Mitigating Multimodal Hallucinations via Gradient-based Self-Reflection

Supplementary Material

1. First Order Taylor Expansion

Let $\mathbf{z}_m^\star \in \mathbb{R}^{|\mathcal{V}|}$ denote the step-m logits $\mathbf{z}_m^\star = \pi_{\theta^\star}(\mathbf{t}^v, \mathbf{t}^p, \mathbf{y}_{<m})$. Around a reference point $(\mathbf{t}^{v(0)}, \mathbf{t}^{p(0)}, \mathbf{y}_{<m}^{(0)})$, the detailed first-order Taylor expansion of the logit $\mathbf{z}_m^\star$ is

$$
\begin{aligned}
\mathbf{z}_m^\star \;\approx\; & \underbrace{\mathbf{z}_m^{\star(0)}}_{\pi_{\theta^\star}(\mathbf{t}^{v(0)}, \mathbf{t}^{p(0)}, \mathbf{y}_{<m}^{(0)})} + \sum_{s=1}^{S} \mathbf{g}_{ms}^{v} \Big(\mathbf{t}_s^v - \mathbf{t}_s^{v(0)}\Big) \\
& + \sum_{n=1}^{N} \mathbf{g}_{mn}^{p} \Big(\mathbf{t}_n^p - \mathbf{t}_n^{p(0)}\Big) \;+\; \sum_{i=1}^{m-1} \mathbf{g}_{mi}^{y} \Big(\mathbf{y}_i - \mathbf{y}_i^{(0)}\Big) \\
= & \sum_{s=1}^{S} \mathbf{g}_{ms}^{v} t_s^v + \sum_{n=1}^{N} \mathbf{g}_{mn}^{p} t_n^p + \sum_{i=1}^{m-1} \mathbf{g}_{mi}^{y} y_i \\
& + \underbrace{\mathbf{z}_m^{\star(0)} - \sum_{s=1}^{S} \mathbf{g}_{ms}^{v} t_s^{v(0)} - \sum_{n=1}^{N} \mathbf{g}_{mn}^{p} t_n^{p(0)} - \sum_{i=1}^{m-1} \mathbf{g}_{mi}^{y} y_i^{(0)}}_{Const}, \\
= & \sum_{s=1}^{S} \mathbf{g}_{ms}^{v} t_s^v + \sum_{n=1}^{N} \mathbf{g}_{mn}^{p} t_n^p + \sum_{i=1}^{m-1} \mathbf{g}_{mi}^{y} y_i + Const,
\end{aligned}
\tag{1}
$$

where the (token-wise) Jacobians are

$$
\mathbf{g}_{ms}^{v} \coloneqq \left.\frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{t}_s^v}\right|_{\mathbf{t}^v = \mathbf{t}^{v(0)}}, \qquad \mathbf{g}_{mn}^{p} \coloneqq \left.\frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{t}_n^p}\right|_{\mathbf{t}^p = \mathbf{t}^{p(0)}}, \qquad \mathbf{g}_{mi}^{y} \coloneqq \left.\frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{y}_i}\right|_{\mathbf{y} = \mathbf{y}_{<m}^{(0)}}, \tag{2}
$$

and $\mathbf{z}_m^{\star(0)} = \pi_{\theta^\star}(\mathbf{t}^{v(0)}, \mathbf{t}^{p(0)}, \mathbf{y}_{<m}^{(0)})$. Here $\big|_{\cdot}$ denotes evaluation at the reference point. Each $\mathbf{g}_{ms}^{v}$, $\mathbf{g}_{mn}^{p}$, $\mathbf{g}_{mi}^{y}$ maps a small token perturbation in its corresponding embedding space to a perturbation of the logit vector in $\mathbb{R}^{|\mathcal{V}|}$. And $Const$ denotes all other terms that are constant w.r.t., the $\mathbf{t}^v, \mathbf{t}^p$.

2. Interpreting Contrastive Decoding through KL Divergence

Kullback-Leibler (KL) divergence can be used to interpret contrastive decoding, It measures the divergence between the reference distribution $p_{\theta^\star}(y_{cm}|\mathbf{t}^o, \mathbf{t}^p, y_{<m})$ to the $\mathbf{t}^u$ joint distribution $p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m})$, where $\mathbf{t}^v = \{\mathbf{t}^u, \mathbf{t}^o\}$.

$$
\begin{aligned}
\mathrm{D}_{KL} &= \sum_c p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m}) \log\Big(\frac{p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m})}{p_{\theta^\star}(y_{cm}|\mathbf{t}^o, \mathbf{t}^p, y_{<m})}\Big) \\
&= \sum_c p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m}) (\log(p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m})) - \log(p_{\theta^\star}(y_{cm}|\mathbf{t}^o, \mathbf{t}^p, y_{<m}))) \\
&= \sum_c p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m}) ([\pi_{\theta^*}(\mathbf{t}^v, \mathbf{t}^p)_m]_c - \log(\sum \exp(\pi_{\theta^*}(\mathbf{t}^v, \mathbf{t}^p)_m)) \\
&- [\pi_{\theta^*}(\mathbf{t}^o, \mathbf{t}^p)_m]_c + \log(\sum \exp(\pi_{\theta^*}(\mathbf{t}^o, \mathbf{t}^p)_m))) \\
&= \sum_c p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m}) (\underbrace{[\pi_{\theta^*}(\mathbf{t}^v, \mathbf{t}^p)_m - \pi_{\theta^*}(\mathbf{t}^o, \mathbf{t}^p)_m]_c}_{\text{adjustment term}} + Const),
\end{aligned}
\tag{3}
$$

