VCGD: Visual Cue Guided Decoding with Caption Model for Mitigating Hallucination in Multimodal Large Language Models

Anonymous EMNLP submission

Abstract

Multimodal large language models (MLLMs) demonstrate strong capabilities in multimodal understanding, reasoning, and interaction but still face the fundamental limitation of hallucinations, where they generate erroneous or fabricated information. We propose Visual Clue-Guided Decoding (VCGD), a novel decoding strategy that incorporates precise visual cues generated by a Caption Model during the decoding phase. These cues serve as comparative references for the model's own outputs, effectively mitigating hallucination phenomena. Specifically, VCGD leverages high-quality visual descriptions to guide MLLMs in correcting perceptual biases while generating answers. Furthermore, we introduce a Reinforcement Learning (RL)-based training paradigm for the Caption Model, in which a Reward Agent provides feedback on the quality of visual clues, further enhancing the accuracy of visual information. Extensive experiments across multiple benchmark datasets and state-of-the-art MLLMs demonstrate that VCGD significantly reduces hallucination rates and substantially improves cross-modal consistency. Our method exhibits strong generalizability and scalability, offering an effective decoding enhancement strategy that can be seamlessly integrated into existing multimodal frameworks. Code is available at `https://anonymous.4open.science/r/VCGD-C860`.

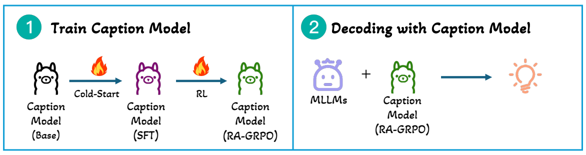


Figure 1: Figure ① illustrates the training procedure of the Caption Model, which consists of two stages: the *Cold-Start Stage* and the *Reinforcement Learning Stage*. Figure ② illustrates the use of the Caption Model to assist the MLLM in contrastive decoding, thereby alleviating hallucination issues.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2023b; Touvron et al., 2023; Achiam et al., 2023; Jiang et al., 2023; Bai et al., 2023a) have marked a pivotal advancement in natural language processing. Building upon their success, researchers have expanded these models into multimodal domains, giving rise to Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Liu et al., 2024d; Team et al., 2023; Bai et al., 2023b). While MLLMs demonstrate remarkable proficiency in tackling a wide array of visual tasks (Zhang et al., 2023; Li et al., 2024; Black et al., 2024), as well as in understanding (Lai et al., 2024) and generating complex content (Brooks et al., 2023; Geng et al., 2024), they are not without limitations. A particularly critical issue is the so-called "*Hallucination*" phenomenon. In practice, MLLMs often produce inaccurate or fabricated responses when interpreting user-provided images and prompts—ranging from irrelevant or nonsensical descriptions to misidentified colors, incorrect object counts, and erroneous spatial relationships within the scene. Such tendencies severely undermine their reliability and present substantial obstacles to their deployment in real-world applications.

The issue of hallucinations in MLLMs originates from the intricacies of their training pipeline, which typically includes an alignment-based projection during pre-training, followed by fine-tuning on a relatively limited amount of instruction-following data. To tackle these hallucination problems, various strategies have been proposed. Some focus on resolving inconsistencies by refining data quality and alignment (Liu et al., 2023a; Wang et al., 2024a), while others emphasize scaling up model architectures (Zhai et al., 2023) or incorporating reinforcement learning-based techniques (Yu et al., 2023; Sun et al., 2024). A distinct line of work involves reactive techniques (Huang et al., 2023;

Deng et al., 2024), which intervene directly during the decoding process to suppress inaccurate outputs. Inspired by contrastive decoding (CD) strategies introduced by Li et al. (2023b), which compare the outputs of expert and amateur models, recent developments in CD-based methods for MLLMs have explored contrasting visual-conditioned generations to suppress hallucinations, taking into account factors such as visual noise (Leng et al., 2023), image-induced bias (Zhu et al., 2024), and detailed visual grounding (Chen et al., 2024), and self-generated description (Kim et al., 2024).

We investigate the following research question: If the model is incapable of accurately perceiving visual content through its own capacity, it becomes highly susceptible to generating hallucinations. Therefore, a key question arises: *can auxiliary models be incorporated to provide precise visual cues during decoding, thereby reducing the likelihood of hallucination*? To address this issue, we propose a novel Visual Cue Guided Decoding strategy. Specifically, we leverage a caption model to provide precise visual cues during decoding, thereby mitigating the inherent biases of the model in visual perception. To further enhance caption quality, we introduce a reinforcement learning algorithm based on a Reward-Agent framework, termed RA-GRPO. This algorithm optimizes the caption generation process by providing reward signals across three dimensions: accuracy, informativeness, and redundancy control. As a result, it facilitates the generation of more precise and pragmatically useful image descriptions. The specific process is briefly illustrated in Figure 1, which includes two main stage: ① *Train Caption Model*, where the Caption Model is obtained through a two-stage training procedure; ② *Decoding with Caption Model*, the Caption Model is employed to assist MLLMs in decoding.

By conducting extensive experiments and analyses on prevailing cutting-edge MLLMs (Liu et al., 2023c; Chen et al., 2023b; Liu et al., 2024c), we corroborate the effectiveness of our method in reducing hallucination in various benchmarks (Li et al., 2023c; Liu et al., 2023d; Tong et al., 2024). Our contribution can be summarized into four-fold as follows:

- We propose **Visual Clue Guided Decoding (VCGD)**, a novel decoding strategy that enhances the model's visual perception by leveraging precise visual clues. These clues are derived from captions generated by an external Caption Model, while suppressing the model's own perception to reduce visual bias.

- We propose an approach for training the Caption Model using reinforcement learning to generate accurate and informative visual clues that effectively guide the decoding process.

- We introduce a reward mechanism for a dedicated **Reward Agent**, which evaluates the quality of the generated captions based on three dimensions: accuracy, informativeness, and redundancy control.

