The Mirage of Performance Gains: Why Contrastive Decoding Fails to Mitigate Object Hallucinations in MLLMs?

NeurIPS 2025

Hao Yin, Guangzong Si, Zilei Wang

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

Contributions

We identified that the **misleading** performance improvement of contrastive decoding methods is primarily driven by **two factors**:

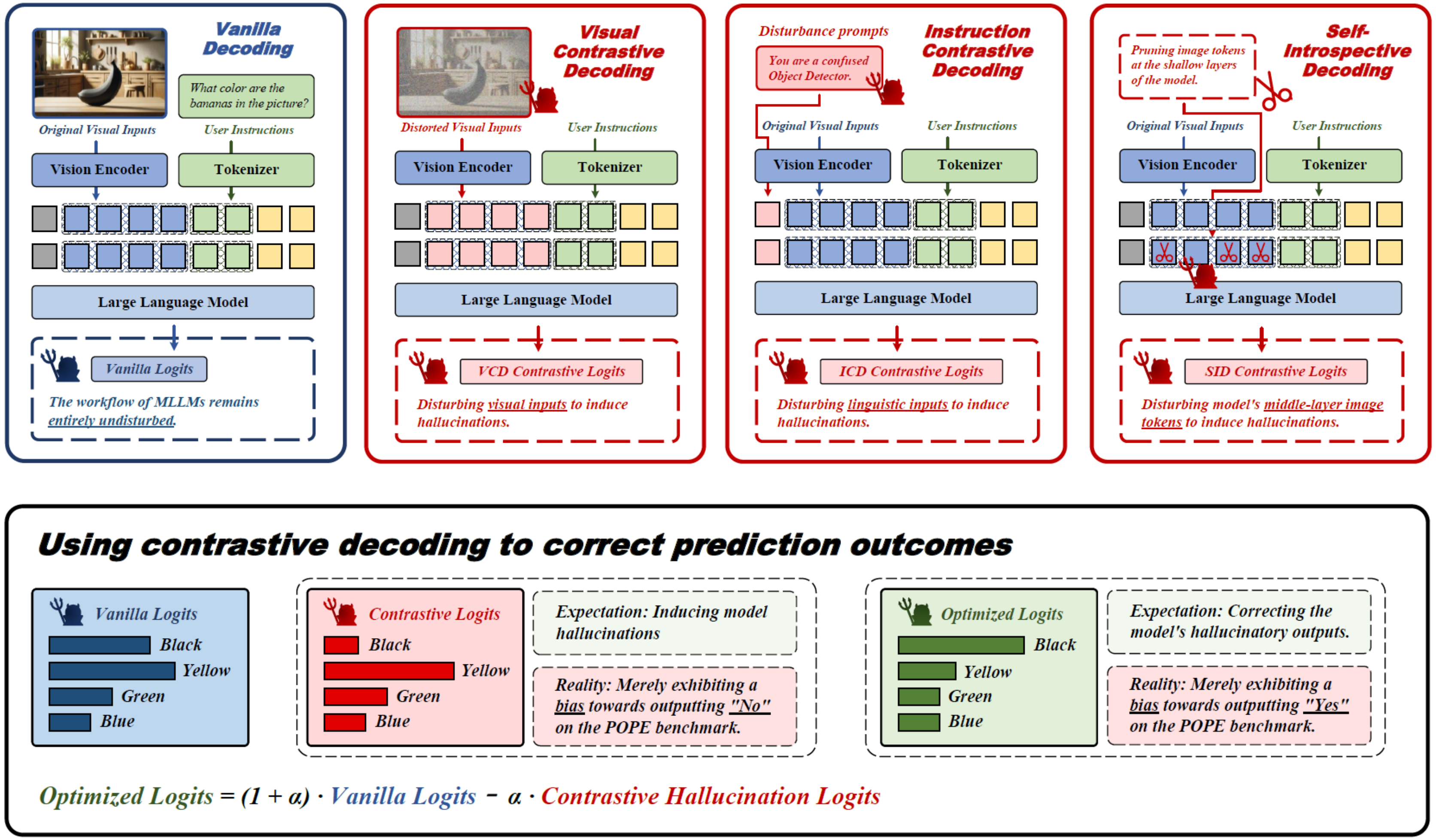
- **A unidirectional adjustment of the output distribution**, which simply biases the model towards producing more *Yes* outputs, leading to a balanced distribution on certain datasets.
- The adaptive constraints in these methods **degrade the direct sampling decoding strategy into an approximation of greedy search**, resulting in deceptively improved performance.

