From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks

Xiaofeng Zhang[1,2]*†, **Yihao Quan**[4]*, **Chen Shen** [3]‡, **Xiaosong Yuan**[3], **Shaotian Yan**[3],

Liang Xie[3], **Wenxiao Wang**[3], **Chaochen Gu** [1,2]✉, **Hao Tang**[5], **Jieping Ye**[3],

[1]Department of Automation, Shanghai Jiao Tong University
[2]Key Laboratory of System Control and Information Processing, MoE
[3]Alibaba Cloud Computing
[4]Beijing Jiaotong University
[5]Peking University

Abstract

Large Vision Language Models (LVLMs) achieve great performance on visual-language reasoning tasks, however, the black-box nature of LVLMs hinders in-depth research on the reasoning mechanism. As all images need to be converted into image tokens to fit the input format of large language models (LLMs) along with natural language prompts, sequential visual representation is essential to the performance of LVLMs, and the information flow analysis approach can be an effective tool for determining interactions between these representations. In this paper, we propose integrating attention analysis with LLaVA-CAM, concretely, attention scores highlight relevant regions during forward propagation, while LLaVA-CAM captures gradient changes through backward propagation, revealing key image features. By exploring the information flow from the perspective of visual representation contribution, we observe that it tends to converge in shallow layers but diversify in deeper layers. To validate our analysis, we conduct comprehensive experiments with truncation strategies across various LVLMs for visual question answering and image captioning tasks, and experimental results not only verify our hypothesis but also reveal a consistent pattern of information flow convergence in the corresponding layers, and the information flow cliff layer will be different due to different contexts. The paper's source code can be accessed from https://github.com/zhangbaijin/From-Redundancy-to-Relevance

1 Introduction

Multimodal models are more prevalent due to the capability of understanding vision-language rather than merely textual information. Most large vision-language models (LVLMs) consist of various visual encoders and large language models (LLMs). When images are fed to LLMs together with language prompts, they are transformed into hundreds or thousands of tokens (Liu et al., 2024b; Bai et al., 2023; Chen et al., 2023b; Zhu et al., 2024; Li et al., 2024; Dai et al., 2024; Zhu et al., 2023; Chen et al., 2024b; Ye et al., 2023; Chu et al., 2023; Zhang et al., 2023; Han et al., 2023; Wu et al., 2023; Wang et al., 2023b; Chen et al., 2023a). Such a transformation leads LVLMs to rely on sequential visual representation. Albeit popular LVLMs exhibit impressive generation capabilities, the black-box design of Transformers decoder-stacked LLMs hinders the interpretability of visual-language models. In this paper, we intend to explore the inner mechanisms during LVLMs reasoning.

Prior works have commenced preliminary exploration into the caption mechanisms of LLMs (Wang et al., 2023a; Zhang et al., 2024; Todd et al., 2023; Dai et al., 2021; Jin et al., 2024). OPERA (Huang et al., 2024) traces the potential causes of hallucinations in LVLMs via attention maps, suggesting that during the inference phase, the model may produce hallucinations by sequentially summarizing previous tokens when processing special tokens such as '-', '?', and other special symbols. OPERA mitigates hallucinations by imposing penalty constraints on attention scores, marking the first work to visualize multimodal hallucinations. In addition, FastV (Chen et al., 2024a) identifies that the computation of attention for visual tokens in the deep layers (near the output) of the LVLMs is extremely inefficient. Although previous studies have made significant strides in enhancing LVLMs perception by leveraging attention-based approaches, they focus less on the dynamic interactions between images and texts. To this end, we aim to deepen the understanding of how images and texts influence each other within reasoning tasks. We define

[†]Work done during an internship at Alibaba Cloud Computing

[*]These authors contributed equally to this work

[✉]corresponding author

[‡]project lead

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2289–2299
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

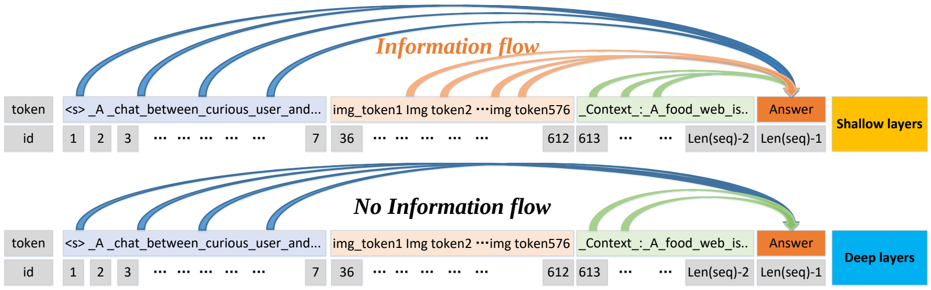

Figure 1: It shows the information flow of tokens, from left to right are system tokens, image tokens, user tokens, and output tokens. There is a convergence of the information flow of the system token, image token, and user token towards the output token at the shallow layers. The convergence of the information flow of the system token and user token is much more obvious than the image token at the deep layers, which we can call the deep layers as information flow cliff layers.

'information flow' as the degree of influence of image, user, and system tokens on answer tokens. Understanding the information flow between image and text tokens is crucial for dissecting multimodal reasoning.

