ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models

Supplementary Material

8. Additional Experimental Results

Sec. 8.1 presents the additional experimental results across all tasks in the MME benchmark. Sec. 8.2 details the experimental outcomes on the three datasets within the POPE benchmark. Sec. 8.3 compares the inference speeds and memory usage of various methods on ScienceQA and Nocaps. Sec. 8.4 highlights case studies of the VAF method on the LLaVA-Bench dataset.

8.1. Detailed Experimental Results on MME

Fig. 9 and Fig. 10 present the performance of the LLaVA model family on perception-related tasks within the MME benchmark. Models utilizing the VAF method demonstrate significantly better performance compared to those employing the VCD method. Notably, VAF achieves consistent leadership across all tasks with the LLaVA-v1.5-13B model, likely due to its ability to balance attention between

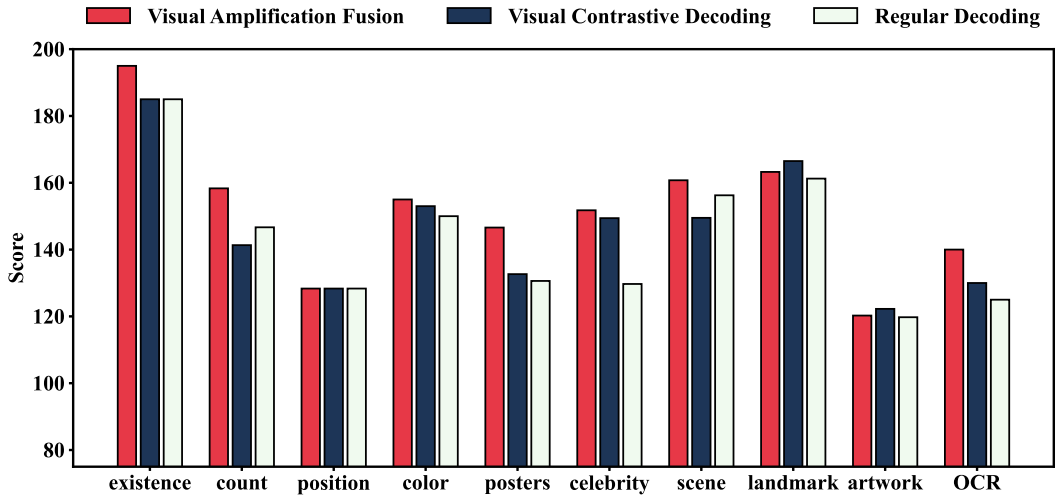


Figure 9. **Performance of LLaVA-v1.5-7B model on perception-related tasks in the MME Benchmark.** VAF consistently achieved the highest scores across nearly all perception tasks.

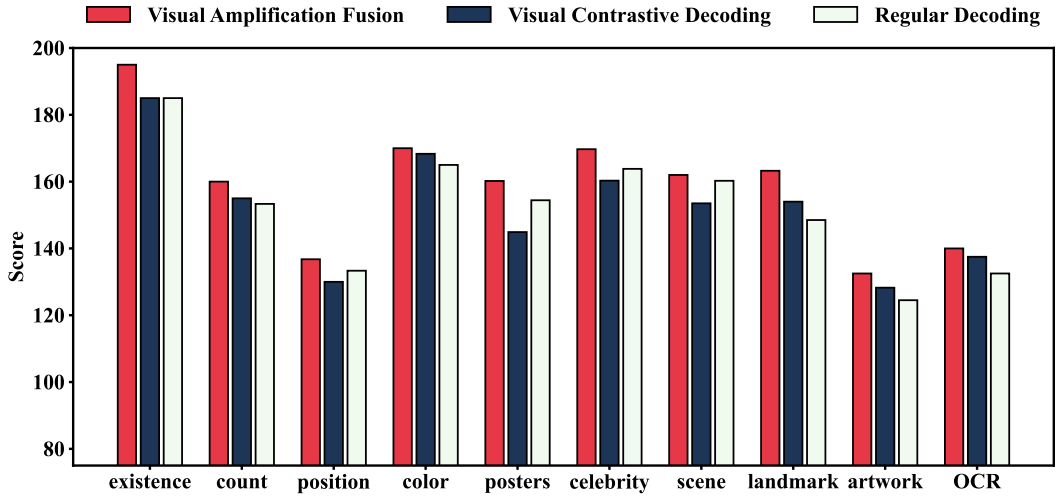


Figure 10. **Performance of LLaVA-v1.5-13B model on perception-related tasks in the MME Benchmark.** VAF consistently achieved the highest scores across nearly all perception tasks.

