Tell Model Where to Look: Mitigating Hallucinations in MLLMs by Vision-Guided Attention

Supplementary Material

A. Detailed Experimental Setup

Evaluated MLLMs. We evaluate the effectiveness of our proposed method, Vision-Guided Attention (VGA), on three representative MLLMs: **LLaVA-1.5** [16], **LLaVA-Next** [17], and **Qwen2.5-VL** [1]. LLaVA-1.5 is available in two sizes: 7B and 13B. Both LLaVA-Next and Qwen2.5-VL are 8B-scale

A.1. Benchmarks

In order to validate the effectiveness of our method in mitigating hallucination, we evaluate its performance on three widely-used multimodal hallucination benchmarks: one generative benchmark (CHAIR [22]), one discriminative benchmark (POPE [13]), and one hybrid benchmark (AMBER [25]).

CHAIR. CHAIR evaluates the proportion of hallucinated objects, which are generated by the model but not present in the reference annotations. Following prior works, we randomly select 500 images from the MSCOCO [15] dataset as the test set. We additionally select another 500 images to construct a validation set for hyperparameter tuning. This benchmark includes two metrics: CHAIRs and CHAIRi, defined as follows:

$$\begin{aligned} \text{CHAIRs} &= \frac{|\text{Hallucinated Objects}|}{|\text{All Objects}|}, \\ \text{CHAIRi} &= \frac{|\text{Hallucinated Captions}|}{|\text{All Captions}|} \end{aligned} \tag{16}$$

POPE. POPE is a widely adopted benchmark for evaluating object hallucinations by prompting LVLMs to identify whether a specific object is present in the image. It comprises three distinct datasets: *MSCOCO* [15], *A-OKVQA* [24], and *GQA* [8]. Each dataset uses three different negative sampling settings: *Random*, *Popular*, and *Adversarial*. Each subset includes 3,000 questions and 500 images. Accuracy and F1 score are used as the primary evaluation metrics.

AMBER. AMBER combines generative and discriminative tasks, and is evaluated on a curated set of 1,004 images. In addition to image captioning, it includes 14,216 questions designed to assess hallucinations in object, attribute, and relation recognition. AMBER contians mulitple metrics: *CHAIR*, *Cover*, *Hal*, *Cog*. It provides an annotated objects list $A_{obj} = obj_1^A, obj_2^A, \cdots, obj_n^A$, and the generated objects are labeled as R'_{obj}. Each metric is calculated as follows:

$$\begin{aligned} \text{CHAIR} &= 1 - \frac{\text{len}(R'_{obj} \cap A_{obj})}{\text{len}(R'_{obj})}, \\ \text{Cover} &= \frac{\text{len}(R'_{obj} \cap A_{obj})}{\text{len}(A_{obj})}, \\ \text{Hal} &= \frac{\{\text{CHAIR} > 0\}}{\{\text{All Caps}\}}, \\ \text{Cog} &= \frac{\text{len}(R'_{obj} \cap H_{obj})}{\text{len}(R'_{obj})}, \end{aligned} \tag{17}$$

where H_{obj} denotes the set of hallucinated target objects generated by the LVLMs, and *All Caps* refers to all generated captions.

A.2. Baselines

We select three existing visual attention-based dehallucination methods for comparative analysis: [2]

- **PAI** [18]: This method directly amplifies the model's attention weights on visual tokens and further enhances visual features using contrastive decoding. To facilitate a fair comparison of visual attention optimization, we denote the variants with and without contrastive decoding as PAI_{CD} and PAI, respectively. We set the visual attention augmentation coefficient α to 0.5 and apply visual attention augmentation starting from the layer specified by our method. PAI_{CD} has a hyperparameter $\gamma = 1.1$ and employs adaptive plausibility constraints with $\beta = 0.1$.
- **VAF** [30]: Similar to PAI, VAF directly scales up visual attention; however, it also suppresses attention on instruction tokens. We set $\beta = 0.1$, $\alpha = 0.15$, and activate this method from layer 9 to layer 14 (counting from 0).
- **TARAC** [28]: This method quantifies the importance of visual tokens based on attention distributions from previous generation steps and utilizes this information to strengthen the model's focus on relevant visual tokens at the current step. We set $\beta = 0.5$, $\alpha = 0.5$, and activate this method from layer 9 to layer 15 (counting from 0).

A.3. Implementation Details

Following the established practice of PAI, we determine the starting layer for VGA based on the *BOS* attention, which

[2] Note: The same parameters may have different meanings across methods. For specific details, please refer to the original papers.

