Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) are increasingly adept at generating contextually detailed and coherent responses from visual inputs. However, their application in multimodal decision-making and open-ended generation is hindered by a notable rate of hallucinations, where generated text inaccurately represents the visual contents. To address this issue, this paper introduces the Instruction Contrastive Decoding (ICD) method, a novel approach designed to reduce hallucinations during LVLM inference. Our method is inspired by our observation that what we call disturbance instructions significantly exacerbate hallucinations in multimodal fusion modules. ICD contrasts distributions from standard and instruction disturbance, thereby increasing alignment uncertainty and effectively subtracting hallucinated concepts from the original distribution. Through comprehensive experiments on discriminative benchmarks (POPE and MME) and a generative benchmark (LLaVa-Bench), we demonstrate that ICD significantly mitigates both object-level and attribute-level hallucinations. Moreover, our method not only addresses hallucinations but also significantly enhances the general perception and recognition capabilities of LVLMs.

1 Introduction

Recent research in large vision-language models (LVLMs) (Liu et al., 2023c,b; Li et al., 2023a) has seen remarkable progress, benefiting from the integration of advanced large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023a,b) known for their robust language generation and zero-shot transfer capabilities. In order to leverage off-the-shell LLMs, it is crucial to facilitate cross-modal alignment. LLaVa (Liu et al., 2023c) employs a linear projection approach, while BLIP-2 (Li et al., 2023a) and InstructBLIP (Liu et al., 2023b) narrow the modality gap using a Q-Former. Although LVLMs have shown promising outcomes, the issue of hallucination remains. This phenomenon occurs when the generated textual content, despite being fluent and coherent, does not accurately reflect the factual visual content.

The object hallucination was initially explored within the realm of image captioning (Rohrbach et al., 2018). As LVLMs harness the sophisticated understanding and generative prowess of LLMs, the scope of hallucination extends beyond mere object existence. It now encompasses more complex elements such as attributes and relationships within the generated content. Consequently, distinguishing discriminative hallucination and the non-hallucinatory portion in the generation has become pivotal in assessing the performance of LVLMs in terms of their fidelity to factual visual information.

The intertwined nature of modalities presents significant challenges in identifying the root causes of hallucinations in LVLMs. Research efforts have begun to uncover the primary contributors to LVLM hallucinations, including statistical biases (You et al., 2023) encountered during the training process and excessive dependence on language priors (Yan et al., 2023; Zhibo et al., 2023). Additionally, multimodal misalignment has been identified as a key factor in the occurrence of hallucinations (Jiang et al., 2023; Liu et al., 2023a). To address dataset bias, annotation enrichment techniques (Gunjal et al., 2024; You et al., 2023; Zhai et al., 2023) have been introduced. Furthermore, to counteract the influence of language priors, post-processing strategies (Yin et al., 2023; Zhou et al., 2023) have been developed, along with comprehensive initiatives aimed at improving multimodal alignment through optimizing alignment with humans (Sun et al., 2023; Jiang et al., 2023). While these interventions have proven to be effective in reducing hallucinations, they demand substantial human involvement and incur significant computational costs for additional training or the integration of supplementary modules.

In this work, we reveal that appending instructions with role prefixes to form disturbance instructions can significantly exacerbate hallucinations. We hypothesize that identifying and subsequently detaching hallucination concepts from the original distribution could effectively reduce such hallucinations. Motivated by this insight, we introduce the Instruction Contrastive Decoding (ICD) method. This approach is novel in that it is training-free and agnostic to the underlying LVLMs. ICD differentiates between two distributions: one from the original instruction and another from the disturbance instruction within the multimodal alignment module. Utilizing their difference, we aim at suppressing hallucinations. Through comprehensive experiments on discrimination hallucination benchmarks such as POPE (Li et al., 2023c) and MME hallucination sets (Fu et al., 2023), as well as the generation hallucination benchmark LLaVa-Bench (Liu et al., 2023c), our method incorporating state-of-the-art LVLMs like miniGPT4 and InstructBLIP, demonstrates significant efficacy in mitigating hallucinations at both object and attribute levels. Furthermore, our approach consistently enhances performance across general perception and recognition tasks. Our main **contributions** are as follows:

- We perform an in-depth analysis of how disturbance in instructions exacerbates hallucinations. This phenomenon is elucidated through statistical bias and language priors, offering a nuanced understanding of underlying causes.
- Drawing on these insights above, we introduce the ICD method. This novel strategy, which emphasizes initial highlight followed by de-emphasize of hallucination, effectively mitigates hallucinations during inference, by adjusting the distributions away from hallucinations that we elicit.
- Through extensive experimentation and analysis, we validate the effectiveness of our proposed ICD method across both discrimination and generation hallucination benchmarks, showcasing its robustness and versatility in enhancing LVLMs performance.

2 Related Work

Large Vision-Language Models: The field of vision-language pre-training (VLP) (Radford et al., 2021; Li et al., 2022; Bao et al., 2022; Wang et al., 2023a) and fine-tuning (Wang et al., 2023b; Wiehe et al., 2022; Alayrac et al., 2022) have seen rapid advancements, propelled by the evolution of large language models (LLMs). As a result, large vision-language models (LVLMs) have emerged, leveraging the strengths of frozen LLMs while emphasizing the facilitating of multimodal alignment modules. Notably, models such as LLaVa and Qwen-VL (Bai et al., 2023) adopt simple linear projections to achieve alignment, contrasting with BLIP-2 and miniGPT4 (Zhu et al., 2023), which introduce a Q-Former. In further work, InstructBLIP integrates task-aware instructions, enriching the understanding of task-aware visual semantics. Our research builds upon these advancements in LVLMs, focusing on the impact of instruction disturbances. We explore how such disturbances increase the uncertainty in multimodal alignment, significantly contributing to the exacerbation of hallucinations.