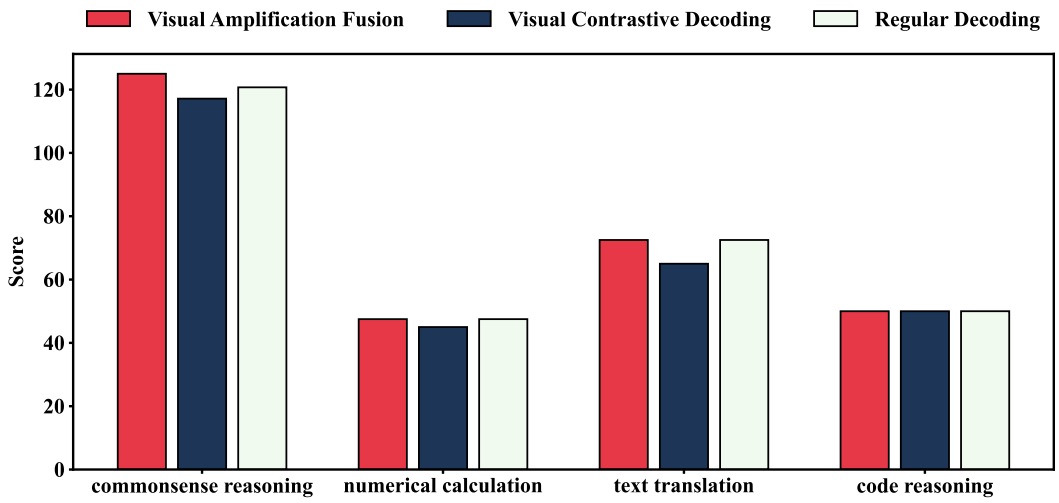


Figure 11. **Performance of the LLaVA-v1.5-7B model on cognition-related tasks in the MME Benchmark.** The VAF method delivers a slight performance improvement compared to the degradation observed with the VCD method.

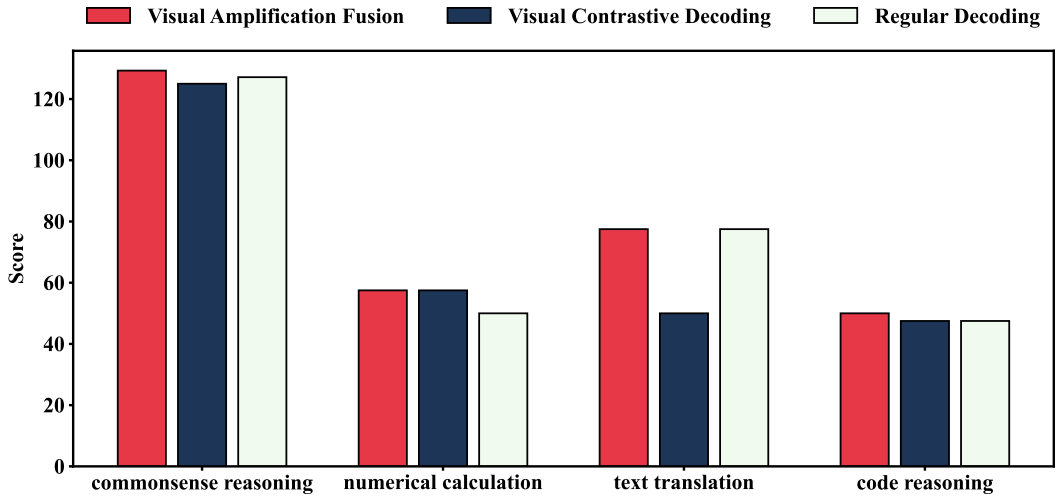


Figure 12. **Performance of the LLaVA-v1.5-13B model on cognition-related tasks in the MME Benchmark.** The VAF method delivers a slight performance improvement compared to the degradation observed with the VCD method.

visual and language modalities, ensuring generated content aligns more closely with visual inputs.

