Shallow Focus, Deep Fixes: Enhancing Shallow Layers Vision Attention Sinks to Alleviate Hallucination in LVLMs

Xiaofeng Zhang[1,2]*†, Yihao Quan[4]*, Chen Shen[3]‡, Chaochen Gu[1,2]✉, Xiaosong Yuan[3], Shaotian Yan[3], Jiawei Cao[1,2], Hao Cheng[1,2], Kaijie Wu[1,2], 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]Rutgers University

Abstract

Multimodal large language models (MLLMs) demonstrate excellent abilities for understanding visual information, while the hallucination remains. Albeit image tokens constitute the majority of the MLLMs input, the relation between image tokens and hallucinations is still unexplored. In this paper, we analyze the attention score distribution of image tokens across layers and attention heads in models, revealing an intriguing but common phenomenon: *most hallucinations are closely linked to the attention sink patterns of image tokens attention matrix, where shallow layers exhibit dense sinks and deep layers exhibit the sparse*. We further explore the attention heads of different layers, finding: *heads with high-density attention sink of the image part act positively in mitigating hallucinations*. Inspired by these findings, we propose a training-free approach called **E**nhancing **V**ision **A**ttention **S**ink (EVAS) to facilitate the convergence of the image token attention sink within shallow layers. Specifically, EVAS identifies the attention heads that emerge as the densest visual sink in shallow layers and extracts its attention matrix, which is then broadcast to other heads of the same layer, thereby strengthing the layer's focus on the image itself. Extensive empirical results of various MLLMs illustrate the superior performance of the proposed EVAS, demonstrating its effectiveness and generality. The code can be accessed in https://github.com/itsqyh/Shallow-Focus-Deep-Fixes.

1 Introduction

Multimodal large language models (MLLMs) have significantly progressed in cross-modal tasks. However, hallucinations remain a challenging problem, particularly in visual question answering and image captioning. Although prior hallucination mitigating strategies such as incorporating external knowledge, retraining with additional data, or training-free methods (Yu et al., 2024; Sarkar et al., 2024; Xiao et al., 2024; Xing et al., 2024a; Ma et al., 2024; Gong et al., 2024; Chen et al., 2024a; Kim et al., 2024; Liu et al., 2024b; Zhou et al., 2023; Zhai et al., 2023; Wang et al., 2023a; Huang et al., 2023; Zhu et al., 2024; Jiang et al., 2024; Zhou et al., 2025; Bai et al., 2025; Suo et al., 2025; Lymperaiou et al., 2025; Wang et al., 2025; Li et al., 2025a; Chen et al., 2025b; Che et al., 2025; Chen et al., 2025a; Tu et al., 2025; Mao et al., 2025; Duan et al., 2025; Yin et al., 2025; Li et al., 2025b; Zhang et al., 2025, 2024c,b; Xue et al., 2025; Yang et al., 2025a,b; Zhuang et al., 2025; Zhang et al., 2024a) can work well in some scenarios, their interpretability is insufficient, especially lacking a clear

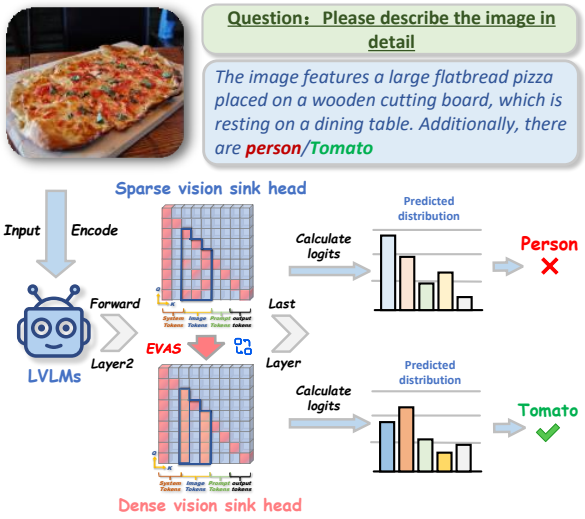

Figure 1: We found a common phenomenon through the attention map: In the range of image token, the attention head of shallow Sparse attention sink is prone to hallucination, while the attention head of Dense attention sink is much less likely to hallucinate.

*These authors contributed equally to this work
✉ corresponding author
†Work done during an internship at Alibaba Cloud Computing
‡project lead

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3512–3534
November 4-9, 2025

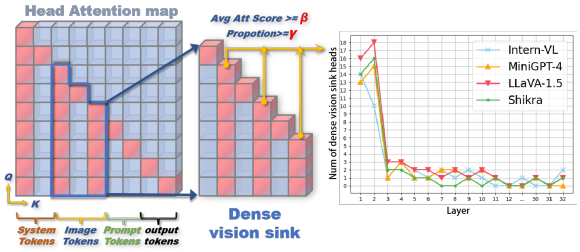

Figure 2: Definition of dense vision sink head and its layer-wise distribution. In this case, β = 0.0015, γ = 15%

explanation of the causes of hallucinations in the autoregressive model.

Current research on attention sink provides new insights for tackling hallucinations. The attention sink is an information flow as introduced in "Label Words are Anchors" (Wang et al., 2023b), depicting how the information flow often converges on a specific user token in LLMs. It's important to note that MLLM's output tokens are generated by the decoder based on logits, whereas input tokens, which constitute most of the input sequence, are more likely to directly reflect the MLLM's internal mechanisms. OPERA, DOPRA, TAME, and Vissink (Huang et al., 2024; Wei and Zhang, 2024; Tang et al., 2025; Kang et al., 2025) further explore the connection between attention sink in vision tokens and output tokens. In our investigation, we observe that when a token has a high attention weight across subsequent tokens, such over-reliance on the token can lead to hallucinations. Albeit these methods clarify the relationship among the attention sink, user tokens, and output tokens, the sparsity and hallucination of the deep and shallow vision attention sinks in the model remain unclear.

Finding 1: Most dense vision sink heads occur in or before layer 2: As aforementioned by FastV (Chen et al., 2024b), the information flow of image tokens is primarily concentrated in the first and second layers. Given this, we conduct experiments on several models and observe their shallow layers, including LLaVA1.5 (Liu et al., 2024a), Minigpt4 (Zhu et al., 2023), MiniGemini (Li et al., 2024d), and Intern-VL (Chen et al., 2024d). As shown in Figure 2, we calculate the average count of dense vision sink heads across these layers to further investigate the distribution of attention sinks across layers.

We define $h_{i,j}$ as the attention head, a "dense vision sink head" as a head (i, j) when the proportion $\alpha^{i,j}$ of columns in the attention map meets the

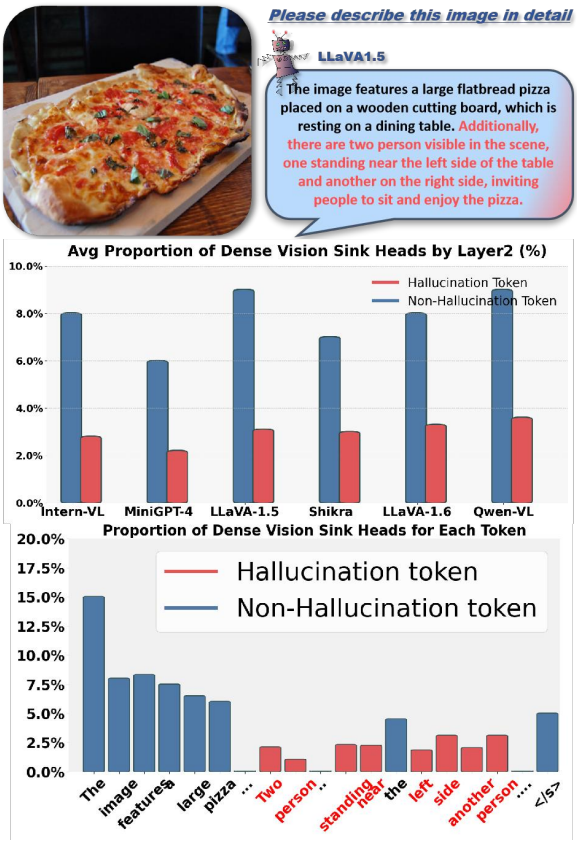

Figure 3: Relationship between text tokens and the average proportion of dense vision sink heads within a single layer by layer2, analyzed across 5,000 randomly selected MSCOCO images using LLaVA1.5-7B.

vision sink threshold γ:

$$visionsink = \frac{\sum_{x=k}^{r} h_{i,j}[x][y] \cdot M}{r-k} > \beta, \quad (1)$$

Where are $k \in [36, 611]$, M is an upper triangle mask matrix. Concretely, we define $\alpha^{i,j}$ as:

$$\alpha^{i,j} = \frac{\#visionsink}{576}, \quad (2)$$

A head is considered as a "dense vision sink head" when:

$$\alpha^{i,j} \geq \gamma. \quad (3)$$

Observations show that most vision attention sinks occur in the first 2 layers.

Finding 2: Fewer dense vision sink heads lead to hallucination output:

$$p = \frac{\#\text{ (dense vision sink heads)}}{32} \quad (4)$$

This proportion p quantifies how many of the total 32 heads are classified as "dense vision sink

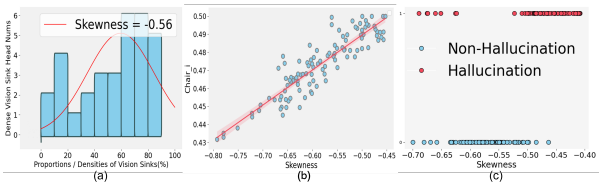

Figure 4: (a) A example of distribution of dense vision head and the corresponding proportions/densities of vision sinks within these heads when model output hallucination token; (b) Relationship between the average skewness and CHAIR_I on 150 randomly selected MSCOCO images, using LLaVA1.5-7B for captioning; (c) Comparative skewness scatter plot for Hallucination and Non-Hallucination classification on 150 randomly selected MSCOCO images, using LLaVA1.5-7B for VQA.

heads", i.e. they have a high proportion of columns that meet the vision sink condition within the image token range. We conduct the image captioning experiment on 5,000 randomly sampled MSCOCO images with LLaVA1.5-7B. When the model generates a new token, we first identify whether it is a hallucination token. Then, we backtrack to layer 2 and analyze the granularity of attention heads. We calculate the proportion of dense vision sink heads in layer 2 relative to the total number of heads (32 heads for 7B models). Such analysis is repeated for layer 1, and the average across layers 1 and 2 is computed later. As shown in Figure 3, we observe that non-hallucination tokens typically activate a larger number of dense vision sink heads, whereas hallucination tokens are generally associated with only a few dense vision sink heads, with the majority of heads being sparse. Through our analysis of different models, e.g., LLaVA1.5 (Liu et al., 2024a), Minigpt4 (Zhu et al., 2023), MiniGemini (Li et al., 2024d), Qwen-VL (Wang et al., 2024) and Intern-VL (Chen et al., 2024d), it suggests that fewer dense vision sinks heads lead to more probable hallucination output.

