



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



模式分析与机器智能
工业和信息化部重点实验室

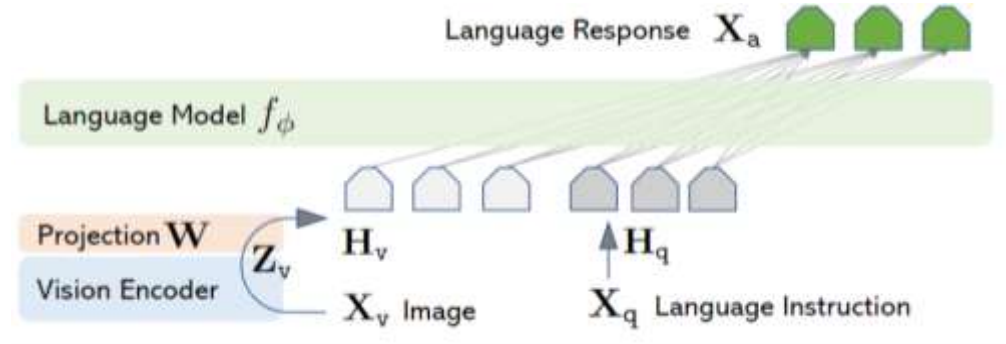
MIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Reducing Hallucinations in LVLMs

Causes

① From language bias

② From visual bias



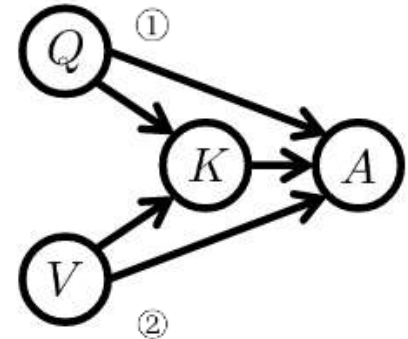
Compared to visual evidence, there are more easily identifiable shortcuts, such as **context/co-occurring elements**, or even the question itself.



What color are the bananas?



What brand is the bottle with red label?