In this paper, as shown in Figure 1, we discover an important phenomenon in a view of information flow: LVLMs tend to depend on the prompt, image tokens do not impact the answers in the middle or deep layers. We refer to this first sparse layer as the information flow cliff layer. As shown in Figure 2, we first visualize the attention scores and maps for system, image, and prompt tokens in relation to the answer. We observe that after the third layer, image tokens account for less than 2% of the total compared to system and prompt tokens, with their weights ranging from 0 to 0.2. This raises two questions:

Q1: *Is there any information flow from the third layer to the deeper layers?*

Q2: *Does the attention score give a complete reflection of the information flow of the image token?*

Though attention scores highlight relevant regions during forward propagation, they don't fully explain the role of image tokens as they lack gradients and can't reveal the model's decision-making process. To address this, we introduce LLaVA-CAM, which captures gradient changes through backward propagation to show how image features contribute to the answer. With LLaVA-CAM, we consistently observe that image token information flow is concentrated in the shallow layers due to differences in the prompt, while deeper image tokens contribute little to the answer. We refer to the deeper layer where the image ceases to contribute to the answer as the **'image information flow (i-f) cliff layer'**.

Given the information flow convergence in shallow layers, we find that the LLaVA-CAM results vary in distinct tasks. For example, in complex reasoning datasets like ScienceQA (Lu et al., 2022), the model focuses on relevant regions in shallow rather than deeper layers. In contrast, the model emphasizes relevant regions from shallow to deep layers for general reasoning tasks such as TextVQA (Singh et al., 2019) and POPE (Li et al., 2023). To validate this hypothesis, we conduct multiple truncation experiments. Empirical results show that when applying fully truncated image tokens based on attention scores, certain layers can maintain the model's inference accuracy and even enhance performance in some cases. Our multi-perspective analysis elucidates the model's inner mechanism and reveals the complex interactions during inference. Notably, we verify in the models LLaVA (Liu et al., 2024b), Intern-VL (Chen et al., 2024b), Qwen-VL (Bai et al., 2023), Shirak (Chen et al., 2023b), InstructbLIP2 (Dai et al., 2024) and LLaVA1.6 (Liu et al., 2024a) that the information flow converges in the shallow or middle layers, and becomes sparse in the deeper layers, and that there are information-flow cliff layers in each of the models, this means that image tokens are highly redundant after the cliff layer. Our methods offer a new perspective for interpretability in LVLMs. In summary, our main contributions are as follows:

- We propose a novel information flow analysis method, which combines LLaVA-CAM and attention score to explore the interaction mechanisms in LVLMs reasoning tasks.

- We find a prevailing pattern in multiple LVLMs: the information flow converges in shallow layers while diverging in deeper layers. This means that image tokens are highly redundant after the cliff layer.

- Results of truncation experiments validate our observation of information convergence in the corresponding layers across various models.

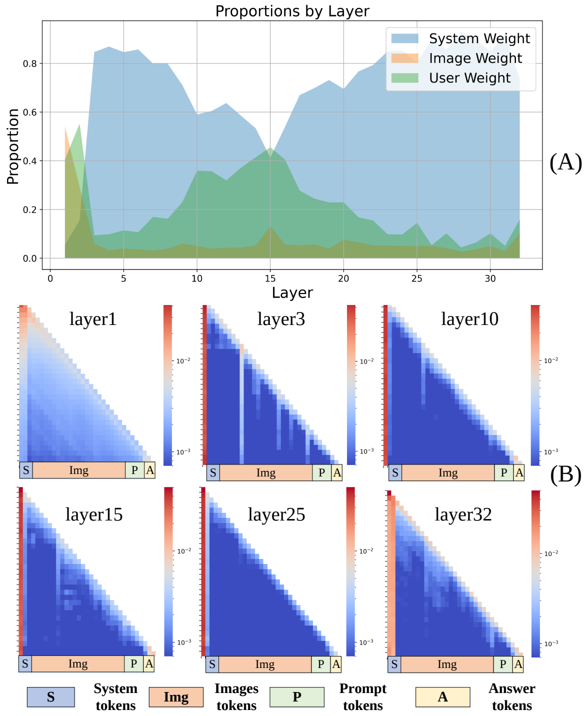

Figure 2: (A) is the percentage of system tokens, image tokens, and prompt tokens on the attention weight of the answer. (B) is the attention map of system tokens, image tokens, prompt tokens, and answer tokens. It can be observed that the attention scores for image tokens decrease rapidly in layers 1-5 and stabilize in layers 6-31. The attention allocated to image tokens is significantly lower throughout these layers than system and user tokens. However, image and user tokens' attention scores increase rapidly at the 32nd layer.

