



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

Sicong Leng^{1,2,*} Hang Zhang^{1,3,*} Guanzheng Chen^{1,3} Xin Li^{1,3,†}
Shijian Lu² Chunyan Miao² Lidong Bing^{1,3}
¹DAMO Academy, Alibaba Group ²Nanyang Technological University
³Hupan Lab, 310023, Hangzhou, China

CVPR 2024

Introduction

大型视觉语言模型（LVLMs）是计算机视觉与自然语言处理的交叉技术，能将图像信息转化为连贯文本，实现“描述图片内容”“回答视觉问题”等任务，在多模态任务中展现出强大能力。

核心问题 —— 物体幻觉

指模型常生成与视觉输入不一致的文本输出，比如错误判断图像中物体的存在或不准确描述物体属性。降低模型可靠性，限制其在高风险场景的应用。



MiniGPT-v2

A collage of pictures of a lion, a giraffe, a bird, a tiger, a monkey and an elephant.

Explanation: Contrary to Model's prediction there is no giraffe, or tiger or a monkey in the original image.



SPHINX

There are five people in the image

Explanation: There's more than five people in the scene, while the model predicts only five of them



Introduction

LVLMM 的自回归生成:

$$y_t \sim p_\theta(y_t | v, x, y_{<t}), \\ \propto \exp \text{logit}_\theta(y_t | v, x, y_{<t})$$

问题根源: 当模型错误地将高概率分配给与图像内容不一致的 token 时, 产生幻觉

1. 统计偏差: 过度依赖训练数据中物体的高频出现或共现 (如 “餐桌” 常与 “叉子” 共现, 模型可能虚构 “叉子”);
2. 语言先验: 依赖语言模型学到的关于语言的固有模式、倾向或知识 (如 “香蕉通常是黄色”, 忽略图像中实际的黑色香蕉)。

解决方案:

基于额外数据的微调方法: 通过引入高质量标注数据, 构建幻觉数据集进行模型微调, 从而引导模型更多关注视觉输入。但该方法需大量人工标注成本, 且更新模型参数需消耗巨额计算资源。

Motivation

Assumption: increased visual uncertainty can amplify the language priors and statistical biases in LVLMs, thus exacerbating object hallucination

Visual Uncertainty Amplifies Statistical Bias

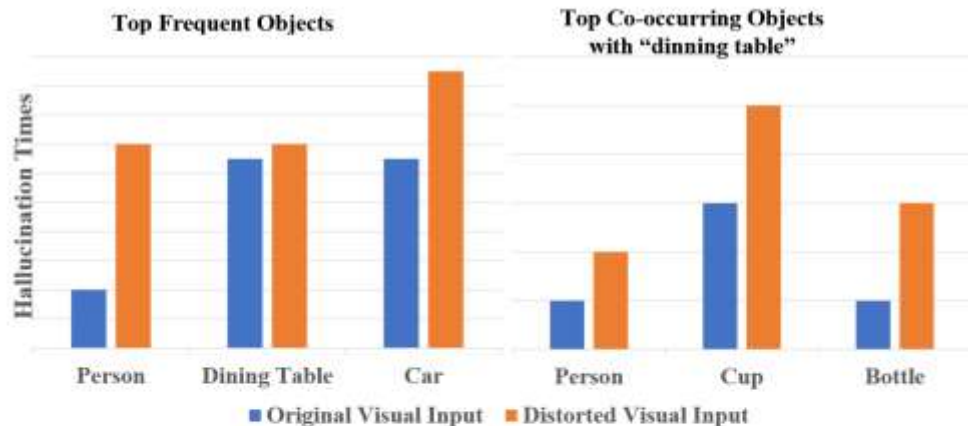


Figure 3. The left subfigure shows the correlation between frequent objects in MSCOCO and their propensity to be hallucinated in the validation set. Objects with a higher occurrence rate in the dataset are more likely to be hallucinated by LVLMs under distorted visual scenarios. The right subfigure charts three objects that often appear alongside "dinning table", where they are also more frequently hallucinated when presented with distorted visual inputs.

Visual Uncertainty Amplifies Language Priors

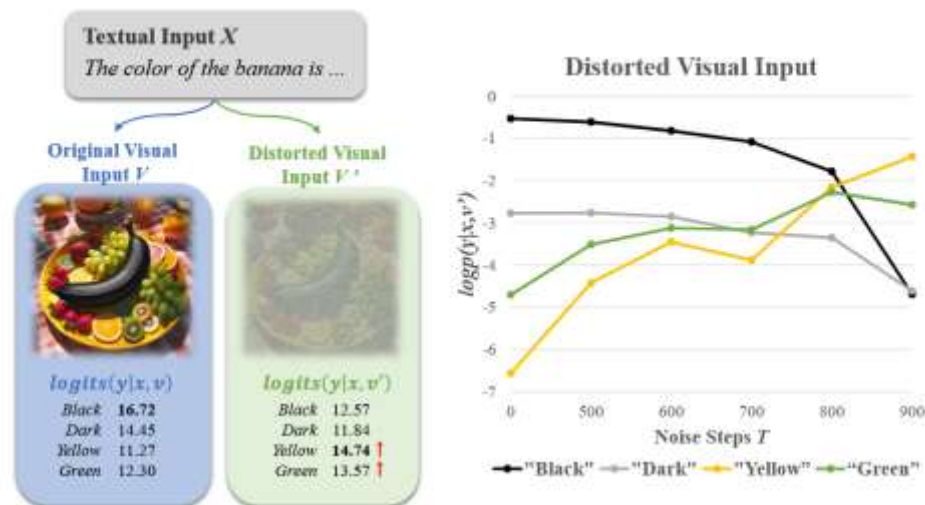


Figure 2. An illustration of visual uncertainty amplifying language priors. Given an image featuring a black banana among other colorful fruits, LVLMs favor more conventional banana colors—such as "yellow" and "green", with increasing visual uncertainty. The ground-truth color "black" diminishes in probability ($\log p(y|x, v')$) as the distortion escalates, making LVLMs over-reliant on the language priors from LLM pre-training that typically associate bananas with being yellow or green.