Finding 3: Lower density of vision sinks and fewer vision sink heads lead to a higher probability of hallucinations: However, the average count of dense vision sink heads across shallow layers does not reveal the individual contributions of each dense vision head, some of which may be negative while others are positive for hallucination. As noted by ITI (Li et al., 2024b), in current LLMs using transformer architecture, only a subset of attention heads plays a more significant role. Effectively optimizing these heads and leveraging them will likely lead to substantial improvements in model efficiency and overall performance. In this case, we conducted a more detailed view for each head, as shown in Figure 4 (a), the sink densities within different vision sink heads vary across the shallow layers (layer1-layer2), with an overall negatively skewed distribution. As shown in Figure 4 (b), for the image captioning task, the average skewness of the distribution of dense vision sink head and its corresponding vision sink densities in layer1 and layer2 is recorded each time a token is output. Once the output token is completed, the CHAIR_I for the entire output is calculated, and the average skewness for all tokens in layer1 and layer2 is obtained. As shown in Figure 4 (c), for the VQA task (with only a single output token), the average skewness of the distribution of vision sink head and its corresponding vision sink densities in layer1 and layer2 is directly recorded for the answer token. It is observed that, regardless of the task (image captioning or VQA), a lower skewness coefficient correlates with a lower hallucination rate. In other words, a higher density of vision sinks within a dense vision sink head and a larger number of vision sink heads lead to a lower probability of hallucination.

These observations highlight the critical role of attention head and vision sink distribution in understanding the attention sink phenomenon, particularly as it relates to alleviating hallucination issues in MLLMs. When the vision sink is sparse, visual tokens concentrate too heavily on specific elements, leading to reduced attention to other parts of the image. Conversely, a dense vision sink helps maintain a global perspective, preventing the model from narrowing its focus too much and minimizing information loss. Our goal is to ensure the model maintains a high-density vision sink within shallow layers. To achieve this, we design a training-free method called **E**nhancing **V**ision **A**ttention **S**inks (EVAS). This plug-and-play approach focuses on each attention head in the early layers, systematically identifying the head with the densest vision sinks. It then broadcasts this attention distribution across the layer, aligning the layer's attention and the head's vision sink distribution with that of the selected head.

We conduct extensive experiments, focusing specifically on hallucination issues, and test mainstream MLLMs to validate the effectiveness of EVAS in reducing hallucinations across various model architectures. Our results demonstrate that

EVAS is a highly effective plug-and-play solution for mitigating hallucinations across various MLLMs. Specifically, our contributions can be summarized as follows:

- This paper investigates how information flow relates to hallucinations in MLLMs. Our analysis reveals a consistent pattern where denser vision sinks and a larger number of vision sink heads in the shallow layers are associated with fewer hallucinations.

- We propose a plug-and-play training-free method called **E**nhancing **V**ision **A**ttention **S**inks (EVAS), which alleviates hallucinations by finding the head with the densest vision sink and broadcasting it to other heads.

- Experiments on multiple models validate the plug-and-play convenience and strong generalization of this method.

2 Related Work

2.1 Attention Sink and Information Flow

While the mechanisms of LLMs and MLLMs remain complex and not fully understood, several approaches focusing on information flow and attention sink patterns provide valuable insights into their operation and offer potential solutions to issues such as hallucinations and inefficiencies.

StreamingLLM (Xiao et al., 2023) first introduces the concept of attention sink. The authors observe an intriguing phenomenon: initial tokens, while seemingly less important for the overall content generation, consistently receive high attention scores. This is visualized in the attention map as columns with notably high attention scores, which is counterintuitive. Furthermore, because of the autoregressive nature of generative models, these initial tokens continue to receive more attention from subsequent tokens, amplifying their impact on the generation process. To address this, StreamingLLM leverages these attention-sink tokens during the pre-training phase to enhance the model's performance.

In the context of MLLMs, OPERA (Huang et al., 2024) introduces a novel perspective by linking the causes of hallucinations with attention sinks. This approach provides new insights into the interpretability of MLLMs. OPERA reveals that during the inference phase, the generation of key tokens

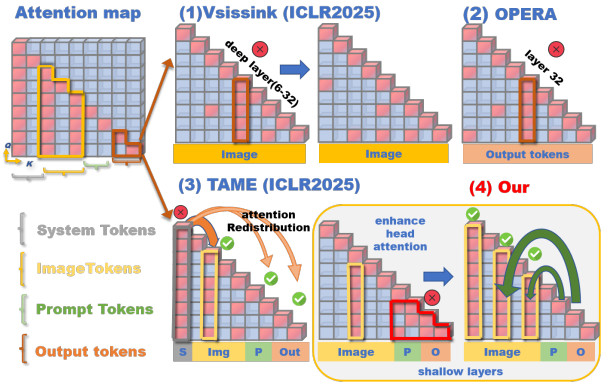

Figure 5: Differences in analysis perspectives between OPERA (Huang et al., 2024), TAME (Tang et al., 2025), Vissink (Kang et al., 2025) and our method.

such as '-', '?', or tokens that summarize previous ones can lead the model to produce hallucinated content. To address this issue, OPERA imposes penalty constraints on the attention scores of these summarization tokens. In light of this, DOPRA (Wei and Zhang, 2024) addresses the over-reliance by improving the strategy of weighted overlay penalties and redistribution in specific layers.

Difference between These Methods:

EVAS differs from existing methods while remaining non-conflicting and even complementary. Existing methods primarily adjust decoding strategies by modifying logits. For example, OPERA (Huang et al., 2024), DOPRA(Wei and Zhang, 2024) and MCA-LLaVA(Zhao et al., 2025) identify that anchor output token can lead to hallucinated token generation and try to penalize anchor tokens' logits. TAME(Tang et al., 2025) focuses on anchor token propagation in all layer, dynamically adjusting these anchor token. Vissink (Kang et al., 2025) found that the vision attention sink in the middle and deep layers converged on some <cls> or image-irrelevant tokens, which was attributed to the massive activation(Sun et al., 2024), so they redistributed the attention of these vision anchor tokens.

3 Method

3.1 Relationship between Vision Sink and Hallucinations

Popular VLMs, such as LLaVA-1.5 (Liu et al., 2024a), Minigemini (Li et al., 2024d), InstructBLIP (Dai et al., 2024), Shikra (Chen et al., 2023), MiniGPT-4 (Zhu et al., 2023), Qwen-VL (Bai et al., 2023), and InternVL (Chen et al., 2024d), consis-

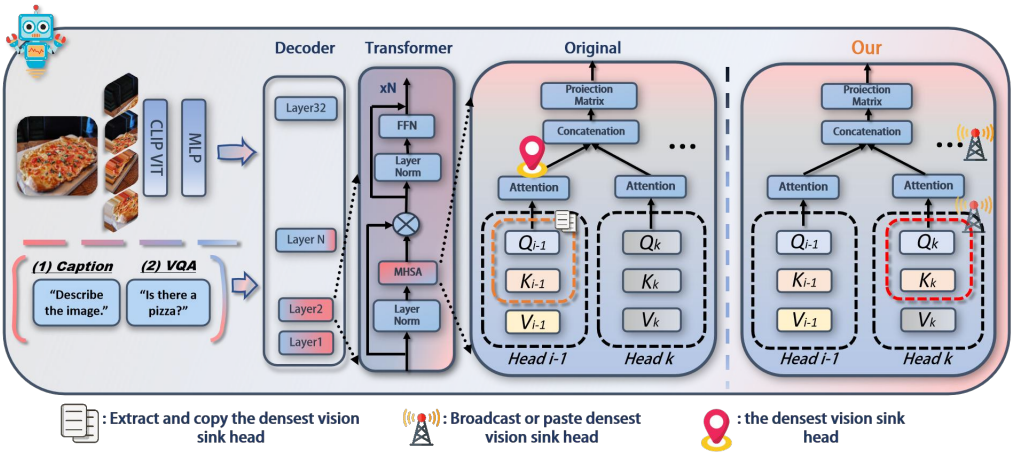

Figure 6: The structure of Enhancing Shallow Layers Vision Attention Sinks.

tently exhibit a notable pattern: vision sinks are densely concentrated within the first and second layers, gradually becoming more sparse in deeper layers. As illustrated in Figure 2, Figure 3, and Figure 4, we conclude that a lower density of vision sinks and fewer vision sink heads correlate with an increased likelihood of hallucinations in model outputs. We hypothesize that maintaining dense attention sinks in shallow layers may help alleviate hallucinations, as concentrated attention in the early layers enhances the transfer of image information to subsequent layers. Therefore, a practical method is proposed to reduce hallucinations by ensuring a dense vision sink of attention heads by layer1 and layer2. Please refer to the **Appendix** for more attention-map visualization results of different LVLMs.

3.2 Vision Sink

Definition of Mask Matrix M. To ignore diagonal elements in the attention map during calculations, we define a mask matrix M as follows:

$$M \leftarrow \text{eye}(r, c) - \text{diag}(1), \tag{5}$$

where $\text{eye}(r, c)$ generates an identity matrix of size (r, c), and we set the diagonal elements to zero.

Definition of Vision Sink. Let $h_{i,j}$ represent the attention map of the j-th head at the i-th layer, with $h_{i,j}[x][y]$ being the element at row x and column y. We define a "vision sink" as the column in the attention map within the image token range (e.g., $k \in [36, 611]$), where the average attention score of one element of the column within the image token range exceeds a threshold β.

For each column y, the "*vision sink*" condition is defined as:

$$\textit{vision sink} = \frac{\sum_{x=k}^{r} h_{i,j}[x][y] \cdot M}{r - k} > \beta, \tag{6}$$

where $k \in [36, 611]$, M is an upper triangular mask matrix.