- We evaluate the proposed decoding approach on various benchmarks using state-of-the-art MLMMs. Experimental results demonstrate that VCGD significantly reduces hallucinations.

2 Related Work

2.1 Multimodal Large Language Models

Recent advancements in MLLMs research are primarily attributed to the evolution of LLMs (Wang et al., 2024c; Zhuo et al., 2024; Lu et al., 2024; Su et al., 2024). With the aid of advanced LLMs like LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023a), a batch of MLLMs such as LLaVA-1.5 (Liu et al., 2024d), Qwen-VL (Bai et al., 2023b), LLaVA-NEXT (Liu et al., 2024c) , InternVL (Chen et al., 2023b) and mPLUGOwl2 (Ye et al., 2024) have emerged, which can comprehend and generate a wide array of content by utilizing information from distinct modalities like texts and images. Despite the success, current MLLMs suffer from serious hallucination problems. Thus, in this paper, we focus on mitigating hallucination problems to promote the use of MLLMs in practical scenarios.

2.2 Hallucinations in MLLMs

Hallucinations in MLLMs have significantly impeded their usage in the real world, especially for tasks that rely on precise captions. Recently, numerous studies focus on the construction of datasets for *evaluating* hallucination phenomena (Rohrbach et al., 2018; Li et al., 2023d; Wang et al., 2023; Sun et al., 2023a; Zhong et al., 2024; Tong et al., 2024; Wu et al., 2024; Cao et al., 2024; Huang et al., 2024b; Mubarak et al., 2024; Jing et al., 2024; Lovenia et al., 2023; Zhai et al., 2023; Wan and Bansal, 2022; Zhang et al., 2024; Min et al., 2023;

Yan et al., 2024). Concurrently, significant attention is directed towards *analyzing* the underlying causes of hallucinations (Tao et al., 2024; Sui et al., 2024; Fadeeva et al., 2024).

Moreover, various approaches have been proposed to *mitigate* hallucinations in MLLMs, including training-free and training-based approaches. *Training-based* approaches seek to mitigate hallucinations in MLLMs via further training, such as Supervised Fine-Tuning (SFT) (Liu et al., 2023b) or preference learning (Sun et al., 2023a; Yu et al., 2024a; Li et al., 2023a; Zhao et al., 2023; Gunjal et al., 2024; Liu et al., 2024a; Yu et al., 2024b; Zhou et al., 2024; Jiang et al., 2024; Jing and Du, 2024). *Training-free* approaches address potential hallucinations by post-processing the outputs of MLLMs (Leng et al., 2024; Huang et al., 2024a; Yin et al., 2023; Manevich and Tsarfaty, 2024; Wang et al., 2024b; Kim et al., 2024). For example, VCD (Leng et al., 2024) aims to address the model's over-reliance on linguistic priors and statistical biases by comparing the output distributions from unaltered and visually perturbed inputs. Woodpecker (Yin et al., 2023) introduces post-processing aimed at mitigating biases from language priors. ICD (Wang et al., 2024b) suppresses hallucinations through disturbance instructions affecting multimodal alignment. Similarly, LCD (Wang et al., 2024b) uses visual noise to guide the decoding process to leverage the language modality to mitigate hallucinations. CODE (Kim et al., 2024) utilize self-generated description as contrasting visual counterpart and correct hallucinatory responses based on the model understanding. Inspired by Kim et al. (2024), our work aligns with CD-based approaches that utilize visual clues from Caption Models to guide decoding. Unlike previous works that focus on manipulating information or self-generated descriptions, we argue that if the model fails to accurately identify the visual content and self-generated description, no amount of correction can resolve hallucinations. We utilize visual clues from Caption Models as a contrasting visual counterpart to correct hallucinatory responses.

3 Methodology

Our proposed VCGD framework consists of two main components: training the Caption Model (Section 3.1) and decoding with the Caption Model (Section 3.2).

3.1 Caption Model

3.1.1 Overview

We design and train a Caption Model. As shown in Figure 1, inspired by the DeepSeek-R1 (Guo et al., 2025) approach, the training of the Caption Model is divided into two stages: *the cold-start stage* and *the reinforcement learning (RL) stage*.

Cold-Start Stage. We construct triplet data consisting of images, questions, and answers based on ShareGPT4V (Chen et al., 2023a) and LLaVA-CoT (Xu et al., 2024). We then utilize GPT-4 to summarize the original questions (Q) and answers (A), extracting Visual Cues that are critical for problem solving, which are subsequently used as training targets.

RL Stage. The Visual Cues that support problem solving, exclude irrelevant details, and remain highly consistent with the image content, and we optimize the Caption Model based on our designed Reward-Agent framework (RA-GRPO), as shown in Figure 2.

3.1.2 Reward Agent

We design a Reward Agent comprising three reward signals as illustrated in Figure 2: the Accuracy Reward, the Matching Reward, and the Image-Consistency Reward.

Accuracy Reward. In order that the Visual Clues (VC) can effectively capture visual information and assist the model in understanding image content, we employ a Large Language Model (LLM), incorporating the VC as contextual input—essentially allowing the VC to serve as the LLM's "eyes". These clues represent the visual perception available to the LLM, based on which it generates an answer to the given question. A reward is then assigned depending on the correctness of the answer. The reward is defined as follows:

$$\text{Reward}_{Acc} = \begin{cases} 1.0, & \text{if Rollout}_{Q+VC} = \text{True} \\ 0.0, & \text{if Rollout}_{Q+VC} = \text{False} \end{cases} \quad (1)$$

Matching Reward. This reward signal is designed to encourage the Caption Model to generate VC that minimize the inclusion of redundant information related to the image content but irrelevant to the given question. To achieve this, we introduce LLaMA as a discriminator J. Specifically, the process is as follows: the discriminator J first analyzes the question Q and identifies a set of key points K_i required to answer Q. Subsequently, the alignment between each VC and the identified key

