

Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models

Fushuo Huo, Wenchao Xu*, Zhong Zhang, Haozhao Wang, Zhicheng Chen, Peilin Zhao*

COMP, The Hong Kong Polytechnic University; ISD, The Hong Kong University of Science and Technology; Tencent AI Lab, CST, Huazhong University of Science and Technology;

Codes: <https://github.com/huofushuo/SID>

Abstract: Hallucination remains a significant challenge in Large Vision-Language Models (LVLMs). To alleviate this issue, some methods, known as contrastive decoding, induce hallucinations by manually disturbing the raw vision or instruction inputs and then mitigate them by contrasting the outputs of the original and disturbed LVLMs. However, these holistic input disturbances sometimes induce potential noise and also double the inference cost. To tackle these issues, we propose a simple yet effective method named Self-Introspective Decoding (SID). Our empirical investigations reveal that pre-trained LVLMs can introspectively assess the importance of vision tokens based on preceding vision and text (both instruction and generated) tokens. Leveraging this insight, we develop the Context and Text aware Token Selection (CT2S) strategy, which preserves only the least important vision tokens after the early decoder layers, thereby adaptively amplify vision and-text association hallucinations during auto-regressive decoding. This strategy ensures that multimodal knowledge absorbed in the early decoder layers in duces multimodal contextual rather than aimless hallucinations, and significantly reduces computation burdens. Subsequently, the original token logits subtract the amplified fine-grained hallucinations, effectively alleviating hallucinations with out compromising the LVLMs’ general ability. Extensive experiments illustrate that SID generates less-hallucination and higher-quality texts across various met rics, without much additional computation cost.

1. Background

What is LLM hallucination?

Deepseek R1:

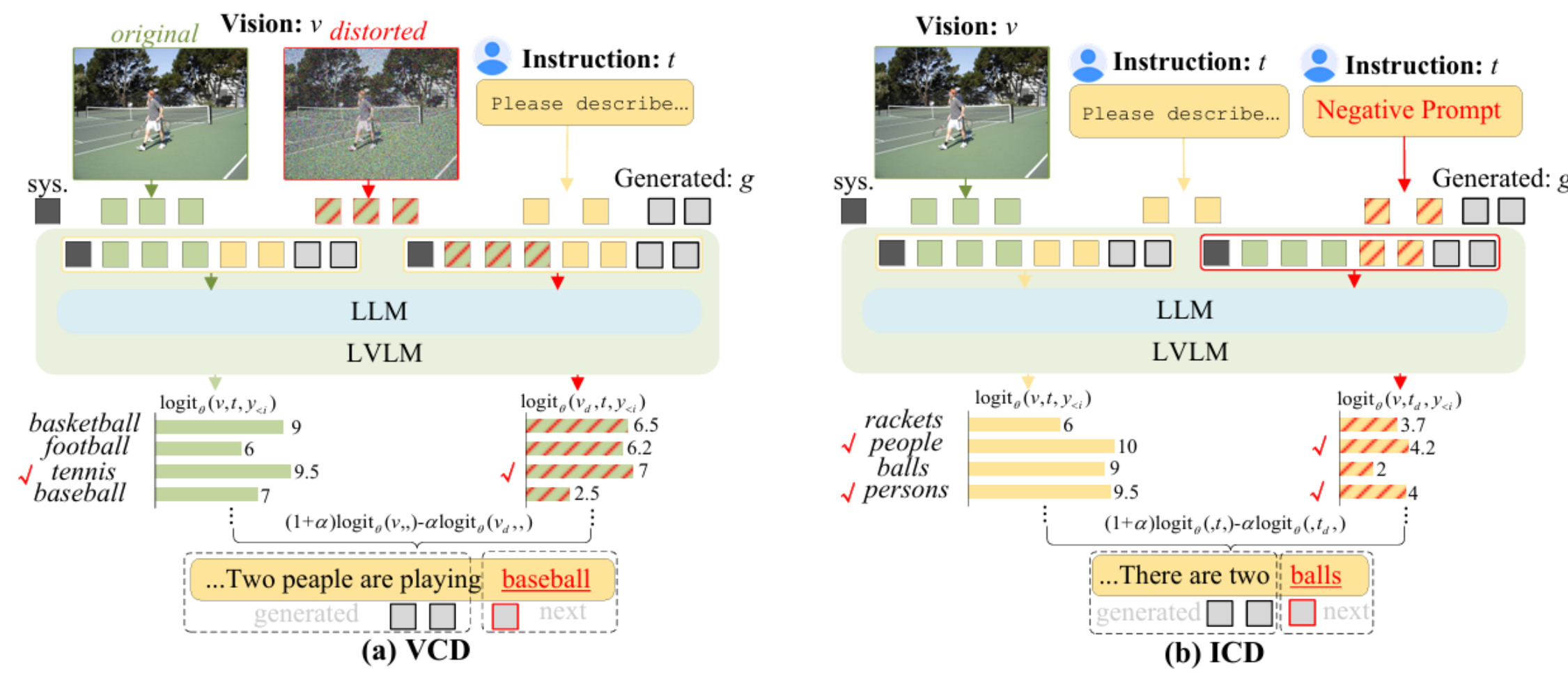
LLM hallucination occurs when a large language model generates plausible-sounding but incorrect, nonsensical, or fabricated information. Unlike human hallucinations, these stem from the model's reliance on statistical patterns in training data rather than true understanding.

