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

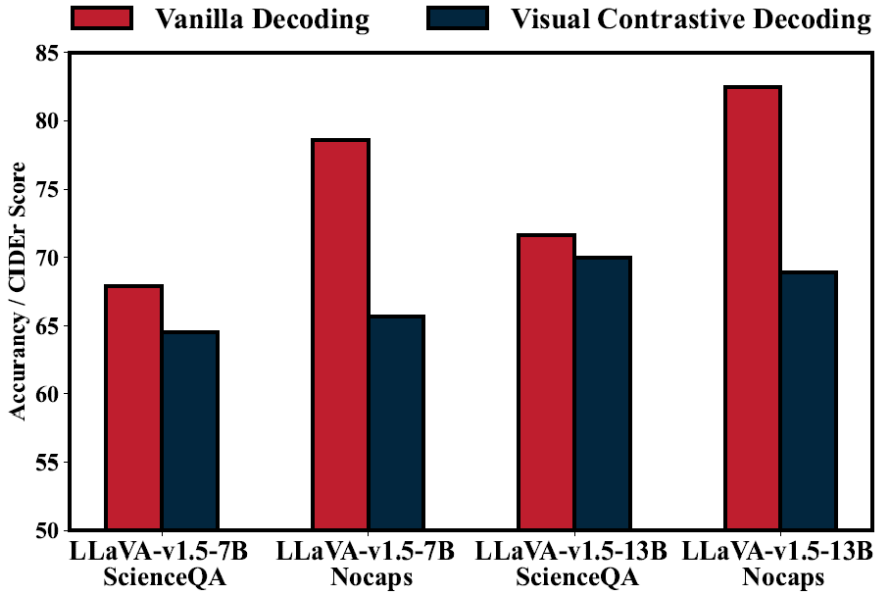
CVPR 2025

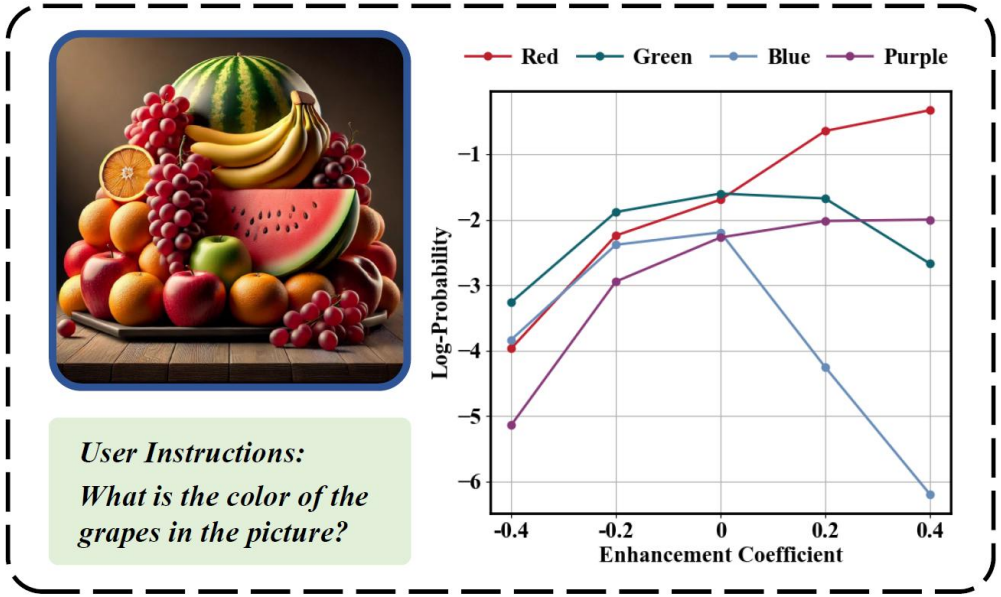
Hao Yin, Guangzong Si, Zilei Wang

中国科学技术大学
University of Science and Technology of China

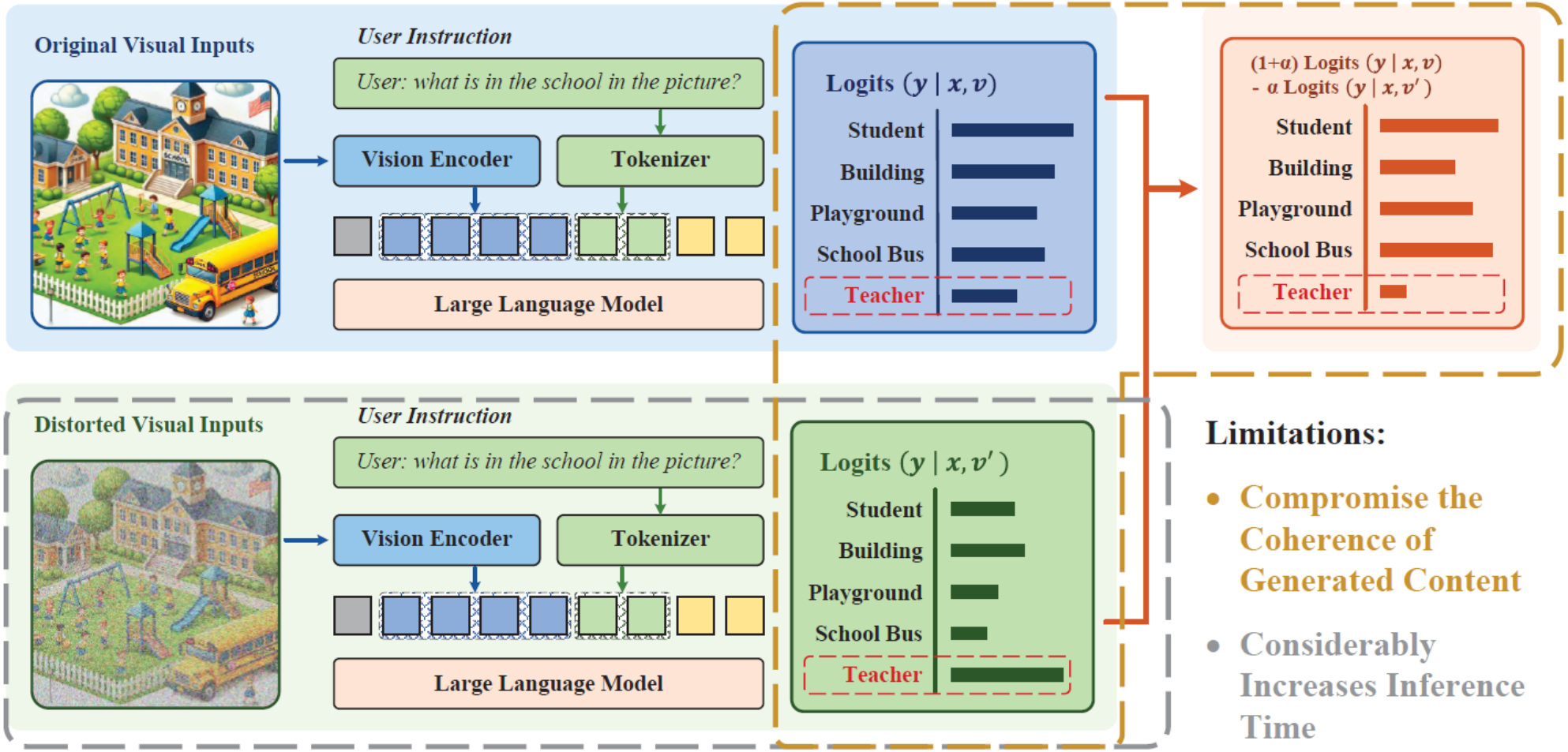
Contributions

- We identify the negative impacts of contrastive decoding methods on both the quality of generated content and model inference speed.
- We analyze the modality fusion mechanism in MLLMs, highlighting its insufficient attention to visual information.