Definition of Dense Vision Sink Head: For a head (i, j), we calculate the proportion of columns that meet the vision sink condition, denoted as $\alpha^{i,j}$. If this proportion of vision sinks within the range of image tokens (e.g., 576) exceeds a preset threshold γ, we classify the head as a "dense vision sink head." which is defined as follows:

$$\alpha^{i,j} = \frac{\text{Num}(\textit{vision sink})}{576} \geq \gamma. \tag{7}$$

where $\alpha^{i,j}$ represents the attention density score of the j-th head at the i-th layer. Num(vision sink) is the count of columns that satisfy the vision sink condition. γ is the threshold for determining whether a head qualifies as a "dense vision sink head."

If $\alpha^{i,j} \geq \gamma$, then the attention head (i, j) is identified as a "dense vision sink head."

3.3 Enhancing Attention Vision Sink

As mentioned above, we introduce a training-free, plug-and-play method called **E**nhancing **V**ision **A**ttention **S**inks (EVAS) to keep attention heads densely concentrated in the early layers. This approach identifies the attention head with the most dense attention sinks and broadcasts its attention map across other heads. This is to reinforce the

attention pattern of a particular head or to broadcast the attention pattern under certain specific conditions (e.g. when a predefined threshold is exceeded).

The algorithmic process is shown in Algorithm 1, let A be a 4D tensor, where $A[i][j]$ denotes the attention matrix of the j_{th} head of the i_{th} layer. Let β be the threshold, image-token-start-index be s, image-token-end-index be e and M be a mask matrix of the same size as $h_{i,j}$. We define $h_{i,j}$ as the attention-map of a head:

Initialization. Set the threshold β and initialize the variables: k as a randomly selected index within the range $[s, e]$, where $s = 36$ is the starting index of the image tokens and $e = 611$ is the end index; also initialize $n = 0$ and $H = [\quad]$ (an empty list).

Step1: Iteration over Heads. Select a specific attention layer i ($i \in \{0, 1, 2\}$), iterate on each head j ($j \in [0, 31]$).

Step2: Calculate Vision Sinks. For each column y in $h_{i,j}$, calculate whether its column is a vision sink based on:

$$\text{id}_{i,j} = \{(x, y) \mid \frac{\sum_{r=k}^{r} h_{i,j}[x, y] \cdot M}{r - k} > \beta\}, \quad (8)$$

where $\text{id}_{i,j}$ stores the indices (x, y) where the average attention score exceeds β.

Step3: Store Count of Vision Sinks for each Attention Head. Compute the count of marked indices for each head:

$$C_{i,j} = \text{count}(\text{id}_{i,j}), \quad (9)$$

then append $(C_{i,j}, j)$ to H:

$$H = H \cup \{(C_{i,j}, j)\}. \quad (10)$$

Step4: Update Head Index n with Maximum Vision Sinks. Find the index n of the head with the maximum count $C_{i,j}$ in H:

$$n = \arg \max_{(C_{i,j}, j) \in H} C_{i,j}. \quad (11)$$

This step dynamically updates n to track the head with the most vision sinks across the layer.

Step5: Enhance Attention Heads across the Layer. For each layer i, set the matrix of head j with the head with the highest number of vision sinks to be the n-th position in $A[i]$:

$$\text{for } j = 0, 1, \ldots, 31: \quad A[i][j] = A[i][n]. \quad (12)$$

Algorithm 1 Attention Process Calculation, $nonzero()$ represents an operation with a non-zero index, and $\sum$ represents a sum, $h_{i,j}$ represents the i layer, j head, k represents the index of token. x and y represent the rows and columns of $h_{i,j}$, respectively.

procedure ATTEN_PROCESS_CAL(A)
  Step 1: $threshold \leftarrow \beta, n \leftarrow 0$ $H \leftarrow [\quad]$
  Step 2: Loop over heads in a specific layer i, $i \in \{0, 1, 2\}$:
  for $j \in \{0, 1, 2, \ldots, 31\}$ **do**:
    $h_{i,j} \leftarrow A[i][j]$
    $s \leftarrow 36, e \leftarrow 611$
    Step 3: Calculate significant token indices:
      $M \leftarrow (\text{eye}(r, c) - \text{diag}(1))$
      $ids \leftarrow \text{nonzero}\left(\frac{\sum_{x=k}^{r} h_{i,j}[x,y] \cdot \text{M}}{r-k} > \beta\right)$
    Step 4: Store the count of vision sinks and its corresponding head index in H:
      $C_{i,j} \leftarrow \text{count}(\text{id}_{i,j}), H \leftarrow H \cup \{(C_{i,j}, j)\}$
    Step 5: Update head index n with maximum vision sinks:
      $n = \arg\max_{(C_{i,j}, j) \in H} C_{i,j}$
  end for
  for $j \in [0, 31]$ **do**
    $A[i][j] \leftarrow A[i][n]$
  end for
  return updated A
end procedure

4 Experiment

Baseline. To demonstrate the broad applicability of our method in LVLM architecture, we applied and evaluated the latest models, including LLaVA-v1.5/1.6 (Liu et al., 2024a), Qwen/2-VL (Wang et al., 2024), Intern-VL (Chen et al., 2024d), MiniGPT4 (Li et al., 2024d), Instructblip (Dai et al., 2024) and MiniGemini (Li et al., 2024d).

Evaluation Benchmarks. We conduct evaluations on image benchmarks. For image benchmarks, we assess three categories: (1) Comprehensive benchmarks (MMBench (Liu et al., 2024c), LLaVA$^{\text{W}}$ (Liu et al., 2024a), MM-Vet (Yu et al., 2023); (2) General VQA benchmarks (VizWiz (Gurari et al., 2018), SEED (Li et al., 2023a) and GQA (Hudson and Manning, 2019); (3) Hallucination benchmarks (POPE(Li et al., 2023b), CHAIR (Rohrbach et al., 2018)).

4.1 Evaluation Results

CHAIR and POPE Evaluations. EVAS on Hallucination BenchmarksIt is shown in Table 1, that the methods to mitigate hallucinations can be broadly classified into three groups. The first group includes OPERA (Huang et al., 2024), DOPRA (Wei and Zhang, 2024), VCD (Leng et al., 2024),

Table 1: **Compare results of SARA with other SOTA methods on POPE, CHAIR and MME datasets**. The best performances within each setting are **bolded**, baseline: LLaVA-1.5-7B. Please note that these results are all reproduced by us.

Method	Venue	POPE		CHAIR				MME				
		F1↑	Acc↑	C_S↓	C_I↓	Recall	length	Exist.↑	Count↑	Pos.↑	Color↑	Total↑
Beam Search	-	85.4	84.0	51.0	15.2	75.2	102.2	175.67	124.67	114.00	151.00	565.34
Dola (Chuang et al., 2023)	ICLR 2024	80.2	83.1	57.0	15.2	78.2	97.5	180.10	127.40	119.30	154.60	594.10
VCD (Leng et al., 2024)	CVPR 2024	85.3	85.0	51.0	14.9	77.2	101.9	184.66	137.33	128.67	153.00	603.66
OPERA (Huang et al., 2024)	CVPR 2024	84.2	85.2	47.0	14.6	78.5	95.3	180.67	133.33	111.67	123.33	549.00
DOPRA (Wei and Zhang, 2024)	MM 2024	84.6	84.3	46.3	13.8	78.2	96.1	185.67	138.33	120.67	133.00	577.67
HALC (Chen et al., 2024c)	ICML 2024	83.9	84.0	50.2	12.4	78.4	97.2	190.00	143.30	128.30	160.00	621.60
CCA-LLaVA (Xing et al., 2024b)	NIPS 2024	86.4	86.5	43.0	11.5	80.4	96.6	190.00	148.33	128.33	153.00	**641.66**
RITUAL (Woo et al., 2024)	Arxiv 2024	85.2	84.3	45.2	13.2	78.3	99.2	187.50	139.58	125.00	164.17	616.25
AGLA (An et al., 2024)	CVPR 2025	84.6	85.5	43.0	14.1	78.9	98.8	**195.00**	153.89	129.44	156.67	635.00
SID (Huo et al., 2025)	ICLR 2025	85.6	85.8	44.2	12.2	73.0	99.4	183.90	132.20	127.80	155.90	599.80
TAME (Tang et al., 2025)	ICLR 2025	85.4	85.7	41.3	12.2	74.4	98.8	183.00	137.33	129.00	154.67	604.00
Vissink (Kang et al., 2025)	ICLR 2025	**86.0**	**86.5**	52.4	14.5	79.1	113.0	190.00	138.33	148.33	155.00	631.33
EVAS	-	85.7	86.0	**36.4**	**9.9**	75.2	97.7	190.00	148.33	128.00	**160.33**	626.66

Table 2: Evaluation results of EVAS on general vision-language benchmarks, baseline: LLaVA1.5-7B, $Layer = 2$, $\beta = 0.002$.

Method	MM-Vet ↑	VizWiz↑	Seed↑	GQA↑	MMB↑
Baseline	31.1	50.1	57.6	62.0	64.2
VCD(Leng et al., 2024)	29.4	50.5	58.3	61.6	61.4
OPERA(Huang et al., 2024)	30.0	52.4	59.4	62.0	64.8
SID (Huo et al., 2025)	31.2	50.8	58.9	62.1	65.0
TAME (Tang et al., 2025)	30.5	51.6	59.4	61.7	65.3
Ours	**31.7**	**53.9**	**60.2**	**62.3**	**65.8**

Table 3: Generalization study of EVAS on other LVLMs models about CHAIR and POPE dataset, metrics are CHAIR_S, CHAIR_I and POPE-$F1$ $-$ $score$.