where $p_{\theta^\star}(y_{cm}|\mathbf{t}^v, \mathbf{t}^p, y_{<m}) = \sigma(\pi_{\theta^*}(\mathbf{t}^v, \mathbf{t}^p)_m)$, $p_{\theta^\star}(y_{cm}|\mathbf{t}^o, \mathbf{t}^p, y_{<m}) = \sigma(\pi_{\theta^*}(\mathbf{t}^o, \mathbf{t}^p)_m)$ and c represents a class in the predefined vocabulary. The adjustment term increases the KL divergence, thereby emphasizing the impact of visual tokens.

3. Derivation of α Computation

We choose α_m such that the influence of $\mathbf{t}^u$ matches the dominant text level. For clarity, we present the derivation for the case where the prompt is dominant; the case where the previous output tokens are dominant is analogous and leads to the same derivative up to a straightforward substitution. Given $\hat{\mathtt{I}}_m^v = (1+\alpha_m)\mathtt{I}_m^v - \alpha_m \tilde{\mathtt{I}}_m^o$ and $\hat{\mathtt{I}}_m^p = (1+\alpha_m)\mathtt{I}_m^p - \alpha_m \tilde{\mathtt{I}}_m^p$, we enforce $\hat{\mathtt{I}}_m^v = \hat{\mathtt{I}}_m^p$, and solve for α_m:

$$\begin{aligned} (1+\alpha_m)\mathtt{I}_m^v - \alpha_m \tilde{\mathtt{I}}_m^o &= (1+\alpha_m)\mathtt{I}_m^p - \alpha_m \tilde{\mathtt{I}}_m^p \\ \alpha_m(\mathtt{I}_m^v - \tilde{\mathtt{I}_m^o} + \tilde{\mathtt{I}}_m^p - \mathtt{I}_m^p) &= \mathtt{I}_m^p - \mathtt{I}_m^v \\ \alpha_m &= \frac{\mathtt{I}_m^p - \mathtt{I}_m^v}{\mathtt{I}_m^v - \tilde{\mathtt{I}_m^o} + \tilde{\mathtt{I}}_m^p - \mathtt{I}_m^p}. \end{aligned} \tag{4}$$

4. MLLMs Architectures

Tab. 1 shows detailed information about the vision encoder and LLM components of the MLLMs used in our experiments.

Table 1. Details of the used MLLM architectures.

MLLMs	Vision encoder	LLM
LLaVA-v1.5 (7B)	CLIP-L-336px	Vicuna-v1.5-7B
LLaVA-v1.5-13B	CLIP-L-336px	Vicuna-v1.5-13B
LLaVA-v1.6	CLIP-L-336px	Vicuna-v1.5-7B
InstructBLIP (7B)	BLIP-2	Vicuna-v1.1-7B
InstructBLIP-13B	BLIP-2	Vicuna-v1.1-13B
mPLUG-Owl2	CLIP-L	LLaMA-2-7B
InternVL2-4B	InternViT-300M-448px	Phi-3-mini-128k-instruct
Qwen2-VL-7B	QwenViT	Qwen2-7B

5. Results on MMBench

We further evaluate our method on MMBench [10]. The results in Tab. 2 indicate that our method improves the overall performance and achieves consistent improvements across MLLMs on Coarse Perception (CP). This outcome aligns with the intended effect of our method, as its focus on increasing visual influence is directly linked to improving coarse perception capabilities. For other metrics, our method yields minor improvements due to the possible reason that certain abilities, such as Logical Reasoning (LR), rely more on the language component of MLLMs and cannot be enhanced solely by increasing visual influence.

Table 2. Results on MMBench Dataset.

MLLMs	Method	Overall	CP	FP-S	FP-C	AR	LR	RR
LLaVA-V1.5	base	**62.3**	68.5	**69.6**	**57.7**	73.1	**29.9**	54.7
	ours	61.8	**73.2**	62.6	53.0	**73.3**	27.8	**57.8**
mPLUG-Owl2	base	63.5	68.1	**69.1**	**55.8**	**78.4**	37.0	57.0
	ours	**65.0**	**72.6**	66.6	53.0	76.0	**41.6**	**63.0**

6. Results on MM-Vet

Table 3. Results on MM-Vet dataset.

MLLMs	Method	Rec	OCR	Know	Gen	Spat	Math	Total
LLaVA-V1.5	base	32.9	20.1	**19.0**	20.1	**25.6**	5.2	28.0
	ours	**38.9**	**24.9**	15.0	15.5	24.9	**7.7**	**28.9**
InstructBlip	base	32.4	14.6	16.5	**18.2**	**18.6**	**7.7**	26.2
	ours	**40.5**	**18.0**	**18.7**	17.4	14.9	3.8	**26.6**
mPLUG-Owl2	base	36.1	19.4	**29.8**	19.4	23.9	**7.7**	27.3
	ours	**45.0**	**26.4**	27.9	**25.9**	**24.8**	3.8	**33.9**

The evaluation on MM-Vet [13] in Tab. 3 shows that our method achieves consistent overall (Total) improvement, along with enhancements in recognition (Rec) and Optical Character Recognition (OCR), indicating its effectiveness in improving visual recognition. However, its performance varies across other metrics, including knowledge (Know), generalization (Gen), spatial awareness (Spat), and math (Math), suggesting that our method, which focuses on token influence balancing, may not effectively enhance the generalization ability of MLLMs.