Limitations of Contrastive Decoding

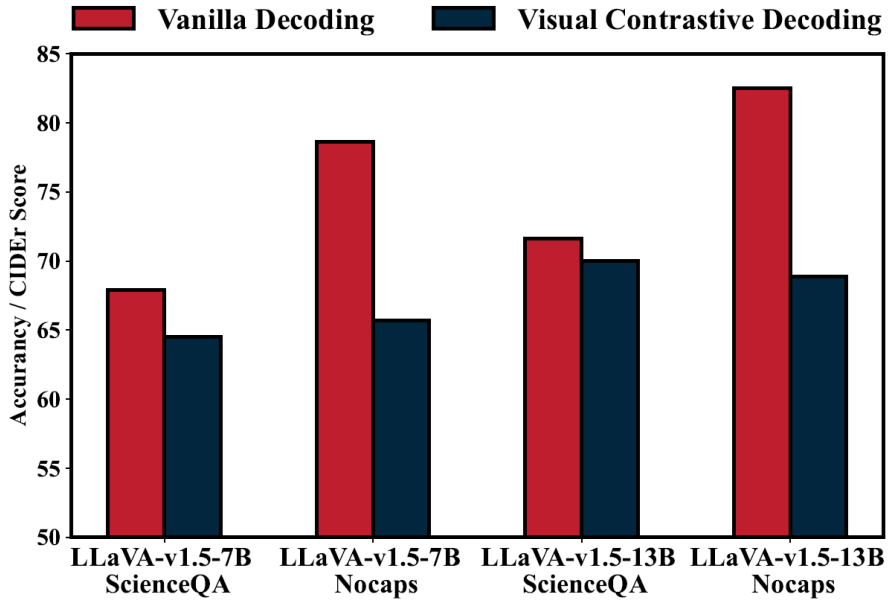


Limitations:

- Compromise the Coherence of Generated Content
- Considerably Increases Inference Time

Limitations of Contrastive Decoding

- While reducing over-reliance on language priors, these methods may **compromise the coherence and accuracy of generated content**.
- Contrastive decoding necessitates separate processing of the original and contrastive inputs, which **considerably increases inference time**.



Model	Method	ScienceQA	Nocaps
LLaVA-v1.5-7B	Regular	0.141s	0.456s
	VCD	**0.293s**	**1.086s**
LLaVA-v1.5-13B	Regular	0.222s	0.602s
	VCD	**0.459s**	**1.372s**

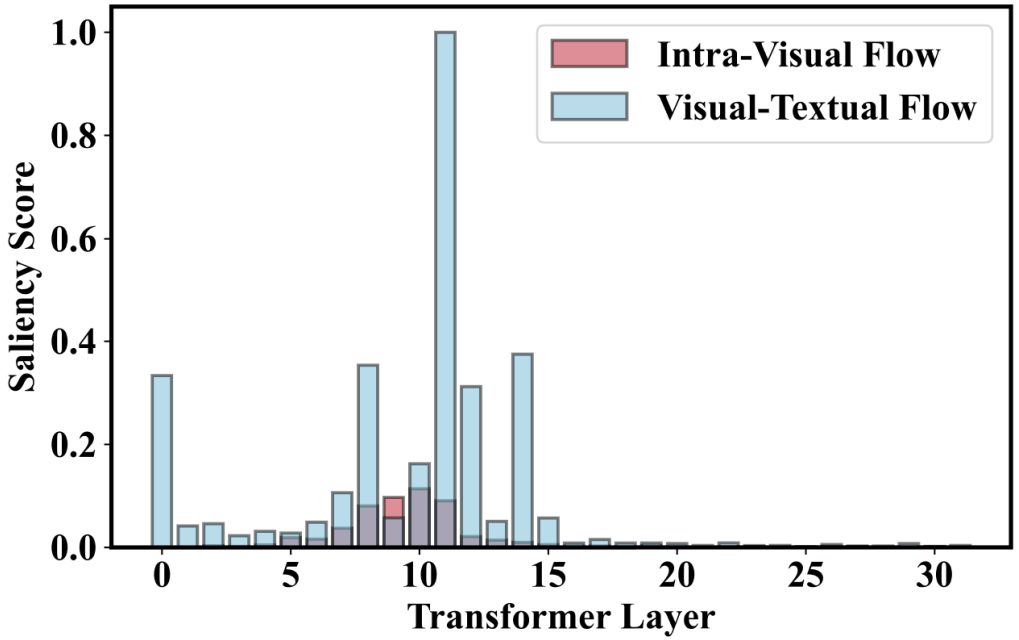
Table 1. **Impact of VCD on Model Inference Speed.** The table shows the average inference time per sample (in seconds) on the ScienceQA and Nocaps benchmarks. Applying the VCD method nearly doubled the inference time of the model.