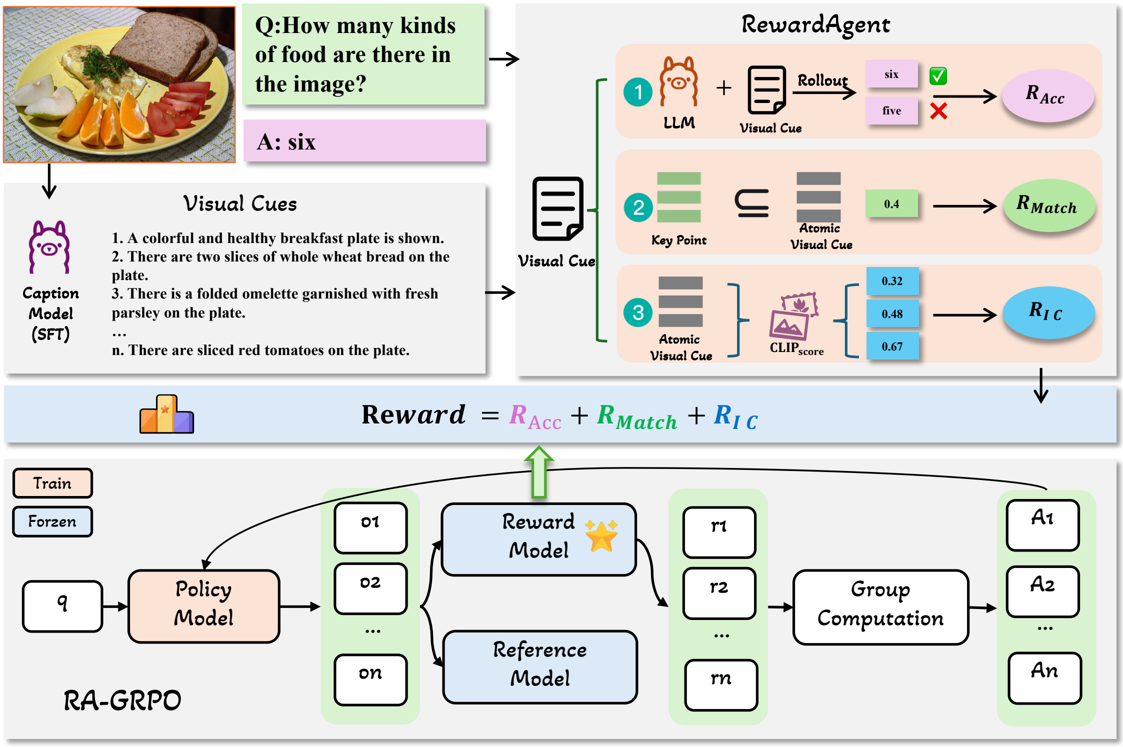

Q:How many kinds of food are there in the image?
A: six
Visual Cues
Caption Model (SFT)
1. A colorful and healthy breakfast plate is shown.
2. There are two slices of whole wheat bread on the plate.
3. There is a folded omelette garnished with fresh parsley on the plate.
...
n. There are sliced red tomatoes on the plate.
Visual Cue
RewardAgent
1
LLM
+
Visual Cue
Rollout
six
five
R_{Acc}
2
Key Point
⊆
Atomic Visual Cue
0.4
R_{Match}
3
Atomic Visual Cue
CLIP$_{score}$
0.32
0.48
0.67
R_{IC}
Reward = R_{Acc} + R_{Match} + R_{IC}
Train
Forzen
q
Policy Model
o1
o2
...
on
Reward Model
Reference Model
r1
r2
...
rn
Group Computation
A1
A2
...
An
RA-GRPO

Figure 2: **Caption Model Train framework:** (1) Reward Agent contains Accuracy Reward, Matching Reward, and Image-Consistency Reward. (2) RL Train: For each data instance, we first employ the Caption Model to sample and generate G Visual Clues (VC). For each VC, the Reward Agent evaluates the corresponding answer candidate and assigns a reward score. The parameters are then optimized using the RA-GRPO algorithm.

points (Key) is computed (Equation (2)) to quantify the amount of useful information contained within the VC. This mechanism suppresses the generation of redundant content that is irrelevant to answering the question.

$$\text{Reward}_{\text{Match}} = \frac{1}{|K|}\sum_{i=1}^{|K|}(K_i \subseteq \text{VC}) \qquad (2)$$

Image Consistency Reward. The image consistency reward aims to reduce hallucination phenomena induced by the VC. Specifically, we employ the FineCLIP (Asokan et al., 2025) model to compute the CLIP-Score for each VC, followed by a normalization process, as formalized in Equation (3). This score quantifies the consistency between the visual clue and the corresponding image content.

$$\text{Reward}_{\text{IC}} = \frac{\sum_{i=1}^{N}\text{CLIP-Score}(\text{VC}_i, \text{Image})}{N} \qquad (3)$$

where $\min$ and $\max$ denote the minimum and maximum CLIP-Scores across the VC, respectively.

Final Reward. The final reward signal produced by the Reward Agent is a composite of the three aforementioned sub-rewards, as follows:

$$\text{Reward} = \text{Reward}_{\text{Acc}} + \text{Reward}_{\text{Match}} + \text{Reward}_{\text{IC}} \qquad (4)$$

3.1.3 Reinforcement Learning

To further enhance caption quality, inspired by Dai et al. (2024), we propose a Reward-Agent framework (RA-GRPO), as shown in Figure 2. RA-GRPO is a reinforcement learning (RL) algorithm that avoids learning a value critic by computing normalized advantages within a group of sampled actions.

Specifically, given a prompt s, we sample G outputs $\{a_1, ..., a_G\} \sim \pi_\theta(\cdot|s)$, evaluate them with a reward function $r(s, a)$ from Reward Agent. We compute the reward as r^i, and repeatedly compute the rewards for all paths from group, i.e., $\{r^1, r^2, ..., r^G\}$.