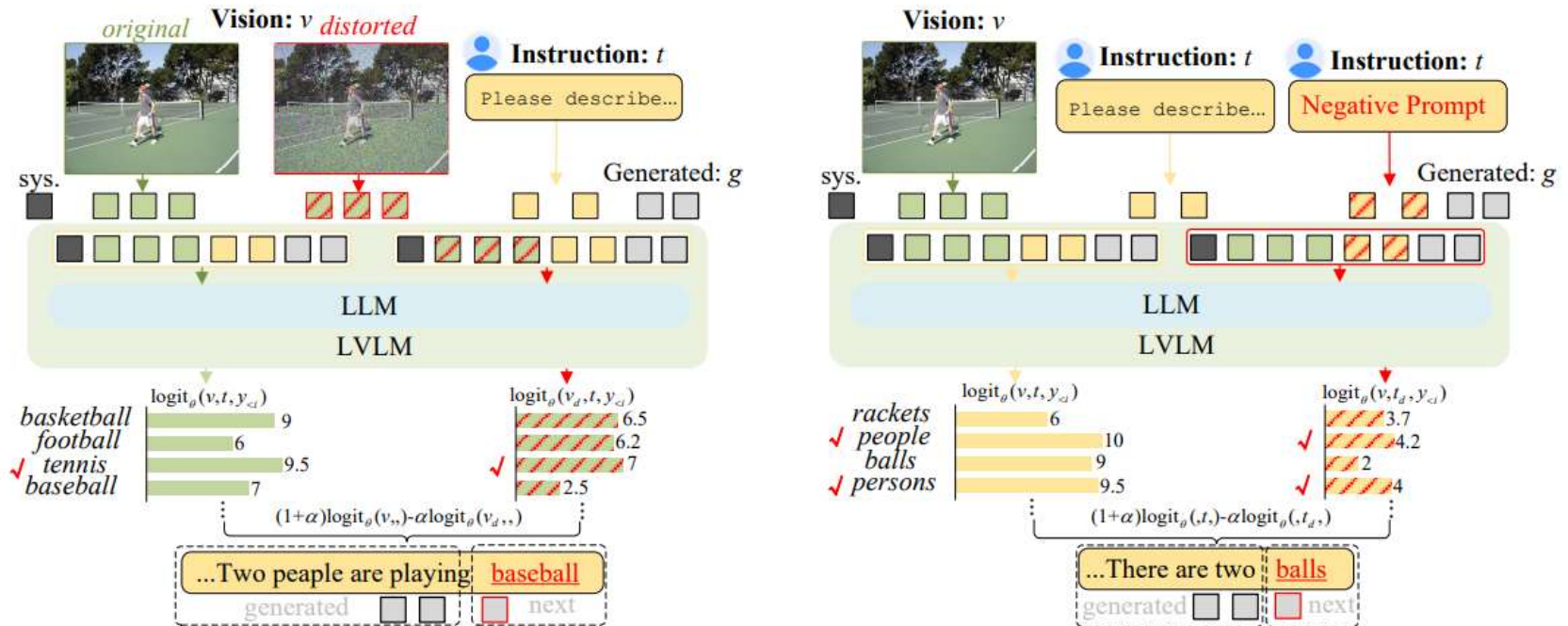


③ From attention sink

Causes

- Regarding Modal Bias

Visual Contrastive Decoding provides an interpolation paradigm **at first**

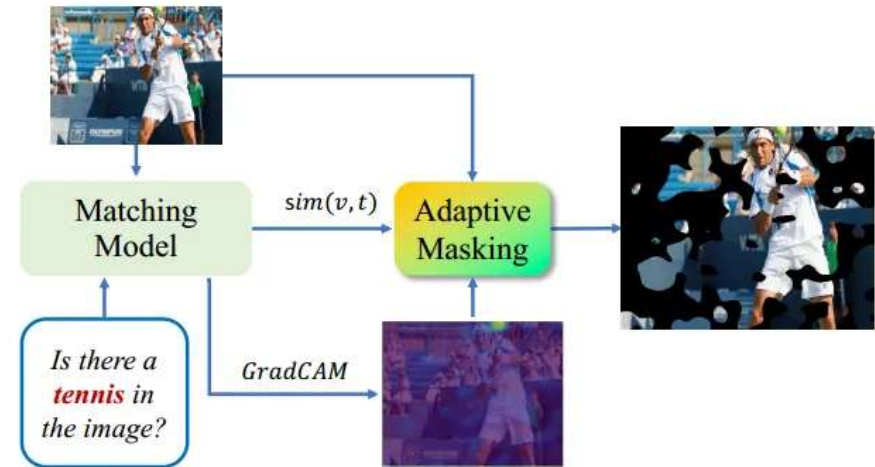
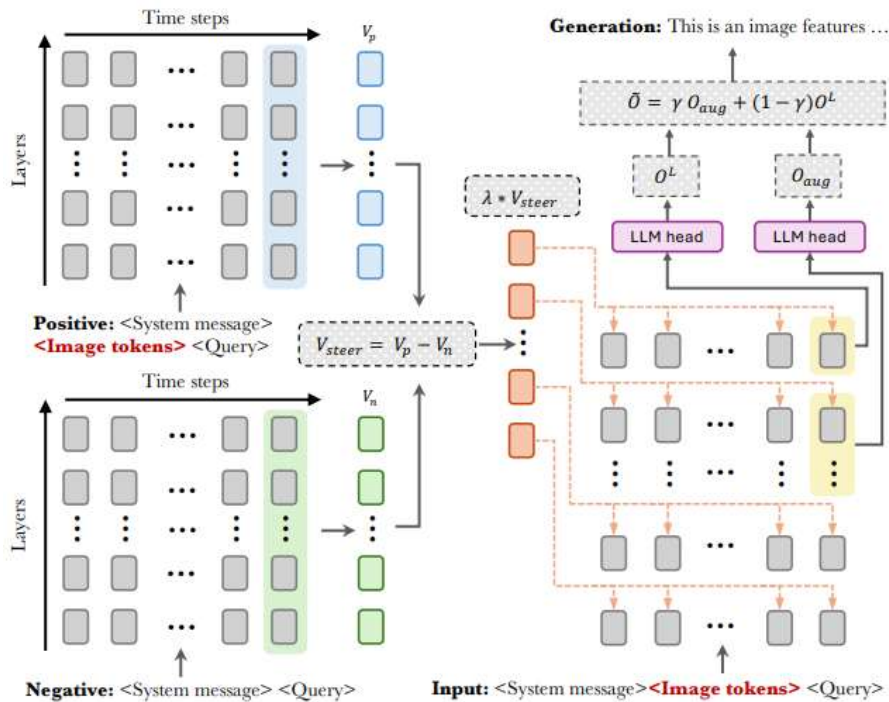
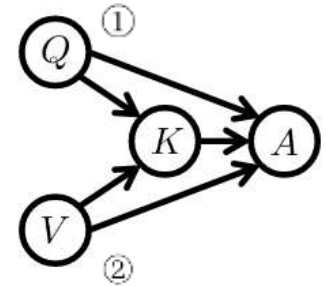