Model	**CHAIR_S**↓	**CHAIR_I**↓	POPE-$F1$↑
Qwen2-VL	25.0	7.3	86.6
+ EVAS	**23.1** (+1.9)	**6.2** (+1.1)	**87.4** (+0.8)
QwenVL-Chat	45.6	12.5	87.0
+ EVAS	**44.6** (+1.0)	**11.9** (+0.6)	**87.8** (+0.8)
MiniGPT-4	31.8	9.9	70.3
+ EVAS	**30.4** (+1.4)	**9.5** (+0.4)	**70.7** (+0.7)
Instructblip	58.8	23.7	84.4
+ EVAS	**56.0** (+2.8)	**15.7** (+8.0)	**85.2** (+0.6)
Shikra	55.8	15.4	82.5
+ EVAS	**47.9** (+7.9)	**13.7** (+1.7)	**83.5** (+1.0)
Mini-Gemini	32.6	8.7	85.6
+ EVAS	**27.8** (+4.8)	**8.5** (+0.2)	**86.8** (+1.2)
LLaVA1.5	47.0	13.8	84.9
+ EVAS	**36.4** (+10.6)	**9.9** (+3.8)	**85.7** (+0.8)
LLaVA1.6	42.6	14.4	86.5
+ EVAS	**34.3** (+8.3)	**10.2** (+4.2)	**87.5** (+1.0)
InternVL	45.8	12.9	86.4
+ EVAS	**32.4** (+13.4)	**9.0** (+3.9)	**87.8** (+1.4)

HACL (Chen et al., 2024c), RITUAL (Woo et al., 2024)and SID (Huo et al., 2025), which address hallucinations by altering the decoding process. The second group, represented by SFT methods such as CCA-LLaVA (Xing et al., 2024b), adjusts the logits of the end-of-sequence (EOS) symbol to control its positioning, allowing the model to terminate earlier, thus reducing hallucinations. The third group includes Vissink (Kang et al., 2025), TAME (Tang et al., 2025) and EVAS, which aim to adjust attention heads to enhance the truthfulness of the model's output during inference. Compared to Vissink (Kang et al., 2025) and TAME (Tang et al., 2025), EVAS's CHAIR performance is more prominent. TAME allocates the attention on the system token to other tokens, but still ignores the visual information, while Vissink only intervenes with the visual attention sink and ignores the contextual association of the text output.

MME and Other Benchmarks/Models Evaluations. It is shown that in Table 2 and Table 3, compared to the baseline model LLaVA1.5, our EVAS method achieves non-negligible gains on all benchmark datasets without introducing additional computation during inferencing. Such performance improvements highlight the potential of EVAS in enhancing LVLM's general visual perception capabilities.

In contemporary MLLMs, images are processed by a CLIP model, mapped through different projectors, and integrated with LLMs. We hypothesize that the convergence of information flow in the early layers is affected by how different projectors—such as Linear, MLP, Cross-attention, and Q-former—map images to tokens. As shown in Table 3, to test this hypothesis, we apply the EVAS method to various models. Notably, Shikra, LLaVA, Intern-VL, Qwen-VL, and Mini-Gemini use greedy search for decoding, while InstructBLIP uses beam search with a beam size of 5. Despite the different decoding strategies and projectors, all models exhibit a consistent pattern of dense attention sink in the shallow layers and sparse attention sink in the deeper layers. Applying EVAS to these models consistently improves performance, demon-

Table 4: Results for ablation study of the hyperparameter on CHAIR(Rohrbach et al., 2018) and POPE(Li et al., 2023b) dataset, Threshold: β, $\mathcal{N}$: broadcast top $\mathcal{N}$ head, baseline: LLaVA1.5-7B, metrics: CHAIR_S=51.0, CHAIR_I=15.2 and POPE-*F1-score=84.9*.

$\mathcal{N}$	β	Layer 1			Layer 2			Layer 3		
		$C_s\downarrow$	$C_I\downarrow$	**P-F1**↑	$C_s\downarrow$	$C_I\downarrow$	**P-F1**↑	$C_s\downarrow$	$C_I\downarrow$	**P-F1**↑
1	0..15	39.4	9.9	85.4	40.6	11.9	85.2	49.0	14.3	85.0
1	0..20	41.8	10.4	85.3	**36.4**	**9.9**	**85.7**	46.8	13.8	84.9
2	0..15	41.2	10.3	85.3	40.6	11.7	85.5	49.8	14.1	84.9
2	0..20	42.2	10.9	85.2	42.4	11.9	85.5	49.6	14.0	84.9
3	0..15	41.6	11.2	85.2	44.4	12.0	85.3	48.4	13.8	84.9
3	0..20	42.0	11.0	85.1	44.2	12.5	85.2	48.2	13.7	84.8

strating its effective plug-and-play capability and broad applicability. In the **Appendix**, we provide several attention maps for different LVLMs.

4.2 Ablation study

Effect of Hyper-parameter. Table 4 presents the ablation study results for the parameters Threshold: β, Layer: L, Top Head: $\mathcal{N}$. The experimental results indicate that the configuration with layer=2, β=0.002, and $\mathcal{N}$=top1, yields the best performance, achieving $\mathbf{C}_S$ of 36.6 and $\mathbf{C}_I$ of 9.9.

The improvement achieved by broadcasting the top attention head primarily benefits from the centralization of attention. Since most vision sinks are concentrated in the first and second layers, broadcasting the attention map of the head with the densest vision sink in these layers to the other heads helps unify each head's focus on visual information, forming a "consensus" attention pattern. Ultimately, the high-density vision sink pattern enables the model to capture key information from the image, effectively reducing hallucinations.

Effect of EVAS on Video Understanding Benchmark. As shown in the Table 5, We conducted validation experiments of the effectiveness of EVAS on LLaVA-onevision (Li et al., 2024a) and VideoLLaMA2 (Cheng et al., 2024). The video understanding and linguistic-related tasks including EgoSchema (Mangalam et al., 2023), MVbench (Li et al., 2024c) and VideoMME (Fu et al., 2024). The results demonstrate that EVAS not only mitigates hallucination and enhances factual consistency, but in turn amplifies these gains to produce significant improvements in all benchmark tests.

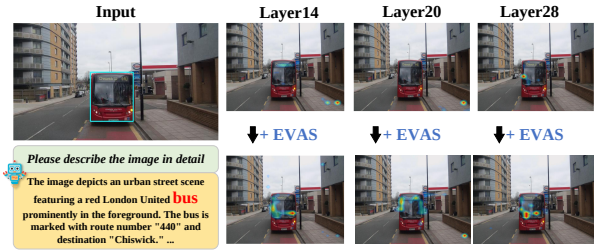

Figure 7: The attention-map results of LLaVA1.5 and LLaVA1.5 after adding EVAS, which is visualization of attention maps over image for 'bus'.

Table 5: Generalization study of EVAS on video understanding dataset including EgoSchema (Mangalam et al., 2023), MVBench (Li et al., 2024c) and VideoMME (Fu et al., 2024).

Model	EgoSchema↑	MVBench↑	VideoMME↑
LLaVA-onevision	60.1	56.7	58.29
+ EVAS	**63.9** (+3.8)	**59.9** (+3.2)	**61.16** (+2.87)
VideoLLaMA2	42.2	45.5	62.40
+ EVAS	**45.7** (+3.5)	**49.9**(+4.4)	**66.88** (+4.48)

4.3 Visualization of Attention Maps with EVAS

As shown in Figure 7, which is visualization of attention map with EVAS, We can find that some attention layers/heads that originally do not focus on the correct region will also gradually focus on the correct region when the EVAS is added. EVAS makes the original model pay more attention to the area of objects such as *"bus"* etc.

This result demonstrates that the EVAS method, which enhances attention heads in shallow/deep layers, improves the model's generalization ability. It enables the model to focus more on essential regions in the image, strengthening the flow of information and enhancing its overall capabilities.

5 Conclusion

In this paper, we introduce a method named **E**nhancing **V**ision **A**ttention **S**inks to alleviate hallucinations in LVLMs. EVAS is designed to enhance the densities and distribution of image token attention sinks in the shallow layers, thereby mitigating hallucinations. Our extensive benchmark tests on hallucination and generalization experiments demonstrate the effectiveness of EVAS as a training-free approach.

6 Acknowledgments

This work was supported by the National Natural Science Foundation NO. 62273235, National Ma-

jor 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.

7 Limitations

The results of this paper validate ITI's (Li et al., 2024b) conclusion that only a subset of attention heads plays a significantly more prominent role. Effectively optimizing these key attention heads is likely to yield substantial improvements in model efficiency and overall performance. To address hallucination issues more fundamentally, we believe that improved alignment of projectors and advanced training methods, such as RLHF, is necessary for more effective resolution.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Jiaqi Bai, Hongcheng Guo, Zhongyuan Peng, Jian Yang, Zhoujun Li, Mohan Li, and Zhihong Tian. 2025. Mitigating hallucinations in large vision-language models by adaptively constraining information flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23442–23450.

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*.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Eazy: Eliminating hallucinations in lvlms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2025a. Attention hijackers: Detect and disentangle attention hijacking in lvlms for hallucination mitigation. *arXiv preprint arXiv:2503.08216*.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025b. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. 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. 2024b. 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*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024d. 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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models.

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025a. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*.

Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, et al. 2025b. Mitigating object hallucinations in mllms via multi-frequency perturbations. *arXiv preprint arXiv:2503.14895*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024d. 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. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

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

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Maria Lymperaiou, Giorgos FIlandrianos, Angeliki Dimitriou, Athanasios Voulodimos, and Giorgos Stamou. 2025. Halcece: A framework for explainable hallucination detection through conceptual counterfactuals in image captioning. *arXiv preprint arXiv:2503.00436*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Shunqi Mao, Chaoyi Zhang, and Weidong Cai. 2025. Through the magnifying glass: Adaptive perception magnification for hallucination-free vlm decoding. *arXiv preprint arXiv:2503.10183*.

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

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. *arXiv preprint arXiv:2503.00361*.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

Chao Wang, Weiwei Fu, and Yang Zhou. 2025. Tpc: Cross-temporal prediction connection for vision-language model hallucination reduction. *arXiv preprint arXiv:2503.04457*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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*.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*.

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

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024a. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024b. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. 2025. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Haolin Yang, Feilong Tang, Ming Hu, Qingyu Yin, Yulong Li, Yexin Liu, Zelin Peng, Peng Gao, Junjun He, Zongyuan Ge, et al. 2025a. Scalingnoise: Scaling inference-time search for generating infinite videos. *arXiv preprint arXiv:2503.16400*.

Haolin Yang, Feilong Tang, Linxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Boqian Wang, Yifan Lu, Xiaofeng Zhang, et al. 2025b. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2503.13107*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halleswitch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024a. From redundancy to relevance: Information flow in lvlms across reasoning tasks. *arXiv preprint arXiv:2406.06579*.

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

Xiaofeng Zhang, Fanshuo Zeng, and Chaochen Gu. 2024c. Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation. *Neural Networks*, page 107059.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2025. Enhancing multimodal large language models complex reason via similarity computation. *AAAI*.