2 Related work

2.1 Information Flow and Interpretability in LVLMs

Recent work by Wang et al. (Wang et al., 2023a) has made significant strides in understanding information flow through the concept of label words as anchors. This approach focuses on how information aggregates, using a combination of attention scores and gradients to measure the saliency of specific tokens. In their attention-score matrix, the convergence of key tokens is evident along the ordinate axis. Label words will converge on the user prompt first in the Zero-shot ICL/CoT case. In their findings, Wang et al. observed that each shot will converge on the last token in the few-shot ICL/CoT case. This insight is crucial for analyzing truncation strategies in models for Visual Question Answering (VQA) and image captioning tasks. OPERA (Huang et al., 2024) introduces a novel method for identifying hallucinations in LVLMs by examining attention maps. It detects hallucinations during the sequential summarization of tokens following key symbols such as '-' and '?'. By imposing penalty constraints on attention scores, OPERA pioneers the visualization of multimodal hallucinations. Complementing this, DOPRA (Wei and Zhang, 2024) investigates the information flow of the output of each layer of the transformer.

Similarly, FastV (Chen et al., 2024a) identifies inefficiencies in attention mechanisms within LVLMs, particularly in models such as LLaVA-1.5 (Liu et al., 2024b), QwenVL (Bai et al., 2023). The authors observed that the computation of attention for visual tokens in the deep layers of these models is extremely inefficient. To address this, they proposed an image-token pruning strategy at specific layers, aiming for a sparser approach than textual data processing.

Although previous studies have made significant strides in enhancing LVLMs' perception and inference speed through leveraging attention-based mechanisms, they tend to focus less on the dynamic interplay of image-text interactions. Recognizing this gap, our research aims to deepen the understanding of how images and texts influence each other within complex reasoning.

3 The Proposed Method

To visualize the information flow, LLaVA-CAM, and attention score are used to provide a comprehensive view of the information flow. Attention scores illuminate the model's forward propagation, while LLaVA-CAM delves into the backward propagation, reflecting how various input elements contribute to the final prediction. This dual-method approach not only quantifies the significance of each input element but also provides a visual representation of their influence on the model's predictive decisions.

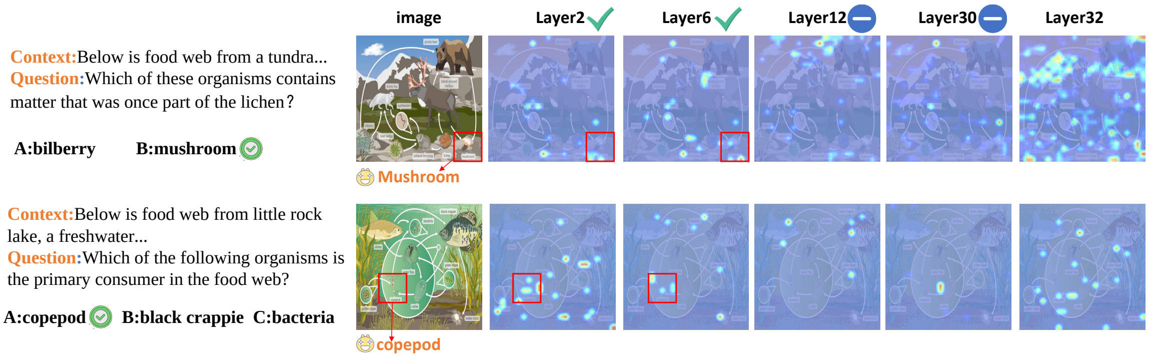

Figure 3: The LLaVA-CAM results of LLM on ScienceQA dataset(Complex reasoning). The information flow of the image converges to the correct region in the early layers and diverges in the deeper layers, and then the information flow cliff layer begins to appear.

3.1 LLaVA-CAM for Large Vision Language Models

Our method is inspired by Smooth-CAM (Omeiza et al., 2019). Complex reasoning is still essentially a text generation task, and the answer generation is a sentence made up of the classification results for each word. The network outputs probabilities for the n tokens denoted as $z = [z_i, ..z_n]$, to visualize the model's output answer using LLaVA-CAM:

$$z_{\text{answer}} = \sum_{i}^{n} z_i, \tag{1}$$

to obtain the logits representing the overall output of the model. The bias G_k is applied to all the feature maps A_k of the layer output by z:

$$G_k = \frac{\partial z_c}{\partial A_k}, \tag{2}$$

all attention mappings A_k in the last layer of the image encoder/LLM decoder, solve for z_{answer}, where A_k represents the feature map at the coordinate point of the n^{th} channel, α is the weight vector, and the resultant derivative feature map G, $c = 1, 2, 3..n$. Then transform the sequence inputs to the two-dimensional shape of the H×W. For sequence inputs, identify the sequence ID associated with image tokens. For instance, in LLaVA1.5, id (35-611) corresponds to image tokens. Then, multiply the α weight vector by the relevant channels of the feature map A to generate a two-dimensional activation map:

$$Heat_{\text{cam}} = \text{ReLU}(\sum_{k} \alpha_k A_k), \tag{3}$$

where the ReLU function creates a heat map highlighting the positive correlation between activation values and the correct class. Next, Gaussian noise is added to the image to generate multiple perturbation samples:

$$\mathbf{x}_{\text{noisy}}^{(i)} = \mathbf{x} + \mathbf{N}\left(\mu = 0, \sigma^2 = noise_s\right), \tag{4}$$

where $noise_s$ represents the standard deviation of the Gaussian noise added to the input image and i represents the first i generation of the image with noise. The CAM plots are averaged by generating the perturbation samples multiple times:

$$\text{LLaVA-CAM} = \frac{1}{N}\sum_{i=1}^{N} Heat_{\text{cam}}\left(\mathbf{x}_{\text{noisy}}^{(i)}\right). \tag{5}$$

Finally, we apply maximum normalization to the cam map. The heat map is then color-mapped and overlaid on the original image, providing a clear visual representation of the model's attention distribution.