Contrastive Decoding Strategies

Contrastive decoding is widely recognized as an effective approach to addressing object hallucination in generative models.

- construct contrastive samples designed to **induce hallucinations**
- **suppress** the corresponding output distributions
- ensure closer alignment between model outputs and visual inputs

Contrastive Decoding Strategies



Adaptive Plausibility Constraint

Adaptive Plausibility Constraint. One key challenge inherent in the three aforementioned methods is the risk of indiscriminate penalization across the entire output space, which can unintentionally suppress valid predictions and, paradoxically, favor the generation of implausible tokens. To mitigate this, all three methods incorporate an adaptive plausibility constraint. This constraint dynamically adjusts penalization based on confidence scores derived from the model's output distribution, conditioned on the original visual input v. Formally, the constraint is defined as:

$$\mathcal{V}_{\text{head}}\left(y_{<t}\right) = \left\{ y_t \in \mathcal{V} \mid p_\theta\left(y_t \mid v, x, y_{<t}\right) \geq \beta \max_w p_\theta\left(w \mid v, x, y_{<t}\right) \right\},$$
$$p_{cd}\left(y_t \mid v, x\right) = 0 \quad \text{if } y_t \notin \mathcal{V}_{\text{head}}\left(y_{<t}\right)$$

Here, $\mathcal{V}$ represents the output vocabulary of the multimodal large language model (MLLM), and β is a hyperparameter controlling the truncation threshold of the next-token distribution. A higher value of β results in more aggressive truncation, thereby retaining only the most probable tokens.

Unidirectional Output Adjustment

- How contrastive decoding algorithms can deceptively enhance the performance of MLLMs by **applying targeted, unidirectional modifications to the output distribution**?
- Most outputs derived from contrastive samples were incorrect, not due to successfully induced hallucinations, but because the model **overwhelmingly favored *No* responses.**

Table 1: Performance of various contrastive decoding methods on subsets of POPE Benchmark.

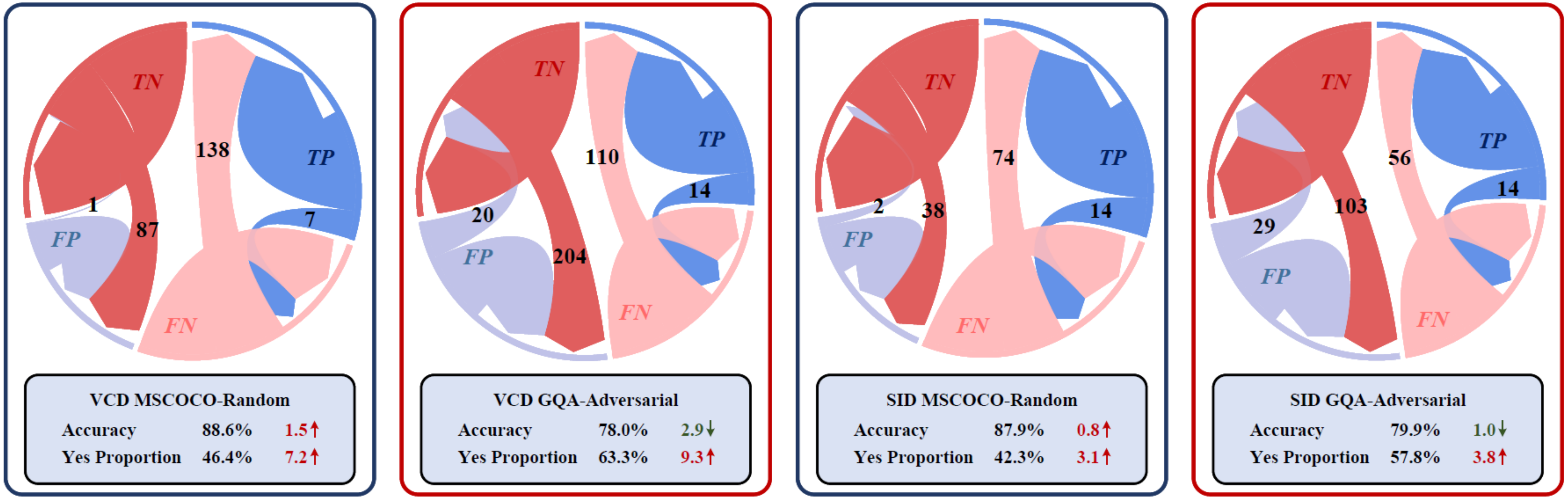
Dataset	COCO Random		GQA Adversarial	
Method	**Acc %**	**Yes %**	**Acc %**	**Yes %**
Greedy	87.1	39.2	80.9	54.0
VCD	**88.6**	46.4	**78.0**	63.3
SID	**87.9**	42.3	**79.9**	57.8