Hallucination in VLMs: Hallucination manifests as detailed, fluent, and coherent responses that inaccurately reflect the visual context, including erroneous objects, attributes, and relations (Liu et al., 2024). Various strategies have been proposed to curb hallucinations. Annotation enrichment techniques like M-HalDetect (Gunjal et al., 2024) and GRIT (You et al., 2023), as well as approaches such as HACL (Jiang et al., 2023) and LLaVA-RLHL (Liu et al., 2023a), seek to improve alignment with human instructions through additional annotations. Similarly, Woodpecker (Yin et al., 2023) introduces a post-processing aimed at mitigating biases from language priors. While these methods have shown promise in reducing hallucinations, they often require extensive data annotation, fine-tuning, and supplementary modules, complicating their implementation. In contrast, our method directly addresses hallucinations during inference. Additionally, (Leng et al., 2023) introduced a visual contrastive decoding (VCD) approach that contrasts with the distributions of distorted visual inputs, a concept that bears resemblance to our method. However, our ICD method suppresses hallucinations through disturbance instructions affecting multimodal alignment.

3 Method

3.1 Inference in LVLMs

Large Vision-Language Models (LVLMs) are comprised of three pivotal components: a visual encoder, a fusion module, and a language model. For

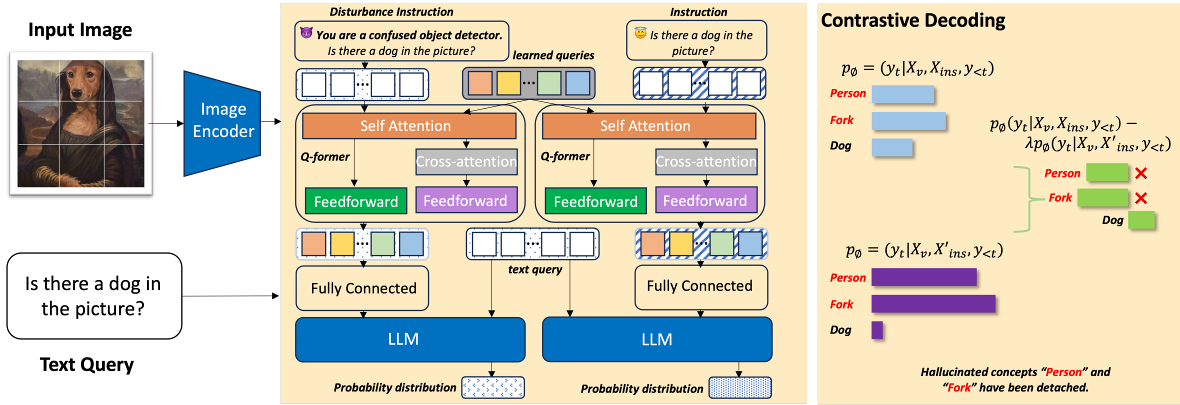

Figure 1: **An illustration on inference framework and contrastive decoding process of ICD method.** At the core (middle orange box), the framework integrates a frozen image encoder, LLM, and query vectors (gray box) within the Q-Former, focusing solely on adjusting the standard and disturbance instructions. The latter, exemplified by adding role prefixes like **'You are a confused object detector,'** aims to increase multimodal alignment uncertainty. This results in two distinct distributions: one from the standard instruction and another influenced by the disturbance. The contrastive decoding method (right orange box) highlights how disturbance instructions amplify hallucinated concepts ('person and fork'), which are then corrected by subtracting probabilities derived from the standard instruction, ensuring accurate recognition of the correct concept 'dog'.

processing an input image, a pre-trained visual encoder, such as ViT-L/14 from CLIP (Radford et al., 2021), is employed to extract visual features, denoted as $\mathbf{X_V}$. The fusion module facilitates multimodal alignment. For instance, InstructBLIP introduces an instruction-aware querying transformer. Q-Former, a lightweight transformer architecture, utilizes K learnable query vectors $\mathbf{Q_K}$ to refine the extraction of visual features, thereby enhancing multimodal alignment. It allows the instruction $\mathbf{X_{ins}}$ to interact with the query vectors, fostering the extraction of task-relevant image features:

$$Z_V = Q_\theta(X_V, Q_K, X_{ins}), \tag{1}$$

where, $Z_V = Q_\theta(\cdot)$ represents the fused visual features, conditioned on the instructions. Given its sophistication and effectiveness in multimodal alignment, we advocate for the adoption of the instruction-aware Q-Former architecture.

For text queries $\mathbf{X_q}$, a large language model, parameterized by ϕ, such as Vicuna (Chiang et al., 2023), processes the query, leveraging the derived visual features to formulate responses:

$$Y_R = LLM_\phi(H_V, X_{ins}), \tag{2}$$

where $H_V = g(Z_V)$ is the transformation ensuring the same dimensionality as the word embedding of the language model. By default, the instruction is the same as text query for both Q-Former and LLM as $\mathbf{X_{ins}} = \mathbf{X_q}$.

Mathematically, in the decoding phase, the response $\mathbf{R}$ can be defined as a sequence of length $\mathbf{L}$, sampled from a probability distribution:

$$p(Y_R|X_V, X_q) = \prod_{t=1}^{L} p_\phi(y_t|H_V, X_q, y_{<t}), \tag{3}$$

where $y_{<t}$ represents the sequence of generated tokens up to the time steps $(t-1)$. In the decoding phase of LVLMs, hallucinations often emerge when probable tokens lack grounding in the visual context. (Jiang et al., 2023; Liu et al., 2023a) indicates that multimodal misalignment is a critical factor contributing to the generation of hallucinations. Thus, we conduct an in-depth analysis of the fusion module, specifically focusing on multimodal alignment. Our work first demonstrates that instructions within the multimodal alignment module can exacerbate hallucinations. To address this, we introduce instruction disturbance and propose an instruction contrastive decoding method, employing a **highlight and then detach** strategy.