3.2 LLaVA-CAM for Exploring Information Flow in LLM

The abundant features from the encoder are passed into the LLM decoder which interacts with the text. We use LLaVA-CAM to visualize the information flow of image tokens and their impact on answer tokens in the LLM decoder. As shown in Figure 3 and Figure 4.

Pattern: In tasks involving complex reasoning, common reasoning, or image captioning, information flow converges in the early or middle layers and starts to disperse in the deeper layers. It is plausible that the 32nd layer functions as the retrospective layer for deep reasoning, potentially integrating both the image and user prompt.

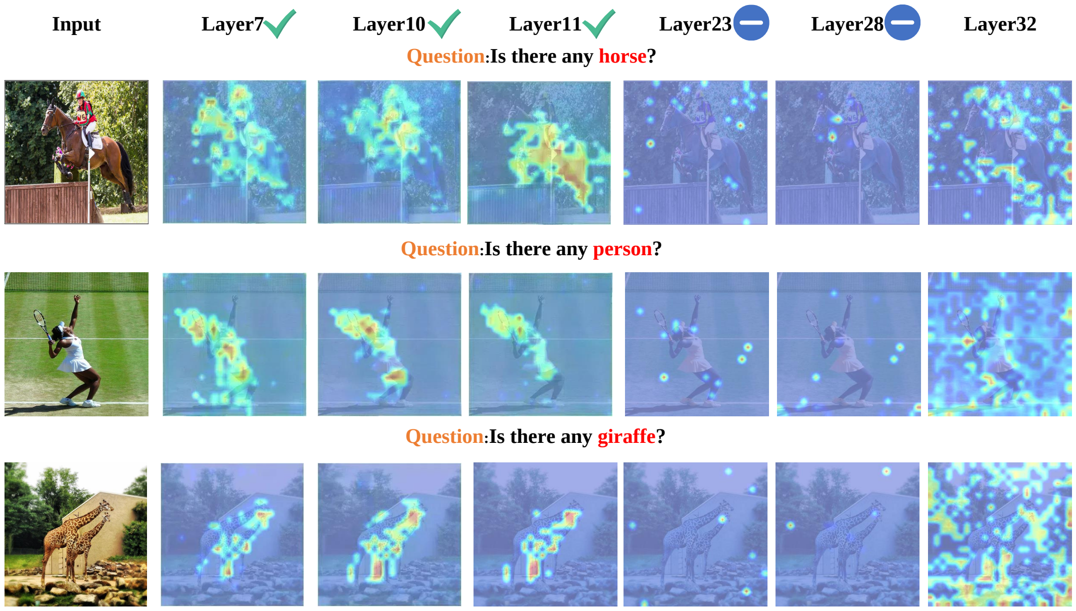

Figure 4: The LLaVA-CAM results of POPE (Li et al., 2023) and TextVQA (Singh et al., 2019) (Common reasoning). It can be analyzed from the LLaVA-CAM diagram of VQA that when the model recognizes the confirmed recognition objects such as "horse, people", etc., It will focus on the corresponding areas from the first layer to the deeper layers until a cliff layer occurs and causes information flow to be sparse.

(1) **Complex reasoning:** In the context of complex reasoning tasks like ScienceQA (Lu et al., 2022), when the prompt explicitly references specific regions within an image, such as those depicted in the first two lines of Figure 3 (*mushroom, copepod*), the model swiftly converges on the relevant areas in the early layers. The deeper layers appear to serve as the foundation for deep reasoning, with the complexity of the prompt dictating the layer at which information dispersal initiates, as illustrated in Figure 5, the information flow cliff layer appears at layer 12 (LLaVA1.5).

(2) **Common reasoning:** For common reasoning tasks such as TextVQA (Singh et al., 2019) and POPE (Li et al., 2023), the model consistently allocates attention to the key regions from the first layer to the deep layers. This consistency may suggest that fewer layers are necessary for deep reasoning, leading to the information flow diverging at deeper layers, specifically at the 18th and 24th layers for TextVQA and POPE, respectively, as indicated (A) in Figure 5. Furthermore, simpler tasks like image captioning, which require sustained attention to global image features, lead to a delayed dispersion of information flow and consequently require fewer layers for reasoning.