7. Results on ScienceQA and Vizwiz

We evaluate our method on two complementary multimodal benchmarks. ScienceQA [11] integrates images, textual context, and curriculum knowledge, requiring models to perform structured multimodal reasoning. VizWiz [3], in contrast, consists of visual questions collected from blind users and features real-world challenges such as low-quality images, conversational queries, and unanswerable cases. These datasets jointly assess both reasoning under structured multimodal contexts and robustness in unconstrained real-world settings. As shown in Table 4, our approach consistently improves over the LLaVA-1.5 baseline. These gains demonstrate the effectiveness of our hallucination mitigation strategy in enhancing visual grounding across both knowledge-driven and real-world VQA tasks.

Table 4. Comparison of LLaVA-1.5 and our method on ScienceQA and VizWiz datasets.

MLLMs	Method	ScienceQA(%) ↑	VizWiz(%) ↑
LLaVA-V1.5	base	66.2	48.7
	Ours	**68.7**	**52.8**

8. Results on MME

Our evaluation on MME [2] dataset is presented in Tab. 5. Our method achieves better overall (Total) results and equal or improved performance in existence and counting, demonstrating its effectiveness in object recognition. However, it does not improve position accuracy and exhibits varying behavior on color. This diversity may stem from the inherent capabilities of MLLMs, which cannot be solely enhanced through token influence balancing.

Table 5. Result on MME Dataset.

MLLMs	Method	Existence ↑	Count ↑	Position ↑	Color ↑	Total ↑
LLaVA-v1.5	base	190.0	140.0	128.3	155.0	613.3
	ours	190.0	**153.3**	128.3	**163.3**	**634.9**
IntructBLIP	base	180.0	55.0	50.0	130.0	415.0
	ours	**185.0**	55.0	50.0	130.0	**420.0**
mPLUG-Owl2	base	170.0	145.0	73.3	**158.3**	546.6
	ours	170.0	**150.0**	73.3	150.0	**548.3**

9. Other Results of POPE

Table 6. More Results on POPE [6].

Dataset	Setting	Method	LLaVA-v1.5		InstructBLIP		mPLUG-Owl2	
			Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
MSCOCO	Random	base	87.1	85.4	87.1	85.7	86.0	84.4
		ours	**87.4**	**86.0**	**87.9**	**86.8**	**87.9**	**87.1**
	Popular	base	85.9	84.4	84.2	83.6	84.6	83.2
		ours	**86.2**	**84.8**	**85.0**	**84.3**	**86.4**	**85.7**
A-OKVQA	Random	base	88.0	**87.6**	88.5	88.5	86.5	85.7
		ours	**88.1**	87.4	**88.8**	**88.8**	**88.4**	**88.1**
	Popular	base	85.5	85.1	81.9	83.1	82.4	82.2
		ours	85.5	85.1	**82.3**	**83.4**	**85.1**	**85.3**
	Adversarial	base	79.1	79.9	74.8	77.9	74.7	76.9
		ours	**79.5**	**80.1**	**75.3**	**78.2**	**78.2**	**79.9**
GQA	Random	base	88.9	88.2	87.2	87.1	85.2	84.0
		ours	88.9	88.2	87.2	**87.2**	**86.1**	**85.0**
	Popular	base	84.1	84.1	78.6	80.4	78.7	78.5
		ours	**84.2**	84.1	**78.8**	80.4	**81.0**	**80.5**
	Adversarial	base	80.8	81.3	75.9	78.4	76.4	76.8
		ours	**81.1**	**81.6**	**76.1**	**78.5**	**79.2**	**79.1**

We report our experimental results on the POPE dataset, in addition to MSCOCO and adversarial settings, in Tab. 6. The results indicate that our method improves performance across all baseline MLLMs, with more significant gains observed in

the adversarial setting. This discrepancy likely arises because adversarial scenarios require models to rely more heavily on visual inputs, aligning with our method's focus on enhancing visual influence. Conversely, for popular and random objects, textual data often provides sufficient statistical information, reducing the necessity for increased visual input reliance.

10. Question Category Results on the AMBER Dataset

We report discriminative results across different question categories in Tab. 7. Our method improves performance in nearly all categories across all MLLMs. The improvement in InternVL2's object existence is minor, likely due to its already high visual influence ratio. For LLaVA-v1.5 and mPLUG-Owl2, which have lower original visual influence ratios, our method achieves more substantial gains in existence, attribute, and state categories.

Table 7. Results on the Question Categories of Discriminative Task on AMBER Dataset.