[1] Visual Contrastive Decoding (VCD) (Leng et al., 2024 CVPR)
 [2] Instruction Contrastive Decoding (ICD) (Kim et al., 2024 ICLR)

Causes

- Other interpretable variants

Counterfactual Approach: De-linguistic and De-contextualized

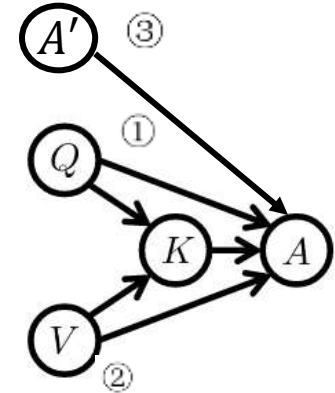
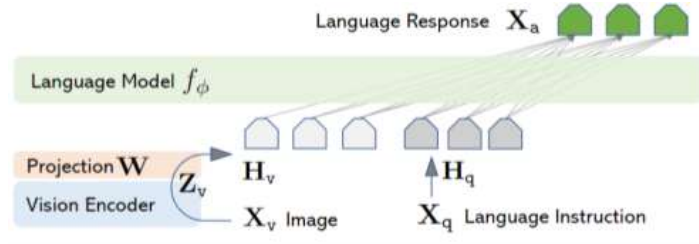


[3] Visual Information Steering (Li et al., 2025 ICML)

[4] Assembly of Local Attention (An et al., 2025 CVPR)

Causes

- ① From language bias
- ② From visual bias
- ③ From attention sink/collapse



Originating from intrinsic structure and modal imbalance

 Input Image



 Please describe this image in detail.

Human

 LLaVA-1.5

This image shows a stream in snowy woods with snow-covered trees. } Existent objects ✓

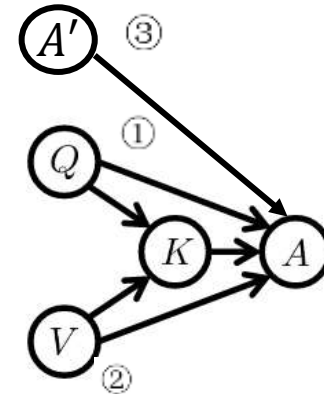
A small wooden bridge crosses the stream, connecting the snowy banks. } Initial Hallucination ✗

The bridge's handrails have carvings of pinecones and leaves, adding rustic charm to the winter scene. } Snowball Hallucination ✗

Causes

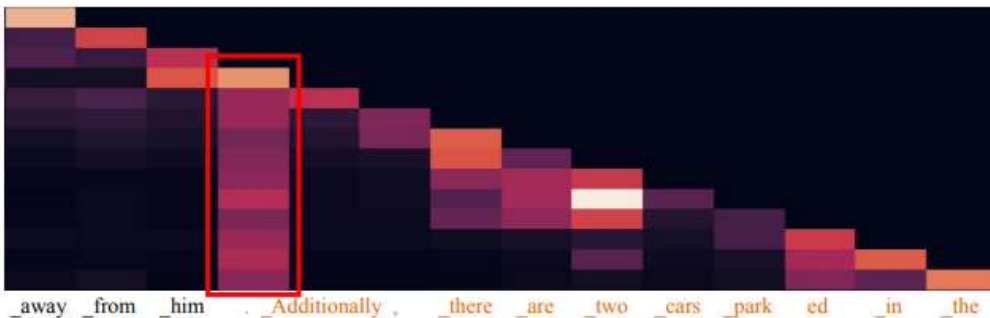
③ From attention sink/collapse

- from intrinsic structure and modal imbalance
- system or BOS、summary token (, .)
- gate attention etc.