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Xu-Yao Zhang, and Da-Han Wang. 2025. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. *The 33rd ACM International Conference on Multimedia*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

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*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

Xinlin Zhuang, Feilong Tang, Haolin Yang, Ming Hu, Huifa Li, Haochen Xue, Yichen Li, Junjun He, Zongyuan Ge, Ying Qian, et al. 2025. Towards efficient medical reasoning with minimal fine-tuning data. *arXiv preprint arXiv:2508.01450*.

A Appendix

A.1 Generalization Study of EVAS on LLM Models

As mentioned above, we find a similar pattern in LLMs and LVLMs. To verify the feasibility of EVAS on LLMs. As shown in Table.6, we chose four models including LLaMA3.1-instruct (Dubey et al., 2024), Ministral-8B-Inst (Jiang et al., 2023), Qwen-2-7B-Inst (Yang et al., 2024) and Qwen-2.5-7B-Inst (Team, 2024). The datasets are GSM8K (Cobbe et al., 2021) and TruthfulQA (Lin et al., 2021), respectively. The results are shown in Table.6, which demonstrates that EVAS can produce consistency gains in LLM as well.

Model	Dataset	Metric	Baseline	w/ EVAS
Llama-3.1-8B-Inst(Dubey et al., 2024)	GSM8K	Acc ↑	85.29	**87.25**((+1.96)
	Truthful-QA	Acc ↑	49.27	**53.17**((+3.90)
Ministral-8B-Inst (Jiang et al., 2023)	GSM8K	Acc ↑	90.00	**91.36**((+1.36)
	Truthful-QA	Acc ↑	47.80	**51.22**((+3.42)
Qwen-2-7B-Inst (Yang et al., 2024)	GSM8K	Acc ↑	88.63	**89.22**((+0.59)
	Truthful-QA	Acc ↑	45.85	**46.34**((+0.49)
Qwen-2.5-7B-Inst (Team, 2024)	GSM8K	Acc ↑	92.72	**93.63**((+0.91)
	Truthful-QA	Acc ↑	52.68	**56.10**((+3.42)

Table 6: Generation study of EVAS on LLM models.

A.2 Why EVAS in Q/K before V

The reason for intervening before V is that the attention matrix attn_weights (i.e., attention_map) represents the weight distribution between different queries and keys. EVAS (Enhancing Attention Heads) specifically modifies these weights to adjust the attention distribution. If the intervention happens before V, it allows direct control over the attention concentration during the soft weight allocation stage, making attn_weights more focused on the relevant image information. The adjusted attn_weights, when multiplied with V, will more effectively filter out important information.

If the intervention occurs after V, the effect of EVAS will be limited to the final attn_output value, rather than modifying the attention matrix itself. This makes it harder to effectively control the attention on specific tokens.

Therefore, intervening at the attn_weights stage before V allows for a more direct impact on the model's focus on different tokens, thereby improving performance.

A.3 Comparison of Generation Time

In Figure 9, we compare the generation time of EVAS with existing methods for alleviating hallucinations. Both EVAS and OPERA (Huang et al., 2024) are methods that require attention intervention, and we utilize a standard self-attention implementation. In contrast, other methods such as Greedy, DoLA, VCD, and HALC (Chen et al., 2024c) do not necessitate attention intervention. All methods were tested on a single A100-80GB GPU. Our observations indicate that EVAS achieves a decoding time similar to that of VCD (Leng et al., 2024). It is slightly longer than the Greedy and DoLA (Chuang et al., 2023) methods due to our intervention in the attention weights at layer 1.2 during inference. In comparison, the other methods inevitably introduce additional computational overhead.

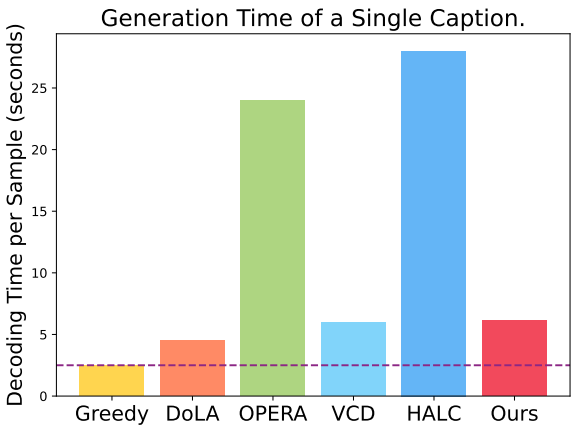

Figure 9: Generation time of a single response.

A.4 Experiment results on other hallucination benchmark

In our ablation study, we utilized the Hallusion-Bench dataset (Guan et al., 2023), which is specifically designed to evaluate hallucination phenomena in multimodal large models. Using LLaVA1.5 as the baseline, we incorporated our proposed Enhancing Attention Heads (EVAS) method. The results demonstrated a significant performance improvement after applying EVAS, particularly in reducing hallucinations. Compared to the baseline model, EVAS effectively concentrated the visual attention in shallow attention heads, enhancing the model's ability to capture relevant regions in images and thereby reducing the likelihood of hallucinations. This indicates that EVAS can significantly improve the robustness and reliability of multimodal reasoning tasks.

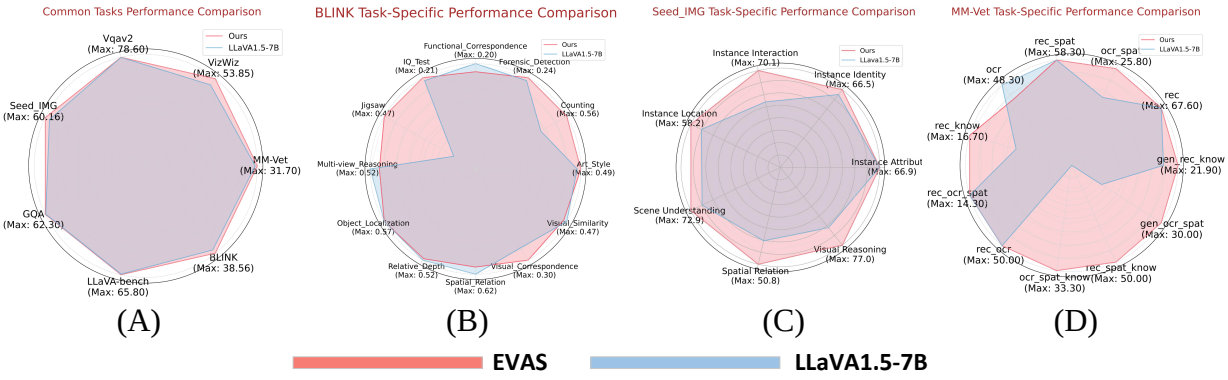

Figure 8: The generalization results on the datasets for the seven multimodal models (MMEs) demonstrate that EVAS combined with LLaVA1.5 enhances the metrics compared to LLaVA1.5 alone. This not only highlights the exceptional performance of EVAS in addressing hallucinations but also reinforces the notion that EVAS can generally enhance the model's overall performance.

HallusionBench (Guan et al., 2023)				
split	**method**	**aAcc ↑**	**fAcc ↑**	**qAcc↑**
Overall	LLaVA1.5-7B	35.54154	17.63006	11.20879
	w/ EVAS	44.37434	17.91908	14.50549

A.5 Qualitative Experiment of Thresholds and Layers

Table 7 presents the results of our ablation experiments, which assess the impact of various thresholds and layers on model performance. We test different thresholds (0.0006, 0.0008, 0.0015, 0.002) and layers (1, 2, 3, 4, 16, 32) to observe their effects on the model's performance with the CHAIR dataset.

The results show that the model achieves its highest performance on the CHAIR_s and CHAIR_I metrics, with scores of 36.6 and 10.0, respectively, when applying EVAS at the second layer with a threshold of 0.002. However, both metrics significantly decrease as the number of layers increases. This supports our hypothesis that information flow converges in early layers and diverges in deeper layers, and keeping the attention sink dense in shallow layers will effectively alleviate the hallucination. As the depth increases, both CHAIR_s and CHAIR_I values rise and exceed the baseline, suggesting that the likelihood of the model generating hallucinations at deeper layers increases. This occurs because attention sinks become more sparse in deeper layers and the differences between attention heads diminish. Therefore, even if the most densely concentrated attention-sink head is identified and broadcasted to other heads, its impact may still be limited.

Table 7: Qualitative experiment of layer and threshold on CHAIR dataset (baseline: LLaVA-1.5-7B, head=top1).

Layer	Threshold	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑	Avg. Len
1	0.0006	46.4	13.9	76.9	96.1
1	0.0008	46.0	12.9	77.1	101.2
1	0.0015	39.4	**9.9**	72.2	108.6
1	0.002	41.8	10.4	72.9	112.4
2	0.0006	41.4	12.3	76.4	95.6
2	0.0008	43.0	11.6	74.9	100.9
2	0.0015	40.6	11.9	74.9	102.7
2	0.002	**36.4**	9.9	73.9	97.7
3	0.0006	49.0	14.3	78.0	98.6
3	0.0008	49.4	14.3	78.3	98.4
3	0.0015	49.0	14.3	77.9	98.5
3	0.002	46.8	13.8	**78.5**	98.7
4	0.0006	46.4	13.6	76.9	96.1
4	0.0008	50.0	14.7	78.3	97.6
4	0.0012	49.6	14.6	78.4	97.7
4	0.002	49.4	14.5	78.2	97.5
16	0.0006	53.0	14.8	77.9	100.9
16	0.0008	49.2	14.8	77.4	100.5
16	0.0015	52.6	15.0	78.0	100.9
16	0.002	47.2	14.1	77.3	98.7
32	0.0006	50.8	14.4	78.1	98.7
32	0.008	50.8	14.4	78.1	98.7
32	0.0015	47.0	13.8	76.9	95.8
32	0.002	53.0	14.8	77.9	100.9

A.6 Qualitative Experiment of Heads Number

As demonstrated in the previous section, the first two layers contain the most attention sinks. Therefore, we focus on applying the EVAS strategy to these layers. In Table 8, we present a qualitative experiment to assess the impact of increasing the number of attention heads affected. We test this by broadcasting the densest attention head across 4, 8, 16, and 32 heads. For instance, when broadcasting to 4 heads, the attention map from the densest head is duplicated across these 4 heads, while the

Table 8: Qualitative experiment of enhancing head number on CHAIR dataset (baseline: LLaVA-1.5-7B).