Fig. 11 and Fig. 12 illustrate the performance of LLaVA model family on cognition-related tasks within the MME benchmark. The application of the VCD method significantly impaired the model's performance on these tasks, likely due to its disruptive effect on linguistic priors. In contrast, VAF method not only avoided such negative impacts but also resulted in a slight performance improvement. This improvement is attributed to VAF's ability to precisely resolve the model's tendency to overlook visual features during the critical fusion stage, facilitating better integration of visual information while preserving its effective use of linguistic information.

8.2. Detailed Experimental Results on POPE

Tab. 6 and Tab. 9 summarize the experimental results of the LLaVA-v.15 model family on the MSCOCO, A-OKVQA, and GQA datasets within the POPE benchmark. The results highlight that our approach consistently delivers more stable and significantly improved hallucination suppression compared to the VCD method. This advantage stems from our direct enhancement of attention to visual features during the modality fusion process, enabling balanced outputs across both visual and linguistic modalities. In contrast, the VCD method relies on suppressing language priors to indirectly enhance attention to visual information. Decoding method employed in all experiments utilizes greedy search.

Dataset	Category	Method	Accurancy	Precision	Recall	F1-score
MSCOCO	Random	Regular	88.2	94.2	81.5	87.4
		VCD	88.5	94.4	81.8	87.6
		VAF	**89.8**	**92.9**	**86.2**	**89.4**
	Popular	Regular	86.1	89.9	81.5	85.5
		VCD	86.3	90.0	81.7	85.8
		VAF	**87.5**	**88.6**	**86.2**	**87.4**
	Adverserial	Regular	82.3	82.9	81.3	82.1
		VCD	82.3	82.9	81.6	82.4
		VAF	**83.4**	**86.8**	**78.9**	**82.6**
A-OKVQA	Random	Regular	87.6	87.6	87.7	87.6
		VCD	87.7	87.8	87.6	87.8
		VAF	**89.4**	**91.7**	**86.6**	**89.1**
	Popular	Regular	81.9	78.4	87.7	82.8
		VCD	82.1	78.5	87.9	83.1
		VAF	**84.2**	**82.6**	**86.6**	**84.6**
	Adverserial	Regular	74.3	68.8	87.7	77.1
		VCD	72.4	68.0	87.4	76.7
		VAF	**77.2**	**72.9**	**86.6**	**79.2**
GQA	Random	Regular	88.0	87.1	89.3	88.2
		VCD	88.6	87.4	89.5	88.8
		VAF	**89.5**	**90.8**	**88.0**	**89.4**
	Popular	Regular	79.4	74.4	89.3	81.1
		VCD	79.9	74.6	89.5	81.7
		VAF	**81.8**	**78.3**	**88.0**	**82.9**
	Adverserial	Regular	76.3	70.6	89.3	78.9
		VCD	75.2	70.2	89.9	78.3
		VAF	**79.7**	**75.4**	**88.0**	**81.2**

Table 6. **Experimental results of LLaVA-1.5-7B model on POPE.** VAF method achieves the most effective hallucination suppression across all three datasets. For emphasis, the **highest scores** in each setting are highlighted in **red**.

Model	Method	Accurancy	Total Time	GPU-Memory	Latency/Example
LLaVA-v1.5-7B	Regular	88.2	5:32	14.5G	0.111s
	VCD	88.5	**10:31**	**15.7G**	**0.210s**
	VAF	89.8	5:48	14.5G	0.116s
LLaVA-v1.5-13B	Regular	88.4	8:39	26.7G	0.173s
	VCD	88.6	**19:38**	**27.8G**	**0.392s**
	VAF	90.2	8:45	26.7G	0.175s

Table 7. **A comparison of inference speed and GPU memory usage for different methods applied to the LLaVA-v1.5 model family on POPE benchmark.** Results with the slowest inference speed and highest memory usage are highlighted in **red**.

