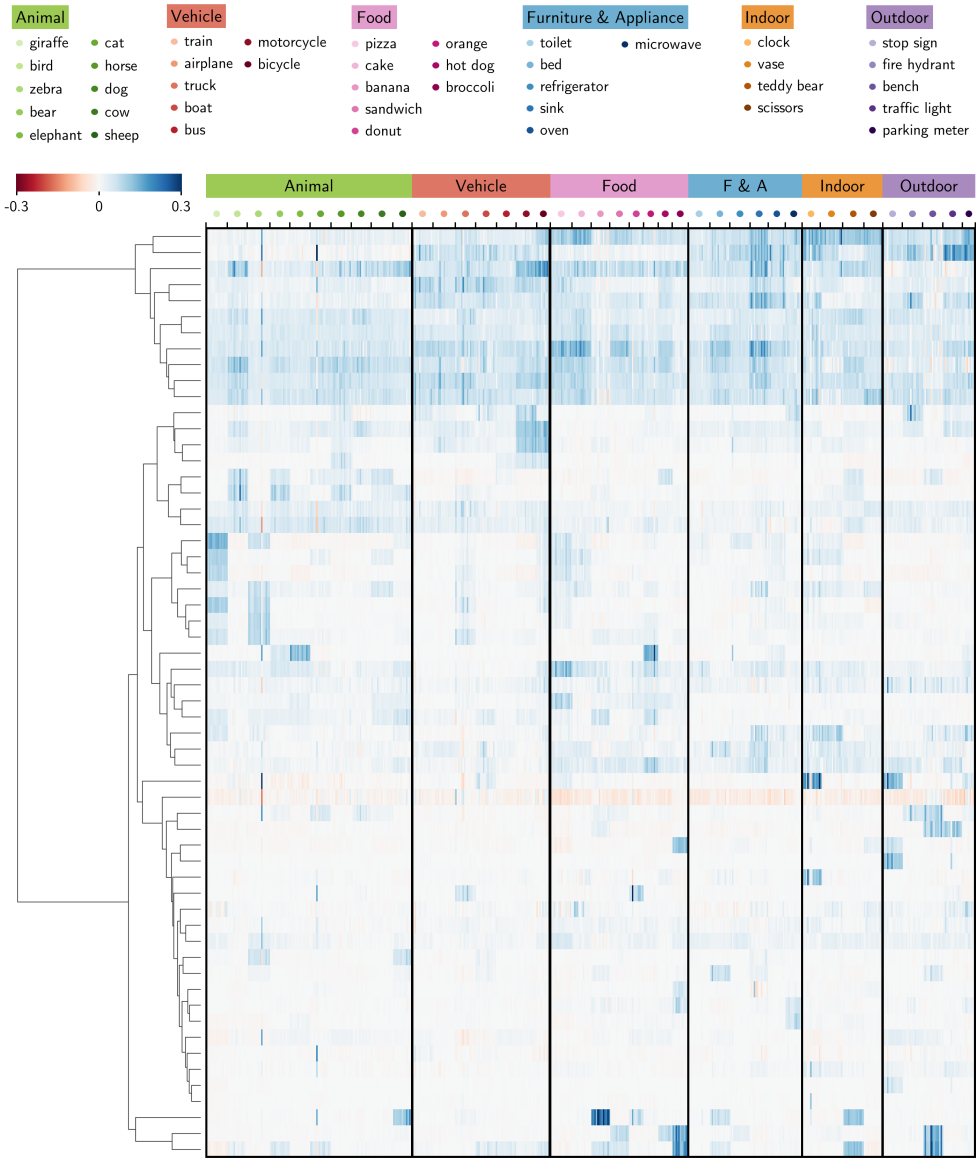

Figure 10: Visualization of the head vectors $\boldsymbol{h}_n$ of attention heads with $\theta_n > 0.1$ in at least one sample in LLaVA-1.5-7B. The head vectors are hierarchically clustered. The x-axis represents each sample, and the y-axis represents each attention head. See §C.5 for the interpretation.

D Additional Results for Other Models

In this section, we provide additional results for other models.