3.2 Instruction Can Amplify Hallucination

Prior studies have attributed the occurrence of hallucinations in LVLMs to statistical biases within multimodal training datasets (You et al., 2023) and an over-reliance on language priors (Yan et al., 2023; Zhibo et al., 2023). Extending this line of observation, we introduce the concept of instruction disturbance in this section. A prefix appended to instructions affects multimodal alignment, thereby exacerbating statistical biases and the over-reliance on language priors.

Introduction of instruction disturbance: We introduce the concept of instruction disturbance, which entails appending a *role prefix* to the original instructions delineated in Section 3.1. This disturbance aims to modulate the multimodal alignment uncertainty within LVLMs. As illustrated in Figure 1, the base instruction *"Describe this photo in detail"* is combined with learned query vectors in the Q-Former. To implement instruction disturbance, we append either positive or negative prefixes to the base instruction. Positive prefixes aim to increase the LVLM's confidence in multimodal alignment. Conversely, negative prefixes are designed to reduce the model's alignment confidence.

$$X_{ins} = \begin{cases} [X_d, X_q] & \textit{if disturbance} \\ X_q & \textit{otherwise} \end{cases}, \tag{4}$$

where X_d denotes the role prefix, and X_q represents the original instruction. Through this method, we strategically influence the LVLM's confidence level in multimodal alignment by either encouraging a more definitive understanding or introducing ambiguity.

Instruction disturbance amplifies statistical biases and language priors: Figure 1 presents the response from InstructBLIP, revealing that the LVLMs generate hallucinated tokens such as "*fork and person*." To further explore this phenomenon, we undertake two specific analyses: the frequent hallucinated object occurrence and the co-occurrence of object hallucinations. Our study utilizes MSCOCO validation set (Lin et al., 2014), a common dataset for LVLM pre-training, to perform hallucination detection across three distinct scenarios: the baseline LVLM, LVLM with a positive disturbance, and LVLM with a negative disturbance. Our analysis focuses on calculating the hallucination ratio, specifically identifying instances where the hallucinated objects are absent from the provided images.

Figure 2 demonstrates that introducing instruction disturbance significantly amplifies the occurrence of hallucinations. Under the influence of negative disturbance, LVLMs are more likely to hallucinate objects that frequently co-occur, such as "*person and dining table*," and show an increased tendency to hallucinate objects that typically co-occur with those actually present in the image, for example, "*fork and person*." This suggests that in-

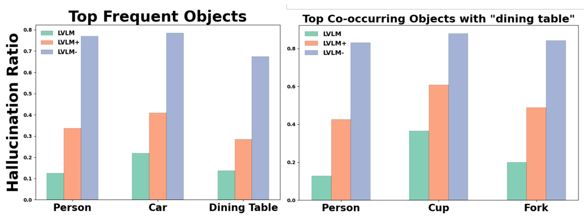

Figure 2: The left figure shows the **top frequent objects hallucination ratio** and the right depicts the **ratio of co-occurring object hallucinations** with *dining table*.

struction disturbances, whether positive or negative, intensify the hallucination effect, exacerbating the issues of imbalanced object distribution and correlation patterns inherent in the training dataset.

3.3 Instruction Contrastive Decoding

3.3.1 Contrastive Decoding with Disturbance

Our analysis reveals that instruction disturbances exacerbate hallucinations by increasing multimodal alignment uncertainty. This uncertainty predisposes LVLMs to more readily adopt biased co-occurrence concepts from pretraining datasets, as reflected in the learned query vectors. As these hallucinations accumulate, LVLMs increasingly over-rely on language priors. Notably, disturbances involving negative prefixes significantly intensify these hallucinations. We hypothesize that by initially emphasizing the probabilities of hallucinated concepts and subsequently detaching these from the original probability distribution, hallucinations may be reduced. Inspired by this insight, we introduce an Instruction Contrastive Decoding method (ICD) aimed at mitigating hallucinations during LVLM inference.

Motivated by the language contrastive decoding (Sennrich et al., 2024) in reducing hallucinations within machine translation frameworks—where it prevents potentially accurate translations that, however, deviate from the desired target language—we adopt a similar approach to our model. Given the extraction of visual features $\mathbf{X_V}$ from the visual encoder and a textual query $\mathbf{X_q}$, our model calculates two distinct token distributions: one conditioned on the original instructions, and the other on instructions with disturbance $\mathbf{X_d}$ as Equation 4. Contrary to the conventional approach of selecting the token that maximizes the probability, our strategy involves choosing the token that concurrently maximizes $p_\phi(y_t|X_V, X_{ins})$ and minimizes $p_\phi(y_t|X_V, X'_{ins})$, the latter representing the probability of tokens that are more likely to be hallucinations. To adjust the balance between these probabilities, we introduce a hyperparameter λ, which regulates the intensity of the contrastive penalty. Formally, this process is described as follows:

$$p_{icd}(Y_R|X_V, X_q) = \prod_{t=1}^{L} \Big(p_\phi(y_t|X_V, X_{ins}, y_{<t}) - \lambda p_\phi(y_t|X_V, X'_{ins}, y_{<t})\Big), \quad (5)$$

where larger λ indicates a more decisive penalty on the decision made by LVLMs with disturbances.