Method-Visual Contrastive Decoding

Contrasting the Predictions

$$p_{vcd}(y | v, v', x) = \text{softmax} [(1 + \alpha) \text{logit}_{\theta}(y | v, x) - \alpha \text{logit}_{\theta}(y | v', x)], \quad (3)$$

Adaptive Plausibility Constraints(自适应合理性约束)

$$\begin{aligned} \mathcal{V}_{\text{head}}(y_{<t}) &= \{y_t \in \mathcal{V} : \\ p_{\theta}(y_t | v, x, y_{<t}) &\geq \beta \max_w p_{\theta}(w | v, x, y_{<t})\}, \\ p_{vcd}(y_t | v, v', x) &= 0, \text{ if } y_t \notin \mathcal{V}_{\text{head}}(y_{<t}), \end{aligned} \quad (4)$$

Full formulation

$$\begin{aligned} y_t &\sim \text{softmax} [(1 + \alpha) \text{logit}_{\theta}(y_t | v, x, y_{<t}) \\ &\quad - \alpha \text{logit}_{\theta}(y_t | v', x, y_{<t})], \\ &\text{subject to } y_t \in \mathcal{V}_{\text{head}}(y_{<t}) \end{aligned} \quad (5)$$

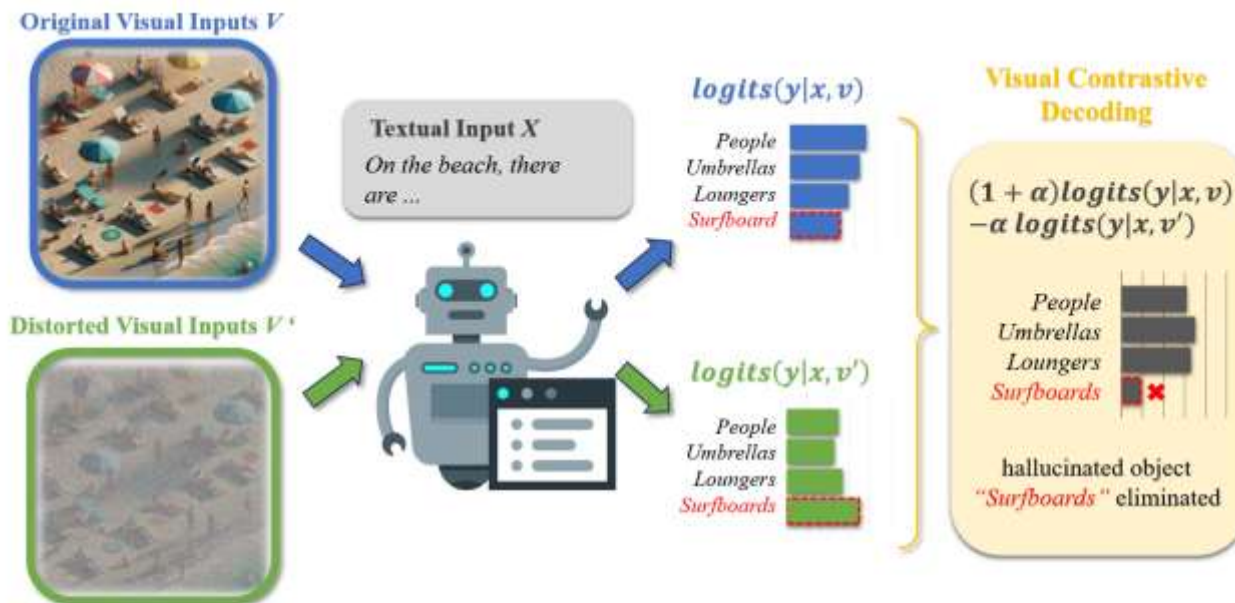


Figure 1. An illustration of Visual Contrastive Decoding. The hallucinated object “Surfboards” is highlighted in red, and it is eliminated during the generative process by contrasting with the output distribution that favors hallucinations.



Experiment

Datasets & Evaluation Metrics:

POPE-the Polling-based Object Probing Evaluation. (Accuracy, Precision, Recall, and the F1 score)

MME-assess LLMs across multiple dimensions. It comprises ten perception-related subtasks and four cognition-focused ones. (MME Score \uparrow)