In §4.1, we fit a linear regression model to estimate the final logit based on which attention heads remain intact. We report the quantitative results of the regression model, including the explained variance (R^2) and the Pearson correlation coefficient (ρ), in Table 2. High R^2 and ρ values indicate that the contribution of attention heads to the final logit is well captured by the linear regression model.

In §4.2, we validate the head attribution method by comparing the number of heads required to achieve the same level of faithfulness and completeness across three other methods: Random, Attention, and Causal. We report the full results in Figure 11. The head attribution method consistently requires fewer heads to achieve the same level of faithfulness and completeness, demonstrating its effectiveness in measuring the importance of attention heads.

In §4.3, we analyze the head attribution results from three perspectives: the distribution of important attention heads (Figure 3C), their relation to image attention weights (Figure 3D), and systematic patterns in attention heads (Figure 3E). Figure 12 shows the distribution of important attention heads across all models. Generally, important heads are concentrated in mid-to-late layers. Interestingly, LLaVA-1.5 and LLaVA-NeXT, which share the same backbone large language model, exhibit similar distributions of important attention heads. This result suggests that, to some extent, the role of attention heads is predetermined by the backbone large language model. We leave further investigation of this phenomenon for future work. Figure 14 visualizes the relationship between attribution coefficients $\boldsymbol{\theta}$ and image attention weights. No clear correlation is observed between the two, supporting the claim that image attention weights are not reliable indicators of attention head importance. Figure 13 presents the t-SNE visualization of the attribution coefficients $\boldsymbol{\theta}$ in LLaVA-NeXT-7B, InternVL2.5-8B, and Qwen2-VL-7B. We select representative models from each family due to the computational cost of collecting additional samples. The t-SNE visualization reveals that LVLMs systematically utilize attention heads in response to the semantic content of the input, as samples belonging to the same super-category are typically clustered together.

In §5.1, we identify the text tokens that receive image information by measuring the causal effect of blocking image information flow to each text token. As shown in Figure 15, these tokens are primarily role tokens and the final token in the text. Furthermore, we validate these findings by measuring the logit when retaining only the role tokens and the final token. More specifically, we retain tokens with $\Delta\pi_i > 0.05$ (see §A.4 for the definition), which are highlighted in red in Figure 15. As shown in Table 3, the logit remains well preserved, indicating that receiving image information in the role tokens and the final token is essential for the model to predict the correct answer.

In §5.2, we qualitatively visualize the attribution coefficients $\boldsymbol{\theta}$ and image attention weights (see §A.4 for technical details). Additional qualitative results are presented in Figure 16 and Figure 17. The results indicate that only a subset of object tokens and a few background tokens contribute to the final logit, while the remaining tokens have minimal impact. Notably, the number of important tokens is smaller than the number of tokens with high image attention weights, suggesting that further token reduction is possible without significantly affecting the model's performance.

Table 2: Quantitative results of head attribution for all models. The table presents the explained variance (R^2) and the Pearson correlation coefficient (ρ) between $\pi(\boldsymbol{x})$ and $\hat{\pi}(\boldsymbol{x})$. These results indicate that the contribution of attention heads to the final logit is well captured by the linear regression model. Related to §4.1.

Model	R^2	ρ
LLaVA-1.5-7B	0.82 ± 0.05	0.91 ± 0.03
LLaVA-1.5-13B	0.79 ± 0.06	0.89 ± 0.03
LLaVA-NeXT-7B	0.83 ± 0.05	0.91 ± 0.03
LLaVA-NeXT-13B	0.82 ± 0.04	0.91 ± 0.02
InternVL2.5-1B	0.92 ± 0.03	0.96 ± 0.02
InternVL2.5-2B	0.79 ± 0.05	0.89 ± 0.03
InternVL2.5-4B	0.79 ± 0.06	0.89 ± 0.03
InternVL2.5-8B	0.77 ± 0.06	0.88 ± 0.03
Qwen2-VL-2B	0.90 ± 0.04	0.95 ± 0.02
Qwen2-VL-7B	0.82 ± 0.05	0.91 ± 0.03

Table 3: The logit relative to $\pi(\boldsymbol{x}^*)$ when retaining only the important text tokens ($\Delta\pi_i > 0.05$, highlighted in red in Figure 15). These results suggest that the role tokens and the final token are essential for the model to predict the correct answer. Related to §5.1.