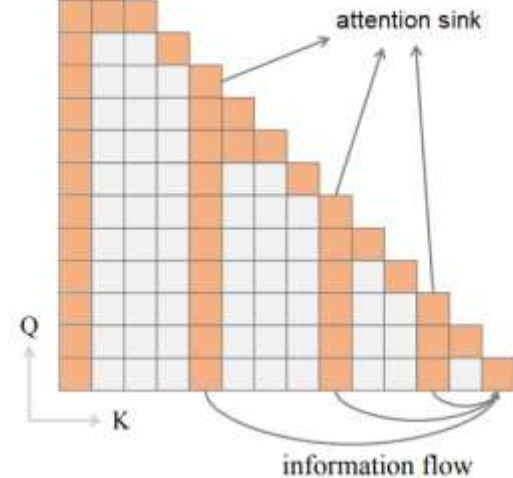
InstructBLIP:

In the image, a young man is performing a skateboarding trick in a residential area. He is riding a skateboard and jumping over a tree trunk that is lying on the ... away from him. Additionally, there are two cars parked in the area, likely belonging to residents or visitors of the neighborhood.



Scaled Product Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$



[5] Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free (Qiu et al., 2025 NIPS oral)

Grounding Language with Vision: A Conditional Mutual Information Calibrated Decoding Strategy for Reducing Hallucinations in LVLMs

**Hao Fang^{*1}, Changle Zhou^{*2}, Jiawei Kong^{*1,2}, Kuofeng Gao¹,
Bin Chen^{†2,3}, Tao Liang⁴, Guojun Ma⁴, Shu-Tao Xia^{1,3},**

¹Tsinghua Shenzhen Internation Graduate School, Tsinghua University,

²Harbin Institute of Technology, Shenzhen, ³Pengcheng Labortary ⁴ByteDance

fang-h23@mails.tsinghua.edu.cn

Innovations

- **Information-theoretic view**

Formulates hallucination mitigation as maximizing conditional mutual information

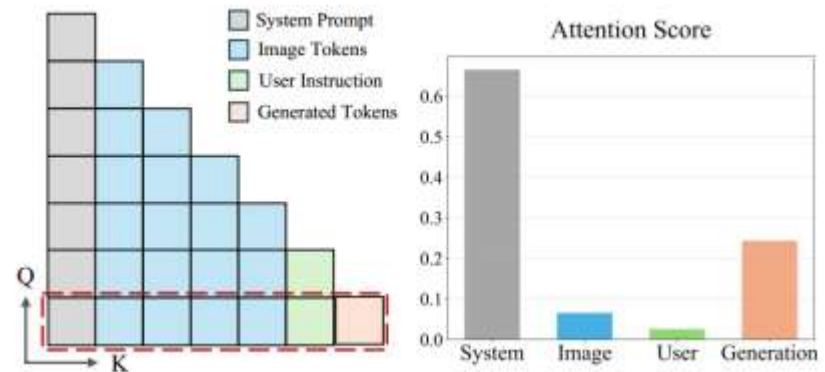
$$I(Y; V | X) = \mathbb{E}_{p(x,v,y)} \left[\log \frac{p(y|x,v)}{p(y|x)} \right]$$

- **CMI-Calibrated Decoding**

Contrastive scoring : $\log p_{\theta}(y_t|x, v, y_{<t}) - \tau \log p_{\theta}(y_t|x, y_{<t})$

- **Visual Purifier**

Enhances the quality of visual information



- **Standard decoding**

$$y_t \sim p_\theta(\cdot | v, x, y_{<t}) = \text{softmax}(f_\theta(\cdot | v, x, y_{<t}))$$

$$q_\theta(y | v, x) = \prod_{t=0}^{l-1} p_\theta(y_t | v, x, y_{<t}) = \prod_{t=0}^{l-1} \text{softmax}(f_\theta(\cdot | v, x, y_{<t}))_{y_t}$$

- **Conditional Mutual Information**

$$I(Y; V | X) = H(Y | X) - H(Y | X, V)$$

$$\max_{v, y} \text{C-PMI}_\theta(V = v, Y = y | X = x) = \max \left(\log \frac{p_\theta(v, y | x)}{p_\theta(v | x) p_\theta(y | x)} \right)$$

$$\rightarrow y^* = \arg \max_y [\log p(y|x, I) - \log p(y|x)]$$

Penalize language patterns that can be generated even without images, so that the generated results become more dependent on images.