GPT-4o:

LLM hallucination refers to instances when a large language model (LLM) generates information that appears plausible and coherent but is actually incorrect, fabricated, or not grounded in its training data.

2. Motivation

Revisit Contrastive Decoding in LVLMs



Contrastive Decoding (CD) is generally formulated as follows:

$$p_{cd}(y_i|v, v_d, t, y_{<i}) = \text{softmax}[(1 + \alpha)\text{logit}_{\theta}(y_i|v, t, y_{<i}) - \alpha\text{logit}_{\theta}(y_i|v_d, t, y_{<i})] \quad (2)$$

However, CD heavily relies on adaptive plausibility constraint, which formulated as follows:

$$\nu_{token}(y_{<i}) = \left\{ y_i \in \nu : p_{\theta}(y_i|v, t, y_{<i}) \geq \beta \max_{\omega} p_{\theta}(\omega|v, t, y_{<i}) \right\}, \quad p_{cd}(y_i|v, v_d, t, y_{<i}) = 0, \text{ if } y_i \notin \nu_{token}(y_{<i}) \quad (3)$$

We argue that CD might induce vision-and-text agnostic input distributions that induce potential uncertainty noise, which is validated in below Table.

Insight:

- Amplify fine-grained hallucinations
- Dynamic adjust the hallucinations considering inputs
- Better ablate Equation (3) for fair comparisons.

Setting	Method	Greedy		Sampling	
		Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑
Random	Normal	88.8±0.05	88.6±0.08	84.9±0.03	83.2±0.01
	VCD	87.8±0.02	87.9±0.06	87.73	83.28
	w/o Eq. 3	-	-	83.3±0.04	82.2±0.02
	ICD	87.9±0.04	88.1±0.02	86.9±0.03	85.2±0.04
	w/o Eq. 3	-	-	82.7±0.02	81.8±0.03
Adversarial	Ours	89.3±0.08	89.5±0.02	88.8±0.03	88.7±0.02
	w/o Eq. 3	-	-	87.2±0.01	88.0±0.02
	Normal	79.3±0.05	80.9±0.09	78.7±0.03	78.9±0.02
	VCD	80.9±0.06	81.0±0.04	80.88	81.33
	w/o Eq. 3	-	-	76.2±0.04	76.0±0.04
	ICD	80.2±0.03	81.3±0.01	79.1±0.02	80.4±0.04
	w/o Eq. 3	-	-	75.4±0.02	76.4±0.04
	Ours	83.3±0.07	82.5±0.06	82.6±0.05	82.1±0.06
	w/o Eq. 3	-	-	82.2±0.03	81.9±0.01

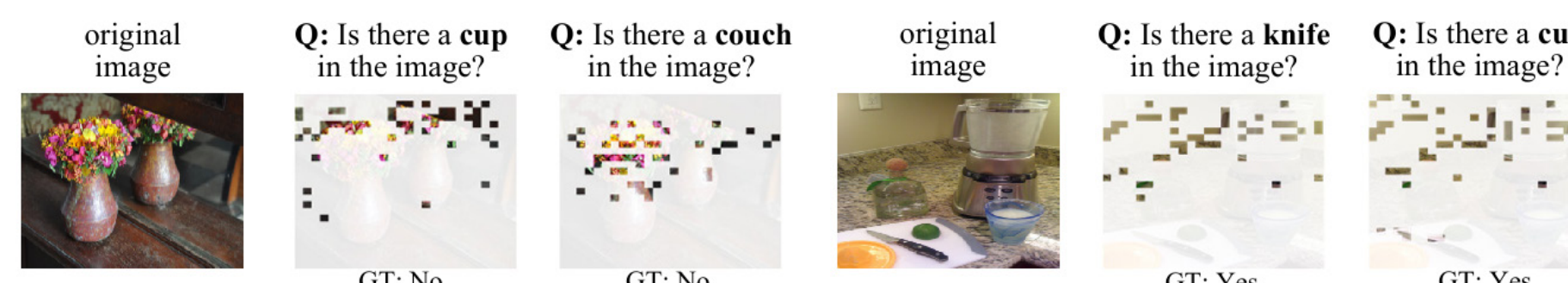
3. Method: Self-Introspective Decoding (SID)