To estimate the advantage of each trajectory, we normalize its reward relative to the group as follow:

$$\hat{A}^i = \frac{r^i - \text{mean}(\{r^1, r^2, ..., r^G\})}{\text{std}(\{r^1, r^2, ..., r^G\})}, \qquad (5)$$

where the mean group reward serves as the baseline, and $\hat{A}_i$ measures how much better or worse r_i is compared to other trajectories within the group. Following this, we optimize the policy model with

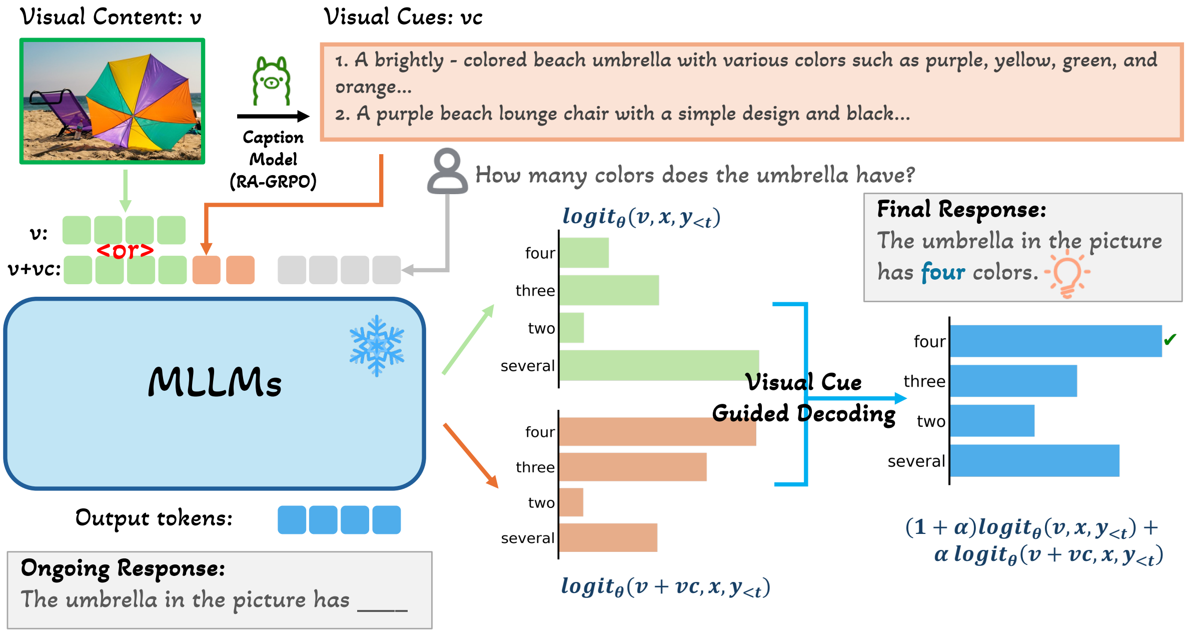

Visual Content: v
Visual Cues: vc
1. A brightly - colored beach umbrella with various colors such as purple, yellow, green, and orange...
2. A purple beach lounge chair with a simple design and black...
Caption Model (RA-GRPO)
How many colors does the umbrella have?
v:
<or>
v+vc:
MLLMs
Output tokens:
Ongoing Response:
The umbrella in the picture has ____
$logit_\theta(v, x, y_{<t})$
four
three
two
several
$logit_\theta(v + vc, x, y_{<t})$
Visual Cue Guided Decoding
Final Response:
The umbrella in the picture has four colors.
$(1+\alpha)logit_\theta(v, x, y_{<t}) + \alpha\, logit_\theta(v + vc, x, y_{<t})$

Figure 3: **The overall decoding procedure of VCGD.** After Caption Model generate Visual Clues for the image, the model recursively outputs logits from each v and $v + vc$. By contrasting between two log-likelihoods, VCGD produces more contextual and correct responses that match the given visual content suppressing inconsistent words (*serveral*→*four*).

the loss defined as:

$$\mathcal{L}_{RA-GRPO} = -\underset{Q\in D_s}{E}[\frac{1}{M}\sum_{i=1}^{M}(\frac{\pi_\theta(\mathbf{c}^i|Q)}{[\pi_\theta(\mathbf{c}^i|Q)]_{\text{no grad}}}\hat{A}^i - \beta D_{KL}(\pi_\theta||\pi_{ref})], \tag{6}$$

where KL divergence is adopted to regularize the policy model, preventing excessive deviation from the reference model. The reference model is typically initialized as the same model as the policy model but remains frozen during RL training. The KL divergence between the policy model and the reference model is estimated as in (Shao et al., 2024):

$$D_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(\mathbf{c}^i|Q)}{\pi_\theta(\mathbf{c}^i|Q)} - \log\frac{\pi_{ref}(\mathbf{c}^i|Q)}{\pi_\theta(\mathbf{c}^i|Q)} - 1. \tag{7}$$

3.2 Visual Clue Guided Decoding

An overview of our proposed VCGD framework is shown in Figure 3.

3.2.1 Problem Setup and Preliminaries

Let M_θ denote a Multimodal Large Language Model (MLLM) parameterized by θ that autoregressively generates responses for the given visual contents v and input textual query x. Then the model maps the logit distribution to the next token prediction output $y_t \in R^{|\mathcal{V}|}$ at time step t in the vocabulary set $\mathcal{V}$ such that $y_t \sim p_\theta(y_t|v, x, y_{<t}) \propto \text{logit}_\theta(y_t|v, x, y_{<t})$, where $y_{<t}$ indicates all previously generated tokens.