Layer	Copy Head	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Avg. Len
1	4	51.0	14.7	78.1	98.9
1	8	49.4	14.5	**78.5**	99.5
1	16	48.4	13.5	77.1	99.8
1	28	40.6	11.8	74.1	103.8
1	32	39.4	**9.9**	72.2	108.6
2	4	50.8	14.4	78.1	98.7
2	8	49.6	13.7	77.6	100.0
2	16	47.4	14.0	77.5	100.5
2	28	42.2	11.9	74.9	98.1
2	32	**36.4**	9.9	73.9	97.7

remaining 28 heads remain unchanged.

The results indicate that broadcasting the densest attention head to 32 heads achieves the best performance. This suggests that using the densest attention pattern from the early layers improves the model's focus on image information, enabling the model to concentrate on global image information rather than allowing attention to converge on specific tokens. This approach significantly helps to alleviate hallucinations. More attention map and results are shown in Figure10, Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, Figure 19.

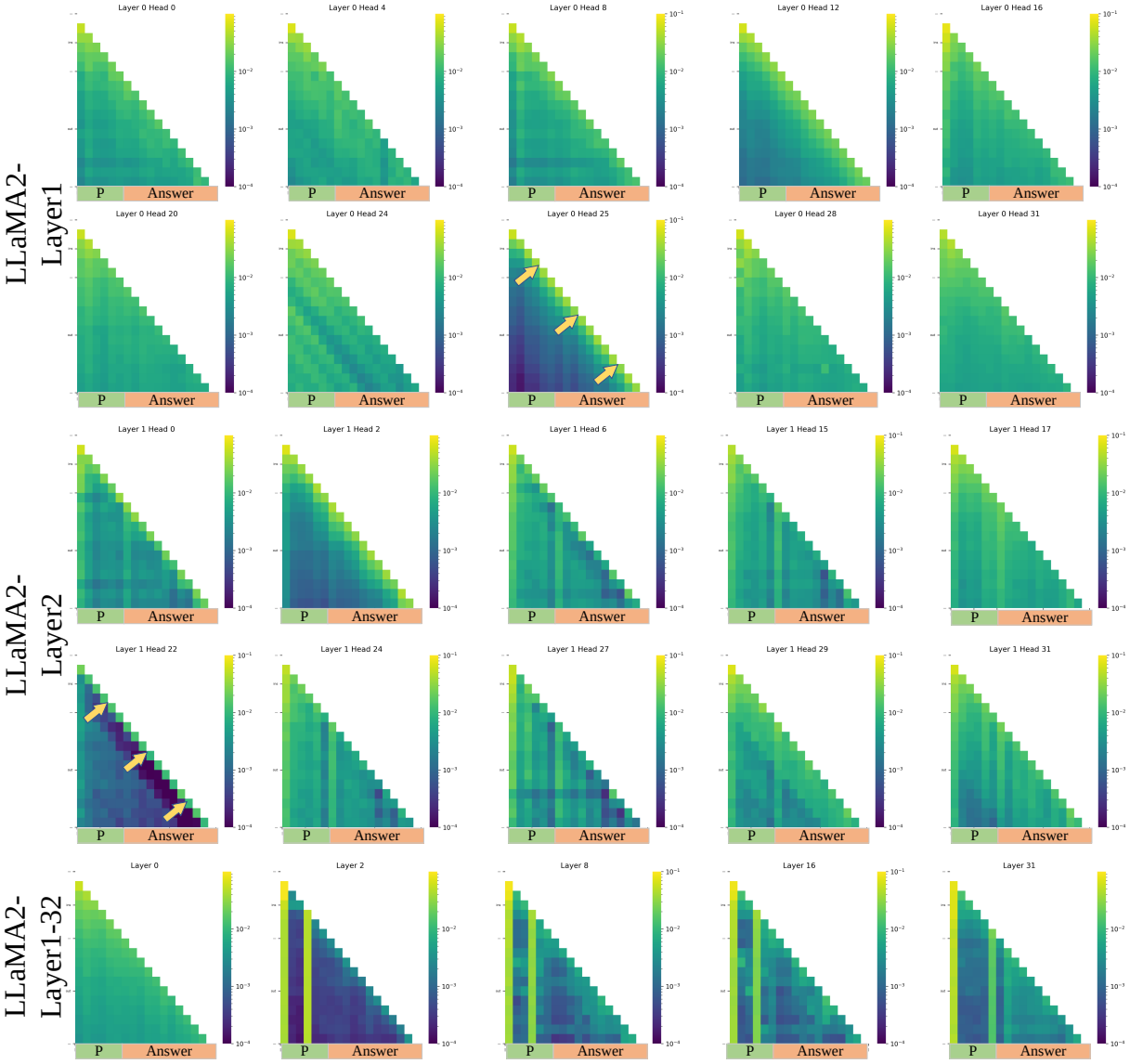

Figure 10: The attention map of LLaMA2.

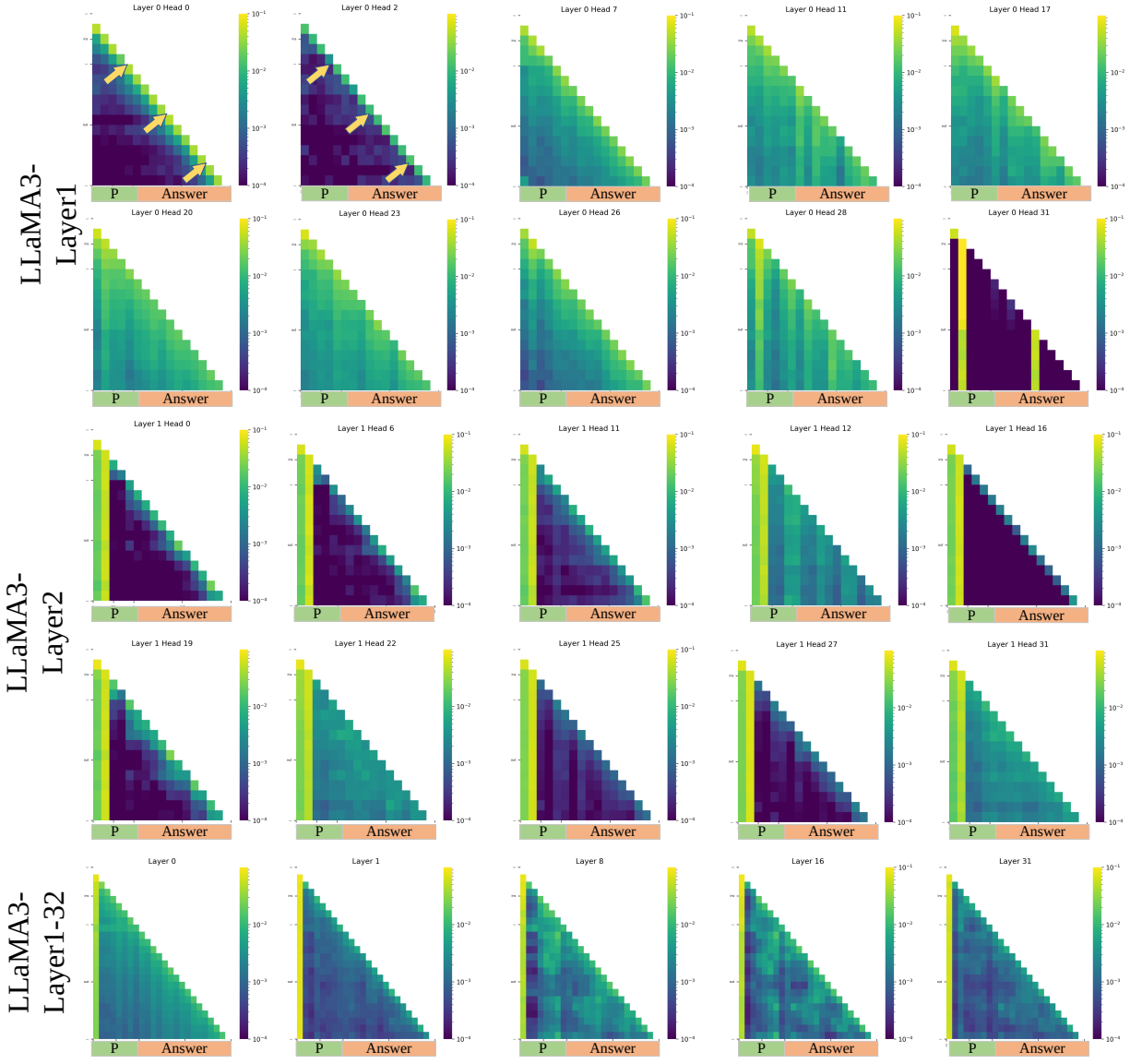

Figure 11: The attention map of LLaMA3.