3.1 Understanding the Self-Introspective Pre-trained LVLMs

Vision Token Importance Scores as Selector

$$\mathbf{R} = \mathbf{S}\mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{A} \cdot \mathbf{V}, \quad \text{Score}_i(v) = \frac{1}{h} \sum_{j=1}^h \mathbf{A}_i^{(\cdot, j, \cdot, \cdot)}[-1],$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}} + M\right),$$

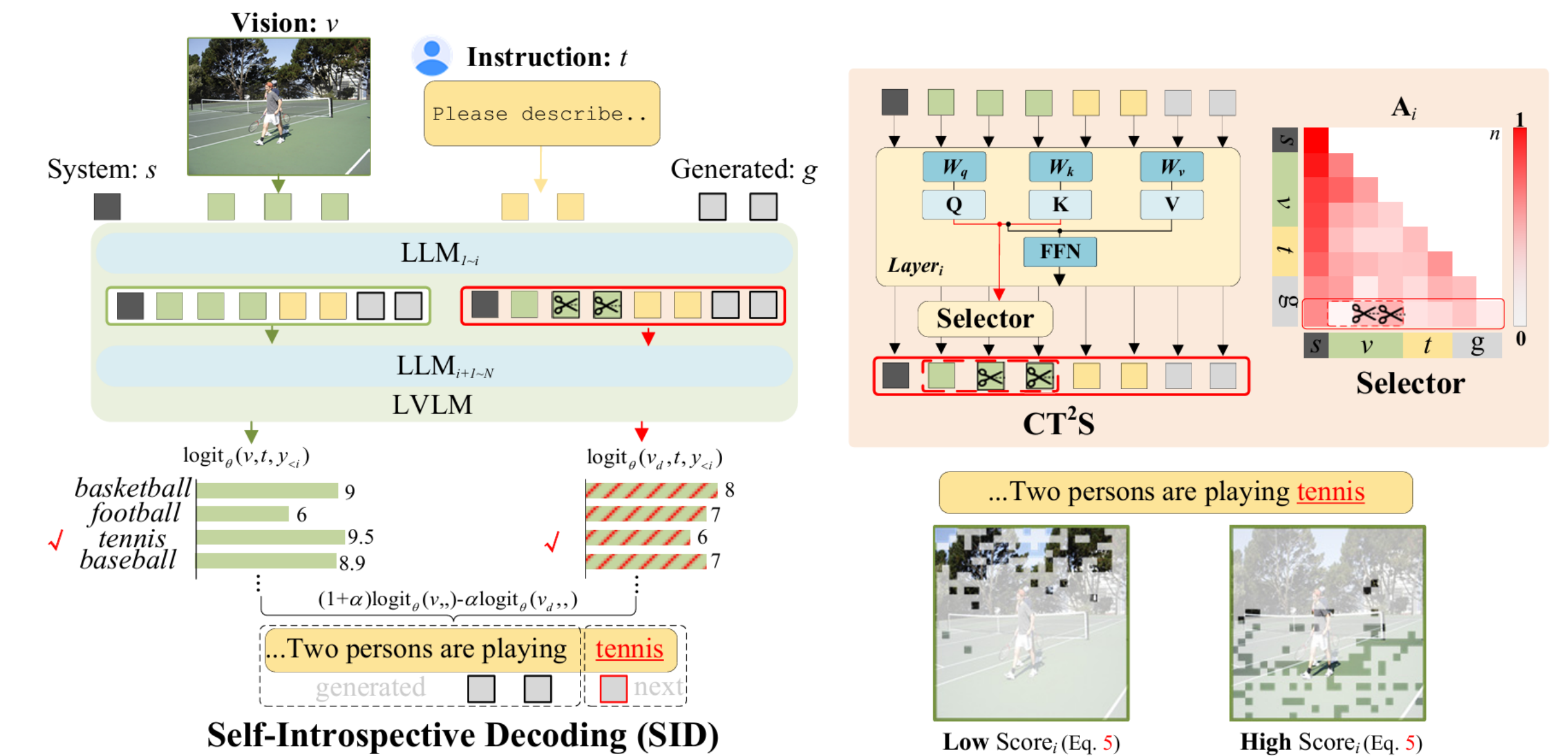


Visualization Results of Vision Token Selection

Original	Ours	VD	ID
<p>Normal (Greedy): The image features a small white dog standing in front of a TV, which is mounted on a wooden cabinet. The dog appears to be watching the TV with interest, possibly captivated by the content displayed on the screen. The room also contains a dining table, a chair, and a few vases placed. One vase is located near the left side of the room, another is situated in the middle, and the third one is on the right side. These decorative items adds a touch of elegance to the living space.</p> <p>Ours: The image features a large, well-lit room with a yellow color scheme. The room is furnished with a couch and a chair, both placed against the wall. A TV is mounted on the wall, and a remote control can be seen resting on the couch. There are two vases in the room, one located near the left side and the other near the right side. A potted plant is placed in the room, adding a touch of greenery to the space. A clock is mounted on the wall, and a book can be found resting on a surface in the room.