Model	Logit relative to $\pi(\boldsymbol{x}^*)$
LLaVA-1.5-7B	0.98 ± 0.02
LLaVA-1.5-13B	0.92 ± 0.04
LLaVA-NeXT-7B	0.95 ± 0.02
LLaVA-NeXT-13B	0.89 ± 0.05
InternVL2.5-1B	0.94 ± 0.15
InternVL2.5-2B	0.93 ± 0.07
InternVL2.5-4B	0.93 ± 0.04
InternVL2.5-8B	0.82 ± 0.14
Qwen2-VL-2B	0.99 ± 0.01
Qwen2-VL-7B	0.82 ± 0.10

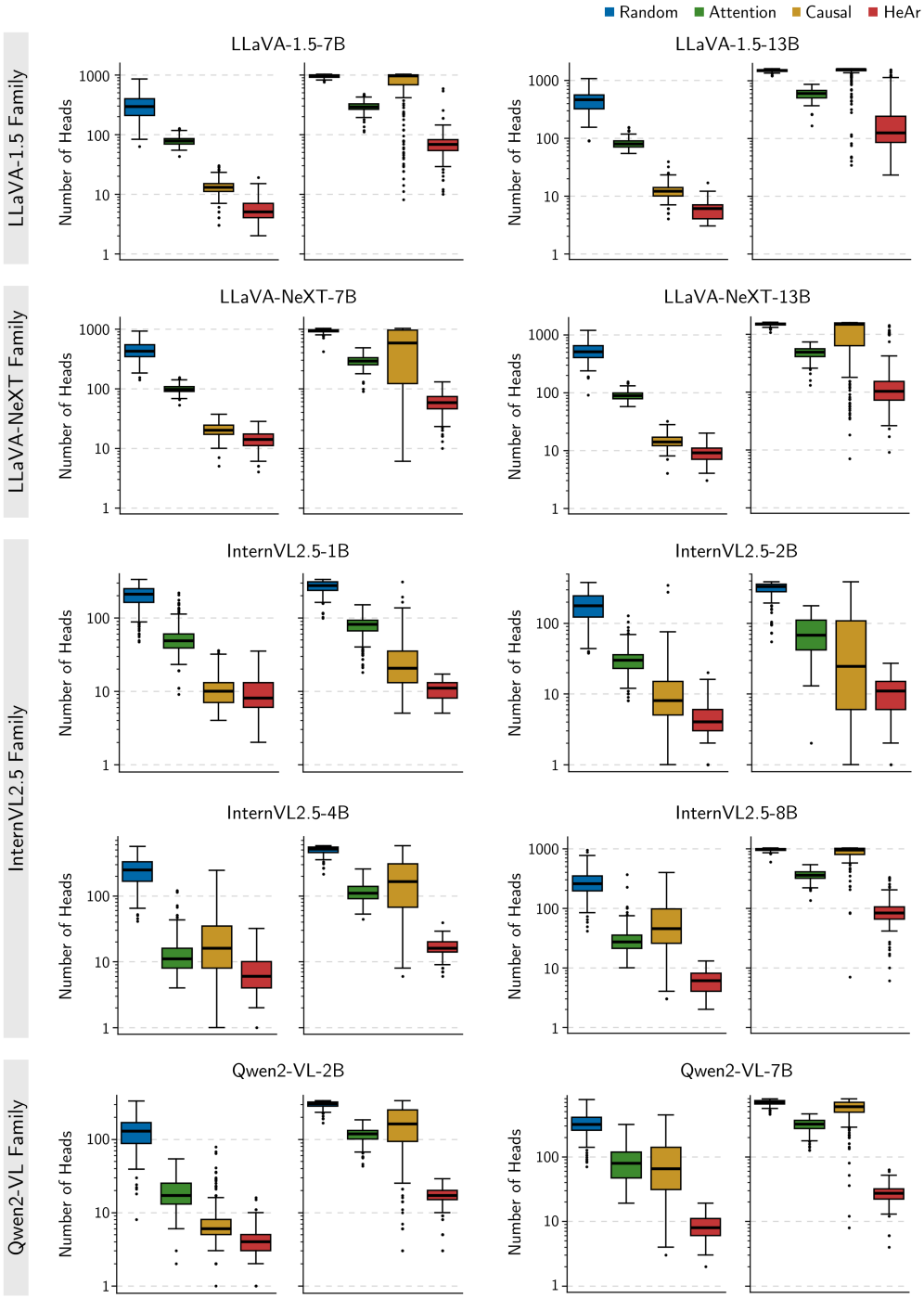

Figure 11: The minimum number of attention heads for faithfulness > 0.8 (Left) and completeness < 0.2 (Right) across all models. The head attribution method (HeAr) needs fewer heads to achieve the same level of faithfulness and completeness compared to the other methods. Related to Figure 3A.

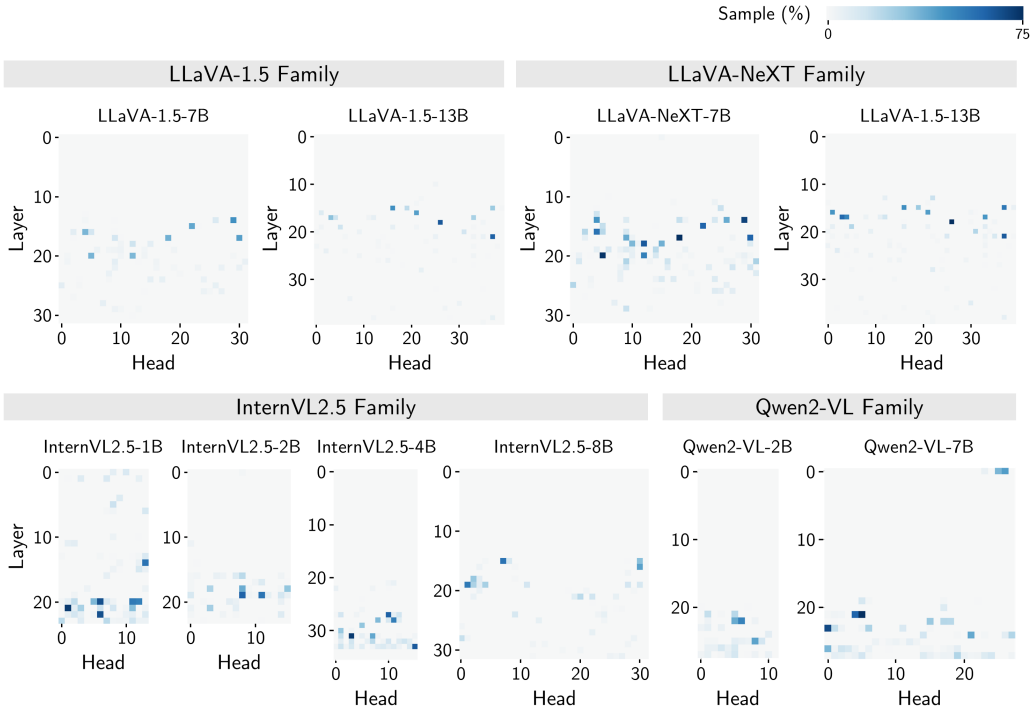

Figure 12: The portion of samples requiring a given head for faithfulness > 0.8 across all models. The important attention heads are located in the mid-to-late layers. Related to Figure 3C.