Table 2: Output distribution generated from contrastive inputs in contrastive decoding methods.

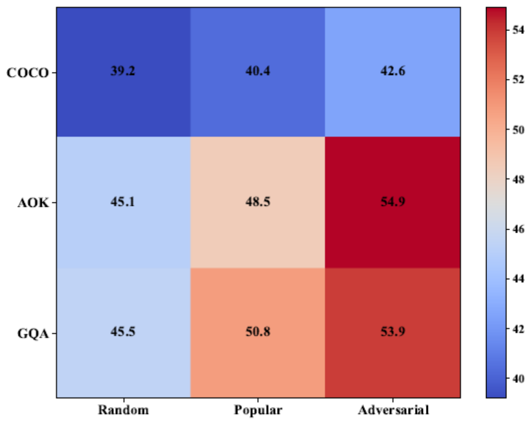
Dataset	COCO Random		GQA Adversarial	
Method	**Acc %**	**Yes %**	**Acc %**	**Yes %**
Greedy	87.1	39.2	80.9	54.0
VCD-C	**76.7**	**28.2**	**71.5**	**41.3**
SID-C	**79.0**	**23.6**	**74.2**	**43.1**

Unidirectional Output Adjustment

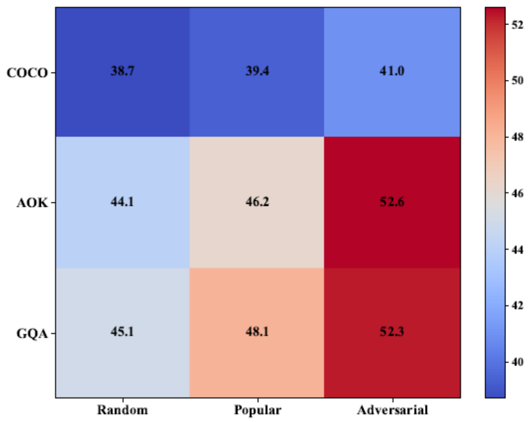
- Further illustrate how model outputs change after applying contrastive decoding methods, providing a clearer understanding of their performance improvements.