8.3. Comparison of Inference Speeds

Tab. 7 and Tab. 8 assess the impact of various methods on the LLaVA-v1.5 model family, focusing on inference speed and GPU memory usage. The results indicate that VCD significantly slows down inference, whereas our proposed method has a minimal effect. Furthermore, our method introduces no additional GPU memory requirements, in con-

trast to VCD, which incurs substantial GPU memory overhead. This efficiency is achieved because our approach eliminates the need for extra processing of contrastive inputs, thereby significantly reducing computational overhead. All experiments were performed on a server equipped with a single A800 80G GPU, employing greedy search as the decoding strategy.

Model	Method	Accurancy	Total Time	GPU-Memory	Latency/Example
LLaVA-v1.5-7B	Regular	68.0	0:36:39	14.5G	0.488s
	VCD	64.5	**1:18:47**	**15.7G**	**1.058s**
	VAF	68.5	0:36:41	14.5G	0.489s
LLaVA-v1.5-13B	Regular	71.6	0:45:20	26.7G	0.604s
	VCD	70.0	**1:46:59**	**27.8G**	**1.426s**
	VAF	71.7	0:48:24	26.7G	0.645s

Table 8. **A comparison of inference speed and GPU memory usage for different methods applied to the LLaVA-v1.5 model family on Nocaps benchmark.** Results with the slowest inference speed and highest memory usage are highlighted in **red**.

Dataset	Category	Method	Accurancy	Precision	Recall	F1-score
MSCOCO	Random	Regular	88.4	94.6	81.6	87.6
		VCD	88.6	95.0	81.8	87.7
		VAF	**90.2**	**94.2**	**85.6**	**89.7**
	Popular	Regular	86.9	91.3	81.6	86.2
		VCD	87.0	91.4	82.0	86.4
		VAF	**88.4**	**90.6**	**85.6**	**88.0**
	Adverserial	Regular	83.4	84.9	81.4	83.1
		VCD	83.7	85.1	81.7	83.1
		VAF	**84.5**	**83.8**	**85.5**	**84.7**
A-OKVQA	Random	Regular	88.0	88.8	87.1	87.9
		VCD	88.2	89.2	87.5	87.9
		VAF	**89.4**	**91.4**	**86.8**	**89.1**
	Popular	Regular	83.9	81.7	87.1	84.3
		VCD	84.2	81.7	87.3	84.3
		VAF	**86.0**	**85.4**	**86.8**	**86.1**
	Adverserial	Regular	76.0	71.0	87.1	78.2
		VCD	76.4	71.2	87.1	78.3
		VAF	**78.2**	**74.1**	**86.8**	**79.9**
GQA	Random	Regular	88.3	87.8	89.0	88.4
		VCD	88.3	88.1	89.3	88.5
		VAF	**89.7**	**87.8**	**92.2**	**89.9**
	Popular	Regular	83.3	79.8	89.0	84.1
		VCD	83.2	80.0	89.2	84.1
		VAF	**85.2**	**83.0**	**88.6**	**85.7**
	Adverserial	Regular	78.5	73.3	89.0	80.4
		VCD	78.7	73.3	88.9	80.3
		VAF	**80.8**	**76.6**	**88.6**	**82.1**

Table 9. **Experimental results of LLaVA-1.5-13B model on POPE.** VAF method achieves the most effective hallucination suppression across all three datasets. For emphasis, the **highest scores** in each setting are highlighted in **red**.