Category	Metric	InstructBLIP		LLaVA-v1.5		LLaVA-v1.6		mPLUG-Owl2		Intern-VL2	
		base	ours	base	ours	base	ours	base	ours	base	ours
Existence	acc	70.0	79.8	70.8	93.2	92.9	93.0	75.2	89.9	90.6	90.6
	P	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	R	70.0	79.8	70.8	93.2	92.9	93.0	75.2	89.9	90.6	90.6
	F1	82.3	88.7	82.9	96.4	96.3	96.3	85.8	94.6	95.0	95.0
Attribute	acc	71.9	78.3	72.3	76.1	75.2	77.1	73.9	78.2	82.3	82.6
	P	76.0	81.7	87.3	74.0	74.6	76.4	86.0	76.9	80.9	80.9
	R	64.3	73.0	52.2	82.7	83.0	83.9	57.1	81.8	84.7	85.2
	F1	69.7	77.1	65.3	78.1	78.5	80.0	68.6	79.3	82.8	83.0
State	acc	73.4	76.4	68.2	73.3	78.6	75.2	70.5	77.9	81.2	81.2
	P	75.1	77.1	86.2	70.3	78.6	74.7	84.9	75.5	79.1	78.7
	R	70.6	75.3	43.3	82.0	78.5	82.9	49.8	83.1	84.8	85.5
	F1	72.8	76.2	57.6	75.7	78.5	78.6	62.8	79.1	81.8	82.0
Number	acc	65.4	80.6	75.0	80.1	80.1	80.2	77.8	76.5	82.6	83.3
	P	75.4	93.1	86.9	79.1	79.2	78.6	86.0	77.0	83.0	84.0
	R	45.8	66.2	59.5	82.4	81.7	84.4	66.9	77.0	82.0	82.3
	F1	57.0	77.4	70.6	80.7	80.4	81.4	75.3	77.0	82.5	83.1
Action	acc	79.7	83.7	83.6	82.3	81.9	80.4	84.0	84.1	88.4	88.6
	P	82.5	88.5	92.9	85.9	79.4	81.2	90.9	85.9	86.5	86.8
	R	75.3	77.5	72.7	87.4	86.0	88.6	75.5	85.9	90.9	91.2
	F1	78.7	82.6	81.6	86.6	82.6	84.7	82.5	85.9	88.6	88.9
Relation	acc	62.7	71.9	71.8	61.5	64.5	65.7	70.5	76.9	72.1	77.0
	P	56.2	64.0	65.9	51.9	54.0	56.6	61.0	67.9	60.0	65.1
	R	48.6	73.4	66.3	97.7	95.1	87.2	79.5	83.9	98.3	95.6
	F1	52.1	68.4	66.1	67.8	68.9	68.6	69.0	75.1	74.5	77.5

11. Different Sampling Strategies

Tab. 8 presents an ablation study on sampling strategies (non-greedy vs. greedy). We follow the non-greedy sampling setting of VCD [5], where both top-p and temperature are set to 1. As shown, our method consistently improves performance across both sampling strategies.

Table 8. Ablation Study on Sampling Strategies on POPE MSCOCO Adversarial Dataset.

strategy	Method	LLaVA-v1.5		InstructBLIP		mPLUG-Owl2	
		Acc	F1	Acc	F1	Acc	F1
non-greedy	base	$79.0_{\pm 0.51}$	$81.1_{\pm 0.53}$	$71.6_{\pm 0.49}$	$74.7_{\pm 0.46}$	$71.5_{\pm 0.30}$	$76.6_{\pm 0.28}$
	ours	$\mathbf{82.3}_{\pm 0.27}$	$81.1_{\pm 0.31}$	$\mathbf{82.2}_{\pm 0.29}$	$\mathbf{81.8}_{\pm 0.25}$	$\mathbf{83.2}_{\pm 0.27}$	$\mathbf{82.9}_{\pm 0.26}$
greedy	base	80.9	81.6	79.8	81.4	72.5	77.5
	ours	**83.5**	**82.1**	**82.5**	**82.1**	**84.2**	**83.7**

12. Different Model Size

We evaluate our method on different model sizes, 7B and 13B, for LLaVA-v1.5 and InstructBLIP, as shown in Tab. 9. The results indicate consistent improvements across various model sizes. In each model series, the smaller model gets a larger performance boost. With our method, we can achieve high accuracy and detection rates with a smaller 7B model, outperforming a 13B model at its original performance level.

Table 9. Ablation Study on Model Size on LLaVA-QA90 Dataset.

Mothod	LLaVA-v1.5				InstructBLIP			
	7B		13B		7B		13B	
	Acc	Det	Acc	Det	Acc	Det	Acc	Det
base	3.23	3.54	4.78	4.2	3.84	4.07	5.67	4.88
ours	**6.20**	**5.14**	**7.36**	**6.5**	**6.28**	**4.77**	**6.42**	**5.99**

13. Revisiting the Accuracy–Informativeness Trade-off

In the main paper, we report recall and output length alongside CHAIR scores, since our objective is to evaluate models under a balance of *accuracy* and *informativeness*. This choice is deliberate: our early stopping mechanism can be tuned to shorten responses, which naturally reduces CHAIR scores but at the expense of recall and content richness. Consequently, the trade-off introduced by early stopping is an explicit design choice, and it can be adjusted depending on the requirements of a specific application.