Dataset	Setting	Model	Decoding	Accuracy \uparrow	Precision	Recall	F1 Score \uparrow
MSCOCO	Random	LLaVA1.5	Regular	83.29 $_{(\pm 0.35)}$	92.13 $_{(\pm 0.54)}$	72.80 $_{(\pm 0.57)}$	81.33 $_{(\pm 0.41)}$
			VCD	87.73 $_{(\pm 0.40)}$	91.42 $_{(\pm 0.55)}$	83.28 $_{(\pm 0.42)}$	87.16 $_{(\pm 0.41)}$
		Qwen-VL	Regular	84.73 $_{(\pm 0.36)}$	95.61 $_{(\pm 0.45)}$	72.81 $_{(\pm 0.38)}$	82.67 $_{(\pm 0.41)}$
			VCD	88.63 $_{(\pm 0.10)}$	94.64 $_{(\pm 0.25)}$	81.91 $_{(\pm 0.19)}$	87.81 $_{(\pm 0.11)}$
	InstructBLIP	Regular	80.71 $_{(\pm 0.73)}$	81.67 $_{(\pm 0.67)}$	79.19 $_{(\pm 1.14)}$	80.41 $_{(\pm 0.80)}$	
		VCD	84.53 $_{(\pm 0.38)}$	88.55 $_{(\pm 0.54)}$	79.32 $_{(\pm 0.44)}$	83.68 $_{(\pm 0.40)}$	
	Popular	LLaVA1.5	Regular	81.88 $_{(\pm 0.48)}$	88.93 $_{(\pm 0.60)}$	72.80 $_{(\pm 0.57)}$	80.06 $_{(\pm 0.05)}$
			VCD	85.38 $_{(\pm 0.38)}$	86.92 $_{(\pm 0.53)}$	83.28 $_{(\pm 0.42)}$	85.06 $_{(\pm 0.37)}$
		Qwen-VL	Regular	84.13 $_{(\pm 0.18)}$	94.31 $_{(\pm 0.43)}$	72.64 $_{(\pm 0.45)}$	82.06 $_{(\pm 0.23)}$
			VCD	87.12 $_{(\pm 0.07)}$	91.49 $_{(\pm 0.10)}$	81.85 $_{(\pm 0.19)}$	86.40 $_{(\pm 0.09)}$
	InstructBLIP	Regular	78.22 $_{(\pm 0.84)}$	77.87 $_{(\pm 1.03)}$	78.85 $_{(\pm 0.52)}$	78.36 $_{(\pm 0.76)}$	
		VCD	81.47 $_{(\pm 0.42)}$	82.89 $_{(\pm 0.64)}$	79.32 $_{(\pm 0.44)}$	81.07 $_{(\pm 0.39)}$	
Adversarial	LLaVA1.5	Regular	78.96 $_{(\pm 0.52)}$	83.06 $_{(\pm 0.58)}$	72.75 $_{(\pm 0.59)}$	77.57 $_{(\pm 0.57)}$	
		VCD	80.88 $_{(\pm 0.33)}$	79.45 $_{(\pm 0.29)}$	83.29 $_{(\pm 0.43)}$	81.33 $_{(\pm 0.34)}$	
	Qwen-VL	Regular	82.26 $_{(\pm 0.30)}$	89.97 $_{(\pm 0.33)}$	72.61 $_{(\pm 0.50)}$	80.37 $_{(\pm 0.37)}$	
		VCD	84.26 $_{(\pm 0.39)}$	85.84 $_{(\pm 0.45)}$	82.05 $_{(\pm 0.39)}$	83.90 $_{(\pm 0.39)}$	
InstructBLIP	Regular	75.84 $_{(\pm 0.45)}$	74.30 $_{(\pm 0.63)}$	79.03 $_{(\pm 0.68)}$	76.59 $_{(\pm 0.40)}$		
	VCD	79.56 $_{(\pm 0.41)}$	79.67 $_{(\pm 0.59)}$	79.39 $_{(\pm 0.50)}$	79.52 $_{(\pm 0.38)}$		
Random	LLaVA1.5	Regular	83.45 $_{(\pm 0.48)}$	87.24 $_{(\pm 0.68)}$	78.36 $_{(\pm 0.54)}$	82.56 $_{(\pm 0.50)}$	
		VCD	86.15 $_{(\pm 0.23)}$	85.18 $_{(\pm 0.34)}$	87.53 $_{(\pm 0.14)}$	86.34 $_{(\pm 0.21)}$	
	Qwen-VL	Regular	86.67 $_{(\pm 0.48)}$	93.16 $_{(\pm 0.55)}$	79.16 $_{(\pm 0.59)}$	85.59 $_{(\pm 0.53)}$	
		VCD	89.22 $_{(\pm 0.14)}$	90.77 $_{(\pm 0.04)}$	87.32 $_{(\pm 0.34)}$	89.01 $_{(\pm 0.16)}$	
InstructBLIP	Regular	80.91 $_{(\pm 0.34)}$	77.97 $_{(\pm 0.59)}$	86.16 $_{(\pm 0.88)}$	81.86 $_{(\pm 0.32)}$		
	VCD	84.11 $_{(\pm 0.27)}$	82.21 $_{(\pm 0.35)}$	87.05 $_{(\pm 0.53)}$	84.56 $_{(\pm 0.28)}$		
A-OKVQA	Popular	LLaVA1.5	Regular	79.90 $_{(\pm 0.33)}$	80.85 $_{(\pm 0.31)}$	78.36 $_{(\pm 0.54)}$	79.59 $_{(\pm 0.37)}$
			VCD	81.85 $_{(\pm 0.44)}$	78.60 $_{(\pm 0.58)}$	87.53 $_{(\pm 0.14)}$	82.82 $_{(\pm 0.36)}$
		Qwen-VL	Regular	85.56 $_{(\pm 0.35)}$	90.44 $_{(\pm 0.56)}$	79.53 $_{(\pm 0.84)}$	84.63 $_{(\pm 0.42)}$
			VCD	87.85 $_{(\pm 0.30)}$	88.10 $_{(\pm 0.36)}$	87.53 $_{(\pm 0.47)}$	87.81 $_{(\pm 0.31)}$
	InstructBLIP	Regular	76.19 $_{(\pm 0.80)}$	72.16 $_{(\pm 0.69)}$	85.28 $_{(\pm 0.79)}$	78.17 $_{(\pm 0.73)}$	
		VCD	79.78 $_{(\pm 0.47)}$	76.00 $_{(\pm 0.52)}$	87.05 $_{(\pm 0.53)}$	81.15 $_{(\pm 0.42)}$	
	Adversarial	LLaVA1.5	Regular	74.04 $_{(\pm 0.34)}$	72.08 $_{(\pm 0.53)}$	78.49 $_{(\pm 0.38)}$	75.15 $_{(\pm 0.23)}$
			VCD	74.97 $_{(\pm 0.39)}$	70.01 $_{(\pm 0.40)}$	87.36 $_{(\pm 0.15)}$	77.73 $_{(\pm 0.29)}$
Qwen-VL		Regular	79.57 $_{(\pm 0.31)}$	79.77 $_{(\pm 0.34)}$	79.23 $_{(\pm 0.73)}$	79.50 $_{(\pm 0.38)}$	
		VCD	81.27 $_{(\pm 0.09)}$	77.79 $_{(\pm 0.20)}$	87.53 $_{(\pm 0.34)}$	82.38 $_{(\pm 0.10)}$	
InstructBLIP	Regular	70.71 $_{(\pm 0.76)}$	65.91 $_{(\pm 0.74)}$	85.83 $_{(\pm 0.80)}$	75.56 $_{(\pm 0.57)}$		
	VCD	74.33 $_{(\pm 0.67)}$	69.46 $_{(\pm 0.73)}$	86.87 $_{(\pm 0.27)}$	77.19 $_{(\pm 0.47)}$		

Random	LLaVA1.5	Regular	83.73 $_{(\pm 0.27)}$	87.16 $_{(\pm 0.39)}$	79.12 $_{(\pm 0.35)}$	82.95 $_{(\pm 0.28)}$	
		VCD	86.65 $_{(\pm 0.45)}$	84.85 $_{(\pm 0.59)}$	89.24 $_{(\pm 0.34)}$	86.99 $_{(\pm 0.41)}$	
	Qwen-VL	Regular	80.97 $_{(\pm 0.32)}$	88.07 $_{(\pm 0.34)}$	71.64 $_{(\pm 0.57)}$	79.01 $_{(\pm 0.40)}$	
		VCD	85.59 $_{(\pm 0.38)}$	86.88 $_{(\pm 0.44)}$	83.84 $_{(\pm 0.36)}$	85.33 $_{(\pm 0.38)}$	
InstructBLIP	Regular	79.65 $_{(\pm 0.34)}$	77.14 $_{(\pm 0.43)}$	84.29 $_{(\pm 0.36)}$	80.56 $_{(\pm 0.18)}$		
	VCD	83.69 $_{(\pm 0.11)}$	81.84 $_{(\pm 0.42)}$	86.61 $_{(\pm 0.48)}$	84.16 $_{(\pm 0.01)}$		
GQA	Popular	LLaVA1.5	Regular	78.17 $_{(\pm 0.17)}$	77.64 $_{(\pm 0.26)}$	79.12 $_{(\pm 0.35)}$	78.37 $_{(\pm 0.18)}$
			VCD	80.73 $_{(\pm 0.47)}$	76.26 $_{(\pm 0.68)}$	89.24 $_{(\pm 0.34)}$	82.24 $_{(\pm 0.35)}$
		Qwen-VL	Regular	75.99 $_{(\pm 0.33)}$	78.62 $_{(\pm 0.41)}$	71.40 $_{(\pm 0.38)}$	74.84 $_{(\pm 0.34)}$
			VCD	81.83 $_{(\pm 0.27)}$	80.45 $_{(\pm 0.47)}$	84.09 $_{(\pm 0.32)}$	82.23 $_{(\pm 0.22)}$
	InstructBLIP	Regular	73.87 $_{(\pm 0.58)}$	69.63 $_{(\pm 0.54)}$	84.69 $_{(\pm 0.68)}$	76.42 $_{(\pm 0.52)}$	
		VCD	78.57 $_{(\pm 0.14)}$	74.62 $_{(\pm 0.22)}$	86.61 $_{(\pm 0.48)}$	80.17 $_{(\pm 0.16)}$	
	Adversarial	LLaVA1.5	Regular	75.08 $_{(\pm 0.33)}$	73.19 $_{(\pm 0.49)}$	79.16 $_{(\pm 0.35)}$	76.06 $_{(\pm 0.24)}$
			VCD	76.09 $_{(\pm 0.43)}$	70.83 $_{(\pm 0.45)}$	88.75 $_{(\pm 0.56)}$	78.78 $_{(\pm 0.36)}$
Qwen-VL		Regular	75.46 $_{(\pm 0.63)}$	77.92 $_{(\pm 0.73)}$	71.07 $_{(\pm 0.97)}$	74.33 $_{(\pm 0.71)}$	
		VCD	80.01 $_{(\pm 0.27)}$	77.80 $_{(\pm 0.24)}$	83.85 $_{(\pm 0.35)}$	80.75 $_{(\pm 0.27)}$	
InstructBLIP	Regular	70.56 $_{(\pm 0.53)}$	66.12 $_{(\pm 0.32)}$	84.33 $_{(\pm 1.05)}$	74.12 $_{(\pm 0.58)}$		
	VCD	75.08 $_{(\pm 0.13)}$	70.59 $_{(\pm 0.16)}$	85.99 $_{(\pm 0.10)}$	77.53 $_{(\pm 0.08)}$		