Method

$$\underbrace{\max I(Y; V|X)}_{\text{outer problem}} \Rightarrow \begin{cases} (1) \text{ maximize visual grounding strength: } p(y|x, v) \\ (2) \text{ penalize over-confident language prior: } p(y|x) \end{cases}$$

- **Outer-level — Global CMI Maximization**

$$y_t \sim p_c(\cdot | v, x, y_{<t}) = \text{softmax} \left[(1 + \lambda) f_\theta(\cdot | v, x, y_{<t}) - \lambda f_\theta(\cdot | x, y_{<t}) \right]$$

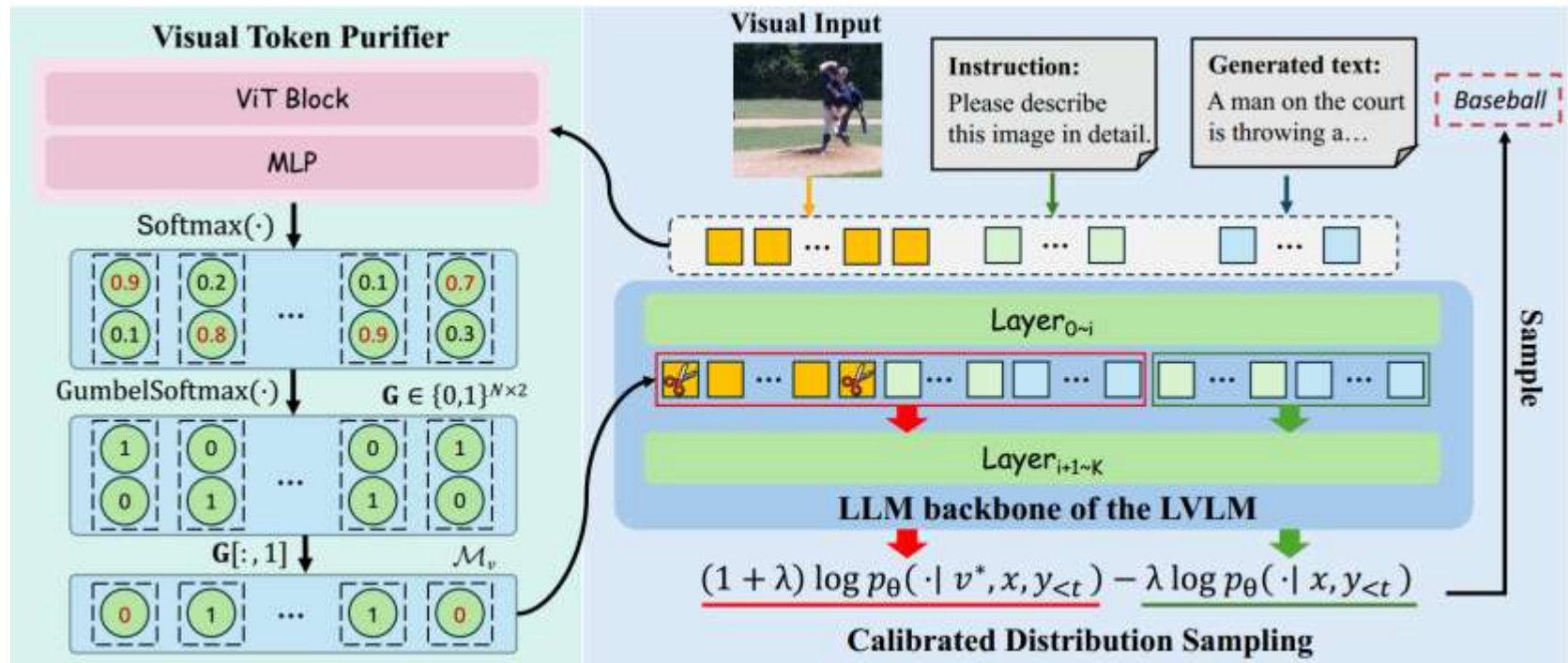
- **Inner-level — Token-wise Calibration**

$$\text{Attn}_i(v) = \frac{1}{H} \sum_{v_j \in v} \sum_{k=0}^{H-1} A_i^{(k, :, :)}[-1][v_j], \quad \text{where } A_i = \text{softmax} \left(\frac{Q_i K_i^\top}{\sqrt{d_k}} + \mathcal{M}_c \right)$$

Unified Bi-level Optimization:

$$\begin{aligned} & \max_y \sum_{t=0}^{l-1} \left[(1 + \lambda) \log p_\theta(y_t | v^*, x, y_{<t}) - \lambda \log p_\theta(y_t | x, y_{<t}) \right], \\ \text{s.t. } & v^* = \arg \max_v \left[\alpha \cdot \text{Attn}_i(v) + \log p_\theta(y_t | v, x, y_{<t}) - \log p_\theta(y_t | x, y_{<t}) \right], \end{aligned}$$