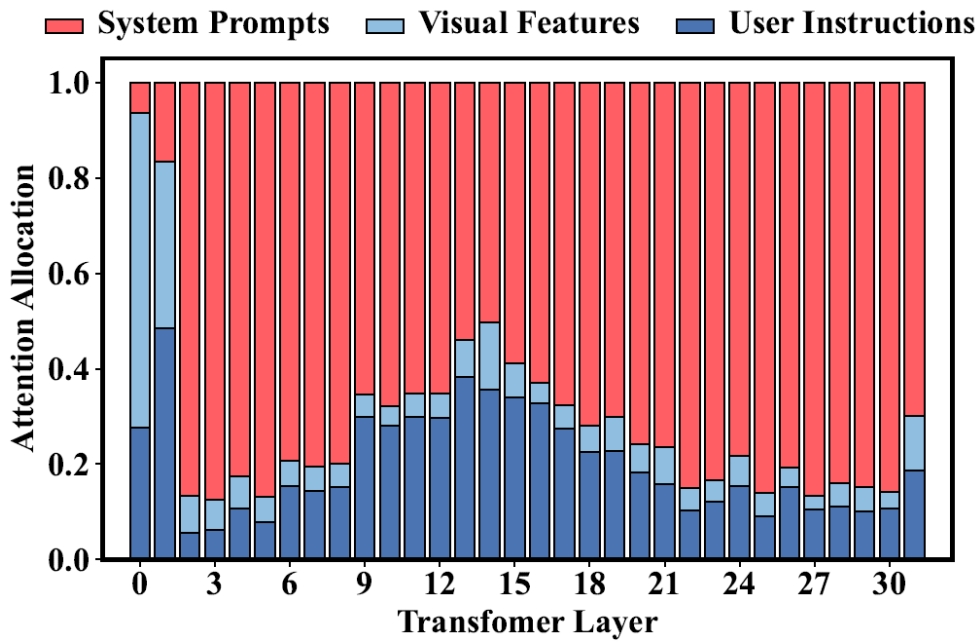
Rethinking Hypothesis of Hallucination

- Methods for hallucination mitigation via contrastive decoding often trace object hallucinations to **excessive dependence on linguistic priors**.
- However, we propose an alternative perspective: object hallucinations primarily stem from the model's **insufficient attention to visual signals**.

Visual Neglect in Modal Fusion

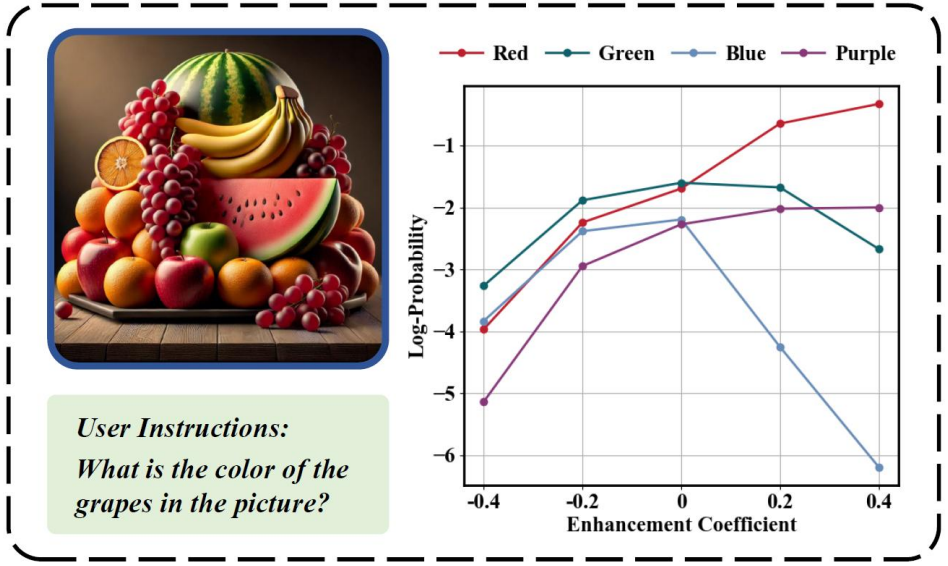
- The model performs the crucial fusion of visual and textual modalities **in the middle layers**, creating cross-modal semantic representations that drive the final predictions.
- During this critical fusion process, the model demonstrates **inadequate attention to the visual modality**.