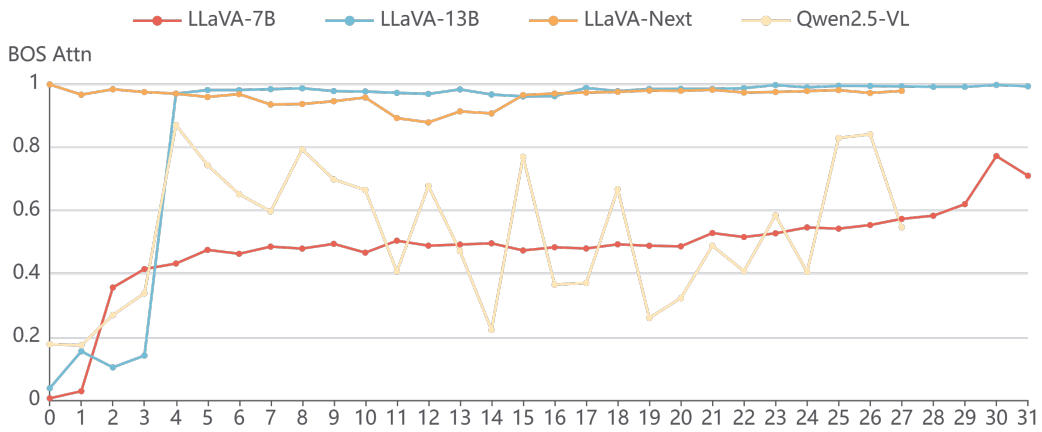

Figure 10. The model's attention to the BOS token in each layer.

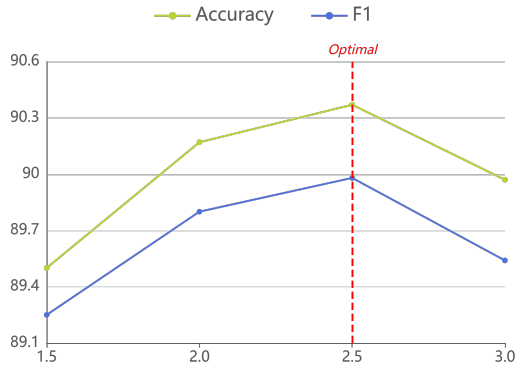

Figure 11. Effectiveness of β in MSCOCO's Random set

is discussed in Section C. Specifically, the starting layers are set as follows (with layers numbered from 0): layer 2 for LLaVA-1.5, layer 0 for LLaVA-Next, and layer 4 for Qwen2.5-VL. We apply VGA up to an intermediate layer of the model: specifically, layer 24 for LLaVA-1.5-13B and layer 16 for all other models. We adopt the default settings of $\beta = 0.2$ and $\lambda = 0.02$. For the larger model (LLaVA-1.5-13B) and visually simpler tasks (POPE), we increase β to 0.25 to apply stronger vision guidance. Stanza [21] is utilized to extract objects mentioned in the question. We employ greedy decoding for next token prediction and set the maximum generation length to 512. All experiments are conducted using a single NVIDIA A800-40G GPU.

B. Discussion of Main Results

POPE. The experimental results on the POPE benchmark are summarized in Table 1. Our proposed method, VGA, achieves significant improvements, clearly demonstrating its advantage in mitigating existence hallucinations. Moreover, VGA consistently delivers positive gains across all tested MLLMs, highlighting its strong generalizability. While the baselines enhance visual understanding by strengthening the model's attention to visual content, they often lack precise localization of key objects, which leads to suboptimal performance. The superior performance of VGA indicates that providing explicit visual guidance effectively enhances the model's ability to perceive and discriminate objects within the image.

CHAIR. The experimental results on the CHAIR benchmark are presented in Table 2. Our method achieves the overall lowest hallucination rate across all evaluated MLLMs, while maintaining no significant drop in F1 score. This result underscores VGA's ability to strike a better balance between generating a richer output and ensuring generation accuracy. Although PAI enhanced by contrastive decoding (PAI_{CD}) further refines visual information in the logits and achieves strong performance, contrastive decoding requires two forward passes, which effectively doubles the inference cost. In stark contrast, VGA achieves superior performance with only a single forward pass, demonstrating both higher efficiency and effectiveness.

AMBER. AMBER is a comprehensive hallucination benchmark that enables a more thorough evaluation of model hallucinations. The results, shown in Table 3 and Table 5, demonstrate that VGA still achieves significant dehallucination performance, further validating the importance of precise visual guidance. Methods like TARAC guide visual attention based on historical attention distributions; however, the localization capability derived solely from visual attention is inherently limited. In contrast, VGA leverages VSC to achieve more accurate visual grounding without relying on external tools, thus enabling a more effective suppression of hallucinations.

Table 5. Results on AMBER with LLaVA-13B. The AMBER metric is calculated as $(1 - \text{CHAIR} + \text{F1})/2$.

MLLM	Method	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	AMBER ↑
LLaVA-13B	Vanilla	6.8	51.3	31.1	3.4	71.5	**96.0**	59.5	73.5	33.85
	PAI	5.4	51.6	27.3	2.1	74.2	95.6	64.0	76.7	36.15
	PAI_{CD}	5.2	**52.0**	28.3	**1.9**	75.0	95.3	65.5	77.6	36.70
	VAF	6.9	51.2	32.0	3.6	69.5	95.6	56.7	71.2	32.65
	TARAC	5.8	49.3	28.5	2.9	73.0	**96.0**	61.9	75.3	35.25
	VGA	**4.6**	49.9	**23.6**	2.0	**78.5**	95.0	**71.3**	**81.5**	**38.95**

Table 6. Ablation study on CHAIR.