Direct comparison with SOTA methods that omit recall and generation length is therefore not entirely fair. Our analysis confirms that CHAIR scores can drop substantially when outputs are truncated, underscoring the importance of jointly reporting recall and length to present a complete view of performance. Without these complementary metrics, lower CHAIR values may simply reflect shorter, less informative responses rather than genuine improvements in visual grounding. To enable a fairer comparison with prior work, we adjust our early stopping threshold to 12%. Under this setting, our method achieves lower CHAIR scores while maintaining competitive recall, thereby outperforming both approaches. This demonstrates that our framework not only mitigates hallucination effectively but also preserves informativeness. Moreover, the adjustable nature of the early stopping mechanism ensures that users can flexibly select the optimal balance between accuracy and informativeness for their specific use cases.

Table 10. Comparison with SOTA Methods with 12% Early Stopping Threshold.

MLLM	Method	$C_s \downarrow$	$C_i \downarrow$	R $\uparrow$	Len $\uparrow$
LLaVA-v1.5	base	48.8	13.4	**78.6**	**99.8**
	PAI [9]	24.8	6.9	-	-
	Middle [4]	25.0	6.7	-	-
	Ours_ES_12%	**23.5**	**6.5**	55.0	54.1

14. Grouping Multiple Influential Tokens with Respect to the Anchor Object

We also explored top-2 variants as a straightforward extension, but they did not yield consistent improvements over the top-1 design. One possible reason is that, for some predicted tokens, supervision is effectively dominated by a single visual token, so forcing a top-2 aggregation can dilute this primary contribution rather than help. As a result, the top-2 scheme may only benefit a subset of objects that are genuinely supported by multiple visual tokens, leading to limited overall gains.

Table 11. Top-k Anchor-Object Influential Token Strategies on the POPE MSCOCO Adversarial Dataset

Norm	LLaVA-v1.5		InstructBLIP		mPLUG-Owl2	
	Acc	F1	Acc	F1	Acc	F1
top-1	83.5	82.1	82.5	82.1	84.2	83.7
top-2	83.1	82.8	82.9	82.5	82.3	82.6

15. Hyper parameter Study

Maximum α_m serves primarily as a *precautionary* upper bound. Because Eq. 13 in the main text naturally bounds α, this maximum threshold is rarely triggered, as evidenced by the low clipping rates reported in Tab. 12.

Table 12. Clipping Rate of α_m at Maximum Value

Task	Discriminative @ POPE Adversarial Setting	Generation @ MSCOCO Subset
GACD (LLaVA-v1.5)	0.04%	0.07%

Nevertheless, to determine the optimal value for α_m and assess its impact on model performance, we conducted a hyperparameter search using LLaVA-v1.5. For discriminative tasks, we evaluated $\alpha_m \in \{1, \ldots, 6\}$ on the POPE dataset. For open-ended generation tasks on a subset of MSCOCO [1], we observed garbled text outputs when $\alpha_m \geq 5$; thus, we restricted our search space to $\alpha_m \in \{1, \ldots, 4\}$. As shown in Tab. 13 and Tab. 14, performance on the POPE discriminative task is relatively insensitive to variations in α_m. Conversely, performance on the generative task initially improves as α_m increases but degrades at higher values. Based on these findings, we set the optimal maximum amplification factor α_m to 5 for discriminative tasks and 3 for generative tasks across all our experiments.

Table 13. α_m Study For Discriminative Task On POPE [6] in MSCOCO Adversarial Setting.

Maximum α_m	1		2		3		4		5		6	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
LLaVA-v1.5	83.4	82.0	83.4	82.0	83.4	82.0	83.4	82.0	83.5	82.1	83.4	82.1

Table 14. Maximum α_m Study For Generation Task on the MSCOCO Subset.

Maximum α_m	1				2				3				4			
	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	Len	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	Len	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	Len	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	Len
LLaVA-v1.5	44.0	11.8	76.2	86.1	41.4	11.1	77.4	84.8	41.0	10.9	77.3	85.0	41.4	10.9	77.3	84.9

Early Stopping Threshold. Our study follows a systematic, data-driven protocol. We conducted a grid search on a subset of the MSCOCO dataset subset following [1]. Recognizing that the visual influence ratio varies across models, we first compute the mean and variance of the EOS visual ratio and the subsequent hallucination rate on a small calibration set. This statistical window defines the range and step size for this study. By searching over the corresponding range, we show results in Tab. 15. These results demonstrate that varying the ES threshold primarily mediates the trade-off between recall and hallucination rate. Our goal is to have balanced recall (R) and instance-level hallucination (C_I), leading us to select thresholds of 7% for LLaVA-v1.5 and LLaVA-v1.6, 25% for InstructBLIP, 2.5% for mPLUG-Owl2 and 10% for InternVL2. We additionally ran an experiment to measure the ES activation rate using LLaVA-v1.5 with ES threshold 7%. As shown in Tab. 16, ES fires on only 8.7% of the test samples, and when it does, the generated responses are on average just 0.7 tokens shorter. This indicates that ES rarely, and only minimally, truncates outputs.