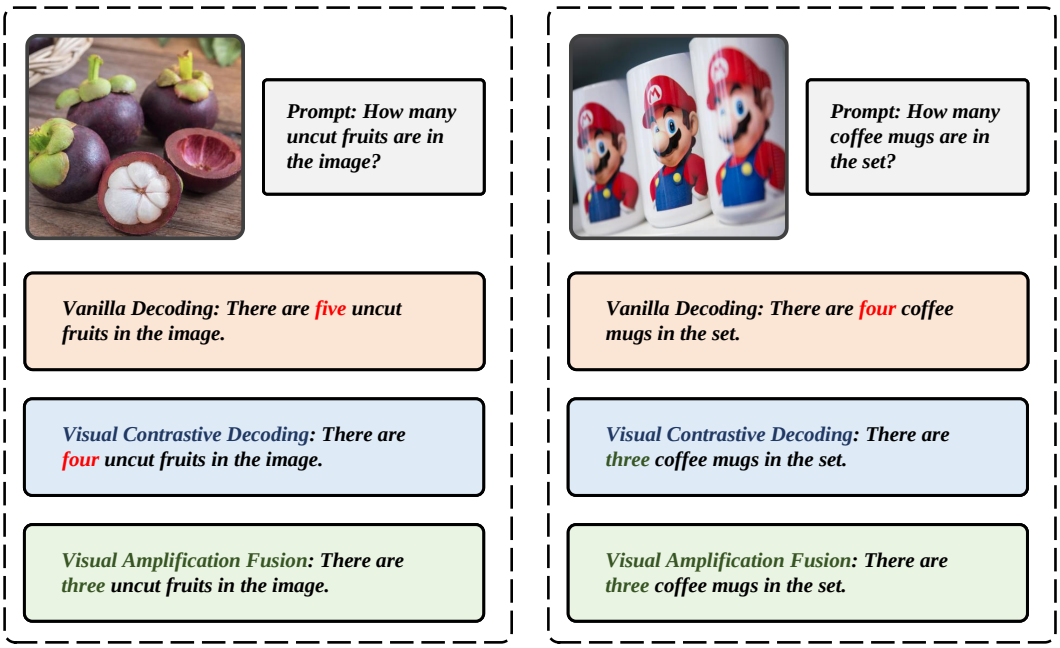


Figure 13. **An illustration of VAF correcting hallucinations on LLaVA-Bench, with a focus on numerical perception tasks.** **Correct** outputs are highlighted in **green**, while **incorrect** ones are marked in **red**.