3.3.2 Adaptive Plausibility Constrains

The ICD objective is designed to favor tokens preferred by the LVLM output while imposing penalties on tokens influenced by instruction disturbances. However, this approach might inadvertently penalize accurate predictions—those tokens that, under both standard and disturbance instruction conditions, are confidently identified and are well-grounded in the visual context (such as objects, verbs, attributes, and relations) due to their simplicity and high likelihood. Conversely, it might erroneously reward tokens representing implausible concepts. To address this issue, we draw inspiration from adaptive plausibility constraints utilized in open-ended text generation (Li et al., 2023b). Consequently, we refine the ICD objective to incorporate an adaptive plausibility constraint:

$$\begin{aligned} y_t \sim softmax\Big(&logit_\phi(y_t|X_V, X_{ins}, y_{<t}) \\ &- \lambda logit_\phi(y_t|X_V, X'_{ins}, y_{<t})\Big) \\ &subject\ to\ y_t \in \mathcal{V}_{head}(y_{<t}) \end{aligned} \quad (6)$$

$$\mathcal{V}_{head}(y_{<t}) = \Big\{y_t \in \mathcal{V} : p_\phi(y_t|X_V, X_{ins}, y_{<t}) \geq \alpha \max_{token} p_\phi(token|X_V, X_{ins}, y_{<t})\Big\}, \quad (7)$$

here, α acts as a pivotal hyperparameter that modulates the truncation of the probability distribution, effectively tailoring the LVLM's response to its confidence level. This is particularly crucial for mitigating the influence of implausible tokens, especially when LVLMs exhibit high confidence and are accurately anchored in visual semantics.

ICD serves as a self-corrective mechanism, which successfully identifies hallucinations in LVLMs and then de-emphasizes them through contrastive decoding. Moreover, the integration of adaptive plausibility constraints further hones the contrastive distribution by considering the confidence levels of LVLMs, thereby narrowing the decision-making process to a more reliable candidate pool. This method not only significantly reduces hallucinations within LVLMs but also curtails the generation of implausible tokens, showcasing the efficacy of our proposed method in enhancing model reliability and output validity.

4 Experiment

In this section, we explore the evaluation of our ICD method for mitigating hallucinations. Our examination is twofold: firstly, through the lens of hallucination discrimination, and secondly, via the generation of non-hallucinatory content. More precisely, we assess the efficacy of ICD in alleviating object-level hallucination symptoms utilizing the POPE benchmark. Furthermore, we extend our analysis to include both object and attribute-level symptoms through the MME benchmark. Finally, the performance of our method in generating non-hallucinatory content is evaluated using the LLaVa-Bench dataset.

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics

POPE: The Polling-based Object Probing Evaluation (POPE) stands as a popular benchmark in discerning hallucination at the object level. POPE employs a binary question-answering format, inquiring LVLMs to determine the presence or absence of a specified object within a given image. This benchmark is structured around three distinct subsets—MSCOCO, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019)—each comprising 500 images alongside six questions per image. POPE introduces three settings within each subset: *random* (selecting absent objects at random), *popular* (choosing the most frequently occurring objects in the dataset as absent), and *adversarial* (selecting absent objects that often co-occur with ground-truth objects). We adopt Accuracy, Precision, Recall, and F1 score as the evaluation metrics.

MME: MME benchmark serves as a comprehensive tool for assessing the capabilities of LVLMs across both perception and cognition, spanning a total of 14 tasks. Among these, tasks focusing on *existence, count, position, and color* are specifically designed as hallucination discrimination benchmarks.

Dataset	Setting	Method	miniGPT4 Backbone				InstructBLIP Backbone			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MSCOCO	Random	*default*	67.04	69.06	66.54	67.77	80.71	81.67	79.19	80.41
		+vcd	69.60	72.76	66.73	69.62	84.53	88.55	79.32	83.68
		+icd	**73.51**	**74.36**	**76.87**	**75.60**	**86.43**	**92.01**	**80.73**	**85.61**
	Popular	*default*	60.89	61.34	65.74	63.46	78.22	77.87	78.85	78.36
		+vcd	62.91	63.69	64.81	64.24	81.47	82.89	79.32	81.07
		+icd	**67.61**	**66.69**	**76.87**	**71.42**	**82.93**	**84.45**	**80.73**	**82.55**
	Adversarial	*default*	59.42	59.64	64.45	61.95	75.84	74.30	79.03	76.59
		+vcd	62.07	62.15	66.76	64.37	79.56	79.67	79.39	79.52
		+icd	**64.36**	**63.68**	**75.11**	**68.93**	**80.87**	**80.95**	**80.73**	**80.84**
A-OKVQA	Random	*default*	64.79	65.26	65.73	65.50	80.91	77.97	86.16	81.86
		+vcd	66.68	66.47	68.21	67.33	84.11	82.21	87.05	84.56
		+icd	**69.04**	**68.50**	**77.04**	**72.52**	**85.82**	**83.80**	**88.94**	**86.29**
	Popular	*default*	60.75	60.67	68.84	64.50	76.19	72.16	85.28	78.17
		+vcd	62.22	**62.23**	68.55	65.24	79.78	76.00	87.05	81.15
		+icd	**62.81**	61.62	**75.78**	**67.97**	**81.64**	**78.50**	**88.77**	**83.32**
	Adversarial	*default*	58.88	58.56	68.50	63.14	70.71	65.91	85.83	75.56
		+vcd	60.67	**60.56**	68.47	64.28	74.33	69.46	86.87	77.19
		+icd	**60.71**	59.27	**77.68**	**67.24**	**74.42**	**70.24**	**88.93**	**78.48**
GQA	Random	*default*	65.13	65.38	66.77	66.07	79.75	77.14	84.29	80.56
		+vcd	67.08	68.30	69.04	68.67	83.69	81.84	**86.61**	84.16
		+icd	**72.24**	**75.08**	**79.54**	**77.24**	**85.10**	**84.21**	86.40	**85.29**
	Popular	*default*	57.19	58.55	60.81	59.66	73.87	60.63	84.69	76.42
		+vcd	62.14	**61.14**	72.26	66.24	78.57	74.62	86.61	80.17
		+icd	**62.84**	61.09	**80.54**	**69.48**	**78.80**	**75.15**	**87.53**	**80.87**
	Adversarial	*default*	56.75	56.26	67.99	61.57	70.56	66.12	84.33	74.12
		+vcd	57.78	57.70	69.82	63.18	75.08	70.59	85.99	77.53
		+icd	**59.64**	**58.21**	**76.81**	**66.23**	**75.17**	**70.59**	**86.27**	**77.65**