3.3 Attention-score for Exploring Information Flow in LLM

In the above section, we employ LLaVA-CAM to visualize the contribution of the image token to the output token. We found that the models converge in the shallow and middle layers and diverge in the deep layer. In this section, as shown in Figure 2, to validate the observed phenomenon of information flow convergence highlighted in LLaVA-CAM, we employ attention-score computation to help us understand the model's reasoning mechanisms, clarifying their contribution to the final output. We can denote the attention score of all the tokens in the output token as the influence rate, denoted by w. This measure can be aggregated for different types of input tokens. For the output token of the reasoning task, in the n-th layer, we define $\mathcal{G}$ as the indices set of all tokens and $\mathcal{G}$ can be divided into three parts that represent the indices set of system, image, and user tokens:

$$\mathcal{G} = \mathcal{S} + \mathcal{I} + \mathcal{U}, \tag{6}$$

where $\mathcal{S} = \{1, \ldots, N_{\text{sys}}\}$ represents the index of system token, N_{sys} represents the length of system token, $\mathcal{I} = \{N_{\text{sys}} + 1, \ldots, N_{\text{sys}} + N_{\text{img}}\}$ represents the index of image token, N_{img} represents the length of image token, and $\mathcal{U} = \{N_{\text{sys}} + N_{\text{img}} + 1, \ldots, N_{\text{sys}} + N_{\text{img}} + N_{\text{user}}\}$ represents the index

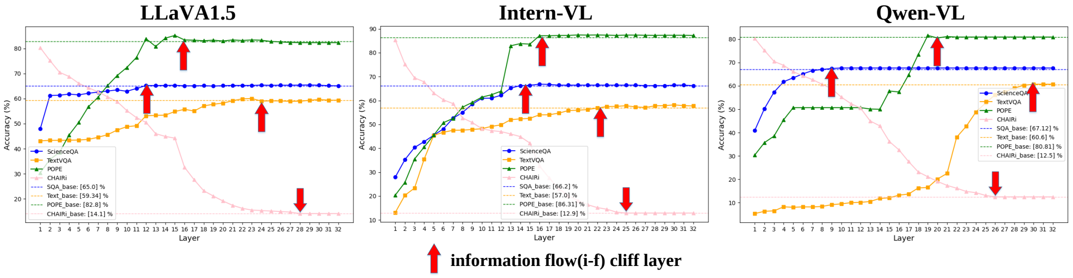

Figure 5: The truncating 576 image tokens experiments on three VQA datasets include POPE/TextVQA/ScienceQA/ and a caption dataset CHAIR dataset, where the red arrow represents the information flow cliff layer. LLaVA1.5-7B (Liu et al., 2024b), Intern-VL 7B (Chen et al., 2024b), and Qwen-VL 7B (Bai et al., 2023) all conform to the pattern of information flow convergence in the early layer and dispersion in the deep layer. Deeper layers can exhibit cliff layers, where truncating image tokens no longer affects the model's accuracy.

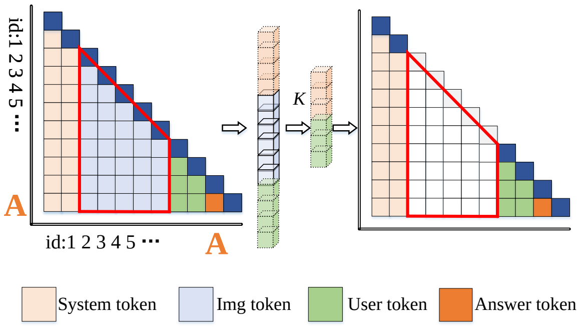

Figure 6: The diagram of attention-score truncation, A represents the answer token id index, K is set 0.

of user token, N_{user} represents the length of user token. $A_{i,j}$ is defined as the total attention score of the output token's attention on different types of tokens. For the i-th query token, the attentions from system, image, and user tokens are summed as 1:

$$\sum_{j \in \mathcal{S}} A_{i,j} + \sum_{j \in \mathcal{I}} A_{i,j} + \sum_{j \in \mathcal{U}} A_{i,j} = 1, \tag{7}$$

to ensure that the sum of attention scores for each token is 1, it is necessary to normalize the above summation results to calculate the total attention score for the image token:

$$\lambda_{\text{img}}^{j} = \sum_{j \in \mathcal{I}} A_{i,j}. \tag{8}$$

There are 576 image tokens in LLaVA1.5 and 256 image tokens in Qwen-VL (Bai et al., 2023). From Figure 2, we can see that image tokens account for most of the input tokens, but they receive significantly less attention; instead, the system prompt, which provides the least semantic information, attracts the most attention scores.

Pattern: Figure 2 and Figure 3 represent the attention score and LLaVA-CAM, respectively, and they show the same pattern, the information flow of image tokens converged in the shallow layers and dispersed in the deep layers. Notably, at layer 32, the attention score of the prompt and image tokens increases, and LLaVA-CAM shows that a large number of features contribute to this layer. This might indicate that layer 32 acts as a retrospective layer and the model potentially refocuses on the image and prompts before making its final decision.