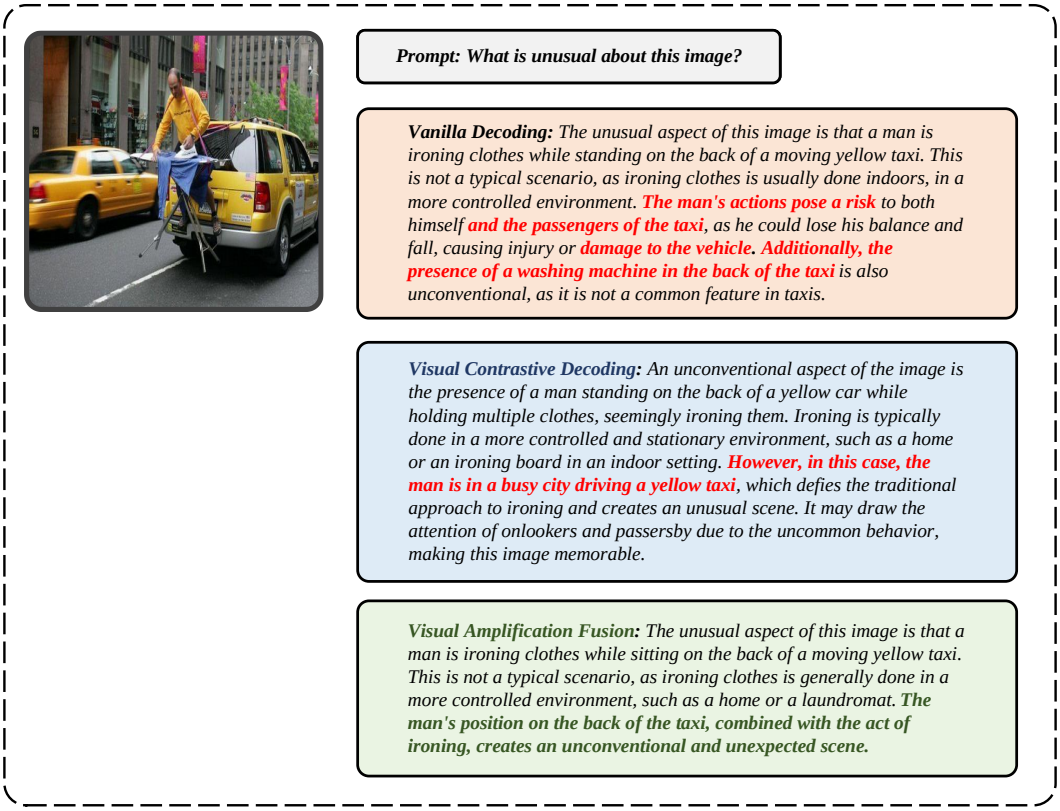


Figure 14. **An illustration of VAF correcting hallucinations on LLaVA-Bench, with a focus on complex reasoning tasks.** **Correct** outputs are highlighted in **green**, while **incorrect** ones are marked in **red**.

Prompt: Describe this photo in detail.

Vanilla Decoding: *The image features a painting of a dog wearing a Renaissance-style dress, giving it a human-like appearance. The dog is sitting in a chair, looking directly at the viewer. The painting is quite detailed, capturing the dog's facial expression and the intricate design of the dress.* ***In the background, there is a person partially visible on the left side of the painting.*** *The overall scene is quite captivating and unique, as it combines the artistic style of the Renaissance with the subject matter of a dog.*

Visual Contrastive Decoding: *The image features a unique painting on a canvas, depicting a portrait of a bulldog in Renaissance style, giving it a Renaissance era appearance. In this painting, the bulldog is presented as a woman, complete with a veil on its head, looking directly to its right.*
The painting also includes several elements of landscape and architecture, giving a sense of a real setting. Apart from the central dog figure, ***there are three smaller figures or figures in the background, adding more depth and life to the painting. In the forefront, a chair is visible, positioned to the left of the dog,*** *further enhancing the painting's depiction of a realistic scene.*

Visual Amplification Fusion: ***The image features a painting of a dog wearing a Renaissance-style dress, giving it a human-like appearance.*** *The dog is* ***positioned in the center of the painting, and appears to be staring at the viewer.***

Figure 15. **An illustration of VAF correcting hallucinations on LLaVA-Bench, with a focus on image description tasks.** **Correct** outputs are highlighted in **green**, while **incorrect** ones are marked in **red**.

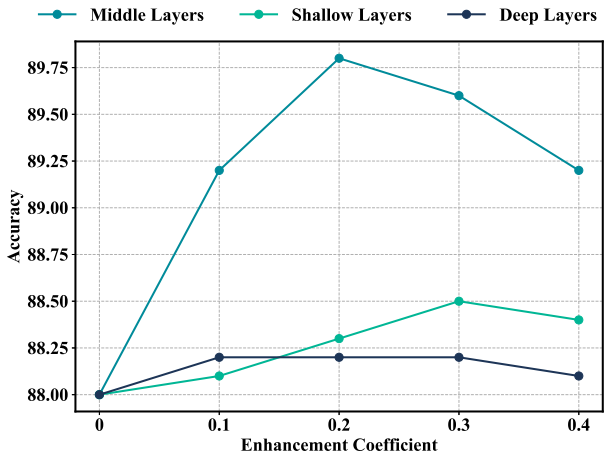
Prompt: Describe this photo in detail.