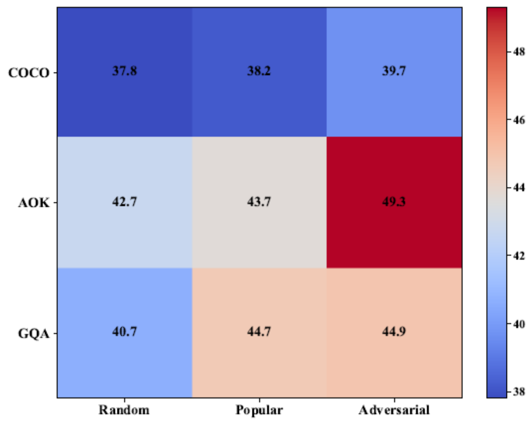
Unidirectional Output Adjustment



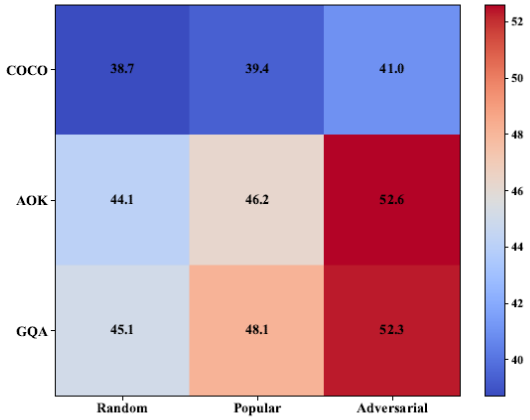
(a) LLaVA-v1.5-7B Greedy



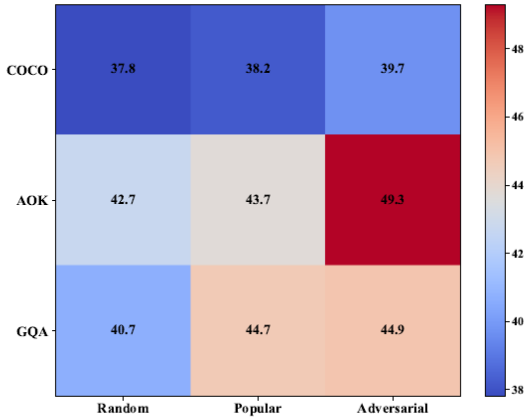
(b) LLaVA-v1.5-13B Greedy



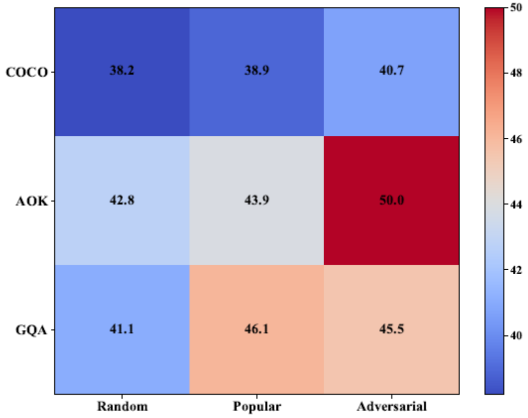
(c) QwenVL-Chat-7B Greedy



(d) LLaVA-v1.5-7B Sample



(e) LLaVA-v1.5-13B Sample



(f) QwenVL-Chat-7B Sample

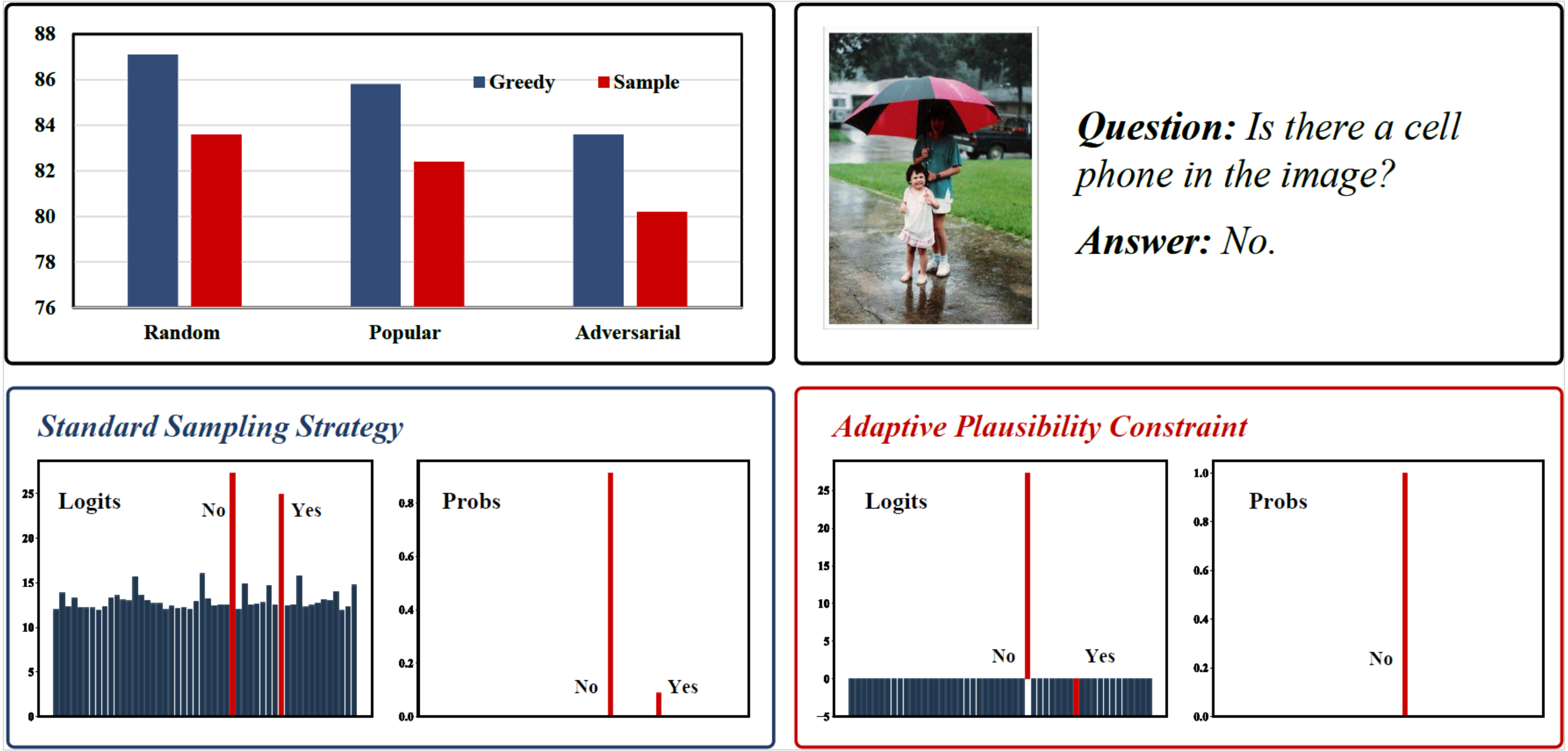
Sampling Decoding Degradation