Table 1. Results on POPE. *Regular* decoding denotes direct sampling, whereas *VCD* refers to sampling from our proposed contrastive distribution p_{vod} . The best performances within each setting are **bolded**.

Experiment



Model	Decoding	Object-level		Attribute-level		Total Scores \uparrow
		Existence \uparrow	Count \uparrow	Position \uparrow	Color \uparrow	
LLaVA1.5	Regular	175.67 (± 7.51)	124.67 (± 19.59)	114.00 (± 9.32)	151.00 (± 10.45)	565.33 (± 32.92)
	VCD	184.66 (± 6.81)	138.33 (± 15.68)	128.67 (± 7.21)	153.00 (± 7.58)	604.66 (± 18.76)
Qwen-VL	Regular	155.00 (± 3.54)	127.67 (± 13.36)	131.67 (± 7.73)	173.00 (± 9.75)	587.33 (± 31.06)
	VCD	156.00 (± 6.52)	131.00 (± 6.19)	128.00 (± 3.61)	181.67 (± 5.14)	596.67 (± 11.61)
InstructBLIP	Regular	141.00 (± 13.97)	75.33 (± 14.16)	66.67 (± 3.91)	97.33 (± 16.94)	380.33 (± 40.20)
	VCD	168.33 (± 11.55)	92.33 (± 8.47)	64.00 (± 6.73)	123.00 (± 11.27)	447.67 (± 13.36)

Table 2. Results on the hallucination subset of MME. Regular decoding denotes direct sampling, whereas VCD refers to sampling from our proposed contrastive distribution p_{vod} . The best performances within each setting are **bolded**.

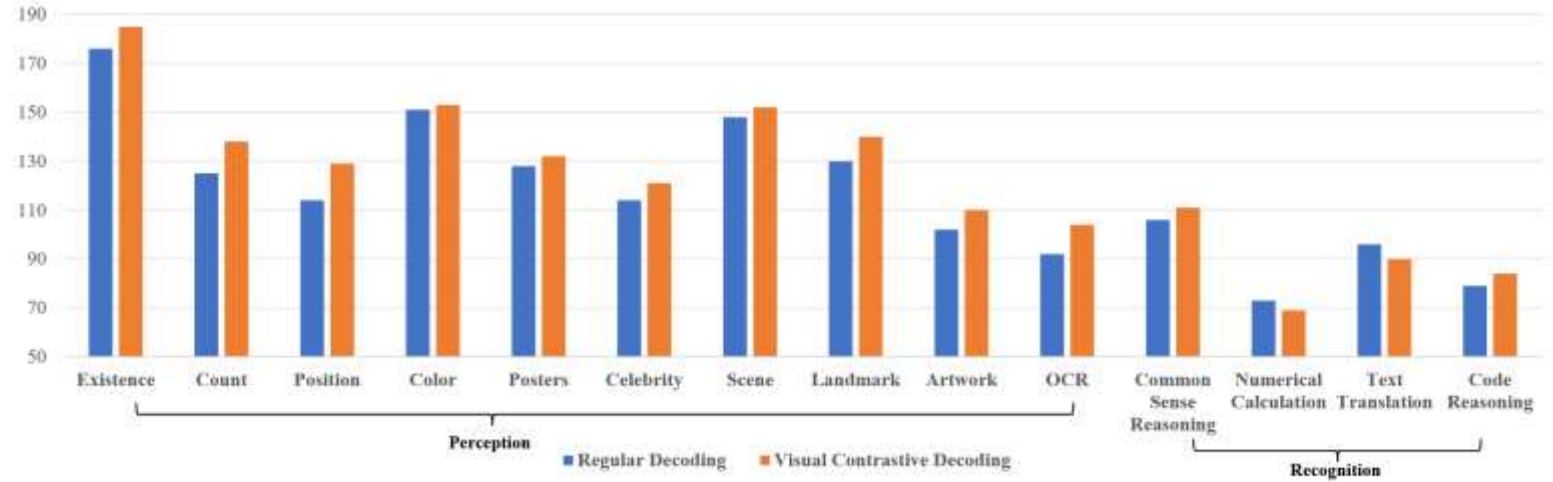


Figure 4. MME full set results on LLaVA-1.5. VCD leads to consistent enhancement in LVLMS' perception capacities while preserving their recognition competencies.

Further Discussions

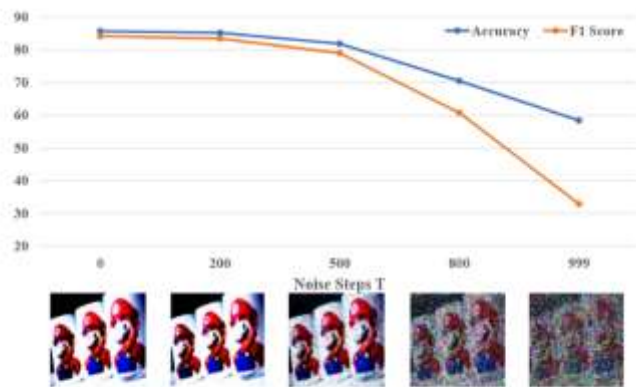


Figure 6. Performance of LLaVA-1.5 on the POPE benchmark across varying noise levels with regular decoding. We visualize the distorted visual inputs subjected to different levels of Gaussian noise at the bottom.

Model	Decoding	Accuracy \uparrow	Detailedness \uparrow
LLaVA-1.5	Regular	3.23	3.54
	VCD	4.15	3.85
InstructBLIP	Regular	3.84	4.07
	VCD	4.23	4.69
Qwen-VL	Regular	4.76	3.46
	VCD	6.69	4.46

Table 3. Results of GPT-4V-aided evaluation on open-ended generation. Accuracy measures the response’s alignment with the image content, and Detailedness gauges the richness of details in the response. Both metrics are on a scale of 10.