Table 1: **Results on discrimination hallucination benchmark POPE.** The default under methods denotes the standard decoding, whereas VCD represents visual contrastive decoding (Leng et al., 2023), and ICD is our instruction contrastive decoding. The best performances within each setting are **bolded**. Comparable ($\pm$1.0) but not the best performances between VCD and ICD methods are underlined.

These tasks aim to scrutinize both *object-level* and *attribute-level* hallucination symptoms. MME similarly utilizes a question-answering format to facilitate this evaluation. Consequently, task scores are reported as the evaluation metric for measuring performance.

LLaVa-Bench: The LLaVa-Bench is designed to quantify the extent of hallucinated content produced during the open-ended generation tasks performed by LVLMs. This benchmark encompasses a varied collection of 24 images, accompanied by 60 questions that cover a wide range of scenarios, including indoor and outdoor scenes, memes, paintings, and sketches. Unlike discriminative benchmarks, where accuracy serves as the evaluation metric, generative benchmarks, such as this, currently do not have well-established metrics specifically devised for the detailed analysis of hallucinations (Liu et al., 2024). Therefore, we utilize case studies on this dataset as a means to qualitatively evaluate the effectiveness of our ICD method (see in appendix B).

4.1.2 LVLM Baselines

We employ two state-of-the-art LVLMs as backbone frameworks. Specifically, we implement our ICD on InstructBLIP and miniGPT4, which utilize the Vicuna 7B as their underlying LLM and the sophisticated Q-Former architecture for fusion modules, respectively. Additionally, we explore the use of LLaVa-1.5 (Liu et al., 2023b), which incorporates linear projection for its fusion module alongside InstructBLIP, to identify optimal practices in applying the ICD method (see in appendix D). Finally, we compare our method against the visual contrastive decoding approach (Leng et al., 2023), designed to mitigate hallucinations arising from visual uncertainties. We posit that our method, being LVLM-agnostic, can be conveniently integrated into various off-the-shelf LVLMs.

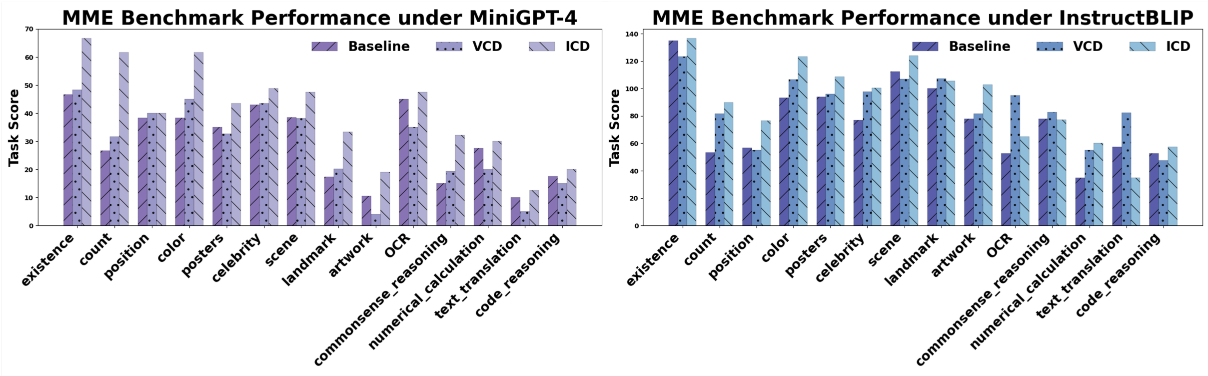

Figure 3: **Performance on MME full benchmark**. The left figure in purple is the results based on miniGPT4, while the right figure in blue is the results based on InstructBLIP.

4.2 Experimental Results

4.2.1 Results on POPE

The experimental results on POPE, summarized in Table 1, demonstrate the efficacy of our instruction contrastive decoding method across three distinct subsets within the POPE benchmark—MSCOCO, A-OKVQA, and GQA settings. Notably, our ICD method consistently outperforms the foundational LVLMs, miniGPT4, and InstructBLIP. Specifically, the ICD method exceeds the performance of miniGPT4 and InstructBLIP, showing a substantial improvement of **10.5%** and **6.0%**, respectively, across all metrics (**7.0%** in accuracy, **8.5%** in precision, **8.7%** in recall, and **7.9%** in F1 score for both models). This significant enhancement as per four metrics on POPE underscores the effectiveness of our *highlight and then detach* strategy.