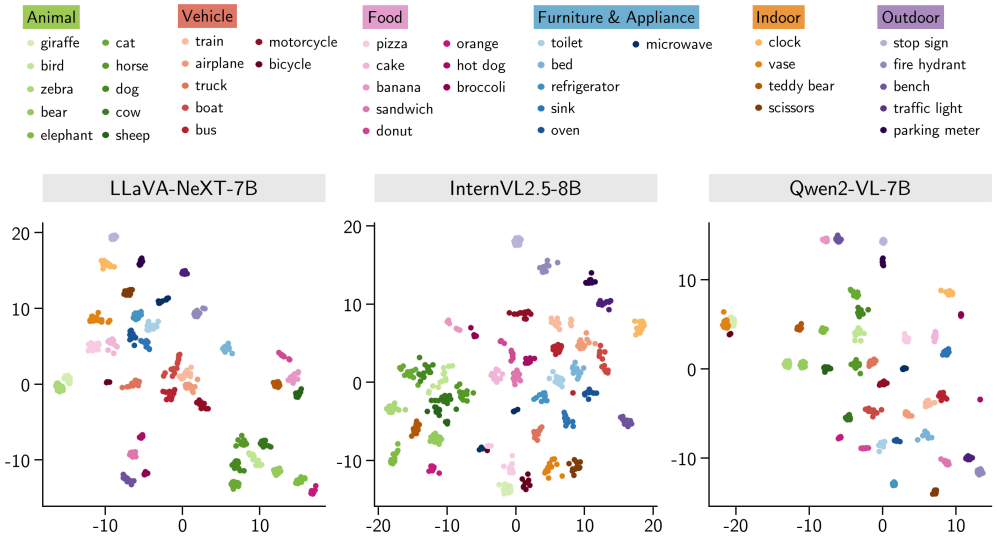

Figure 13: t-SNE visualization of attribution coefficients $\boldsymbol{\theta}$ in LLaVA-NeXT-7B, InternVL2.5-8B, and Qwen2-VL-7B. Objects with similar semantic meanings are typically clustered together, indicating that LVLMs systematically utilize attention heads to process the input image based on their semantic contents. Related to Figure 3E.