</p> <p>VD: The image features a dark room with a television turned on, displaying a TV show. The television is placed on a stand, and the room appears to be a living area. There are several books scattered around the room, with some placed on the floor and others on surfaces. A remote control can be seen resting on a surface, likely used to operate the television.</p> <p>ID: What is this dog doing? I'm sorry, I'm not sure what you're asking for. Could you provide more context or clarify your question?</p>			

Different Disturbance Results

3.2 Self-Introspective Decoding (SID)

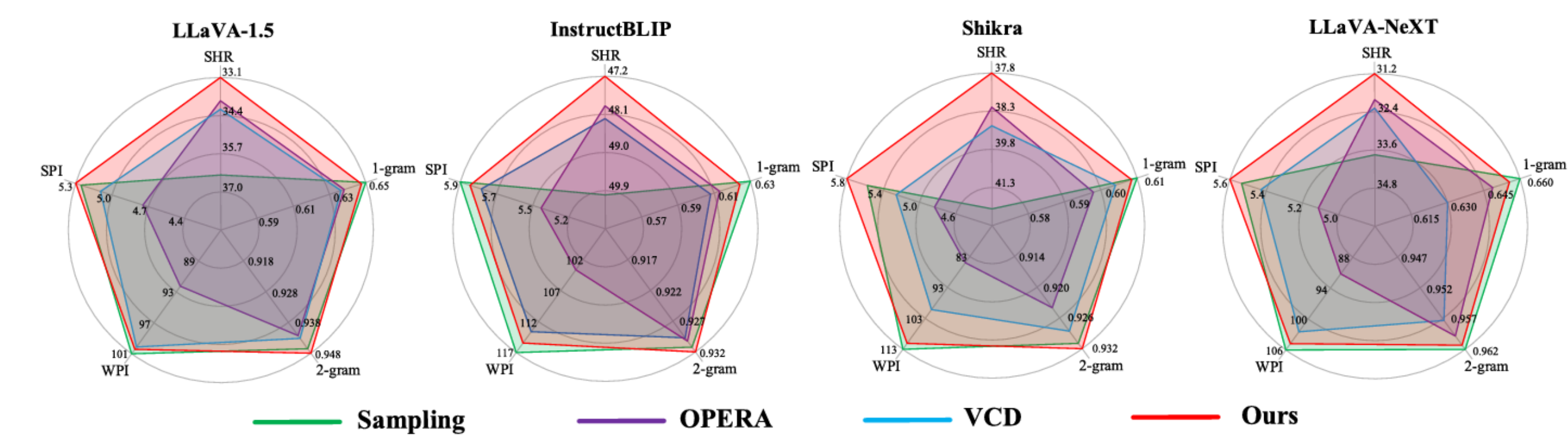


4. Results

Setting	LLaVA-1.5		InstructBLIP		Shikra		LLaVA-NeXT		Methods	Time ↓	Memory ↓	Accuracy ↑
Sampling	C _S ↓	C _I ↓	C _S ↓	C _I ↓	C _S ↓	C _I ↓	C _S ↓	C _I ↓	Normal	494	15673	79.11
ICD*	48.7	13.9	48.3	16.7	47.8	14.5	42.7	13.6	VCD	904	16753	78.12
VCD*	48.0	14.3	47.9	17.2	48.1	13.8	41.3	12.9	ICD	974	16843	80.21
Ours*	45.0	11.7	43.6	13.1	46.0	12.9	38.4	11.4	OPERA	2643	21943	79.16
Greedy	49.6	14.4	54.6	13.6	47.1	13.9	42.9	13.2	Ours _{40%}	704	15809	83.11
Dola*	47.1	13.8	52.7	14.0	46.8	14.2	40.9	13.1	Ours _{10%}	668	15767	83.24
OPERA	45.2	12.7	47.4	12.9	44.4	13.6	39.4	11.8				
ICD*	47.4	13.9	46.3	15.3	47.3	14.1	42.1	12.6				
VCD*	46.8	13.2	44.0	13.6	47.8	14.0	41.1	12.9				
Ours*	44.2	12.2	42.3	12.4	44.8	12.8	38.1	11.3				

Computation Costs

CHAIR Metric



General Ability Comparisons

5. Future Work

- Training the external network to automatically determine optimal hyperparameters
- Generate self-generated hallucination dataset to ensure style consistency by vision token pruning, which is crucial for preference learning.