3.4 Image Token Truncation by Attention-score

As discussed previously, our observations indicate that the information flow of image tokens converges at shallow layers and diverges at deeper layers. To verify whether the information flow diverges at deeper layers, we design a truncation strategy for image tokens as shown in Figure 6. In the attention matrix, each row represents the attention score of a query token to all key tokens. Let H denote the total number of attention heads, and let S denote the sequence length. The computation of the average attention matrix A across the different heads for the attention map O at the l_{th} layer is formulated as follows:

$$A = \frac{1}{H} \sum_{h=1}^{H} O_h. \tag{9}$$

From this matrix, we identify the indices of tokens with the highest attention scores within specific ranges, defining the corresponding attention matrix segments as:

$$A_{\text{img}} = A_{N_{\text{sys}}+N_{\text{img}},\mathcal{I}}, \tag{10}$$

where A_{img}represents the attention matrix segment for image tokens. Further, we define:

$$A_{i,\mathcal{K}} = [A_{i,k_1}, A_{i,k_2}, \cdots] \in \mathbb{R}^{|\mathcal{K}|}, \tag{11}$$

where $\mathcal{K} = \{k_1, k_2, \cdots\}$ and $A_{i,\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}|}$ represents the vector composed of elements whose row index is i and column indices belonging to the set $\mathcal{K}$. Here, $|\mathcal{K}|$ denotes the number of elements in the set $\mathcal{K}$.

To pinpoint the top $\mathcal{K}$ contributing elements, we utilize the function $\text{argtop}(\cdot, k)$, which retrieves the indices of the top $\mathcal{K}$ elements. To validate the information-flow cliff layers, here k is set to 0.

$$\mathcal{I}' = N_{\text{sys}} + \text{argtop}(A_{\text{img}}, k). \tag{12}$$

Finally, we combine the above formula with Eq. (6) to obtain:

$$\mathcal{G}' = \mathcal{S} + \mathcal{I}' + \mathcal{U}. \tag{13}$$

4 Experiment

4.1 Dataset and Implementation Details

The ScienceQA (Lu et al., 2022) is the Chain-of-thought dataset for complex reasoning, containing 21,208 multimodal questions (both visual and textual), along with corresponding answers, background knowledge (lectures), and explanations. POPE (Li et al., 2023) evaluates hallucination by having the model answer true/false questions about the presence of objects in an image (e.g., ‘Is there a car in the image?’). TextVQA (Singh et al., 2019) involves identifying text-related questions, detecting text areas, converting image text to text representations, focusing on relevant text areas, and determining if the final answer requires reprocessing. CHAIR (Rohrbach et al., 2018) metric is utilized to assess the phenomenon of object hallucination in the domain of image captioning. It quantifies the ratio of objects referenced in the caption that are not present in the actual image. This metric is available in two forms: CHAIR$_I$, which evaluates hallucination at the individual object instance level, and CHAIR$_S$, which does so at the sentence level.

Our experiments were conducted on an A100 GPU. It should be noted that we utilize *model* in the replication of the model, defaulting to greedy search to avoid interference from other parameters.

4.2 Truncation Experiments for Verifying Information Flow Cliff Layer

As shown in Figure 5, we introduce an early truncation strategy to investigate the information flow of image tokens in shallow layers. This strategy involves truncating all the image tokens to prevent the information flow within LLMs.

When handling complex reasoning tasks such as ScienceQA (Lu et al., 2022), which present lengthy and complex prompts with multiple choice options, the model’s image tokens tend to converge on relevant regions in shallow layers. As shown in Figure 5, starting from layer 12, the information flow begins to diverge. It seems that more layers might be necessary for deep reasoning in intricate tasks. This early divergence could also suggest a possible need for deeper processing in complex reasoning tasks. In contrast, for simpler tasks such as TextVQA and POPE, the model maintains its focus on the correct reasoning regions from the shallow to the deep layers. Divergence in the information flow is observed at layers 18 and 22, respectively, indicating that fewer layers are needed for deep reasoning in these tasks.

Compared to the ScienceQA and TextVQA tasks, image captioning is less complex, requiring the model to focus on the global features of the image. It seems that the requirement for deep reasoning layers is comparatively reduced, with the divergence of information flow occurring at later layers, such as around layer 30 in LLaVA1.5.

Therefore, we can conclude that a consistent pattern emerges in these different types of datasets: the information flow converges at shallow layers and diverges at deeper layers. One possible explanation is that the more complex the reasoning task is, the more layers of reasoning are needed, which means that the layer where the information flow diverges appears more forward. The 32nd layer seems to be like a retrospective layer; after multiple layers of thought and reasoning, the retrospective image and user prompt output the final answer.

4.3 Truncation Experiment in Different LVLMs

As shown in Table 1, to verify the generalization of the information flow convergence phenomenon across several models, we conduct experiments on Qwen-VL (Bai et al., 2023), LLaVA1.5 (Liu et al., 2024b), Intern-VL (Chen et al., 2024b), Shikra (Chen et al., 2023b), InstructBlip (Dai et al., 2024) and LLaVA1.6 (Liu et al., 2024a). Current LVLMs (Large Vision-Language Models) generally process images through a CLIP model, project them using various projectors, and then combine them with LLMs (Large Language Models) for fine-tuning.

Table 1: Generation study of truncation strategy on LLaVA (Liu et al., 2024b), Intern-VL (Chen et al., 2024b), Qwen (Bai et al., 2023), InstructBLIP2 (Dai et al., 2024) and shikra (Chen et al., 2023b). The experimental results prove that this phenomenon of information flow convergence in the shallow layer and dispersion in the deep layer has widely existed. The i-f cliff layer represents the information flow cliff layer.