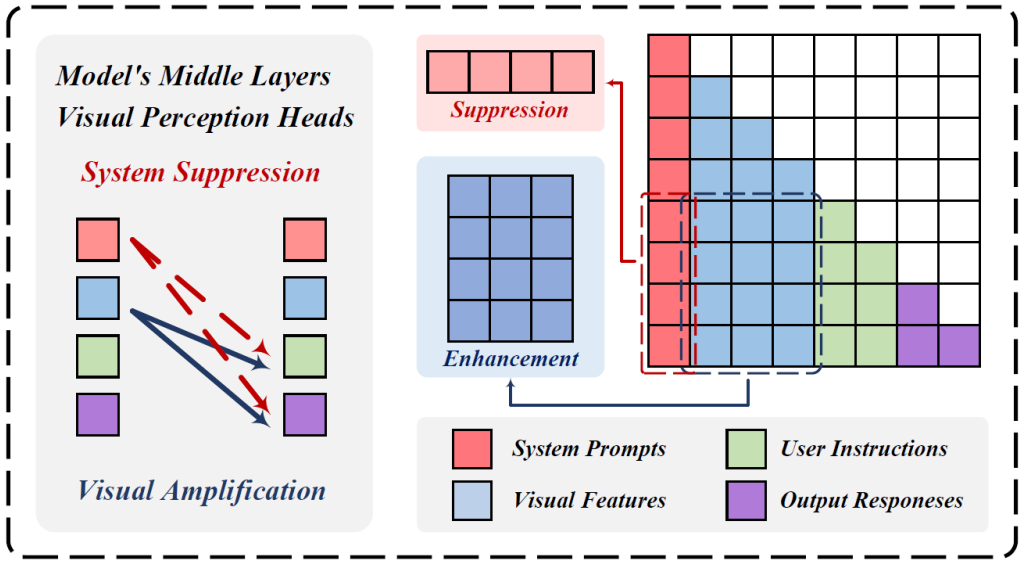




Method: Visual Amplification Fusion

- Visual Amplification Fusion specifically **amplifies visual signals at these middle layers**, enabling the model to **capture more distinctive visual features during fusion**
- This technique not only strengthens the model's visual representations but also retains the beneficial influence of language priors, thus **preserving content quality**





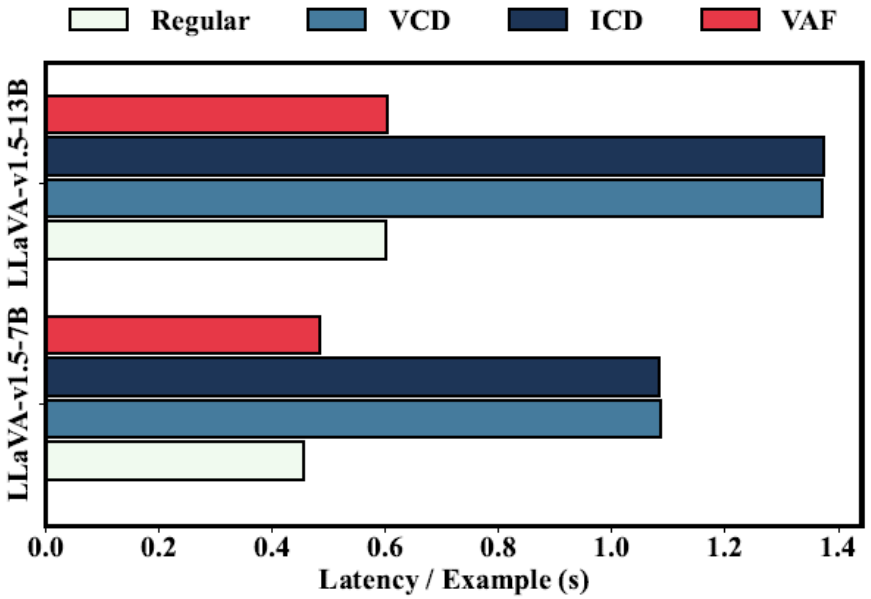
Results: Hallucination Mitigation

- Visual Amplification Fusion demonstrates **comparable effectiveness** to existing methods in alleviating object hallucinations.