Vanilla Decoding: *The image features a beautiful display of* ***three ripe and juicy passion fruit,*** *cut in half to reveal their exquisite white meat inside.* ***The ripe fruit is placed in a wooden bowl, making them the focal point of the image.*** *Each fruit half rests in a different location on the table, showcasing the various sizes of these unique fruit-like berries. The close-up view of the cut open fruits and the table setting creates an appetizing scene, highlighting the natural color of the passion fruit and its appeal to anyone who enjoys a fresh and tangy experience with a hint of exotic taste.*

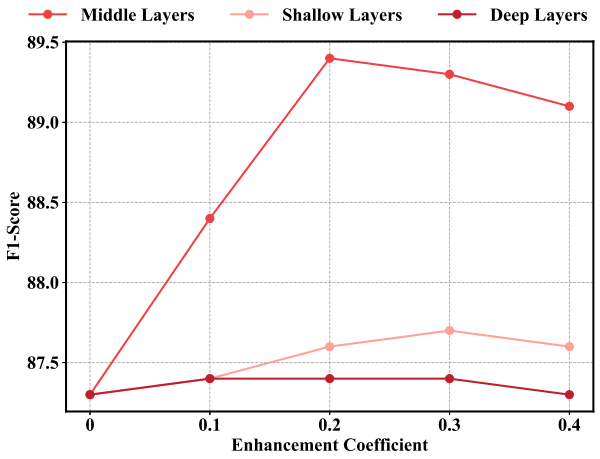
Visual Contrastive Decoding: *The image features a wooden table with a basket of fruit, including a variety of fruits such as apples, oranges, and bananas. There are three apples on the table, with one placed near the left side, another in the middle, and the third one on the right side. In addition to the apples, there are two oranges on the table, one located in the middle and the other towards the right side.* ***A banana can be seen on the left side of the table, and a cluster of grapes is also present, with some grapes placed near the center and others on the right side of the table. The assortment of fruits creates a colorful and appetizing display.***

Visual Amplification Fusion: *The image features a wooden table with a basket of* ***four purple fruits****, possibly eggplants or figs, placed on it. The fruits are cut open, revealing their innards.* ***The fruits are arranged in a way that they are sitting on top of each other****, with one fruit on the left side, another in the middle, and the third on the right side of the table. The fourth fruit is placed at the edge of the table.*

Figure 16. Additional example of VAF correcting model hallucinations on LLaVA-Bench, primarily focusing on image description tasks. **Correct** outputs are highlighted in **green**, while **incorrect** ones are marked in **red**.



(a) Accuracy Metric



(b) F1-Score Metirc

Figure 17. **The Effect of Enhancing Visual Attention at Different Layers on Prediction Accuracy.** This experiment, conducted with the LLaVA-v1.5-7B model on the COCO-Random dataset within the POPE Benchmark, demonstrates that enhancing attention to visual features in the model's middle layers significantly reduces hallucinations.

8.4. Case study on LLaVA-Bench

Fig. 13, Fig. 14, Fig. 15, and Fig. 16 illustrate the effectiveness of various methods in mitigating model hallucinations on LLaVA-Bench. Across tasks such as numerical perception, image description, and complex reasoning, our approach demonstrates consistently superior performance in suppressing hallucinations. Experiments are conducted using LLaVA-v1.5-7B model.

9. Additional Ablation Studies

In Sec. 9.1, we examine how enhancing attention to visual features at different levels affects hallucination suppression. In Sec. 9.2, we analyze the influence of varying the suppression coefficient β on mitigating hallucinations. Finally, in Sec. 9.3, we evaluate the performance of the VAF method in suppressing hallucinations under various sampling strategies.

9.1. Effect of Enhancement at Different Layers

We enhanced attention to visual features in layers 0-5, 10-15, and 20-25. Fig. 17 demonstrates the impact of enhancing visual attention at different layers. Notably, enhancing attention in the middle layers significantly reduces hallucination, while modifications in the shallow and deep layers have minimal effect on the generation results. As discussed in Sec. 4.1, this is because the model primarily integrates modality information in the middle layers. Thus, enhancing the focus on visual features during this phase is crucial for effectively mitigating hallucination. Experiments are conducted using LLaVA-v1.5-7B model on COCO-Random dataset from the POPE Benchmark.