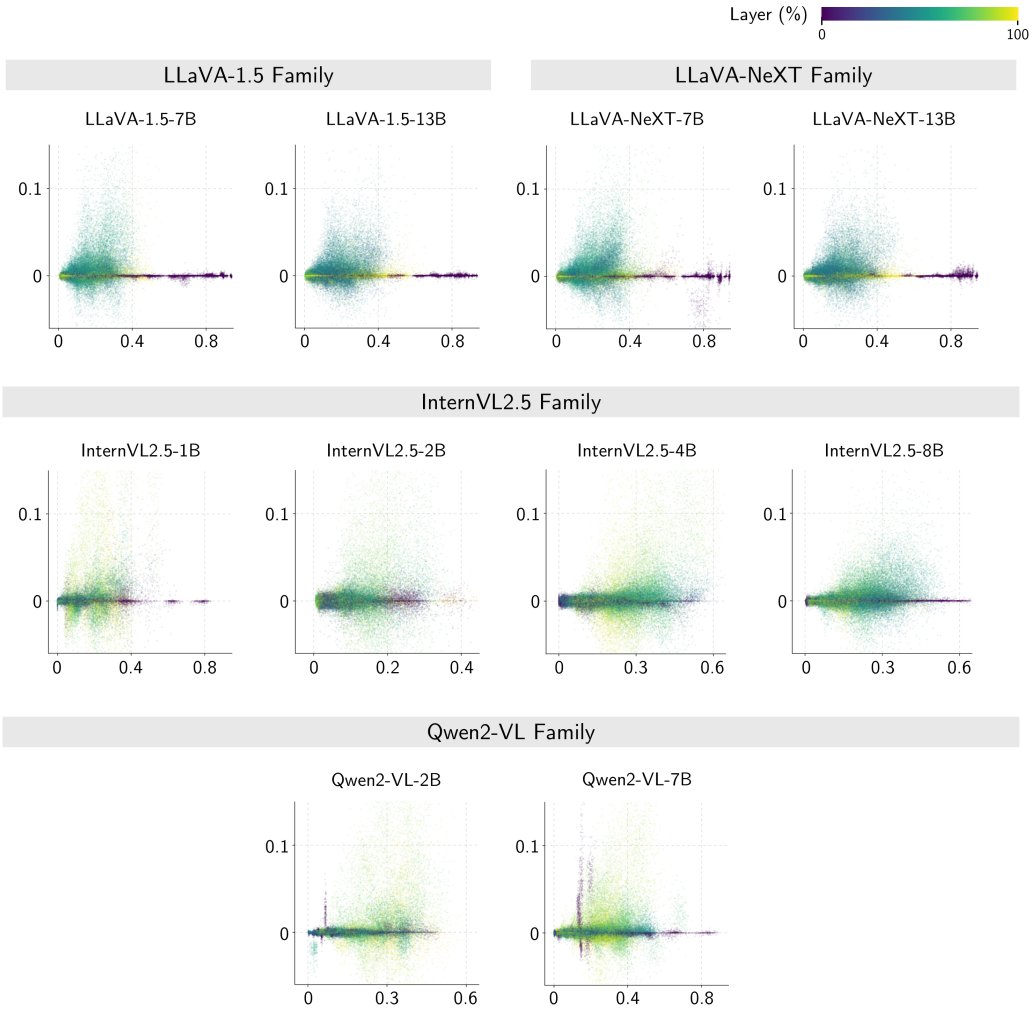

Figure 14: Scatter plot showing the relationship between attribution coefficients $\boldsymbol{\theta}$ and image attention weights across all models. The x-axis represents image attention weights, and the y-axis represents attribution coefficients. The color of each point represents the relative layer depth. No significant correlation is observed between the two, suggesting that image attention weights are not reliable for explaining the importance of attention heads. Related to Figure 3D.