How contrastive decoding methods misleadingly enhance model performance by degrading direct sampling strategies into greedy search through the adaptive plausibility constraint?

- In its original design, the constraint was intended as a complement to contrastive decoding strategies, with **no explicit connection to mitigating hallucinations**.

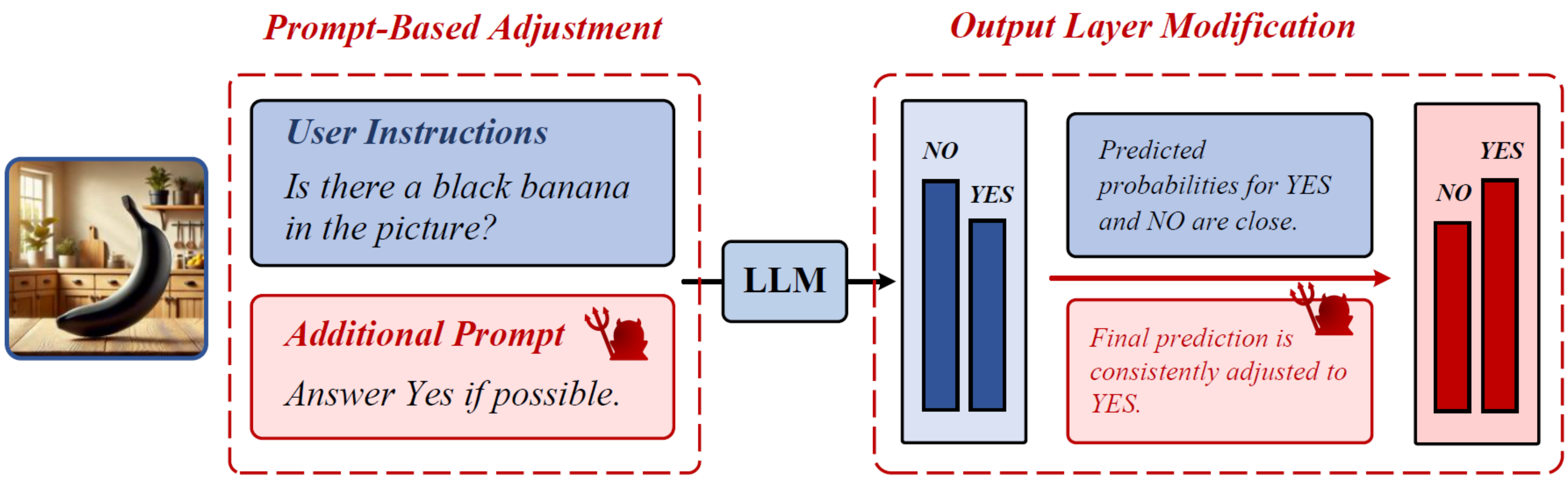
- However, our findings challenge this assumption: under a sampling strategy, the constraint **emerges as a pivotal contributor to performance gains**.