Table 15. Early Stopping Threshold Study on the MSCOCO Subset

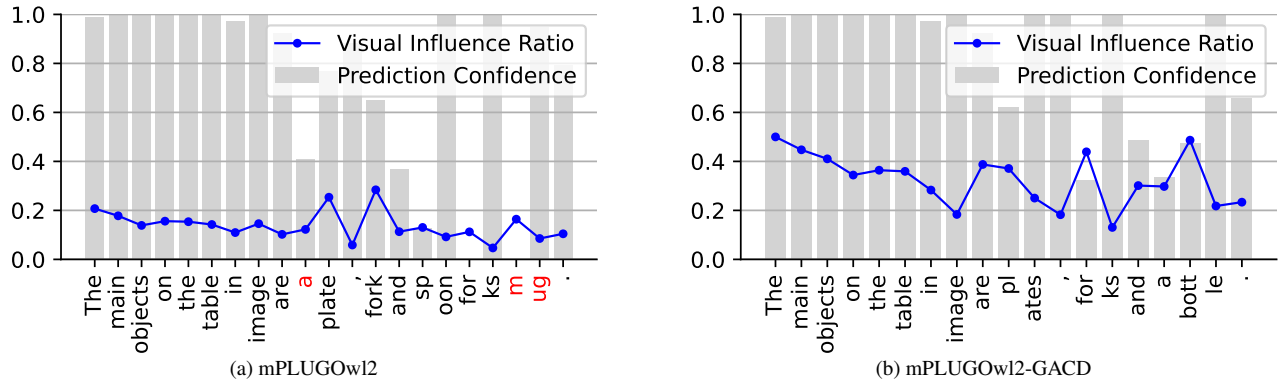
	LLaVA-v1.5				LLaVA-v1.6				IntructBLIP				mPLUG-Owl2				InternVL2			
	6%	7%	8%	9%	5%	7%	9%	11%	15%	20%	25%	30%	2%	2.5%	3%	3.5%	8%	10%	12%	14%
C_S	45.6	41.0	36.6	31.4	29.0	26.0	23.0	17.8	52.6	51.4	47.4	36.0	51.8	45.0	41.2	41.0	37.0	35.2	34.6	32.9
C_I	11.5	10.9	10.2	10.2	8.5	8.1	7.8	7.5	15.0	14.3	13.4	11.7	13.7	12.4	11.0	11.0	8.6	8.1	8.0	7.9
R	79.7	77.3	75.2	70.8	68.5	63.0	58.8	53.0	75.1	74.4	72.3	68.8	77.7	74.9	73.8	73.5	65.8	65.4	65.4	64.0
L_{en}	$92._0$	$85._0$	$75._4$	$63._9$	$119._1$	$101._8$	$81._1$	$62._7$	$107._9$	$103._4$	$93._9$	$74._3$	$89._1$	$83._5$	$78._9$	$77._8$	$180._4$	$175._5$	$170._6$	$162._2$

16. Confidence and Visual Influence

Low confidence often signals potential failure modes in base MLLMs. Here, we demonstrate that our method not only improves accuracy but also increases model confidence. It remains effective even in low-confidence regions for three rea-

Table 16. Activation Rate of the Early Stopping on the MSCOCO Subset

	Methods	Activate Percentage	Average Length
LLaVA-v1.5 (7%)	base	-	85.1
	ours	8.7%	84.4



(a) mPLUGOwl2 (b) mPLUGOwl2-GACD

Figure 1. Comparison of prediction confidence with and without GACD. (a) Without GACD, mPLUGOwl2 exhibits low confidence in hallucinated predictions and near-zero confidence in the initial predictions for 'forks' and 'mug'. (b) With GACD, mPLUGOwl2's confidence increases alongside the visual influence ratio, effectively mitigating hallucinations.

sons: 1) We aggregate token gradients at the component level (Eq. **??**) rather than using individual token gradients which yields robustness against local gradient noise. 2) We adjust influence towards visual tokens which consistently reduces the hallucination likelihood; 3) Empirically, low model confidence does not correlate with noisy gradients. In our experiments, pretrained MLLMs usually maintain meaningful gradient signals even at low confidence levels. Fig. 1 shows an example where the baseline model mPLUG-Owl2 exhibits low confidence in hallucinated predictions and near-zero confidence in the initial predictions for 'forks' and 'mug'. With GACD, prediction confidence increases alongside the visual influence ratio, with the minimum confidence rising to over 30%.

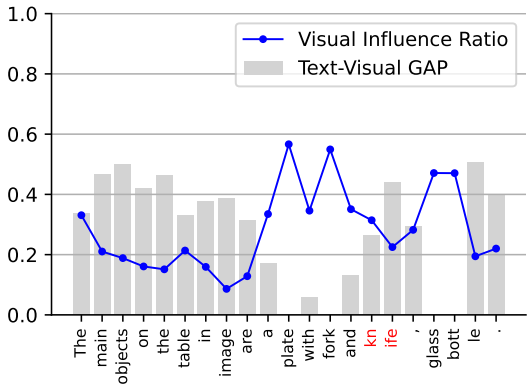
17. Additional Implementation and Experimental Details

We identify noun tokens using the spaCy library due to its lightweight, CPU-only operation; this component is interchangeable with any alternative noun detector. At each step, if the current token is classified as a noun and at least one noun has appeared in previous outputs, we trigger noun-only grouping. In this mode, we inject noun-related visual tokens into the negative guidance; otherwise, negative guidance is computed using only text tokens. For experiments on the Amber dataset [12], we adopt the original data splits and evaluation metrics. In the MSCOCO [7] subset, we follow the data partitioning and evaluation protocol of Deng et al. [1], with splits available in their official repository. For the LLaVA-QA90 [8], MME [2] and POPE [6] datasets, our setup replicates that of Leng et al. [5] and use their provided scoring scripts for LLaVA-QA90. Experiments on MMBench [10], MM-Vet [13] follow the VLMEvalKit_InternVL2_5 repository. All comparison methods are executed using their official code; we only modify them to enforce greedy sampling and a uniform maximum generation length to align with our experimental settings.