3.2.2 Contrastive Decoding with Disturbance

We can obtain a pair of visual content and precise visual clues from the Caption Model (v, vc), where v represents the image information and vc corresponds to $Caption(y|v, x_0)$. By contrasting the logit variations between the set of information during model response generation, we can formulate the next-word prediction using our proposed VCGD method:

$$y_t \sim \text{Softmax}\Big[(1+\alpha_t)\text{logit}_\theta(y_t \mid v, x, y_{<t}) + \alpha_t\text{logit}_\theta(y_t \mid v + vc, x, y_{<t})\Big)\Big]. \tag{8}$$

Inspired by CODE (Kim et al., 2024), we define the dynamic constraint α_t as:

$$1 - \mathcal{D}_{\text{bd}}(P_t^v \mid P_t^{v+vc}), \tag{9}$$

$$\mathcal{D}_{\text{bd}}(P\|Q) = \frac{1}{2}\sum_{i=1}^{n}(p_i + q_i)\log_2(|p_i - q_i|^k + 1), \tag{10}$$

where $\mathcal{D}_{\text{bd}}(P\|Q) \geq 0$ and equals 0, if and only if p=q, and k denotes a smoothing parameter. Here, the upper-bound of the divergence apparently exists, such that $\mathcal{D}_{\text{bd}}(P\|Q) \leq \sum_{i=1}^{n} p_i \log_2 2{=}1$, due to the following condition $|p_i{-}q_i| \leq 1$.

That implements a token-level feedback control mechanism that dynamically adjusts the information weight based on the proximity of two distributions. The primary role of this constraint is to maintain a balanced variation in logits given the predictive discrepancies between the v and $v + vc$ distributions. Specifically, when the two distributions are sufficiently close (*i.e.*, $P_t^v \approx P_t^{v+vc}$), the value of $\mathcal{D}_{\text{bd}}(P_t^v \mid P_t^{v+vc})$ approaches zero, indicating minimal distributional divergence. In this case, α_t approaches 1, thereby allowing greater amplification of the logits' variation when predicting the next token.

3.2.3 Adaptive Information Constraint

The VCGD objective is designed to favor tokens preferred by the MLLMs output while imposing penalties on tokens influenced by instruction disturbances. However, it also might erroneously reward tokens representing implausible concepts. To address this issue, we refine the VCGD objective to incorporate an adaptive plausibility constraint $\mathcal{V}_{\text{head}}$. By comparing prediction distributions between P_t^v and P_t^{v+vc}, we filter out less relevant tokens from the candidate pool as follows:

$$\begin{aligned}\mathcal{V}_{\text{head}}(y_{<t}) = \{y_t \in \mathcal{V} : p_\theta(y_t \mid v, x, y_{<t}) \\ \geq \beta_t \max_{w} p_\theta(w \mid v, x, y_{<t})\},\end{aligned} \tag{11}$$

where β_t dynamically regulate the token candidate pool utilizing the divergence term in Eq. (10), defined as $\beta_t = \mathcal{D}_{\text{bd}}(P_t^v \| P_t^{v+vc})$. This strategy can expand the token searching pool when the next-token prediction, derived from both visual content and comprehensive description, shows a similar distribution yet uncertainty in selecting the candidate token (*i.e.,* false negatives). Finally, we only consider the next-token prediction within $\mathcal{V}_{\text{head}}(y_{<t})$, and for the tokens satisfying $y_t \notin \mathcal{V}_{\text{head}}(y_{<t})$, we set their logits to $-\infty$ to filter out from the candidate pool.

4 Experiment

4.1 Experimental Setup

4.1.1 Caption Model

During the Cold-Start Stage, we sample 160K data points from ShareGPT4V (Chen et al., 2023a) and LLaVA-CoT (Xu et al., 2024) to train Qwen-2.5-VL-7B. In the RL stage, we sample 50K data points from the aforementioned datasets and use Qwen-2.5-7B as Rollout Model, with rollout=8 and K-L=0.

4.1.2 Inference Model

We apply VCGD on three MLLMs in different sizes, LLaVA-1.5-13B (Liu et al., 2024b), LLaVA-NeXT-34B (Liu et al., 2024c) and InternVL-26B (Chen et al., 2023b).

4.1.3 Evaluation Benchmarks

We evaluate the performance of VCGD on four widely used benchmarks, including POPE (Li et al., 2023d), MMVP (Tong et al., 2024), MMHalBench (Sun et al., 2023b), and LLaVA-Bench (In-the-Wild) (Liu et al., 2023d) for MLLMs with a special focus on hallucination. The benchmarks are detailed in Appendix A.

4.1.4 Baselines

We compare our method with six baseline decoding strategies. For conventional decoding strategies, we use greedy decoding, nucleus sampling (Holtzman et al., 2020), and beam search decoding. Additionally, we select the recent state-of-the-art (SOTA) methods, including the OPERA method (Huang et al., 2023), the VCD method (Leng et al., 2023), and the CODE method (Kim et al., 2024) as comparative decoding approaches.

4.2 Main Results

Table 1 presents the primary experimental results. We observe the following points:

Results on MMVP. The MMVP benchmark comprehensively evaluate CLIP blind pairs across 9 different visual modalities. As shown in Table 1, the results indicate a significant improvement in average accuracy after employing VCGD contrastive decoding.

Results on POPE. Our method demonstrates consistent improvements over previous baselines across various settings. The composition of POPE focuses solely on questioning the existence of objects, rather than their absence (*e.g.,* "Is there `{something}` in the image?"). The combinatorial results of a high accuracy and F1 score indicate that our method can boost the existing MLLMs to effectively mitigate hallucination by cautiously confirming *yes* for the existence of objects (*i.e.,* the model does not often *make up* objects).