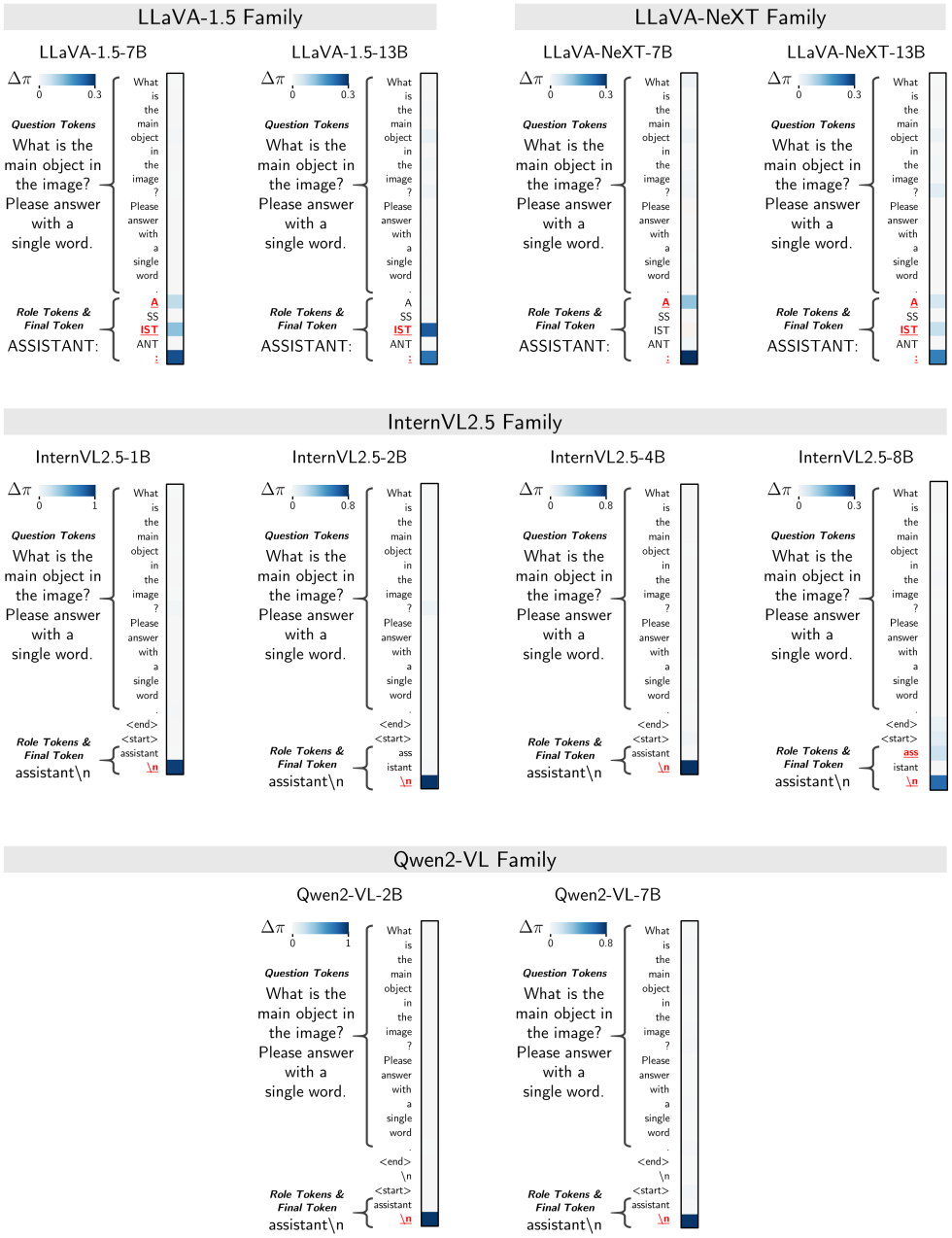

Figure 15: Logit difference relative to $\pi(\boldsymbol{x}^*)$ when blocking each text token across all models. The text tokens with $\Delta\pi_i > 0.05$ are highlighted in red, which are consistently the role tokens and