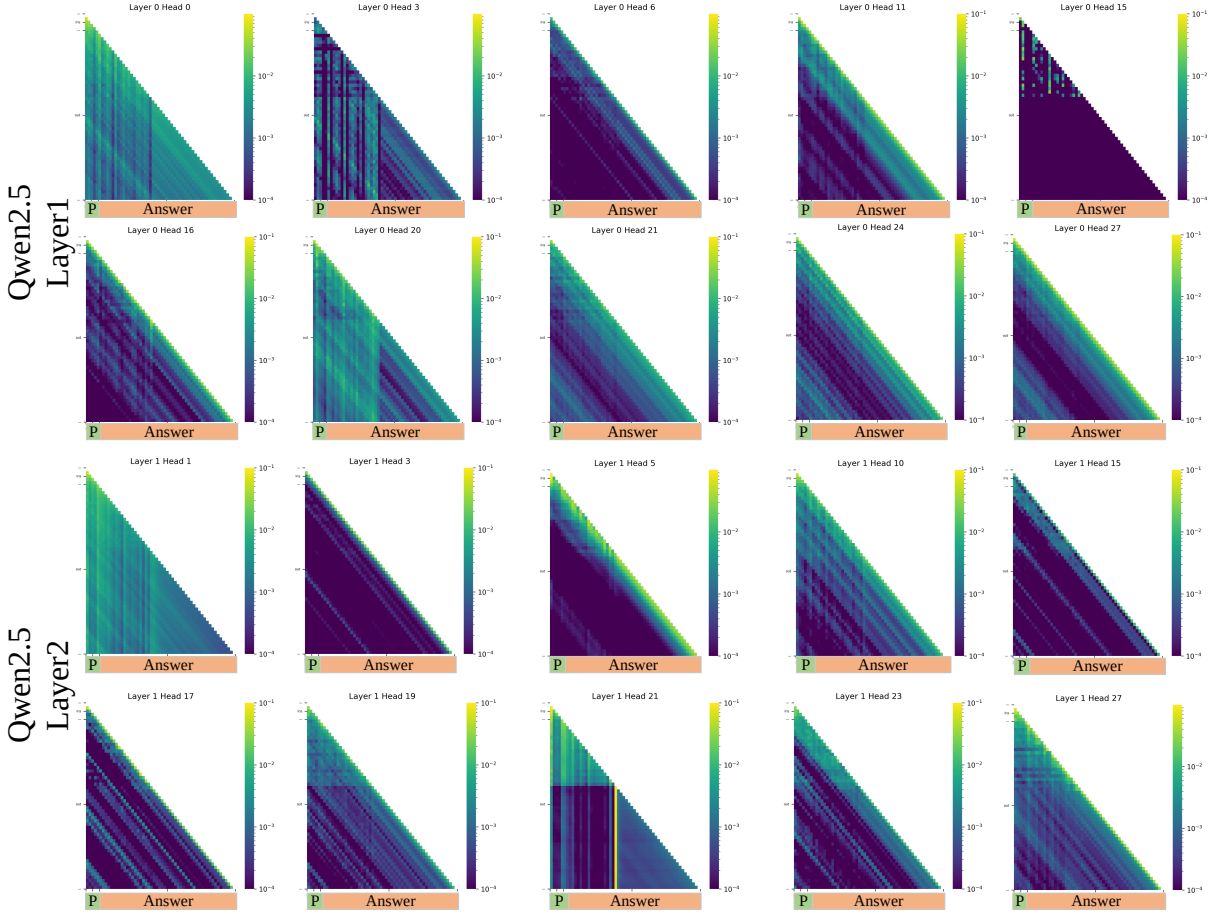

Figure 12: The attention map of Qwen2.5.

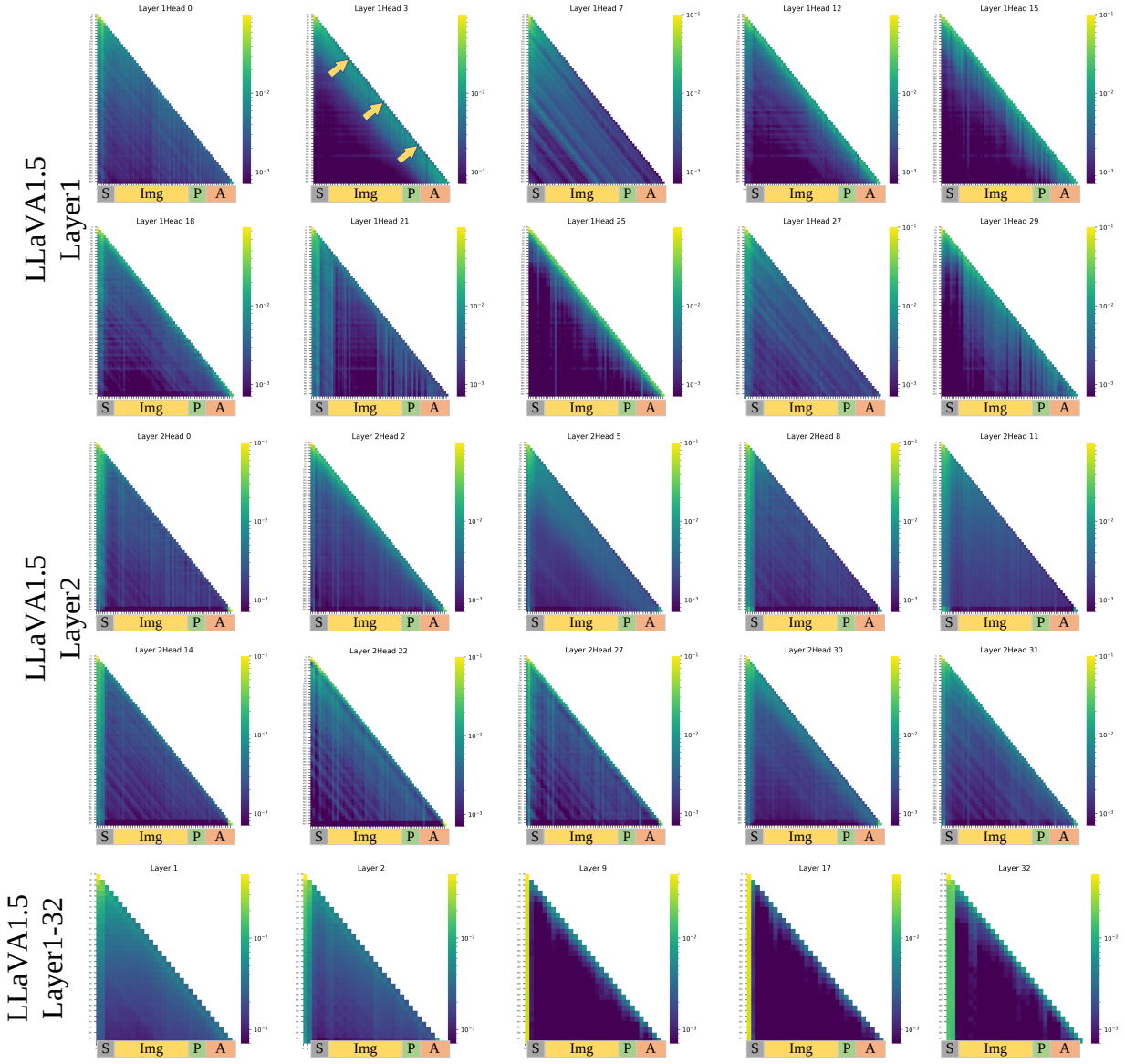

Figure 13: The attention map of LLaVA1.5.

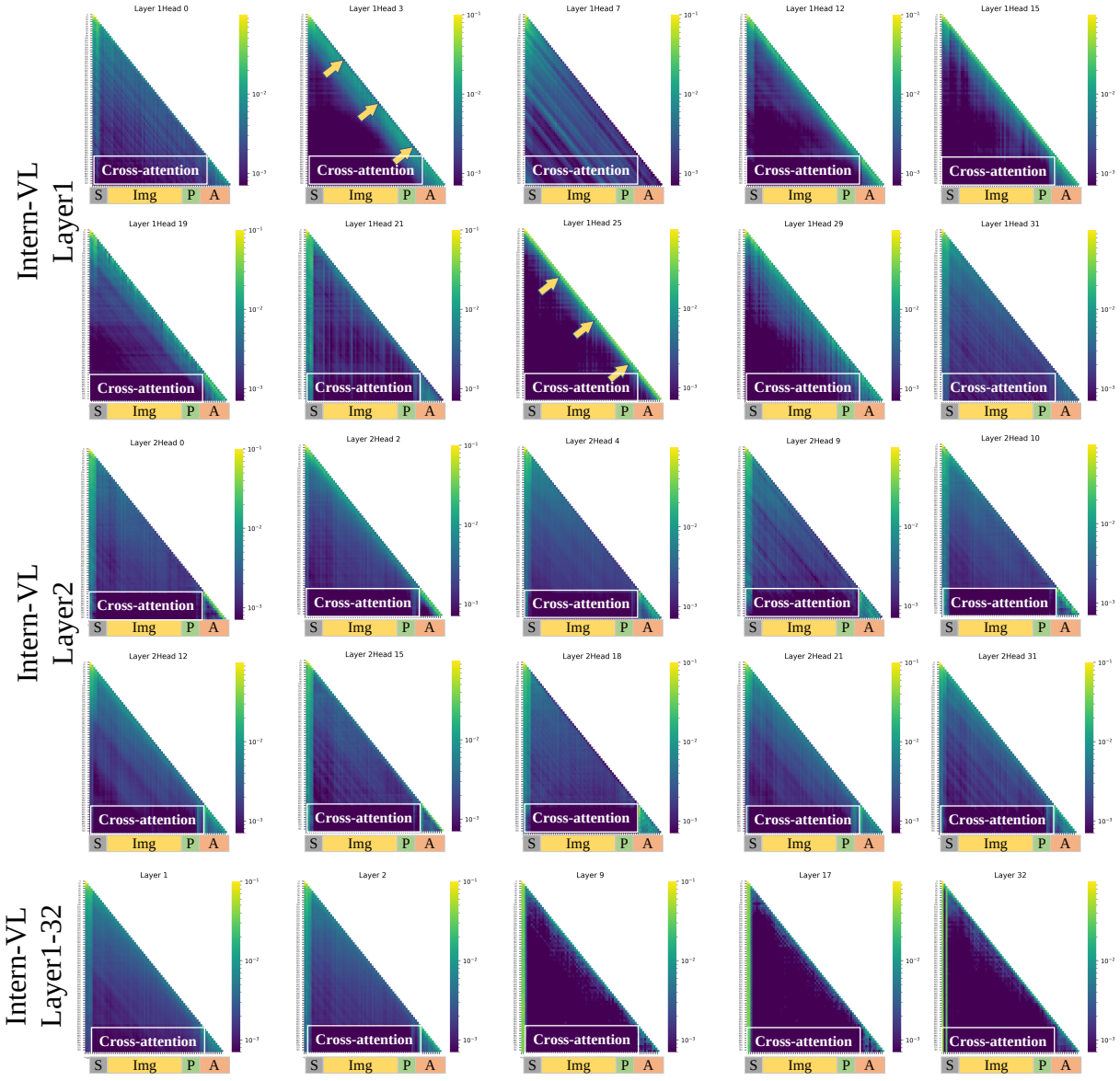

Figure 14: The attention map of Intern-VL.