Furthermore, the progressive movement from *random* to *popular* and then to *adversarial* settings reveals a marked decline in performance, highlighting the growing impact of statistical biases and language prior to contributing to hallucinations in LVLMs. Despite these challenges, our ICD method consistently demonstrates improvements across all settings, affirming our hypothesis that disturbance instruction exacerbates hallucinations by influencing multimodal alignment, thereby deepening errors rooted in statistical bias and over-reliance on language priors, which can be subtracted by contrastive decoding. Our method effectively mitigates these issues and object-level hallucinations.

In comparison to the VCD approach, our ICD method achieves an overall improvement of **3.9%**. While the VCD method aims to ensure that the output distributions are closely aligned with visual inputs and compares distributions derived from distorted images, it requires additional processing to distort images via diffusion models (Ho and Salimans, 2022) and is sensitive to the choice of hyper-parameters in its experimental setup (Leng et al., 2023). Conversely, our ICD method offers a more straightforward and efficient solution, yielding superior results in an end-to-end manner.

LVLM	Method	Object-Level		Attribute-Level		Total Scores
		Existence	*Count*	*Position*	*Color*	
miniGPT4	*default*	46.67	26.67	38.33	38.33	150.00
	+vcd	48.33	31.67	40.00	45.00	165.00
	+icd	**66.67**	**61.67**	**40.00**	**61.67**	**230.01**
InstructBLIP	*default*	135.00	53.33	56.67	93.33	338.33
	+vcd	123.33	81.67	55.00	106.67	366.67
	+icd	**136.67**	**90.00**	**76.67**	**123.33**	**426.67**

Table 2: **Results on the MME hallucination Subset.** The best performances within each setting are **bolded**.

4.2.2 Results on MME

Results on MME Hallucination Subset: The analysis of the POPE benchmark underscores the efficacy of our ICD method in mitigating object-level hallucination symptoms. Given that hallucinations can also manifest at the attribute level (Liu et al., 2024), it becomes imperative to extend our investigation to these dimensions. To this end, we leverage the MME hallucination subset, which encompasses both object-level (*existence and count tasks*) and attribute-level (*position and color tasks*) benchmarks, to conduct a comprehensive evaluation of the ICD method.

As detailed in Table 2, our ICD method significantly surpasses the baseline LVLMs and the VCD method across all four tasks, demonstrating its superior capability in suppressing both object and attribute-level hallucinations with a large margin (**+84.2** and **+62.5** respectively in total scores). Interestingly, while the VCD method experiences a decline in performance on the *position* hallucination task, our method maintains robust performance. This distinction underscores the adaptability and effectiveness of the ICD method in addressing a broader spectrum of hallucination symptoms, making it a more versatile solution in LVLMs.

Results on MME Benchmark: Our method is designed to mitigate hallucinations in LVLMs during inference. We delve deeper into ascertaining whether our approach not only preserves but potentially enhances the fundamental *recognition* and *reasoning* capabilities of LVLMs. To this end, we analyze performance across the full comprehensive MME benchmark, which encompasses 14 subtasks designed to assess *perception* and *recognition*.

Figure 3 illustrates that implementing ICD with both backbone models significantly improves task scores, surpassing the performance of foundation LVLMs and established VCD method. This outcome suggests that our method not only manages hallucinations effectively during inference but also elevates the accuracy of foundational LVLM tasks.

In a more detailed model-specific analysis, our approach consistently outperforms both the backbone miniGPT4 and the VCD method with the same backbone across all 14 subtasks. Conversely, the VCD method exhibits diminished performance in specific areas such as *posters, artwork, OCR, numerical calculation, text translation, and code reasoning* when compared to the baseline LVLM.

Moreover, when InstructBLIP serves as the backbone, the effectiveness of VCD decreases in tasks related to *existence, position, scene, and code reasoning*. We surmise that while leveraging visual uncertainty may anchor predictions more firmly in visual input, it simultaneously introduces drawbacks by fostering an over-reliance on visual cues at the expense of instruction-based grounding. Conversely, our ICD method, by focusing on multimodal alignment, does not compromise the fundamental reasoning capabilities of LVLMs. Notably, our method's performance on the *landmark, OCR, commonsense reasoning, and text translation* tasks under InstructBLIP is weaker than the VCD method, whereas VCD exhibits superior results in these domains. This suggests that these subtasks within the MME benchmark may demand a robust visual discrimination capability.

4.3 Discussions on ICD and VCD

In addressing hallucinations in LVLMs, our ICD method and the baseline VCD both leverage contrastive decoding tailored for open-ended generation (Li et al., 2023b). While our ICD method introduces disturbance instructions to increase multimodal alignment uncertainty, VCD employs distorted images to amplify visual uncertainty. Positing that a synergistic approach could harness the strengths of both methods, we propose to analyze a straightforward integration of these two methods.

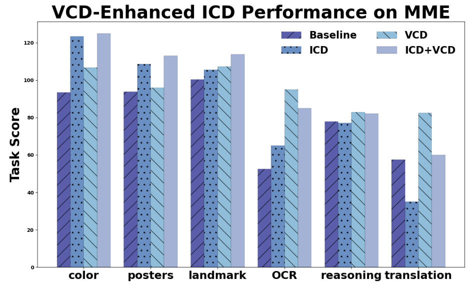

Figure 4: **Performance of the VCD-enhanced ICD method on MME Subset.** The underlying LVLM is InstructBLIP.

Our combined approach begins with the VCD, utilizing standard instructions. This is followed by contrasting the resulting distribution with that of a VCD output generated under disturbance instructions, thereby establishing the final output distribution. Figure 4 showcases the integration method on *color, posters, landmarks, OCR, commonsense reasoning, and text translation*. This approach yields notable enhancements across these subtasks, underlining the importance of discriminative visual features and multimodal alignment as complements in grounding LVLM responses.