Model	Dataset	Metric	Baseline	i-f cliff layer
LLaVA1.5	POPE	Acc ↑	84.70	**85.51**
	POPE	F1-score ↑	85.50	**85.99**
	SQA	Acc↑	65.00	**66.48**
	TextVQA	Acc ↑	59.34	**60.02**
	CHAIR	CHAIR$_s$ ↓	13.80	**13.80**
Qwen-VL	POPE	Acc ↑	80.81	**81.13**
	POPE	F1-score ↑	77.29	**77.82**
	SQA	Acc↑	67.12	**67.67**
	TextVQA	Acc ↑	60.60	**60.73**
	CHAIR	CHAIR$_i$ ↓	12.50	**12.50**
Instruct-Blip2	POPE	Acc ↑	84.85	**85.44**
	POPE	F1-score ↑	85.40	**85.40**
	SQA	Acc↑	60.50	**60.50**
	TextVQA	Acc ↑	50.10	**50.62**
	CHAIR	CHAIR$_s$ ↓	24.50	**22.80**
Shikra	POPE	Acc ↑	75.41	**76.34**
	POPE	F1-score ↑	78.52	**79.16**
	SQA	Acc↑	45.80	**46.70**
	TextVQA	Acc ↑	-	-
	CHAIR	CHAIR$_s$ ↓	12.50	12.50
Intern-VL	POPE	Acc ↑	94.01	**94.58**
	POPE	F1-score ↑	86.31	**87.21**
	SQA	Acc↑	66.20	**66.85**
	TextVQA	Acc ↑	57.00	**58.18**
	CHAIR	CHAIR$_s$ ↓	12.90	12.90

We speculate that the observed convergence of information flow in shallow layers is related to two factors: (1) The convergence is related to the multimodal paradigm. Different projectors, such as Linear, MLP, Cross-attention, and Q-former, map images to tokens in distinct ways. (2) This convergence is also related to the nature of LLMs. Since LVLMs are fundamentally LLMs, images are converted into tokens and aligned with text before being processed by the LLM. Thus, this pattern may be inherited from the underlying LLM structure.

To explore further, we conduct generalization experiments with 3 types of projectors. We select various models: LLaVA (Projector: MLP), Qwen-VL (Projector: Cross-attention), Intern-VL (Projector: MLP), InstructBlip (Projector: MLP), and Shikra (Projector: Linear). As shown in Table 1, all models exhibit a consistent pattern of shallow layer convergence and deep layer dispersion. Moreover, the performance after deep layer truncation is equal to or surpasses the baseline, suggesting that shallow layer convergence is independent of the projector type and is instead related to the LLM.

The SFT+LLM paradigm does not fundamentally alter the reasoning process of the LLM, as noted in references (Wang et al., 2023a; Clark et al., 2019; Hewitt and Manning, 2019), which may come from that shallow layers extract basic features and patterns from the input data, learning essential semantic information. The input takes the form of a 'Prompt + Question', where the information flow initially focuses on specific anchor tokens at a shallow layer. At these shallow layers, the model is processing semantic understanding, while at a deeper layer, it engages in more complex reasoning and thinking. This process is similar in Large Vision Language Models (LVLMs), where the prompt structure follows the format '<image> Prompt + Question'. At the early stage, the model focuses on understanding the image and the prompt, and the middle layers combine these basic features to form more complex representations, while the deep layers perform high-level modeling and reasoning, integrating information from earlier layers to generate the final understanding and decision of the task.

5 Conclusion

In this paper, we employ LLaVA-CAM and attention score to visualize the information flow of reasoning processes layer-by-layer within LVLMs. Our analysis of several models shows the same pattern that information flow converges in the shallow layers and diverges in the deep layers. Comprehensive experiments using truncation strategies across various large-vision language models for VQA and image captioning tasks further validate our findings, confirming a consistent pattern of information flow convergence in the relevant layers across different models. These insights contribute to a deeper understanding of LVLMs and their intrinsic functioning, particularly in the context of complex reasoning tasks.

6 Acknowledgments

This work was supported by the National Natural Science Foundation NO. 62273235, National Major Scientific Research Instrument Development Project (62227811), the Joint Fund of the Ministry of Education NO. 8091B022101, Deep Blue Program Fund Project, Second Institute of Oceanography, Ministry of Natural Resources. This work is supported in part by Alibaba Research Intern Program.

References

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

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *18th European Conference on Computer Vision ECCV 2024*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Bo Jin, Nuno Gonçalves, Leandro Cruz, Iurii Medvedev, Yuanyu Yu, and Jiujiang Wang. 2024. Simulated multimodal deep facial diagnosis. *Expert Systems with Applications*, 252:123881.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2023b. Cogvlm: Visual expert for large language models.

Jinfeng Wei and Xiaofeng Zhang. 2024. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. Prompt highlighter: Interactive control for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13215–13224.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. 2024. Llava-ϕ: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*.

A Appendix

A.1 LLaVA-CAM Structure

The LLaVA-CAM is shown in Figure 7, and more results are shown in Figure 8. The results show that LLaVA-CAM visualization is clearest after the post-attention layer normalization. This is because the post-attention-layer norm stabilizes the output of the self-attention mechanism by normalizing it. This normalization process helps stabilize the feature distribution, reducing the impact of numerical instability and making the features more consistent and reliable, which is crucial for generating high-quality heat maps. In contrast, feature maps from other positions may lack the same level of normalization or feature expression, resulting in lower-quality heatmaps and difficulty in accurately locating the target object. The feature maps after the attention mechanism contain rich contextual information.