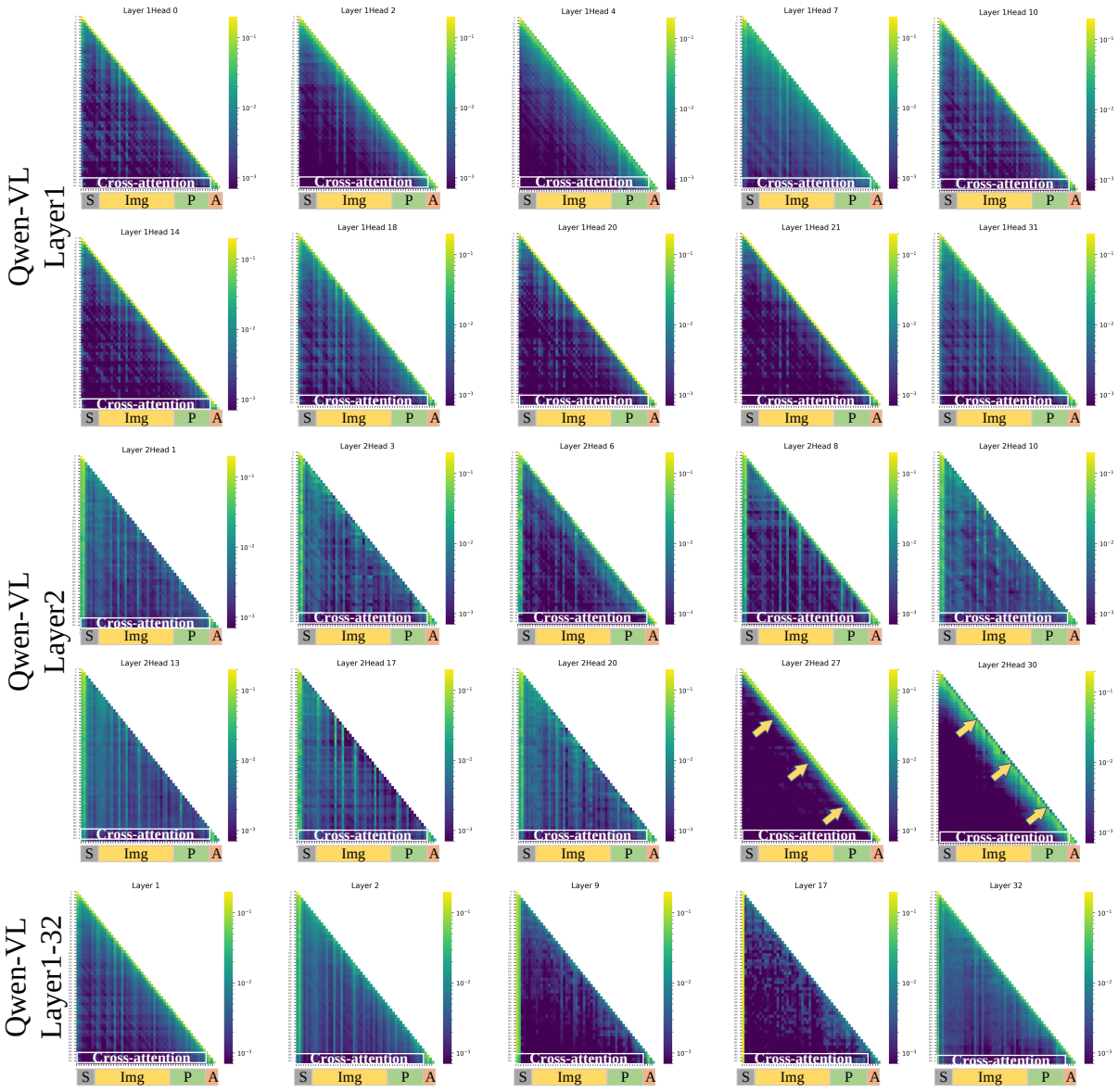

Figure 15: The attention map of Qwen-VL.

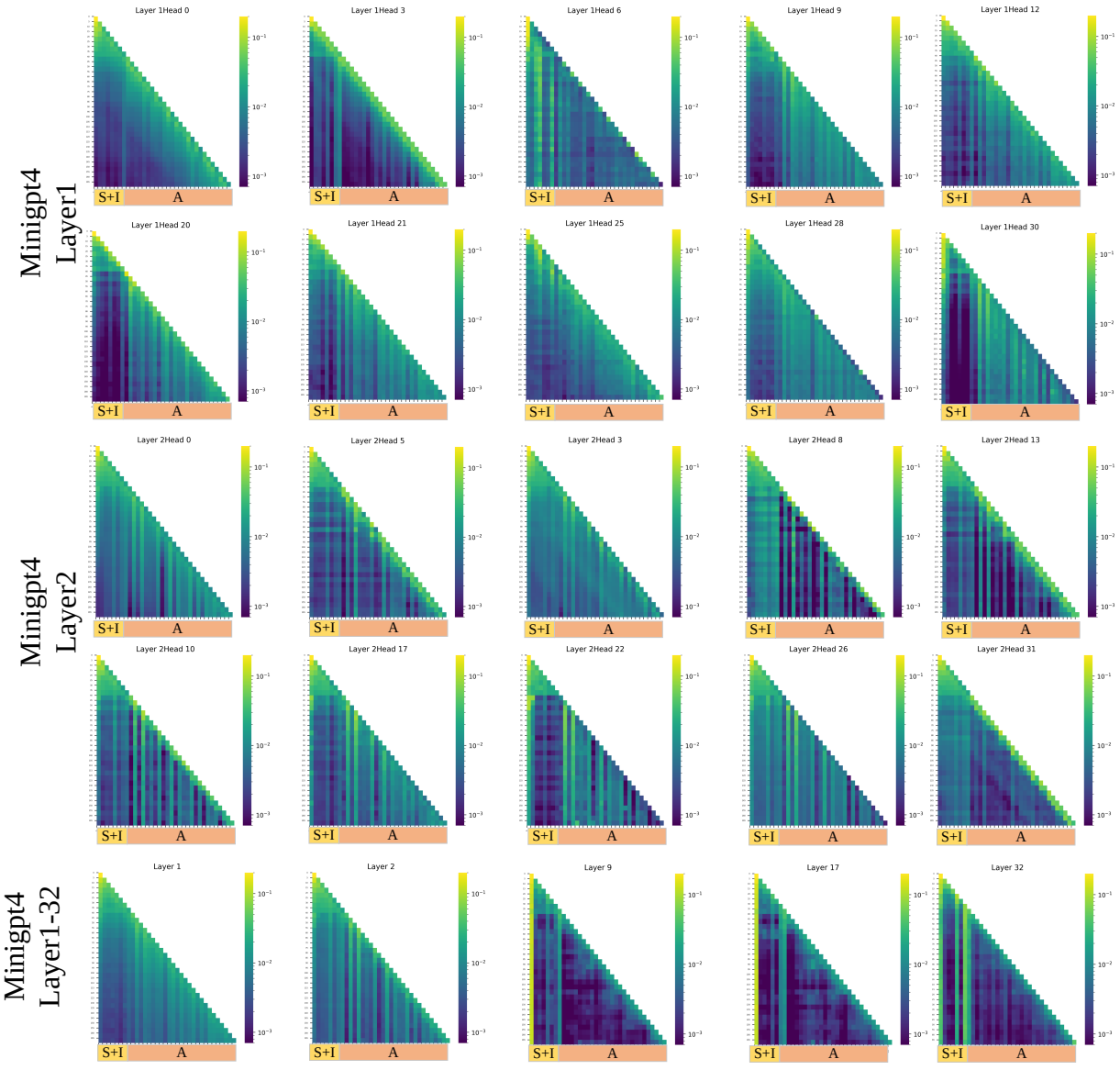

Figure 16: The attention map of Mini-GPT4.

Input image

LLaVA1.5-7B

The image features a bathroom with a sink and a mirror above it. The sink is white, and the mirror is large, covering most of the wall above it. The bathroom has a pink color scheme, giving it a warm and inviting atmosphere.\n\nThere are several items placed around the sink, including a toothbrush, a cup, and a bottle. Above the sink, there are two pictures hanging on the wall, adding a decorative touch to the space. Additionally, there are two towels hanging on the wall, one on the left side and the other on the right side of the mirror.

LLaVA1.5-7B+EAH

The image features a bathroom with a sink and a mirror. The sink is white, and the mirror is placed above it. The bathroom is decorated with a pink color scheme, giving it a warm and inviting atmosphere. There is a picture hanging on the wall above the sink, adding a personal touch to the space.\n\nIn addition to the sink and mirror, there are several items placed around the bathroom. A bottle can be seen on the counter, and a cup is also present. A toothbrush is located near the sink, indicating that the bathroom is well-used and functional. The overall appearance of the bathroom is clean and well-maintained, making it a comfortable and pleasant space for daily routines

Figure 17: Results of LLaVA1.5 with EVAS, EVAS can significantly reduce hallucinations while maintaining the original sentence length.

Input image

InternVL-7B

The image features a young girl riding a white horse near a river. The girl is wearing a red shirt and is seated on the horse, which is positioned close to the water. The scene appears to be a peaceful and enjoyable outdoor activity for the girl.\n\nIn the background, there are several boats visible on the water, adding to the scenic atmosphere. Additionally, there are two other people in the scene, one standing near the left side of the image and another person closer to the right side.

InternVL-7B+EAH

The image features a young girl riding a white horse along a path near a river. The girl is wearing a red shirt and is seated on the horse, enjoying the ride. The horse is positioned on the left side of the image, and the girl is sitting on its back.\n\nIn the background, there are several boats visible on the river, adding to the scenic atmosphere. The scene appears to be a pleasant outdoor activity for the girl and her horse.

Figure 18: Results of Intern-VL with EVAS, EVAS can significantly reduce hallucinations while maintaining the original sentence length.

Input image

Shikra-7B

The image features a large flatbread pizza placed on a wooden cutting board, which is resting on a dining table. The pizza is cut into square slices and is generously topped with cheese and a variety of herbs. Additionally, there are fresh basil leaves scattered across the pizza, adding a touch of greenery to the dish.\n\nThere are two chairs visible in the scene, one on the left side of the table and another on the right side, inviting people to sit and enjoy the pizza. A person can be seen in the background, partially cut off by the left edge of the image.

Shikra-7B+EAH

The image features a large, rectangular pizza placed on a wooden cutting board. The pizza is topped with tomato sauce and cheese, and it appears to be freshly baked. There are several slices of the pizza visible, indicating that it has been cut into smaller portions. The pizza is placed on a dining table, suggesting that it is ready to be served or enjoyed as a meal

Figure 19: Results of Shikra with EVAS, EVAS can significantly reduce hallucinations while maintaining the original sentence length.