This exploration suggests a promising avenue for future research aimed at optimally amalgamating the advantages of both methods. Detailed results and comprehensive analysis of the combined method performance across full MME are provided in the appendix C for further reference.

5 Conclusion

We introduce a novel instruction contrastive decoding approach that effectively detaches hallucinatory concepts by contrasting distributions derived from standard and disturbance instructions where role prefixes are appended to amplify hallucinations. Comprehensive experiments across various benchmarks and different LVLMs demonstrate the capability of our method in mitigating hallucinations and substantially improving the general perception and recognition performance of LVLMs.

Limitations

In this paper, we have concentrated on addressing hallucinations within LVLMs by deploying our novel ICD method. We have validated its efficacy through rigorous evaluation on various hallucination discrimination benchmarks and have also qualitatively assessed its performance on generative benchmarks, which are pivotal for examining hallucinatory content. Despite their importance, generative benchmarks currently lack established metrics for thoroughly analyzing hallucinations, indicating a significant area for future research to enhance open-ended generation performance evaluation with robust automatic metrics.

Ethics Statement

We propose the Instruction Contrastive Decoding method to address hallucination issues in LVLMs, thereby enhancing their safety and reliability within the community. Additionally, the datasets utilized for inferring and evaluating the ICD method are publicly accessible, promoting transparency and reproducibility in our research. Furthermore, we have made our code available to the public, ensuring it is convenient for researchers and practitioners to access and implement.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *LMSYS Organization*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. In *Annual Conference on Neural Information Processing Systems (NeurIPS) DGMDA*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *The European Conference on Computer Vision (ECCV)*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *The European Conference on Computer Vision (ECCV)*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023a. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Xintong Wang, Xiaoyu Li, Liang Ding, Sanyuan Zhao, and Chris Biemann. 2023b. Using self-supervised dual constraint contrastive learning for cross-modal retrieval. In *European Conference on Artificial Intelligence (ECAI)*.

Anton Wiehe, Florian Schneider, Sebastian Blank, Xintong Wang, Hans-Peter Zorn, and Chris Biemann. 2022. Language over labels: Contrastive language supervision exceeds purely label-supervised classification performance on chest x-rays. In *The Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics SRW*.

Hong Yan, Lijun Liu, Xupeng Feng, and Qingsong Huang. 2023. Overcoming language priors with self-contrastive learning for visual question answering. *Multimedia Tools and Applications*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint*.

Ren Zhibo, Wang Huizhen, Zhu Muhua, Wang Yichao, Xiao Tong, and Zhu Jingbo. 2023. Overcoming language priors with counterfactual inference for visual question answering. In *China National Conference on Computational Linguistics (CNCC)*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*.

A Implementation Details

In our experiments, we adopted the contrastive decoding configurations by setting the decisive penalty on the decision made by LVLMs with disturbance $\lambda = 1$ and the hyperparameter $\alpha = 0.1$ that modulates the truncation of the probability distribution, in line with the configurations reported in previous studies (Li et al., 2023b; Leng et al., 2023). For the decoding strategy, we uniformly applied the sampling method across all experiments, incorporating a $top\ p = 1$, a $repetition\ penalty = 1$, and a $number\ of\ beams = 1$ for LLMs. For both VCD and ICD methods, we sample from the modified $softmax$ distribution, as delineated in Equation 7.

B Qualitative Evaluation on LLava-Bench

In this section, we extend our analysis by focusing on the evaluation of generative hallucination. Utilizing LLaVa-Bench, we conduct a qualitative analysis on the task of open-ended generation. Figure 5 showcases two case studies that compare our method with backbone LVLMs using identical input images. The example displayed on the left presents various Asian dishes. While the baseline LVLMs accurately identify and generate concepts such as *spoons, tables, and cups*, they also erroneously introduce the unrelated concept of a "*person*." This error stems from the high frequency of co-occurrence between *"person" and "tables"* in the training data. Furthermore, the example on the right depicts a well-known scene from the movie "Titanic." Here, the baseline LVLMs incorrectly perceive the characters Jack and Rose as *two women*, leading to an inaccurate generation of text regarding *same-sex relationships*. This error is a result of the language prior biases, which contribute to hallucinations in LVLMs.

Contrastingly, our ICD approach produces fluent, coherent text that is closely grounded in the visual context, effectively mitigating the hallucinations caused by statistical biases and the inherent language priors of LVLMs.

C Further Analysis on VCD-Enhanced ICD

We comprehensively analyze the ICD and VCD combined method, detailed in Section 4.3, within the full MME benchmark, utilizing InstructBLIP as the backbone LVLM. Figure 6 illustrates that integrating our ICD method significantly enhances the VCD's performance across various tasks, including *existence, count, color, celebrity, scene, landmark, and artwork*. Similarly, incorporating VCD in ICD yields improvements in *color, posters, landmarks, OCR, commonsense reasoning, and translation tasks*. These findings suggest that addressing both visual and multimodal alignment uncertainties in a complementary fashion effectively mitigates hallucinations. However, we also note a performance decrement in the ICD method for *count, position, artwork, calculation, and code reasoning tasks* when combined with VCD. This observation underscores the necessity for more refined combination strategies to fully harness the potential of integrating these two methods.

Combining the strengths of both the ICD and VCD methods has opened a promising avenue for future investigations. We aim to develop and refine contrastive decoding methods for the seamless integration of both techniques, potentially a new method for mitigating hallucinations in LVLMs.

D Optimal Position to Apply Contrastive Decoding

Upon detailed examination of the inference framework depicted in Figure 1, we identify three potential points for integrating the ICD method: within the Q-Former's instruction, the LLM's instruction, and a combination of both. This analysis, based on the POPE GQA Random sub-dataset, aims to pinpoint the optimal application site for ICD. To ensure a comprehensive comparison, we selected two distinct LVLMs, InstructBLIP and LLaVa, as backbones to represent varied fusion approaches. InstructBLIP employs Q-Former for multimodal alignment, whereas LLaVa utilizes a linear projection.