the final token. Related to Figure 4A.

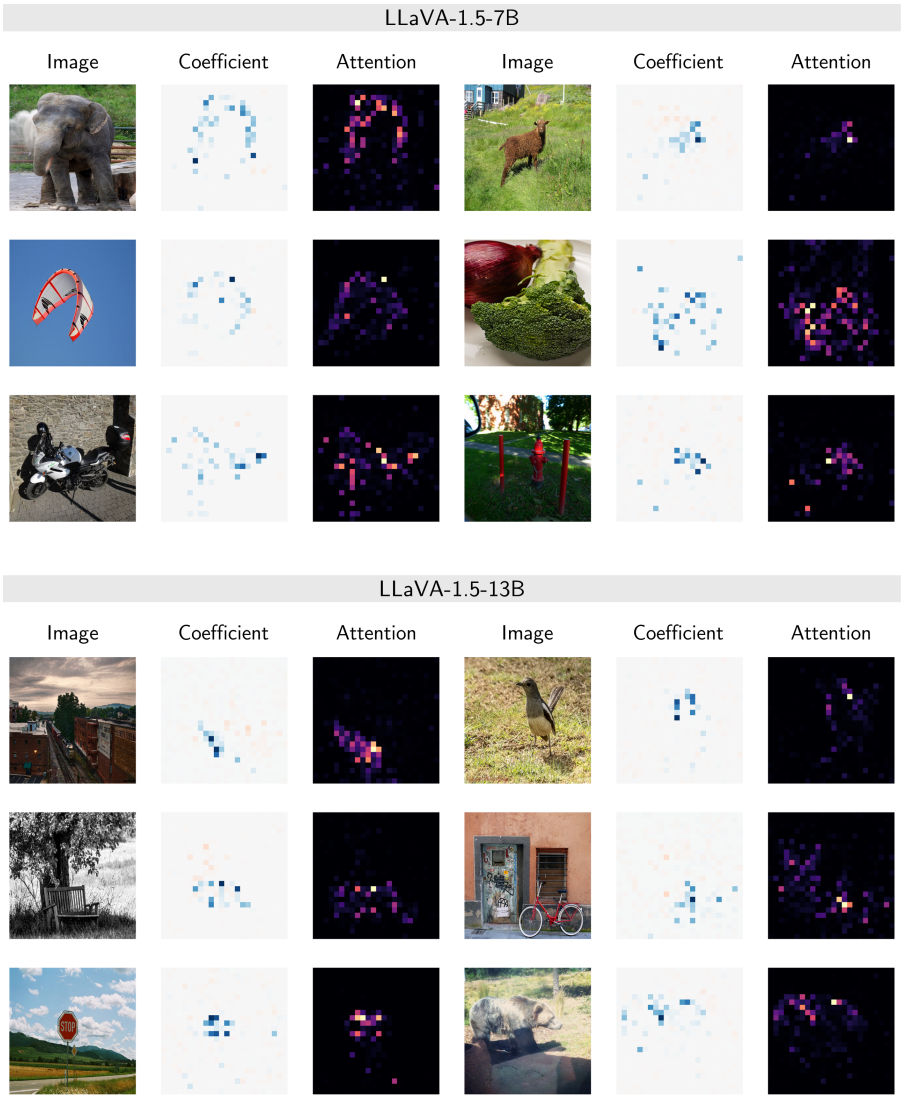

Figure 16: More qualitative results of image, attribution coefficients $\boldsymbol{\theta}$, and image attention weights for LLaVA-1.5-7B and LLaVA-1.5-13B. Related to Figure 4B.