Model	MMVP										POPE		LLaVA QA90	MMHal Bench	
	🧭	🔍	🔄	⇅	📍	🎨	⚙	**A**	📷	Avg↑	Acc_{adv}	F1_{adv}	Overall	Overall	Hal ↓
LLaVA-1.5-13B															
+ Greedy	30.7	27.2	0.0	12.5	10.0	53.3	16.6	50.0	40.0	30.6	84.0	82.6	82.4	2.39	52.0
+ Beam	19.2	27.2	11.1	25.0	10.0	**60.0**	16.6	70.0	35.0	32.6	84.1	82.7	**83.5**	2.33	53.1
+ Nucleus	26.9	27.2	**22.2**	12.5	20.0	33.3	0.0	60.0	20.0	26.6	80.6	79.4	79.3	2.03	60.4
+ Opera	**42.3**	36.3	11.1	25.0	10.0	56.6	16.6	70.0	35.0	33.3	84.0	82.5	80.7	2.22	55.0
+ VCD	34.6	18.1	**22.2**	37.5	**50.0**	43.3	**33.3**	40.0	35.0	34.0	81.0	80.3	79.3	2.28	54.0
+ CODE	19.2	31.8	11.1	25.0	20.0	53.3	16.6	**80.0**	40.0	34.0	84.2	82.8	**83.5**	2.49	51.0
+ VCGD(ours)	40.2	**38.4**	**22.2**	**37.5**	20.0	**60.0**	16.6	**80.0**	**45.0**	**37.2**	**85.0**	**84.7**	82.2	**2.62**	**49.0**
LLaVA-NeXT-34B															
+ Greedy	38.4	**40.9**	16.6	37.5	30.0	60.0	0.0	**80.0**	35.0	40.6	86.5	87.0	90.7	3.30	34.0
+ Beam	38.4	31.8	22.2	37.5	**50.0**	60.0	0.0	**80.0**	30.0	40.6	84.1	82.7	94.5	3.26	35.4
+ Nucleus	34.6	22.7	27.7	25.0	20.0	43.3	0.0	50.0	**45.0**	33.3	84.9	85.3	90.0	3.08	40.6
+ Opera	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+ VCD	42.3	22.7	22.2	37.0	**50.0**	46.6	**16.6**	**80.0**	40.0	39.3	85.2	85.6	92.1	3.16	39.5
+ CODE	34.6	36.3	**33.3**	25.0	**50.0**	**70.0**	0.0	70.0	30.0	42.6	86.9	87.5	**95.3**	3.43	34.0
+ VCGD(ours)	**50.0**	**40.9**	22.2	**62.5**	**50.0**	60.0	**16.6**	60.0	**45.0**	**47.5**	**87.8**	**89.3**	93.8	**3.88**	**31.3**
InternVL-26B															
+ Greedy	42.3	36.3	27.7	25.0	30.0	80.0	33.3	**80.0**	45.0	48.0	85.8	86.4	86.6	3.15	33.3
+ Beam	38.4	45.4	22.2	37.5	50.0	**83.3**	50.0	70.0	45.0	50.6	86.8	86.6	89.3	3.36	31.2
+ Nucleus	**50.0**	31.8	27.7	12.5	**60.0**	70.0	33.3	60.0	25.0	44.0	81.2	81.7	86.4	3.14	37.5
+ Opera	42.3	27.2	16.6	25.0	30.0	76.6	50.0	70.0	50.0	45.3	86.3	86.6	88.7	3.32	32.2
+ VCD	30.7	36.3	11.1	12.5	50.0	66.6	50.0	50.0	**55.0**	42.0	81.7	82.1	88.3	2.94	42.0
+ CODE	42.3	**50.0**	**44.4**	12.5	30.0	**83.3**	50.0	70.0	40.0	51.3	86.9	87.5	92.2	3.52	**30.2**
+ VCGD(ours)	25.0	**50.0**	**44.4**	**50.0**	**60.0**	**83.3**	**56.6**	**80.0**	**55.0**	**55.9**	**88.2**	**89.8**	**94.8**	**3.48**	31.2

Table 1: Experimental results of various hallucination benchmarks on different decoding strategies. The best result for each metric in **each group** is in bold.

	MMVP	POPE		LLaVA QA90	MMHal Bench	
	Avg	Acc	F1	Oa	Oa	Hal ↓
LLaVA-1.5-13B						
VC As Prefix	35.4	84.8	83.6	**85.5**	**2.69**	49.6
VCGD	**37.2**	**85.0**	**84.7**	82.2	2.61	**49.0**
LLaVA-NeXT-34B						
VC As Prefix	45.5	87.2	88.6	**95.2**	3.68	32.5
VCGD	**47.5**	**87.8**	**89.3**	93.8	**3.88**	**31.3**
InternVL-26B						
VC As Prefix	**56.3**	87.5	88.2	93.6	3.38	33.1
VCGD	55.9	**88.2**	**89.8**	**94.8**	**3.48**	**31.2**

Table 2: Comparison of using VC as prompt prefixes versus incorporating them through the VCGD.

Results on LLaVA-QA90. To explore the broader applicability of our method beyond basic multiple-choice formats, we evaluate sentence-level model outputs on the LLaVA-QA90 (Liu et al., 2023d). As shown in Table 1, VCGD achieves competitive performance compared to other contrastive decoding (CD) methods.

Results on MMHal-Bench. Additionally, we compare our models in MMHal-Bench (Sun et al., 2024) specialized to evaluate hallucination effects sourced from more challenging image-question pairs. As in the result, our method generally not only improves overall average score with consistent results among other baseline MLLMs, but also effectively mitigates the hallucination ratio.

4.3 Ablation Analysis

We conduct analysis on VCGD considering the following questions: (**Q1**) To what extent do the Cold-Start Stage and the RL Stage contribute to the performance improvement of the Caption Model? (**Q2**) Are all three reward functions in the Reward Agent necessary? (**Q3**) What would be the effect on performance if the Visual Cues generated by the Caption Model were used directly as prompt prefixes, instead of being processed through the CD method?