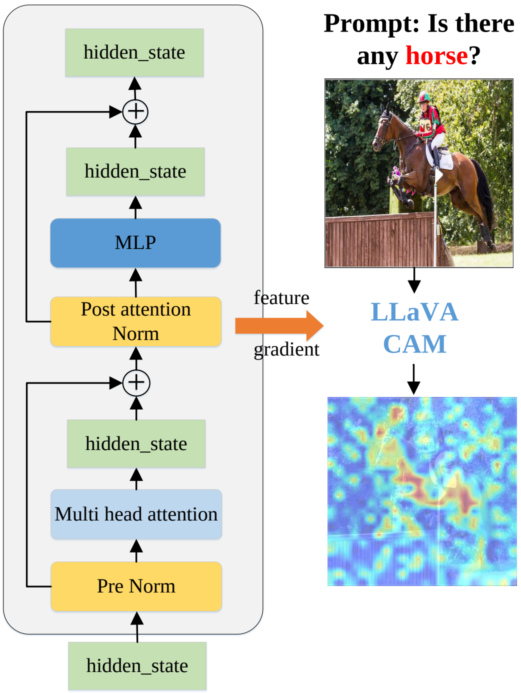

Figure 7: LLaVA-CAM selects features after the post-attention layer normalization to reduce numerical instability and produce clearer heatmaps.

A.2 Redundancy Truncation Experiment in Shallow Layers

To explore the saliency and information flow of image tokens in the shallow layer, we designed an early truncation strategy. We ranked the attention scores and selected the top K image tokens

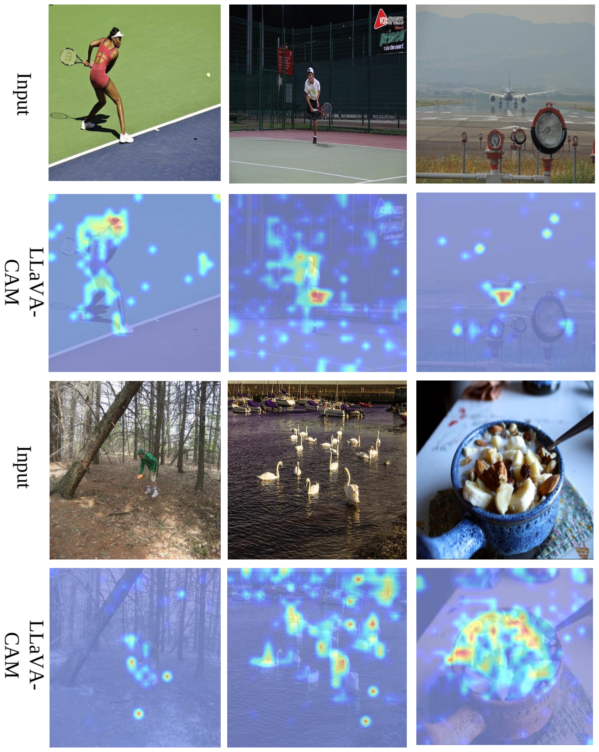

Figure 8: The LLaVA-CAM results.

to pass through, simplifying the inference process by focusing on these key tokens. We applied this attention-score truncation strategy with different values of K: 0.

When the layer=1, there is no image token passed to the LLM, while the LLM can surprisingly achieve an accuracy of 47.5 %, which suggests that in some scenarios the LLM relies only on the text to answer the questions on their own through prior knowledge, such as *'What is the capital of Nebraska?'* in Figure 9. Therefore, LLM does not attend to graphical information when LLM's knowledge can directly answer a question.

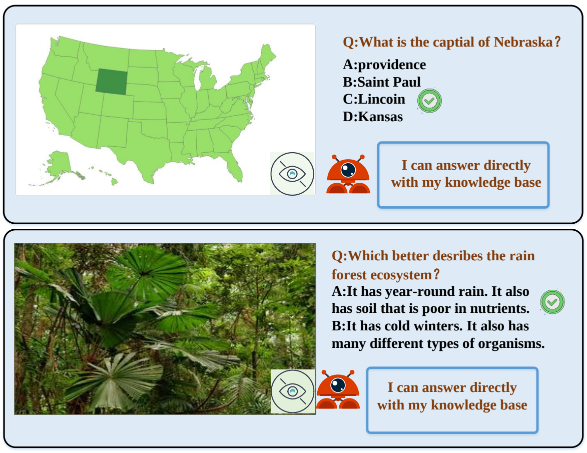
Q:What is the captial of Nebraska?
A:providence
B:Saint Paul
C:Lincoin
D:Kansas
I can answer directly with my knowledge base
Q:Which better desribes the rain forest ecosystem?
A:It has year-round rain. It also has soil that is poor in nutrients.
B:It has cold winters. It also has many different types of organisms.
I can answer directly with my knowledge base

Figure 9: It shows the case lacks visual dependency.