Method



Why introduce the Visual Token Purifier

computationally expensive and non-differentiable

Method

- Visual Token Purifier

Visual mask $\mathcal{M}_v = \left\{ \arg \max_{j \in \{0,1\}} \pi_{ij} \mid i \in \{0, 1, \dots, N-1\} \right\}$

differentiable sampling \downarrow $\mathbf{G} = \text{GumbelSoftmax}(\pi, \tau)$

$$\mathbf{G}[:, 1] = \mathcal{M}_v \in [0, 1]^N$$

Training Objective:

$$\begin{aligned} \mathcal{L}_{total} = & (\log p_{\theta}(y_t | v, x, y_{<t}) - \log p_{\theta}(y_t | x, y_{<t})) \\ & + \alpha \cdot \text{Attn}_t(v) + \beta \cdot \|\text{sum}(\mathcal{M}_v) / N - \gamma\|_F \end{aligned}$$

Input $\mathbf{z} = [z_v, z_x, z_{y_{<t}}]$



Output

$$\pi = \text{softmax}(\mathcal{P}(\mathbf{z})) \in [0, 1]^{N \times 2}$$

- CHAIR metric

Table 1: Comparison of the proposed CMI-VLD with SOTA baselines on the CHAIR metric. We evaluate the performance on MSCOCO. The [†] indicates decoding strategies based on beam search.

Method	LLaVA-1.5		InstructBLIP		Shikra		LLaVA-Next	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
<i>Sampling</i>	52.2	15.8	55.0	25.3	56.2	15.8	34.8	9.4
ICD	51.0	15.2	64.0	20.2	56.6	15.5	33.4	8.7
VCD	50.4	15.6	57.6	19.2	56.2	15.5	36.0	9.3
VTI	37.2	11.4	49.2	21.9	47.0	14.1	32.2	7.8
SID	49.2	14.5	58.0	18.7	54.4	14.4	39.4	9.9
CMI-VLD	30.2	9.3	51.0	16.1	38.2	10.1	30.6	7.6
<i>Greedy</i>	45.0	13.5	52.2	21.8	54.8	15.8	31.6	8.2
ICD	44.8	12.8	48.8	14.1	55.0	14.0	32.8	9.1
VCD	49.4	14.0	46.6	13.3	55.8	15.3	36.8	9.4
HALC [†]	33.2	10.3	61.4	20.0	55.4	14.7	36.7	9.5
OPERA [†]	39.4	10.3	48.2	13.8	36.8	11.7	33.6	8.3
VTI	30.6	10.1	48.3	20.7	44.6	13.7	30.1	7.6
SID	42.8	12.1	56.2	15.8	51.2	13.6	38.0	8.9
VASparse [†]	49.6	14.2	53.6	14.9	51.6	14.8	33.6	9.1
CMI-VLD	29.9	8.9	43.2	12.9	30.6	8.9	27.2	6.8

Experiment



- POPE metric

Table 2: Comparison of the proposed CMI-VLD with SOTA baselines on the POPE metric. The † indicates decoding strategies based on beam search.

Model	Method	Random		Popular		Adversarial	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
LLaVA-Next	<i>Sampling</i>	82.53%	79.19%	81.57%	78.31%	80.30%	77.16%
	ICD	82.77%	79.57%	81.97%	78.81%	81.03%	77.95%
	VCD	83.67%	80.80%	82.17%	79.40%	80.90%	78.25%
	VTI	82.70%	79.45%	81.43%	78.25%	80.10%	77.05%
	SID	84.67%	82.20%	83.57%	81.16%	81.60%	80.27%
	CMI-VLD	85.17%	83.02%	84.10%	82.03%	82.30%	80.40%
	<i>Greedy</i>	83.40%	80.32%	82.60%	79.55%	81.77%	78.77%
	ICD	83.47%	80.41%	82.60%	79.56%	81.90%	78.91%
	VCD	84.43%	81.85%	83.30%	80.77%	82.33%	79.88%
	HALC†	83.34%	80.36%	82.33%	79.48%	81.40%	78.92%
	OPERA†	83.50%	80.46%	82.70%	79.69%	81.87%	78.91%
	VTI	84.70%	82.09%	83.67%	81.11%	82.90%	80.40%
	SID	84.97%	82.53%	83.93%	81.56%	82.97%	80.67%
	VASparse†	83.47%	80.52%	82.24%	79.69%	81.33%	78.88%
	CMI-VLD	86.43%	84.52%	85.07%	83.22%	83.90%	82.14%
	InstructBLIP	<i>Sampling</i>	82.03%	81.30%	78.77%	78.66%	76.37%
VTI		83.50%	82.01%	80.83%	79.70%	79.13%	78.29%
ICD		83.20%	82.29%	79.87%	79.51%	77.63%	77.74%
VCD		83.43%	82.49%	79.70%	79.36%	77.53%	77.65%
SID		85.43%	84.81%	82.43%	82.24%	79.47%	79.84%
CMI-VLD		86.33%	85.41%	84.60%	83.87%	81.57%	81.29%
<i>Greedy</i>		87.27%	85.91%	84.87%	83.72%	82.97%	82.04%
ICD		87.23%	85.82%	84.90%	83.68%	83.13%	82.11%
VCD		86.73%	85.30%	84.37%	83.16%	82.47%	81.49%
HALC†		87.30%	85.96%	84.83%	83.70%	83.00%	82.08%
OPERA†		87.53%	86.26%	85.07%	84.00%	83.07%	82.24%
VTI		85.73%	83.86%	84.13%	82.36%	82.50%	80.89%
SID		88.10%	87.15%	85.87%	85.10%	82.90%	82.52%
VASparse†		87.33%	86.00%	84.87%	83.74%	83.00%	82.09%
CMI-VLD		88.37%	87.50%	86.10%	85.40%	82.87%	82.64%