A1: Both the Cold-Start Stage and the RL Stage are necessary. To validate the improvements in Caption Model performance brought about by the Cold-Start and RL Stages, as shown in Table 3, we employed three versions of Caption Model: the untrained version, the Cold-Start version, and the RL version, for generating Visual Cues during

	MMVP	POPE		LLaVA QA90	MMHal Bench	
	Avg	Acc	F1	Oa	Oa	Hal ↓
LLaVA-1.5-13B						
No Train	32.3	83.8	81.9	80.0	2.36	52.4
Cold-Start	35.3	84.5	82.9	81.4	2.38	52.2
RL	**37.2**	**85.0**	**84.7**	**82.2**	**2.62**	**49.0**
LLaVA-NeXT-34B						
No Train	43.7	87.1	87.9	89.8	3.52	33.6
Cold-Start	44.8	87.2	88.2	90.4	3.58	33.2
RL	**47.5**	**87.8**	**89.3**	**93.8**	**3.88**	**31.3**
InternVL-26B						
No Train	50.8	87.5	88.4	92.9	3.31	**31.0**
Cold-Start	51.9	87.3	88.2	93.2	3.35	32.2
RL	**55.9**	**88.2**	**89.8**	**94.8**	**3.48**	31.2

Table 3: Ablation results of Caption Models at different training stages.

R_A	R_M	R_{IC}	MMVP	POPE		LLaVA QA90	MMHal Bench	
			Avg	Acc	F1	Oa	Oa	Hal ↓
LLaVA-1.5-13B								
✓			34.8	84.7	84.0	82.3	2.53	51.3
	✓		32.2	84.3	83.1	80.8	2.29	52.2
		✓	33.3	83.9	82.5	80.3	2.25	52.7
✓		✓	36.6	84.8	84.2	82.6	2.55	50.6
✓	✓		36.8	84.9	84.4	**82.8**	2.58	50.0
	✓	✓	32.6	84.5	82.9	81.8	2.45	50.2
✓	✓	✓	**37.2**	**85.0**	**84.7**	82.2	**2.62**	**49.0**
LLaVA-NeXT-34B								
✓			46.1	87.5	88.9	93.0	3.74	31.8
	✓		45.3	86.9	88.3	92.8	3.68	32.3
		✓	45.0	87.1	88.2	92.7	3.70	31.9
✓		✓	46.8	87.6	88.9	93.2	3.80	31.5
✓	✓		47.0	87.5	89.0	93.3	3.83	**31.0**
	✓	✓	45.7	87.3	88.3	92.9	3.79	31.7
✓	✓	✓	**47.5**	**87.8**	**89.3**	**93.8**	**3.88**	31.3
InternVL-26B								
✓			53.8	87.9	88.9	93.8	3.44	31.2
	✓		52.8	87.8	88.7	93.4	3.38	31.2
		✓	54.1	88.0	89.2	93.2	3.35	32.5
✓		✓	54.3	88.0	89.3	94.1	3.46	31.0
✓	✓		55.6	88.1	88.4	94.3	3.43	**29.0**
	✓	✓	55.2	87.3	88.7	93.6	3.41	**29.0**
✓	✓	✓	**55.9**	**88.2**	**89.8**	**94.8**	**3.48**	31.2

Table 4: Ablation results of different reward in Reward Agent. R_A denote Reward$_{\text{Acc}}$, R_M denote Reward$_{\text{Match}}$, and R_{IC} denote Reward$_{\text{I-C}}$. The results demonstrate that each component Reward$_{\text{Acc}}$, Reward$_{\text{Match}}$, and Reward$_{\text{I-C}}$ plays an indispensable role in the effectiveness of the Reward Agent.

VCGD. The results indicate that both the Cold-Start and RL Stages contribute positively to the final performance.

A2: All three reward functions in the Reward Agent are necessary. To validate the contribution of each reward functions to Reward Agent, we conduct an ablation study in Table 4. The results demonstrate that all three reward functions are essential for achieving the final objectives. They complement each other, and their combined use is crucial for improving overall model performance. The results also show that the $Reward_{Acc}$ is the most effective.

A3: Visual Cues as prompt prefixes are effective, but less so than the VCGD. As shown in Table 2, we conducted the following ablation experiments, where the Visual Cues generated by the Caption Model were directly used as context prefixes for testing. Specifically, we used the input format "{Original question} \n The following are the Visual Cues of the image, please use these visual cues to answer the question: \n {Visual cues}". The experimental results indicate that this method performs well in benchmark, validating the effectiveness of the Caption Model. However, compared to the results obtained using the VCGD, its performance is somewhat inferior.

4.4 Case Study

To provide a more intuitive demonstration of VCGD's performance in mitigating hallucinations, we conducted case studies of the VCGD. For detailed results, please refer to the Appendix B.

5 Conclusion

In this paper, we propose VCGD, a novel strategy that uses high-quality Visual Cues from Caption Model to guide MLLMs during decoding, reducing hallucinations. The Caption Model is further improved via reinforcement learning, where a Reward Agent evaluates the quality of visual clues. Experiments on multiple benchmarks show that VCGD enhances cross-modal consistency, lowers hallucination rates, and integrates seamlessly into existing multimodal systems.

Limitations

In this paper, we focus on addressing the hallucination problem in MLLMs by introducing our novel VCGD approach. We demonstrate the effectiveness of this method through rigorous evaluations on a variety of hallucination discrimination benchmarks. Furthermore, we qualitatively assess its performance on generative benchmarks, which are essential for detecting hallucinated content. While generative benchmarks play a critical role, there remains a notable lack of established metrics capable of thoroughly analyzing hallucinations. This highlights an important direction for future research: the development of robust automatic metrics to enhance the evaluation of open-ended generative performance.

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 arXiv:2303.08774*.

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. *arXiv preprint arXiv:2504.01916*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2024. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Siyuan Dai, Kai Ye, Kun Zhao, Ge Cui, Haoteng Tang, and Liang Zhan. 2024. Constrained multiview representation for self-supervised contrastive learning. *arXiv preprint arXiv:2402.03456*.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024b. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9614–9631, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

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 arXiv:2311.16922*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

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

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. In *International Conference on Learning Representations*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Zhenyi Lu, Jie Tian, Wei Wei, Xiaoye Qu, Yu Cheng, Wenfeng Xie, and Dangyang Chen. 2024. Mitigating boundary ambiguity and inherent bias for text classification in the era of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6008–6022, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015, Torino, Italia. ELRA and ICCL.