Prompt: Describe this photo in detail.

Regular Decoding: The image features a painting of a brown dog sitting in a chair, with a turkey feather in its mouth, giving it a unique, conspicuous appearance. The dog’s expression is one of mischief as it looks directly into the camera. The scene is set in a room with a **dining table** and a **fork** next to the **table**. A rock, likely from the surrounding landscape, can be seen beside the **table**. A **person’s** face is visible near the bottom left corner of the painting, partially covered by the dog’s furry coat.

Visual Contrastive Decoding: The image features a unique and amusing painting of a dog with a large brush mustache, giving it an artistic human-like appearance. The dog is sitting on a couch, with a red background behind it, adding to the artistic flair of the painting. The couch has an orange and green tint to it. In addition to the dog’s human-like mustache, the painting also includes a monochromatic style with the background, giving the painting a striking visual appeal. The dog appears to be wearing a hat, further emphasizing the artistic interpretation of the scene. Overall, the painting captures the dog’s whimsical and charming personality.



Prompt: Describe this photo in detail.

Regular Decoding: In the image, a man is standing on the back of a Subaru in a city with the color of a Ford SUV parked in the background. He appears to be ironing clothes while standing on a wine rack located under the vehicle. He is wearing a yellow shirt. There is a **white and blue taxi** parked nearby the yellow Subaru. Additionally, there are several buildings located in the cityscape, perhaps in the background of the Subaru and the taxi, and possibly overlapping with the yellow Subaru.

Visual Contrastive Decoding: In the photo, a man is ironing clothes on a table placed on the back of a taxi in a busy city street. The man is wearing a yellow shirt, suggesting he is a taxi driver or a part of the yellow cabs’ team. The yellow vehicle in the background is an SUV-like vehicle, which is not a typical yellow cab. The table on the back of the car is set up with a pile of clothes, which the man is ironing. The man is working hard to get the clothes as wrinkle-free as possible.



Figure 5. Illustration of hallucination correction by our proposed VCD with two samples from LLaVA-Bench. Hallucinated objects from LVLM’s regular decoding are highlighted in red.



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models

**Junzhe Chen^{1,2*}, Tianshu Zhang^{1*}, Shiyu Huang³, Yuwei Niu⁴,
Linfeng Zhang⁵, Lijie Wen^{1†}, Xuming Hu^{2†}**

¹Tsinghua University, ²The Hong Kong University of Science and Technology (Guangzhou),

³Zhipu AI, ⁴Chongqing University, ⁵Shanghai Jiao Tong University.

chenjz24@mails.tsinghua.edu.cn,

wenlj@tsinghua.edu.cn, xuminghu97@gmail.com,

CVPR 2025

Introduction

Problem:

VCD eliminate language priors to reduce the model's over-reliance on text modality

The language prior that identifies Curry as a basketball player is removed. This language prior is valuable for achieving an accurate interpretation.

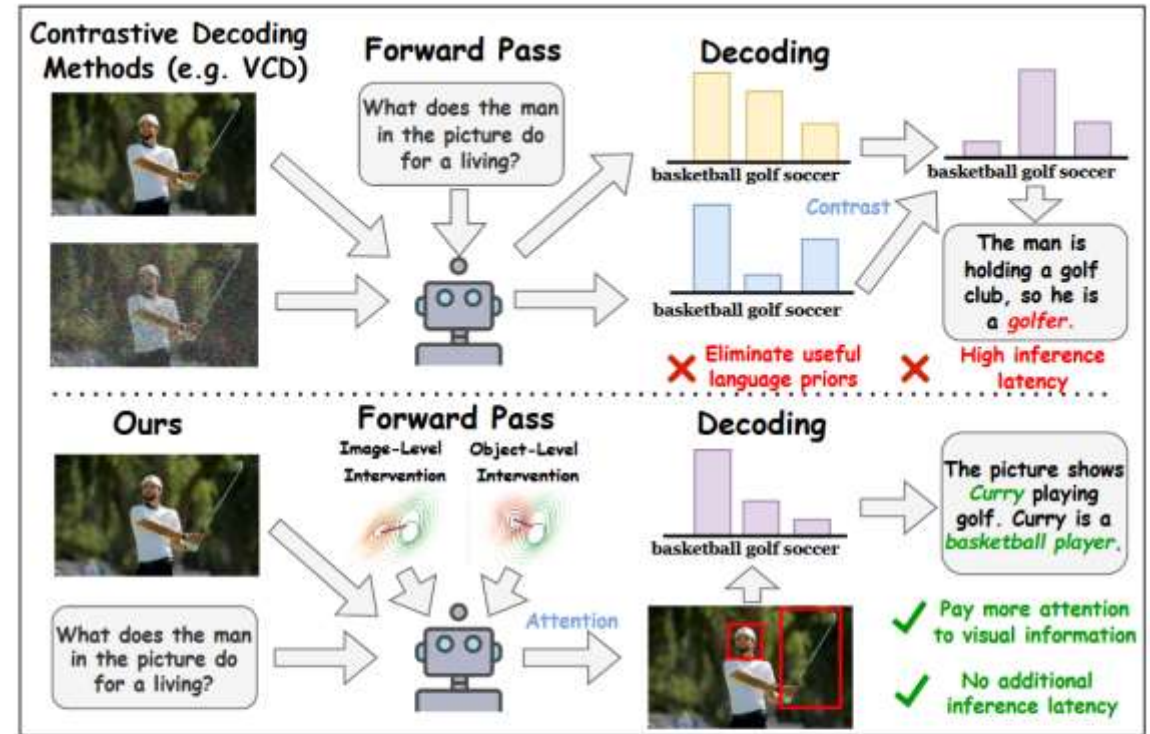


Figure 1. Comparison between Contrastive Decoding (top) and our proposed ICT (bottom). In the top example, Contrastive Decoding indiscriminately removes both beneficial and detrimental language priors, leading to hallucinations. In contrast, our approach enhances the model's attention to visual details while preserving useful language priors, allowing it to correctly identify and describe objects in the image.