Experiment



- GPT-4 assisted evaluation

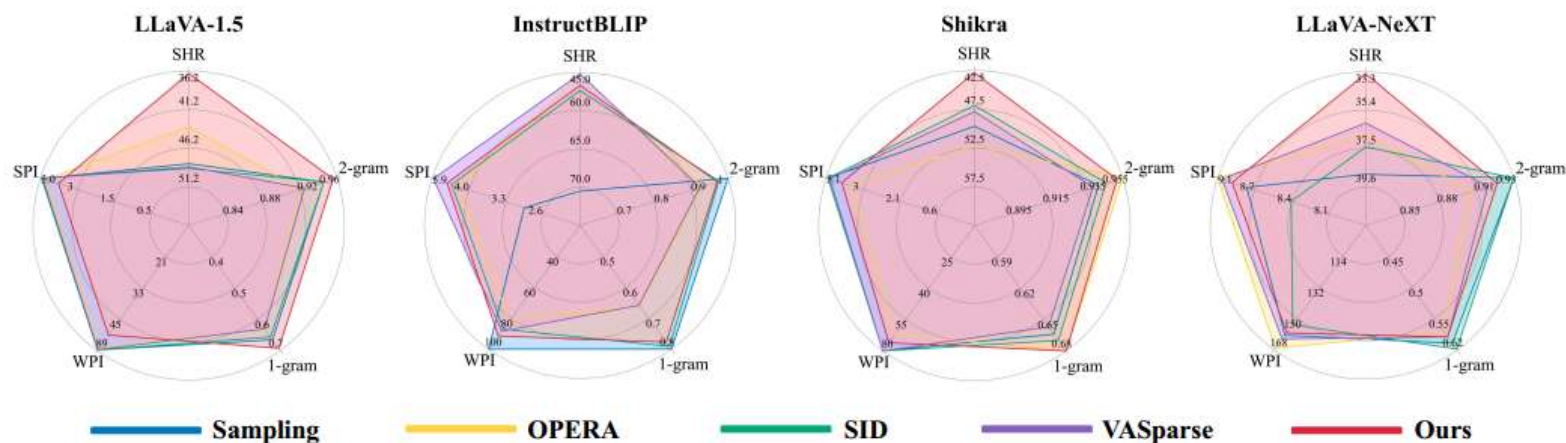


Figure 3: GPT-4o assisted benchmark. We calculate the Sentence-level Hallucination Ratio (SHR) as the major metric for hallucination degree, along with 1&2-gram, the number of sentences per image (SPI), and the number of words per image (WPI). A larger radar area indicates better performance.

Exploration



Q: Please describe the content of this picture.

A: The image features a woman sitting on a chair in a park, accompanied by her dog. The dog is lying on the grass next to the woman, enjoying the outdoor setting. The woman is holding a guitar, possibly playing or preparing to play.

In the scene, there are **several potted plants scattered around the area**, adding to the park's ambiance. A basket can be seen on the ground, possibly containing some items for the woman and her dog to enjoy during their time in the park.

70%



Q: Please describe the content of this picture.