OpenAI. 2023a. ChatGPT. `https://openai.com/blog/chatgpt/`.

OpenAI. 2023b. Chatgpt: optimizing language models for dialogue. openai. 2022. *URL https://openai.com/blog/chatgpt*.

OpenAI. 2023c. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023a. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. 2024. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13050–13056, Torino, Italia. ELRA and ICCL.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. *arXiv preprint arXiv:2211.02580*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024a. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024c. Twin-gpt: Digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53553–53570. PMLR.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Bowen Yan, Zhengsong Zhang, Liqiang Jing, Eftekhar Hossain, and Xinya Du. 2024. Fiha: Autonomous hallucination evaluation in vision-language models with davidson scene graphs. *arXiv preprint arXiv:2409.13612*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

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 arXiv:2310.16045*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halleswitch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2023. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

Linhai Zhuo, Yuqian Fu, Jingjing Chen, Yixin Cao, and Yu-Gang Jiang. 2024. Unified view empirical study for large pretrained model on cross-domain few-shot learning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

A Evaluation Benchmarks

We introduce additional details about the benchmarks we used for evaluation. Benchmarking the evaluation of hallucination phenomena in Multimodal Large Language Models (MLLMs) can generally be categorized into discriminative and generative types. Discriminative benchmarks detect hallucinations by assessing the predicted answers within given options (e.g., multiple-choice or true/false questions), whereas generative benchmarks typically employ more advanced language models (e.g., GPT-based evaluations) to score the descriptions generated by the target model. Under this classification, we have carefully selected four benchmarks to test the baseline models.

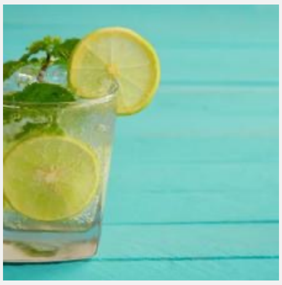
As discriminative benchmarks, we primarily use two datasets for detailed evaluation. Specifically, **POPE** (Li et al., 2023c) is a commonly used benchmark that detects target hallucinations by transforming target label information sourced from the Microsoft COCO dataset (MSCOCO) (Chen et al., 2015). POPE employs binary classification performance on simple true/false questions across three distinct subsets: random, common, and adversarial. **MMVP** (Tong et al., 2024) aims to evaluate the understanding of visual details across nine different visual modes through paired classification accuracy. Given its evaluation design, which involves comparing two similar CLIP-blind image pairs, MMVP requires Multimodal Large Language Models (LMMs) to capture subtle visual differences.

We use two benchmarks as generative benchmarks, extending the evaluation scope to open-ended image description tasks, rather than being limited to evaluating within the context of given answer options. In general, ChatGPT (OpenAI, 2023a) is used to score the quality of sentences generated by the model. **LLaVA-Bench (In-the-Wild)** (Liu et al., 2023d) is a scoring ratio, defined as the sum of the absolute values of model scores divided by the sum of the absolute values of ground truth scores, all of which are evaluated by GPT-4 (OpenAI, 2023c). It includes three types of questions: dialogue, detailed descriptions, and complex reasoning. **MMHal-Bench** (Sun et al., 2024) evaluates the degree of hallucination across eight different question types, including: target attributes, adversarial targets, comparisons, counting, spatial relations, environment, overall descriptions, and others. GPT-4 measures the severity of hallucinations on a scale from 0 to 7, where higher scores indicate fewer hallucinations.

B Case Study

Details of VCGD's performance in mitigating hallucinations are shown in Figure 4 and 5. We present two case studies involving the VCGD: we compared the performance differences of the InternVL model in generative and discriminative tasks under two settings: without and with the VCGD method. Additionally, specific examples of the Caption Model were provided to assist in the analysis.

Image:



Question:

Is there a lemon inside the drink in the cup or are all the lemons outside the drink?
A.There is one inside
B.All are outside

InternVL(Original):

B ✗

Caption Model:

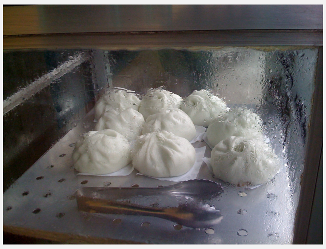
1. One lemon slice is fully submerged inside the drink, clearly visible through the transparent glass.
2. Another lemon slice is placed on the rim of the glass, not inside the liquid.
3. The submerged lemon slice is distorted by the water and appears partially below the liquid level, confirming it is inside the drink.

InternVL(VCGD):

A ✅

Figure 4: **VCGD's performance on discriminative task.** ✓represents a correct answer, while ✗represents an incorrect answer.

Image:



Question:

Describe the image in detail.

InternVL(Original):

The image shows a clear plastic container filled with small white dumplings. The container is placed on a white surface, which could be a table or a counter. There is a fork nearby, possibly used for serving or eating the dumplings. The dumplings are arranged neatly in the container, and the overall presentation appears to be clean and appetizing.

Caption Model:

1. The glass is fogged with condensation, indicating that the food inside is being kept warm in a steamer.
2. Each bun is placed on a small piece of white paper, a common practice for steamed buns to prevent sticking.
3. A metal pair of tongs is placed inside the display, suggesting the buns are ready to be served.

InternVL(VCGD):

This image shows a steamed bun display case, commonly used to keep buns warm and moist. Inside the case, several white steamed buns are placed on small paper liners. The glass walls are fogged up with condensation, indicating a hot and humid interior environment. A pair of metal tongs is placed in the front, used for picking up the buns. The setting appears to be a food stall or a small eatery, possibly serving breakfast or snacks.

Figure 5: **VCGD's performance on generative task.** Text highlighted in **red** denotes hallucinatory descriptions.