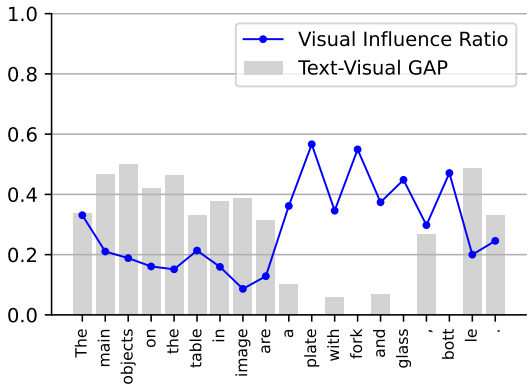
18. Influence Ratio in VQA

Fig. 2 illustrates the visual influence ratio across outputs in VQA tasks, comparing baseline predictions with those obtained after applying GACD. The results confirm that text tokens dominate influence across MLLMs, including InternVL2, which exhibits a relatively high visual influence ratio. As shown in Fig. 3 of the main paper, the overall 60%–100% visual influence ratio across the POPE dataset suggests that visual inputs predominantly determine object existence in VQA tasks. GACD enhances visual influence, effectively balancing text-visual bias. Furthermore, the visualization on InternVL2 demonstrates that the co-occurrence hallucination 'knife' persists despite a high visual influence. GACD successfully eliminates this co-occurrence hallucination.

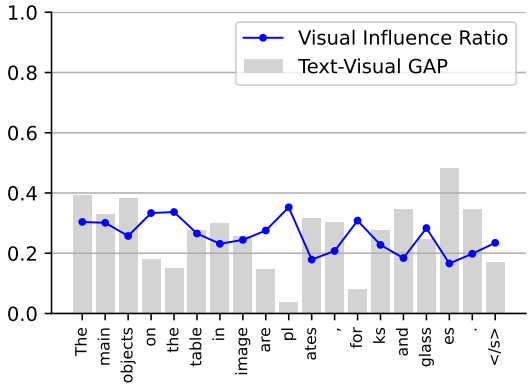
(a) What are the main objects on the table in the image?



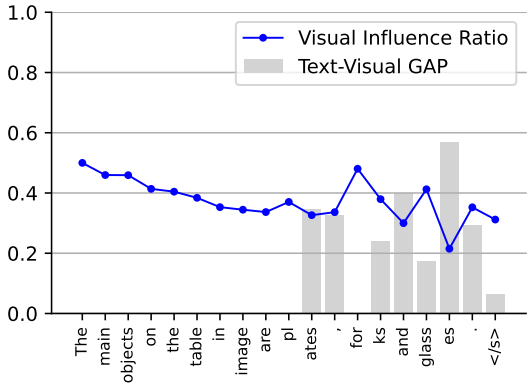
(b) LLaVAv1.5



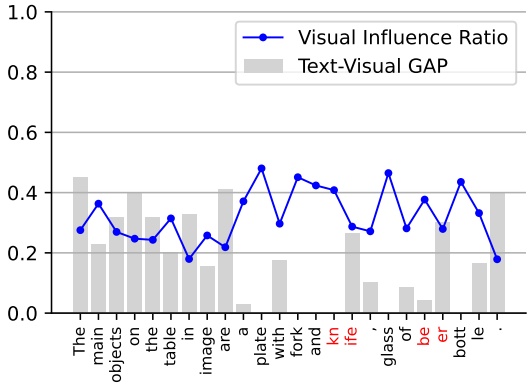
(c) LLaVAv1.5-GACD



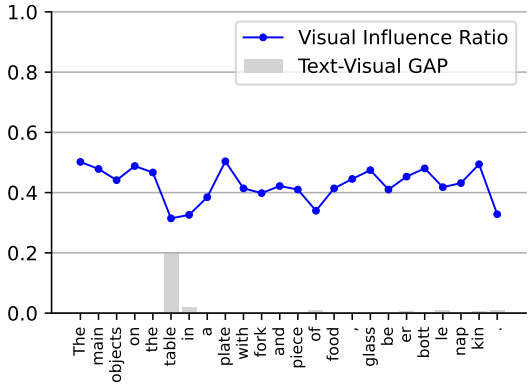
(d) InstructBLIP



(e) InstructBLIP-GACD



(f) InternVL2



(g) InternVL2-GACD

Figure 2. Influence Ratio across Predicted Tokens in VQA: (left) Baseline predictions; (right) Predictions with GACD. GACD effectively mitigate Text-Visual GAP, balancing text-visual bias. (f) The original InternVL2 shows a dominant visual influence ratio at the hallucinated prediction 'knife', indicating a co-occurrence bias that remains unaddressed even with dominant visual influence. (g) GACD successfully eliminates co-occurrence hallucinations, including 'knife'.