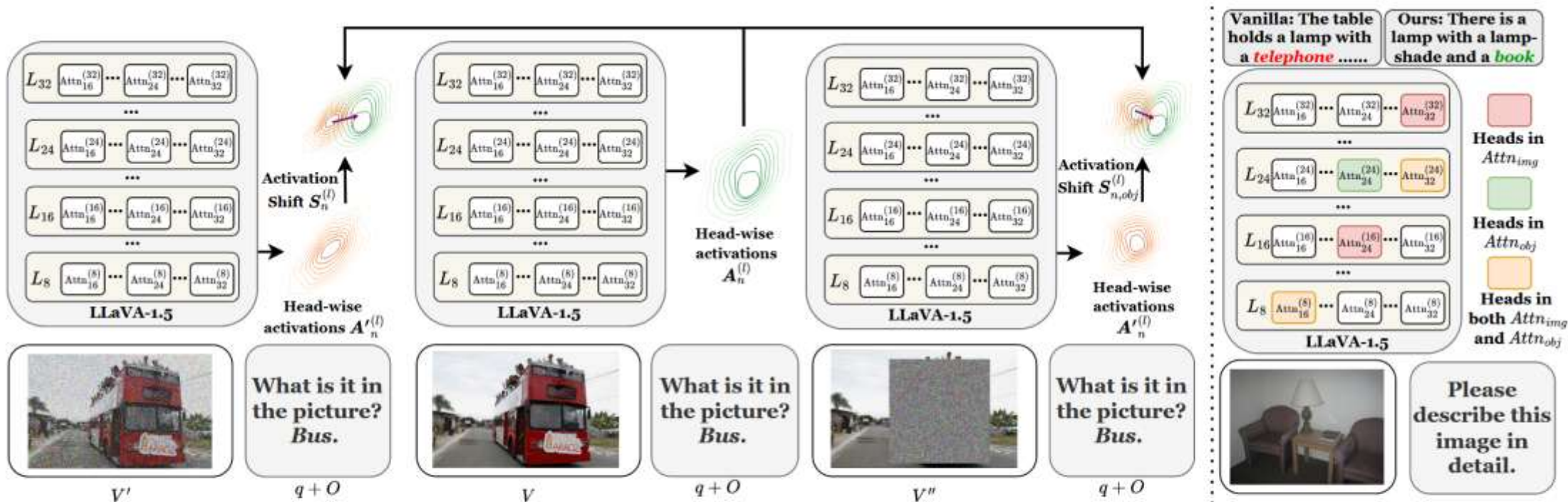


Figure 2. Overview of our proposed ICT method. ICT consists of two levels of intervention: Image-Level and Object-Level. The Image-Level module enhances the model’s focus on the overall scene to reduce reliance on language priors, while the Object-Level module focuses on fine-grained details to mitigate hallucinations. We apply targeted activation shifts to selected attention heads identified through binary classifiers trained to distinguish trusted and untrusted data pairs.

Method--Intervention at Image-Level

高斯噪声扩散生成模糊图像:

$$q(V_i^{(t)} | V_i^{(t-1)}) = \mathcal{N}(V_i^{(t)}; \sqrt{1 - \beta_t} V_i^{(t-1)}, \beta_t \mathbf{I}),$$

$$q(V_i' | V_i) = \prod_{t=1}^T q(V_i^{(t)} | V_i^{(t-1)}), \quad (3)$$

为样例构造可信/不可信数据对:

$$(q + O_i, V_i) \text{ and } (q + O_i, V_i')$$

提取样例所有头的注意力激活值:

$$A_i = \{Attn_n^{(l)} \mathbf{H}_s^{(l)}\}_{n=1, l=1}^{N, L}$$

$$A_i' = \{Attn_n^{(l)} \mathbf{H}'_s^{(l)}\}_{n=1, l=1}^{N, L}$$

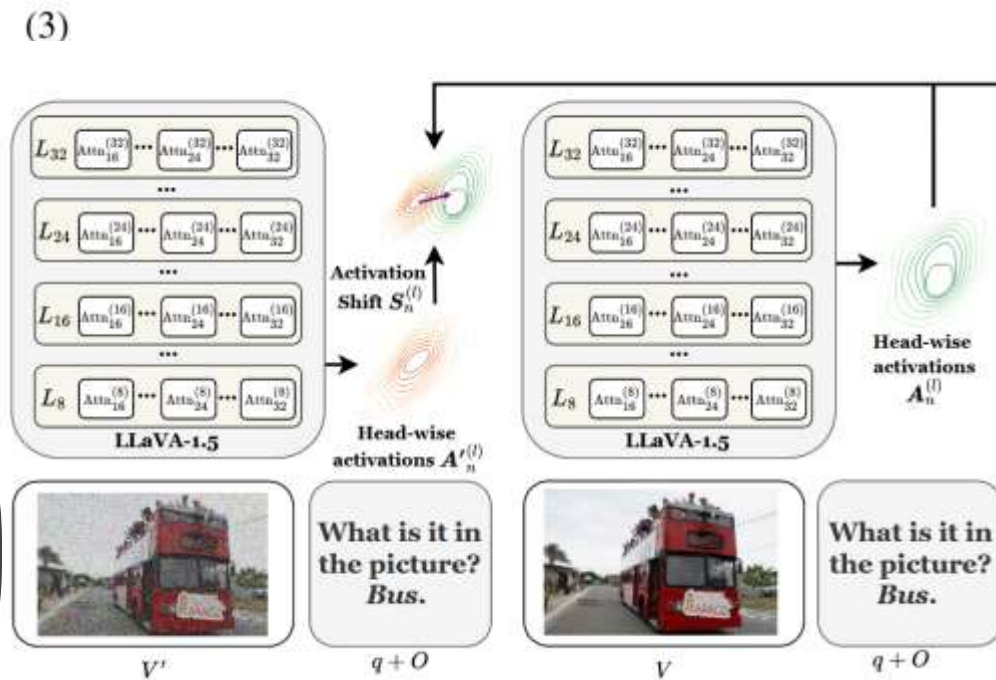
图像级干预的偏移向量计算:

$$\mathbf{S}_n^{(l)} = \frac{1}{B} \sum_{i=1}^B (A_{i,n}^{(l)} - A'_{i,n}^{(l)}). \quad (4)$$

注意力头筛选与更新:

$$Attn_{img} = \{Attn_n^{(l)} | Attn_n^{(l)} \in TopK(\text{Accuracy}(f_n^{(l)}(\cdot)))\},$$

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{n=1}^N (Attn_n^{(l)}(\mathbf{H}^{(l)}) + \mathbb{I}_{img,n}^{(l)} \alpha \mathbf{S}_n^{(l)}) \cdot W_o^{(l)}, \quad (5)$$



两级干预结合

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{n=1}^N (Attn_n^{(l)}(\mathbf{H}^{(l)}) + \mathbb{I}_{img,n}^{(l)} \alpha \mathbf{S}_n^{(l)} + \mathbb{I}_{obj,n}^{(l)} \beta \mathbf{S}_{obj,n}^{(l)}) \cdot W_o^{(l)}. \quad (8)$$