A: The image features a woman sitting on a chair in a park, accompanied by her dog. The dog is lying on the ground, close to the woman, and appears to be enjoying the outdoor setting. The woman is holding a guitar, possibly playing or preparing to play a song.

In the background, there is a basket filled with flowers, adding a touch of color and beauty to the scene. The park setting and the presence of the dog create a relaxed and pleasant atmosphere.

90%



Q: Please describe the content of this picture.

A: The image features a **man** sitting on a chair in a grassy field, holding a guitar. He is accompanied by a dog, which is lying on the ground nearby. The man appears to be enjoying his time outdoors, playing his guitar while the dog lays down, possibly resting or observing the man's performance.

In the background, there is a basket filled with flowers, adding a touch of color and beauty to the scene. The combination of the man, the dog, and the flowers creates a peaceful and relaxing atmosphere.

Blurring images can induce illusions but has benefits.

Blurring simple backgrounds (redundant tokens) promotes focus on details.

Exploration



Q: Please describe the content of this picture.

A: The image features a person sitting in a chair, holding a guitar. The person is positioned in the scene, with the guitar placed in their hands. The chair is placed in the background, and the person appears to be enjoying their time playing the guitar.



Q: What is the gender of the person in the picture?

A: The person in the picture is a woman.

Q: Who's playing the guitar?

A: A woman is playing the guitar.

Q: Who's near the dog?

A: A woman is near the dog.



Q: Person

A: A person sitting on a chair in the park, accompanied by a dog.

Q: Guitar

A: A woman is holding a guitar sitting on a chair, with a dog nearby.

Q: Dog

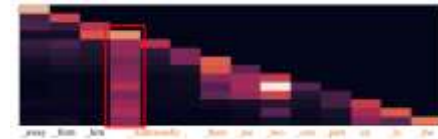
A: A dog is lying on the ground with a woman around it and a basket to the left.

The model possesses recognition capabilities for task-specific problems, but becomes confused when faced with general problems.

Exploration

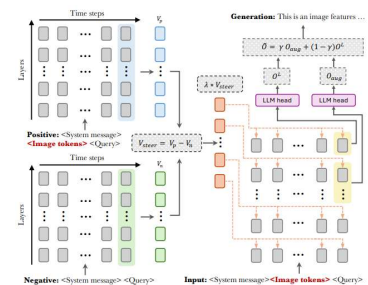
The model possesses recognition capabilities for task-specific problems, but becomes confused when faced with general problems.

- ✓ **Distracted Attention in Challenging Science**
- ✓ **Attention dilution caused by the long context required for general task generation**
- ✓ **The image itself possesses recognition capabilities, and when a token is input individually, it will pay attention to its surrounding areas.**



Plan:

- **Maintain t-1 step image evidence approximation for t-step token usage**
- **Evidence Information from image Re-injection LLM attention block**
- **Separation of image evidence information and text during softmax processing**



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Other insight

A woman sits on a park bench accompanied by her beloved dog. The dog lies on the grass beside her, basking in the outdoor setting. The woman holds a guitar, perhaps playing or preparing to perform a song. Scattered throughout the scene are several potted plants, adding vitality to the park. A basket is visible on the ground, likely holding personal items or belongings. Additionally, two chairs are present in the frame—one positioned near the woman, the other slightly farther away.

Q: Please describe the content of this picture.

A: The image depicts a woman sitting on a park bench, accompanied by her dog. The dog is lying on the grass beside her, and the woman is holding a guitar. The scene appears to be set in an outdoor park, with a chair placed nearby. The woman is preparing to perform a song, and the guitar is placed in front of her, ready for her to play.



Q: Please describe the content of this picture.

A: A **man** is sitting on a park bench, accompanied by his beloved **cat**. **The cat** lies on the ground beside him, enjoying the outdoor environment. The man is holding a guitar, perhaps playing music while spending time with his pet. **There are two more chairs in the scene, one near the man and the other slightly farther away. In the background, potted plants can be seen, adding a pleasant atmosphere to the park. On the ground, there is also a vase, possibly holding fresh flowers or other decorative items.**



Q: What is the gender of the person in the picture?

A: The person in the picture is a woman.

Q: Who's playing the guitar?

A: A woman is playing the guitar.

Q: What animal is near the basket?

A: A dog is beside a basket.

Text within images receives higher attention in LLM models, but certain local issues can be avoided.



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



模式分析与机器智能
工业和信息化部重点实验室

MIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS