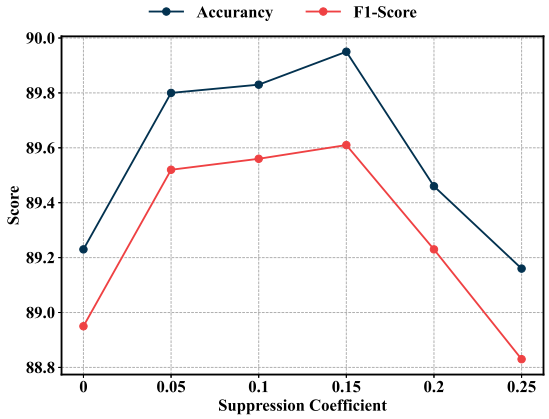


Figure 18. **The effect of the suppression coefficient β on the VAF method's ability to mitigate model hallucinations.** The experiments were performed using the LLaVA-v1.5-7B model on the COCO-Random dataset from the POPE Benchmark.

9.2. Effect of Suppression Coefficient

We assessed the effect of the suppression coefficient β on the performance of the VAF method using the LLaVA-v1.5-7B model on the COCO-Random dataset within the POPE Benchmark. In our experiments, α was fixed at 0.15, while β was systematically adjusted. The results, presented in Fig. 18, reveal that when $0 < \beta < 0.15$, VAF significantly enhanced its ability to suppress hallucinations in the model. This improvement is likely due to VAF reducing redundant attention to system prompts in this range, thereby reinforcing focus on visual features and enabling generated content to better align with the visual input. Conversely,

Sampling Strategy	Method	Accurancy	Precision	Recall	F1-Score
Greedy	Regular	88.2	94.4	81.4	87.4
	VAF	**89.8**	92.9	86.2	**89.4**
Direct Sampling	Regular	82.9	90.4	71.3	80.9
	VAF	**83.9**	90.6	80.9	**85**
Top P	Regular	84.3	92.1	72.5	82.1
	VAF	**85.7**	89.6	82.4	**85.9**
Top K	Regular	83.3	91.9	72.8	81.1
	VAF	**85**	88.3	81.9	**84.9**
Top K + Temp0.5	Regular	85.5	95.1	74.9	84.5
	VAF	**86.7**	91.2	83.4	**87**
Top K + Temp1.5	Regular	80.4	87.1	70.2	77.8
	VAF	**82.1**	86	78.2	**81.9**

Table 10. **Effectiveness of the VAF method in mitigating model hallucination under different sampling strategies.** The **highest score** in each setting is highlighted in **red**. Experiments were conducted using the LLaVA-v1.5-7B model on the COCO-Random dataset within the POPE Benchmark.

when $\beta > 0.15$, the model's performance deteriorated. We hypothesize that this decline stems from excessive suppression of attention to system prompts, which disrupts the delicate balance required for effectively integrating multimodal information, ultimately leading to a degradation in overall performance.

9.3. Effect of Different Sampling Strategies

We evaluated the effectiveness of the VAF method in mitigating model hallucination under different sampling strategies using the LLaVA-v1.5-7B model on the COCO-Random dataset from the POPE Benchmark. The experimental results, shown in Tab. 10, indicate that the VAF method significantly mitigates model hallucination across all sampling strategies.

10. Prompts for Different Tasks

POPE Dataset. In the POPE dataset, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER:** **IMAGE**
> **Is there a cow in the image? Please just answer yes or no.**
>
> **ASSISTANT:**

Nocaps Datasets. In Nocaps and Flickr30k dataset, input template for the model is presented below, with prompts highlighted in **green** and image highlighted in **red**.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER:** **IMAGE**
> **Provide a one-sentence caption for the provided image.**
>
> **ASSISTANT:**

Sci-VQA Dataset. In the Sci-VQA dataset, input template for the model is presented below, with the prompts highlighted in **green** and the image highlighted in **red**.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER:** **IMAGE**
> **Context: Select the best answer.**
> Which property do these three objects have in common?
> A. shiny B. slippery C. opaque
> **Answer with the option's letter from the given choices directly.**
>
> **ASSISTANT:**