Experiment



Dataset	Setting	Method	LLaVA-v1.5		Qwen-VL	
			Accuracy	F1 Score	Accuracy	F1 Score
COCO	Random	Regular	83.29	81.33	84.37	82.67
		VCD	87.73	87.16	88.63	87.81
		OPERA	89.20	88.81	87.31	86.92
		ICT w/o image	89.70 (+6.41)	89.93 (+8.60)	88.76(+4.39)	87.84(+5.17)
		ICT w/o object	88.50 (+5.21)	88.81 (+7.48)	89.13(+4.76)	88.20(+5.63)
		ICT	90.11 (+6.82)	90.03 (+8.70)	89.46(+5.09)	89.20(+6.63)
	Popular	Regular	81.88	80.06	84.13	82.06
		VCD	85.38	85.06	87.12	86.40
		OPERA	86.64	86.62	87.44	86.68
		ICT w/o image	87.43 (+5.55)	87.09 (+7.03)	87.63(+3.50)	86.78(+4.72)
		ICT w/o object	87.30 (+5.42)	86.65 (+6.59)	87.10(+2.97)	86.28(+4.22)
		ICT	87.50 (+5.62)	87.60 (+7.54)	88.16(+4.03)	87.33(+5.27)
	Adversarial	Regular	78.96	77.57	82.26	80.37
		VCD	80.88	81.33	84.26	83.90
		OPERA	81.24	81.38	84.78	83.45
		ICT w/o image	84.13 (+5.17)	83.31 (+5.74)	85.13(+2.87)	84.35(+3.98)
		ICT w/o object	83.63 (+4.67)	82.13 (+4.56)	85.23(+2.97)	84.62(+4.25)
		ICT	84.43 (+5.47)	83.74 (+6.17)	84.96(+2.70)	84.42(+4.05)
AOKVQA	Random	Regular	83.45	82.56	86.67	85.59
		VCD	86.15	86.34	89.22	89.01
		OPERA	88.02	84.59	88.19	88.43
		ICT w/o image	89.60 (+6.15)	89.75 (+7.19)	89.46(+2.79)	89.28(+3.69)
		ICT w/o object	89.00 (+5.55)	88.91 (+6.35)	89.26(+2.59)	88.95(+3.36)
		ICT	89.20 (+5.75)	89.41 (+6.85)	89.46(+2.79)	89.03(+3.44)
	Popular	Regular	79.90	79.59	85.56	84.63
		VCD	81.85	82.82	87.85	<u>87.81</u>
		OPERA	83.22	84.67	87.91	87.13
		ICT w/o image	85.36 (+5.46)	85.34 (+5.75)	87.70(+2.14)	87.68(+3.05)
		ICT w/o object	84.65 (+4.75)	84.32 (+4.73)	87.90(+2.34)	87.48(+2.85)
		ICT	85.73 (+5.83)	85.34 (+5.75)	88.13(+3.57)	87.83(+3.20)
	Adversarial	Regular	74.04	75.15	79.57	79.50
		VCD	74.97	77.73	81.27	<u>82.38</u>
		OPERA	73.82	77.91	80.82	81.54
		ICT w/o image	79.26 (+5.22)	80.14 (+4.99)	81.50(+1.93)	82.10(+2.60)
		ICT w/o object	77.79 (+3.75)	79.53 (+4.38)	81.40(+1.83)	82.18(+2.68)
		ICT	79.60 (+5.56)	80.43 (+5.28)	81.94(+2.37)	82.44(+2.94)

Random	Regular	83.73	82.95	80.97	79.01
	VCD	86.65	86.99	85.59	85.33
	OPERA	88.13	88.91	86.02	85.29
	ICT w/o image	89.03 (+5.30)	88.99 (+6.04)	85.20(+4.23)	85.90(+6.89)
	ICT w/o object	88.97 (+5.24)	89.30 (+6.35)	87.20(+6.23)	86.80(+7.79)
	ICT	89.60 (+5.87)	89.44 (+6.49)	86.38(+5.41)	86.96(+7.95)
Popular	Regular	78.17	78.37	75.99	74.84
	VCD	80.73	82.24	81.83	82.23
	OPERA	79.27	82.11	81.97	82.12
	ICT w/o image	84.23 (+6.06)	83.92 (+5.55)	81.43(+5.44)	80.86(+6.02)
	ICT w/o object	84.70 (+6.53)	<u>84.45 (+6.08)</u>	83.93(+7.94)	82.95(+8.11)
	ICT	84.70 (+6.53)	84.78 (+6.41)	<u>82.63(+6.64)</u>	82.22(+7.38)
Adversarial	Regular	75.08	76.06	75.46	74.33
	VCD	76.09	78.78	80.01	80.75
	OPERA	75.00	78.71	80.24	80.64
	ICT w/o image	81.92(+6.84)	82.11 (+6.05)	80.03(+4.57)	79.90(+5.47)
	ICT w/o object	81.65 (+6.57)	82.00 (+5.94)	81.60(+6.14)	81.53(+7.10)
	ICT	81.50 (+6.42)	82.27 (+6.21)	<u>80.83(+5.37)</u>	80.60(+6.17)

Table 1. Main results on POPE tasks. We evaluate the accuracy of various LVLMs on the POPE task across the MSCOCO, A-OKVQA, and GQA datasets. **Regular** represents the setting where direct sampling is applied. **ICT w/o image** and **ICT w/o object** correspond to the exclusion of image-level and object-level interventions, respectively. The **bold** and underlined values indicate the highest and second-highest metrics under each setting, respectively.

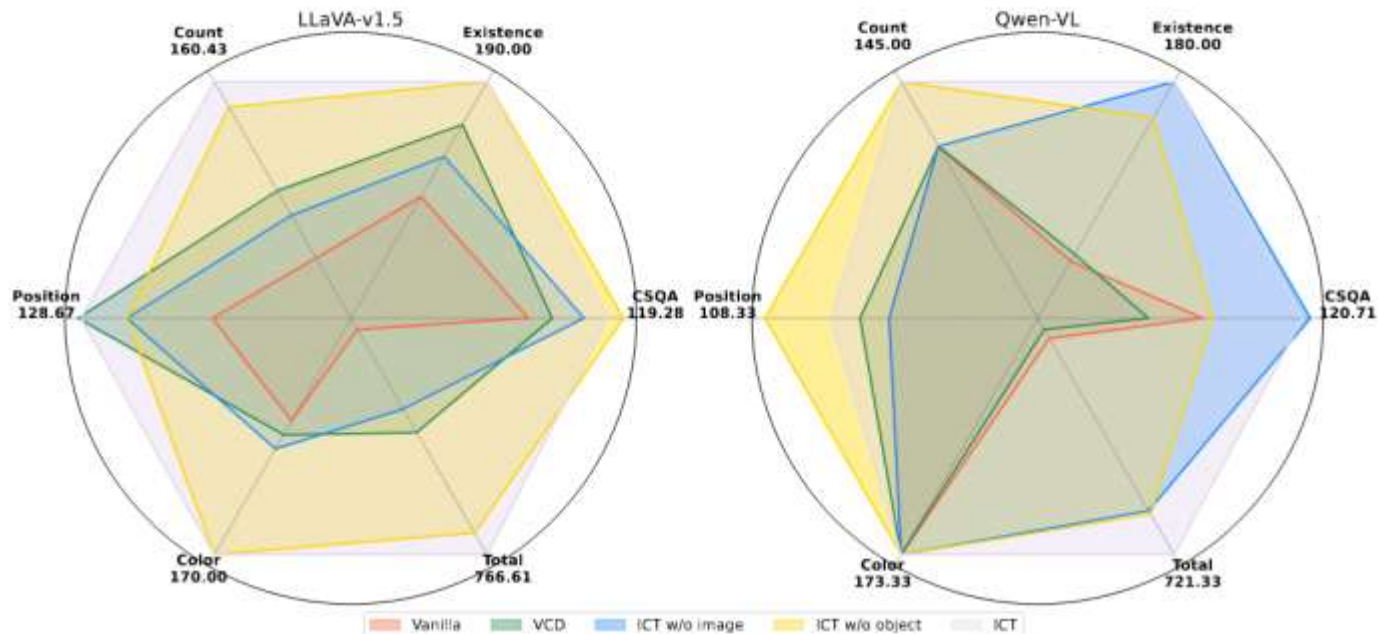


Figure 3. Comparison of ICT with baseline methods (Vanilla and VCD) on the MME benchmark. The radar chart illustrates improvements across various evaluation categories, including existence, position, count, color, and commonsense QA (CSQA).