19. Influence Ratio in Image Caption

We further visualize the influence ratio in the image captioning task. Fig. 3 shows that in the baseline LLaVA-v1.5, the influence gap between previous output tokens and visual tokens widens as more tokens are generated. However, GACD effectively narrows this gap, preventing visual information from being forgotten and thereby reducing hallucinations.

LLaVA: The image features a black dog lying on the floor in a kitchen, resting comfortably on a tiled surface. The dog is positioned near a bowl, possibly enjoying a meal or a snack. In the kitchen, there are several items scattered around. A bottle can be seen on the left side of the dog, while a cup is located closer to the center. A spoon is placed near the right side of the dog, and a bowl is situated on the right side of the scene.

LLaVA-VA: The image features a black dog lying on the floor of a kitchen, resting comfortably on a tile floor. The dog is positioned near a bowl, a toy, and a bag of cat food. The bowl is placed on the floor, while the toy is located closer to the dog.

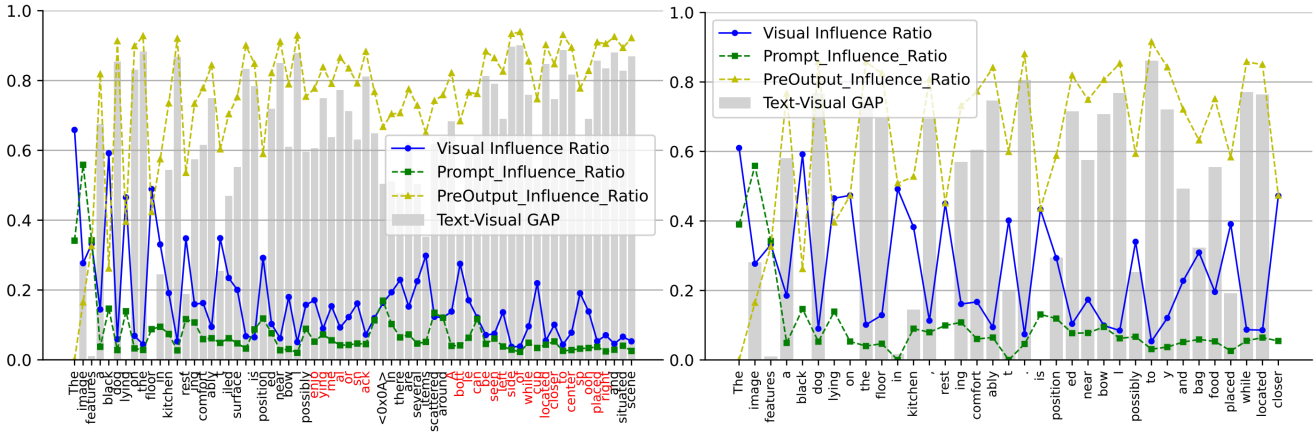


Figure 3. Comparison of influence ratios across predicted tokens with and without GACD. (Left) Without GACD, the influence gap between previous output tokens and visual tokens widens as more tokens are generated. (Right) With GACD, the gap is periodically narrowed to nearly zero, mitigating this trend and reducing hallucination.

20. Qualitative Example on Occluded Images

Image caption w/o Ours: The image features a white truck parked on a city street, with graffiti covering its side. The truck is positioned near a crosswalk, and there are several other vehicles in the scene, including a car and a bus. In addition to the vehicles, there are a few people walking around the area.

Image caption w Ours: The image features a large white truck parked on a city street, with graffiti covering its side. The truck is positioned near a crosswalk, and there is another vehicle visible in the background. Additionally, there are buildings in the scene, suggesting an urban setting.

Figure 4. Example of our method applied to an occluded image.

We include a qualitative example in Fig. 4, where a sedan and a building are partially occluded by a white truck, our method prevents the baseline model from hallucinating of persons and vehicles behind the occluding object. This demonstrates our method remains effective on images consisting of occlusions. Image caption w/o Ours: The image features a white truck parked on a city street, with graffiti covering its side. The truck is positioned near a crosswalk, and there are several other vehicles in the scene, including a car and a bus. In addition to the vehicles, there are a few people walking around the area. Image caption w Ours: The image features a large white truck parked on a city street, with graffiti covering its side. The truck is positioned near a crosswalk, and there is another vehicle visible in the background. Additionally, there are buildings in the scene, suggesting an urban setting.

21. Broader Impacts

Our method enhances the factual reliability of multi-modal language models, not only for vision–language tasks but also for modalities such as video and audio, by mitigating hallucinations at inference time. This improvement has several positive societal implications: it can make systems for visual question answering, assistive technologies for the visually impaired, and automated image captioning more dependable, thereby increasing user trust and safety; it can power educational tools that generate accurate descriptions of complex diagrams or historical media, benefiting learners and instructors; and in critical domains such as medical imaging or remote sensing, it can reduce spurious outputs and support more robust decision-making. Conversely, if deployed within surveillance or facial-recognition systems, stronger multi-modal grounding could facilitate more intrusive inferences about individuals from visual data, exacerbating privacy risks.

References

[1] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024. 6, 7

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 3, 7

[3] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[4] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 5

[5] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 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, 2024. 4, 7

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3, 6, 7

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[9] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024. 5

[10] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 2, 7

[11] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[12] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 7

[13] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 7