Category	Method	LLaVA-v1.5-7B		LLaVA-v1.5-13B		Qwen-VL-Chat-7B	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Random	Regular	87.8 ↑ 0.0	87.5 ↑ 0.0	87.6 ↑ 0.0	87.4 ↑ 0.0	88.2 ↑ 0.0	87.9 ↑ 0.0
	VCD	88.4 ↑ 0.6	87.7 ↑ 0.2	88.9 ↑ 1.3	87.8 ↑ 0.4	89.1 ↑ 0.9	88.4 ↑ 0.5
	ICD	88.1 ↑ 0.3	87.6 ↑ 0.1	88.1 ↑ 0.5	87.6 ↑ 0.2	88.9 ↑ 0.7	88.1 ↑ 0.2
	VAF	**89.6** ↑ 1.8	**89.3** ↑ 1.8	**90.1** ↑ 2.5	**89.9** ↑ 2.5	**90.0** ↑ 1.8	**89.7** ↑ 1.8
Popular	Regular	82.5 ↑ 0.0	83.2 ↑ 0.0	82.7 ↑ 0.0	84.1 ↑ 0.0	82.4 ↑ 0.0	83.1 ↑ 0.0
	VCD	83.1 ↑ 0.6	84.1 ↑ 0.9	83.7 ↑ 1.0	85.1 ↑ 1.0	83.0 ↑ 0.6	84.1 ↑ 1.0
	ICD	82.1 ↓ 0.4	82.9 ↓ 0.3	82.9 ↑ 0.2	84.3 ↑ 0.2	83.2 ↑ 0.8	84.5 ↑ 1.4
	VAF	**84.5** ↑ 2.0	**84.9** ↑ 1.7	**85.2** ↑ 2.5	**86.4** ↑ 2.3	**84.9** ↑ 2.5	**85.1** ↑ 2.0
Adversarial	Regular	77.6 ↑ 0.0	79.4 ↑ 0.0	77.8 ↑ 0.0	79.5 ↑ 0.0	77.2 ↑ 0.0	78.9 ↑ 0.0
	VCD	78.1 ↑ 0.5	79.6 ↑ 0.2	78.2 ↑ 0.4	79.7 ↑ 0.2	78.8 ↑ 1.6	80.1 ↑ 1.2
	ICD	78.5 ↑ 0.9	79.9 ↑ 0.5	79.1 ↑ 1.3	80.1 ↑ 0.6	78.1 ↑ 0.9	79.2 ↑ 0.3
	VAF	**80.1** ↑ 2.5	**81.0** ↑ 1.6	**80.7** ↑ 2.9	**81.7** ↑ 2.2	**80.4** ↑ 3.2	**81.2** ↑ 2.3

Results: Side Effects and Trade-offs

- Alternative approaches often result in **substantial degradation in output quality and inference efficiency**
- Visual Amplification Fusion **maintains both with negligible compromise**



Model	Decoding	ScienceQA	Nocaps
		Accurancy	CIDEr
LLaVA-v1.5-7B	Regular	**68.0**	**78.7**
	VCD	64.5	65.7
	ICD	62.4	62.3
	VAF	**68.5**	**78.8**
LLaVA-v1.5-13B	Regular	**71.6**	**82.6**
	VCD	70.0	68.9
	ICD	69.2	60.3
	VAF	**71.7**	**82.3**

Conclusions

Limitations of Contrastive Decoding

- We identify two key drawbacks of using contrastive decoding to mitigate hallucinations: **reduced quality of generated content** and **slower inference speed**.

Insights into Object Hallucination

- The model performs the crucial fusion of visual and textual modalities **in the middle layers**, creating cross-modal semantic representations that drive the final predictions.
- During this critical fusion process, the model demonstrates **inadequate attention to the visual modality**.