Setting	Cs	Ci	F1
VGA	**29.8**	**8.2**	**76.3**
w/o Head Balancing	34.6	9.4	**76.3**
w/o Early Termination	32.6	8.8	75.4

Table 7. Effectiveness of β and λ in CHAIR's validation set.

Setting	Cs	F1
Vanilla	49.6	76.5
$\beta = 0.15, \lambda = 0.01$	41.6	77.5
$\beta = 0.15, \lambda = 0.02$	43.4	76.8
$\beta = 0.15, \lambda = 0.04$	43.6	77.1
$\beta = 0.20, \lambda = 0.01$	26.2	76.4
$\beta = 0.20, \lambda = 0.02$	32.8	77.7
$\beta = 0.20, \lambda = 0.04$	38.6	77.9
$\beta = 0.25, \lambda = 0.01$	10.0	68.7
$\beta = 0.25, \lambda = 0.02$	11.8	69.1
$\beta = 0.25, \lambda = 0.04$	19.8	71.8

C. The Starting Layer of VGA

In self-attention mechanisms, a "sink" phenomenon [10, 36] exists—where the model abnormally focuses its attention on a few individual tokens. PAI [18] observes that the BOS token carries very limited information yet consistently receives disproportionately high attention. Consequently, when the BOS token receives excessively high attention from the model, it serves as an appropriate trigger to enhance the model's visual attention. Following the approach of PAI, we analyze the BOS attention patterns across different models to determine the optimal starting layer for VGA. We apply max pooling across all attention heads. The analysis results on the image captioning task are shown in Figure 10. Based on our observations, we set the starting layer of VGA for different MLLMs as follows: layer 2 for LLaVA-1.5-7B, layer 4 for LLaVA-1.5-13B, layer 0 for LLaVA-Next, and layer 4 for Qwen2.5-VL.

D. Ablation Study

Considering the functional diversity across attention heads, VGA employs head balancing by applying stronger guidance to heads with weaker visual features and lighter guidance to those that already exhibit strong visual functionality. Additionally, since visual understanding in MLLMs primarily occurs in the early to middle layers, we terminate visual attention guidance at an intermediate layer of the model. We conduct ablation studies on both of these design choices, and the results are presented in Table 6. Head balancing preserves the model's original visual capabilities while incorporating additional vision-guidance, thereby achieving improved performance. Meanwhile, the early exit mechanism prevents visual guidance from being applied during non-visual understanding stages, maintaining consistency with the model's inherent behavioral characteristics. In short, both the head balancing and early termination modules contribute positively to performance.

E. Hyperparameters

β controls the strength of visual guidance, while λ controls the penalty strength on already-generated content in image captioning tasks. We perform a grid search for β on the POPE benchmark, as shown in Figure 11. Additionally, we conduct a grid search over both β and λ on the CHAIR benchmark, with results presented in Table 7. We observe that for tasks like POPE, which require focusing on a single visual region, stronger visual guidance is beneficial. In contrast, for image captioning tasks that demand broader attention across multiple visual regions, overly strong guidance can restrict the model's capacity. Therefore, we set $\beta = 0.25$ on the POPE benchmark and $\beta = 0.2$ on other benchmarks. Furthermore, for image captioning, we select $\lambda = 0.02$ based on a balanced consideration of hallucination rate and F1 score.

F. Beyond Object Hallucination.

The model's comprehension of auxiliary information, such as relations and attributes, fundamentally relies on its under-

Table 8. *Att.* and *Rel.* denote the F1 scores on AMBER's attribute and relation tasks, respectively. *MME* denote perception tasks in the MME benchmark.

MLLM	Method	Att.	Rel.	MME
LLaVA-1.5-7B	Vanilla	64.4	68.5	1456.5
	TARAC	63.5	68.4	1462.1
	VGA	**65.7**	**73.9**	**1465.7**
Qwen2.5-VL-7B	Vanilla	84.4	75.8	1691.1
	TARAC	84.8	76.2	1713.1
	VGA	**86.4**	**76.3**	**1719.7**

standing of objects; thus, object grounding remains critical. Furthermore, for tasks lacking explicit objects (*e.g.*, image captioning), we employ the VSS mechanism to establish informative grounding (see Sec. 3.2). Additional results presented in Table 8 indicate that VGA maintains advantages in mitigating relation- and attribute-based hallucinations. We further validate VGA on general VQA benchmarks. Results on MME [5] demonstrate that VGA performs effectively on object-agnostic tasks, confirming that these performance gains stem from enhanced visual perception via vision guidance.

G. First Token Approximation.

We approximate the visual semantic confidence of an object using the first token of its tokenized representation. Under a single-step prediction setting, utilizing the first token aligns with the autoregressive paradigm; conversely, the probability of any non-first token (without the "_" prefix) is extremely low (always less than $1e-4$), rendering their distribution across visual tokens meaningless. We also experimented with aggregating the probabilities of all tokens ($\log c(O) \approx \sum_i \log c(o_i)$); however, this yielded negligible performance differences. Consequently, we adopt the simplest approach.