Method	20-Token Len	50-Token Len	80-Token Len
LLaVA-v1.5	405.3 $\uparrow \times 1.0$	934.6 $\uparrow \times 1.0$	1440.0 $\uparrow \times 1.0$
+ VCD	988.9 $\uparrow \times 2.4$	2031.7 $\uparrow \times 2.2$	3077.5 $\uparrow \times 2.1$
+ OPERA	1371.7 $\uparrow \times 3.4$	3294.1 $\uparrow \times 3.5$	5717.3 $\uparrow \times 3.6$
+ ICT	415.9 $\uparrow \times 1.0$	931.9 $\uparrow \times 1.0$	1485.5 $\uparrow \times 1.0$

Table 2. Comparison of the efficiency of different methods in generating tokens of varying lengths on an NVIDIA H800 GPU. Inference times are recorded in milliseconds.

Setting	Method	Accuracy	F1 Score
<i>Random</i>	Regular	80.97	79.01
	VCD	85.59	85.33
	ICT	86.96	86.38
	ICT-LLaVA-v1.5	85.10	83.27
<i>Popular</i>	Regular	75.99	74.84
	VCD	81.83	82.23
	ICT	82.63	82.22
	ICT-LLaVA-v1.5	81.50	80.10
<i>Adversarial</i>	Regular	75.46	74.33
	VCD	80.01	80.75
	ICT	80.83	80.60
	ICT-LLaVA-v1.5	79.73	78.68

Table 3. **Cross-model generalization of ICT.** Performance of Qwen-VL on GQA after applying activation shift from LLaVA-v1.5. ICT-LLaVA-v1.5 refers to the results obtained by applying LLaVA-v1.5’s activation shift vectors to intervene in Qwen-VL.

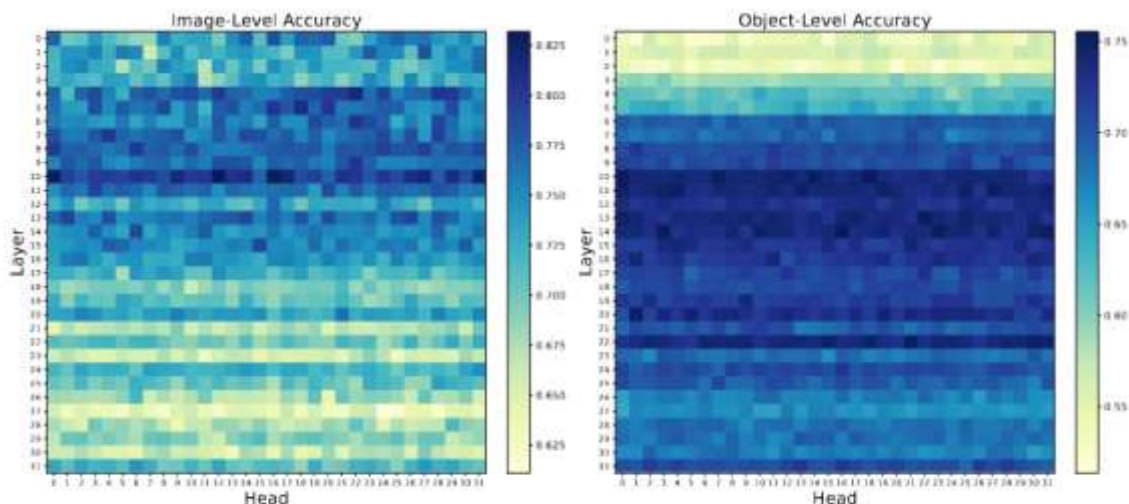


Figure 7. Heatmap of classification accuracy for Image-Level (left) and Object-Level (right) features of LLaVA-v1.5.

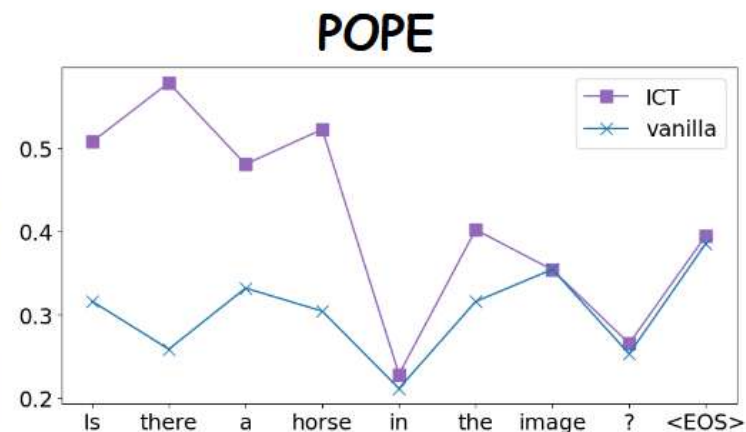
Case Study and Error Analysis



Question: Is there a horse in the image?

VCD: **Yes**, there is.

ICT: **No**, there isn't.



Question: How many uncut fruits are in the image?

VCD: There are **two** uncut fruits in the image.

ICT: There are **four** uncut fruits in the image.

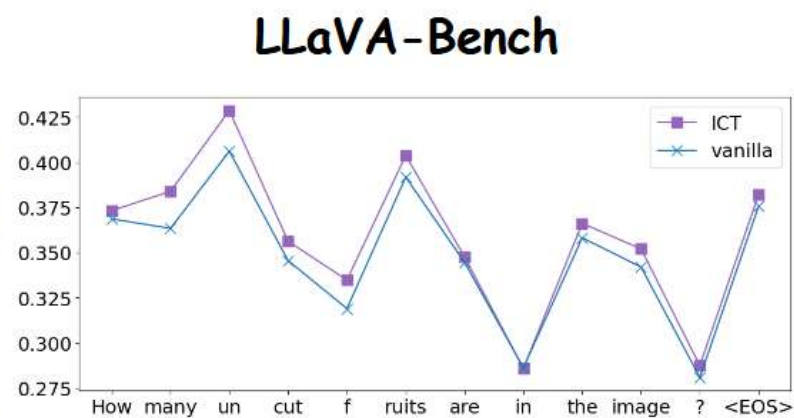


Figure 5. Case Study and Error Analysis of ICT.