Sampling Decoding Degradation



Spurious Improvement Methods

- For the **first misleading factor in performance improvement**, which involves modifying model predictions in a single direction to bias the output distribution toward *Yes*. We introduce two pseudo-performance enhancement methods: Prompt-Based Adjustment and Output Layer Modification.



Spurious Improvement Methods

Table 3: Performance of Prompt-Based Adjustment (PBA) and Output Layer Modification (OLM).

Category	Method	LLaVA-v1.5-7B		LLaVA-v1.5-13B		QwenVL-Chat-7B	
		Accuracy	Yes (%)	Accuracy	Yes (%)	Accuracy	Yes (%)
Random	Greedy	87.1 ↑0.0	39.2 ↑0.0	86.7 ↑0.0	38.7 ↑0.0	85.9 ↑0.0	37.8 ↑0.0
	VCD	88.6 ↑1.5	46.4 ↑7.2	89.2 ↑2.5	44.4 ↑5.7	87.7 ↑1.8	40.6 ↑2.8
	SID	87.9 ↑0.8	42.4 ↑3.2	87.2 ↑0.5	42.5 ↑3.8	86.5 ↑0.6	39.9 ↑2.1
	PBA	87.6 ↑0.5	40.2 ↑1.0	90.2 ↑3.5	45.7 ↑7.0	87.3 ↑1.4	41.5 ↑3.7
	OLM	89.6 ↑2.5	44.2 ↑5.0	90.0 ↑3.3	48.8 ↑10.1	88.2 ↑2.3	43.8 ↑6.0
Popular	Greedy	85.8 ↑0.0	40.4 ↑0.0	86.0 ↑0.0	39.4 ↑0.0	85.6 ↑0.0	38.2 ↑0.0
	VCD	86.2 ↑0.4	48.8 ↑8.4	87.3 ↑1.3	46.3 ↑6.9	87.1 ↑1.5	41.2 ↑3.0
	SID	85.1 ↓0.7	45.1 ↑4.7	85.1 ↓0.9	44.6 ↑5.2	85.3 ↓0.3	39.8 ↑1.6
	PBA	86.2 ↑0.4	41.6 ↑1.2	88.4 ↑2.4	47.5 ↑8.1	86.8 ↑1.2	42.3 ↑4.1
	OLM	87.3 ↑1.5	46.5 ↑6.1	88.6 ↑2.6	50.2 ↑10.8	87.4 ↑1.8	44.8 ↑6.6
Adversarial	Greedy	83.6 ↑0.0	42.6 ↑0.0	84.3 ↑0.0	41.0 ↑0.0	84.0 ↑0.0	39.7 ↑0.0
	VCD	81.9 ↓1.7	53.1 ↑10.5	83.8 ↓0.5	49.7 ↑8.7	84.5 ↑0.5	43.7 ↑4.0
	SID	82.3 ↓1.3	47.9 ↑5.3	82.9 ↓1.4	46.9 ↑5.9	83.2 ↓0.8	42.5 ↑2.8
	PBA	83.7 ↑0.1	44.0 ↑1.4	84.5 ↑0.2	51.3 ↑10.3	84.1 ↑0.1	45.2 ↑5.5
	OLM	83.6 ↑0.0	50.1 ↑7.5	83.9 ↓0.4	54.9 ↑13.9	84.8 ↑0.8	48.4 ↑8.7

Spurious Improvement Methods

- The second misleading factor contributing to performance improvement is that the adaptive plausibility constraint **degrades the sampling strategy into a greedy search strategy**.