Figure 7 reveals that, under the InstructBLIP framework, ICD enhances performance across all implementation sites, with the singular application within Q-Former yielding the most significant improvement. A comparison between the LVLMs indicates that LLaVa also benefits from the ICD method when ICD is applied within LLMs. However, exclusive application of ICD in LLMs produces less pronounced improvements, mirroring the observations with InstructBLIP as the backbone. Consequently, our findings suggest that deploying the ICD method within the Q-Former architecture

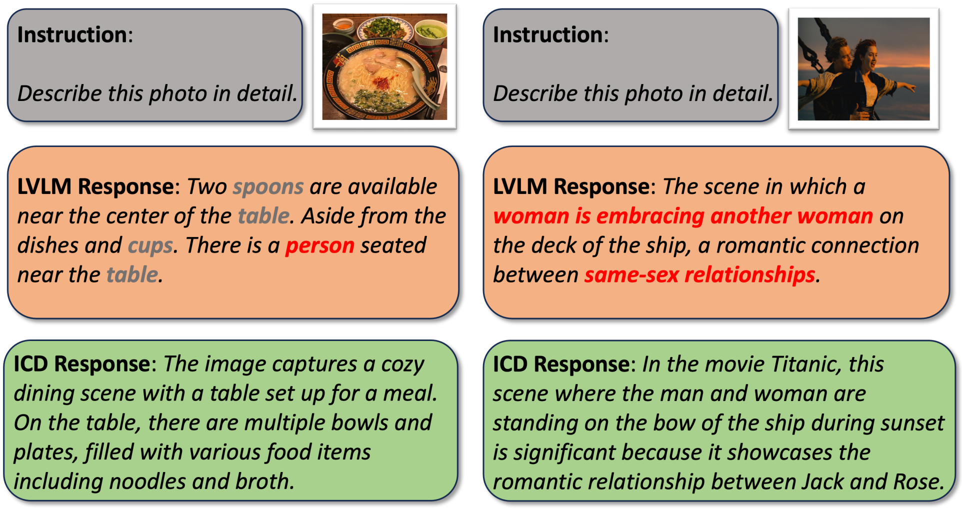

Figure 5: **Qualitative analysis on LLava-Bench.** The left figure highlights the statistical bias, and the right figure shows the language prior that contributes to hallucinations in LVLMs. Hallucinated concepts have been highlighted in red.

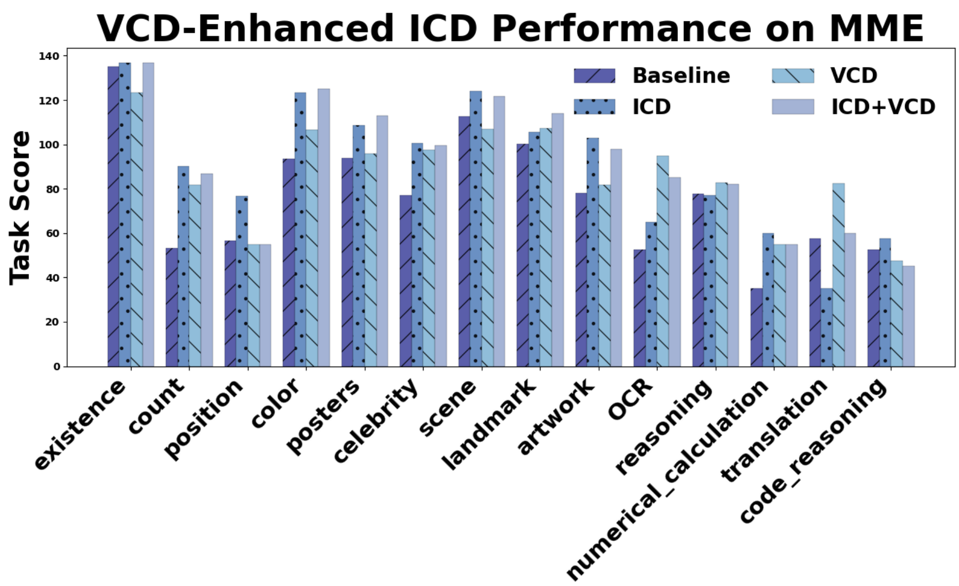

Figure 6: **Performance of the VCD-enhanced ICD method on full MME benchmark.** The underlying LVLM is InstructBLIP. ICD+VCD indicates the combination approach detailed in Section 4.3.

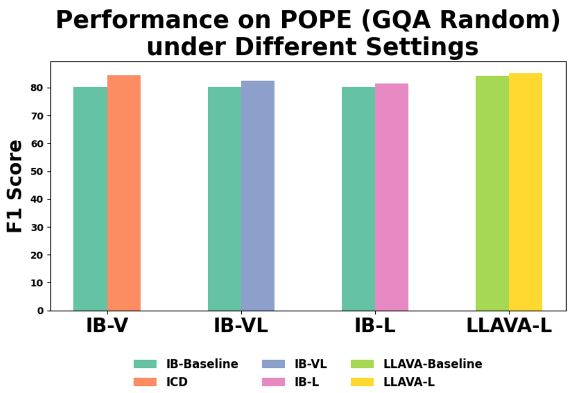

Figure 7: **Performance of the ICD method implemented on difference positions evaluated on POPE (GQA Random) dataset.** The underlying LVLMs are InstructBLIP and LLaVa-1.5.

represents the most effective strategy.