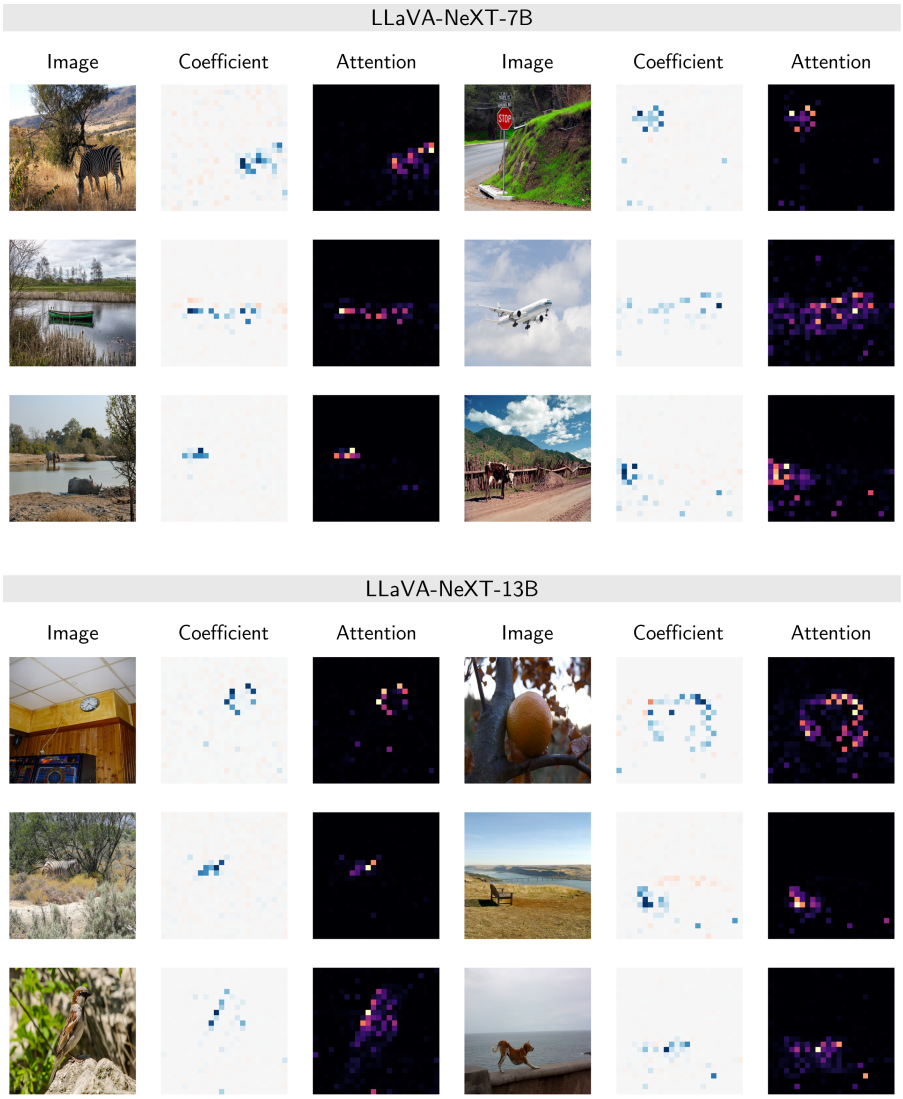

Figure 17: More qualitative results of image, attribution coefficients $\boldsymbol{\theta}$, and image attention weights for LLaVA-NeXT-7B and LLaVA-NeXT-13B. Related to Figure 4B.