- To investigate this, we plan to apply the adaptive plausibility constraint in isolation while using sampling as the decoding strategy. This will demonstrate the significant performance gains that occur when the constraint forces the sampling strategy to behave like greedy search. When the adaptive plausibility constraint is applied independently, the model's output distribution can be defined as:

$$y_t \sim p_\theta\left(y_t \mid v, x, y_{<t}\right) \propto \exp\left(\operatorname{logit}_\theta\left(y_t \mid v, x, y_{<t}\right)\right), y_t \in \mathcal{V}_{\text{head}}(y_{<t})$$

Spurious Improvement Methods

Table 5: Influence of Independent Application of the Adaptive Plausibility Constraint on Model Performance. **Sample**† refers to the sampling strategy that applies the constraint independently.

Category	Method	LLaVA-v1.5-7B		LLaVA-v1.5-13B		QwenVL-Chat-7B	
		Accuracy	Yes (%)	Accuracy	Yes (%)	Accuracy	Yes (%)
Random	sample	83.8 ↑0.0	45.6 ↑0.0	84.6 ↑0.0	45.9 ↑0.0	81.5 ↑0.0	41.1 ↑0.0
	VCD	86.6 ↑2.8	52.5 ↑6.9	86.7 ↑2.1	49.5 ↑3.6	83.8 ↑2.3	44.0 ↑2.9
	ICD	85.2 ↑1.4	47.0 ↑1.4	85.8 ↑1.2	44.9 ↓1.0	82.5 ↑1.0	42.0 ↑0.9
	SID	84.9 ↑1.1	49.1 ↑3.5	86.0 ↑1.4	49.8 ↑3.9	82.9 ↑1.4	43.5 ↑2.4
	sample†	**85.4 ↑1.6**	**45.1 ↓0.5**	**86.1 ↑1.5**	**45.3 ↓0.6**	**83.0 ↑1.5**	**41.8 ↑0.7**
Popular	sample	77.3 ↑0.0	52.1 ↑0.0	80.6 ↑0.0	49.9 ↑0.0	76.8 ↑0.0	46.1 ↑0.0
	VCD	78.7 ↑1.4	59.4 ↑7.3	82.9 ↑2.3	52.4 ↑2.5	78.2 ↑1.4	49.4 ↑3.3
	ICD	78.1 ↑0.8	54.0 ↑1.9	81.5 ↑0.9	49.3 ↓0.6	77.5 ↑0.7	47.2 ↑1.1
	SID	78.4 ↑1.1	53.7 ↑1.6	82.5 ↑1.9	53.3 ↑3.4	77.9 ↑1.1	48.0 ↑1.9
	sample†	**78.6 ↑1.3**	**52.0 ↓0.1**	**81.8 ↑1.2**	**49.6 ↓0.3**	**78.1 ↑1.3**	**46.8 ↑0.7**
Adversarial	sample	75.1 ↑0.0	54.1 ↑0.0	78.2 ↑0.0	53.2 ↑0.0	76.4 ↑0.0	45.5 ↑0.0
	VCD	76.4 ↑1.3	62.5 ↑8.4	80.3 ↑2.1	57.0 ↑3.8	78.6 ↑2.2	49.2 ↑3.7
	ICD	75.8 ↑0.7	54.2 ↑0.1	79.2 ↑1.0	52.8 ↓0.4	76.8 ↑0.4	46.0 ↑0.5
	SID	76.3 ↑1.2	57.5 ↑3.4	78.7 ↑0.5	57.5 ↑4.3	77.2 ↑0.8	47.5 ↑2.0
	sample†	**76.3 ↑1.2**	**54.2 ↑0.1**	**79.5 ↑1.3**	**53.1 ↓0.1**	**77.9 ↑1.5**	**46.2 ↑0.7**

Conclusion

This study demonstrates that the performance improvements of contrastive decoding on the POPE benchmark largely **stem from two misleading factors**:

- **A unidirectional shift in the model's output distribution**, which biases it toward generating *Yes* responses, artificially balancing the distribution in certain datasets
- The adaptive plausibility constraint, which **reduces sampling decoding to greedy search**.

By comparing experimental results from spurious methods and contrastive decoding, we confirm that while contrastive decoding enhances performance, it ultimately **fails to mitigate hallucinations**.