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

Guoqing Chen, Fu Zhang*, Bingqian Liu, Chenglong Lu, Jingwei Cheng

School of Computer Science and Engineering, Northeastern University, China
chenguoqing247@gmail.com; {zhangfu, chengjingwei}@neu.edu.cn

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. Most existing research induces hallucinations by manually perturbing visual or instruction inputs, then uses output differences or model-generated descriptions as references to mitigate hallucinations and improve response-visual consistency. However, these methods are constrained by model capabilities and prone to hallucination propagation. We propose **Visual Clue Guided Decoding** (VCGD), a novel decoding strategy that introduces an auxiliary Caption Model to generate precise visual clues during decoding for guiding model generation. It further incorporates image confidence constraints to critically suppress hallucination propagation during generation, thereby significantly improving content reliability and visual consistency. Specifically, VCGD leverages high-quality visual descriptions to guide MLLMs in correcting perceptual biases while generating answers. Furthermore, we introduce a Reinforcement Learning-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 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.

Introduction

Large Language Models (LLMs) (OpenAI 2023; Touvron 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) (Liu et al. 2024b; Team et al. 2023; Bai et al. 2023b). While MLLMs demonstrate remarkable proficiency in tackling a wide array of visual tasks (Li et al. 2024; Black et al. 2024), as well as in understanding (Lai et al. 2024) and generating complex content (Brooks, Holynski, and Efros 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 applications.

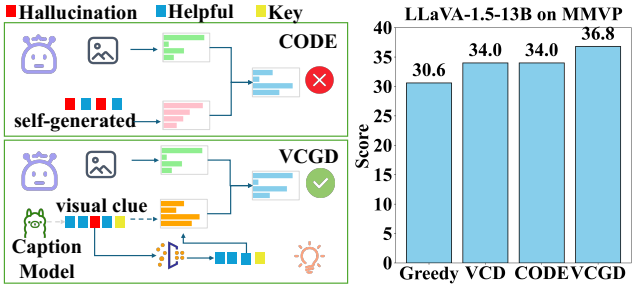

Figure 1: The left figure shows the difference between our VCGD and a state-of-the-art method CODE (Kim et al. 2024). The right shows the performance of LLaVA-1.5-13B on the MMVP benchmark (Tong et al. 2024).

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. 2023). A distinct line of work involves reactive techniques (Huang et al. 2024a; Deng, Chen, and Hooi 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

*Corresponding author.

generations to suppress hallucinations, taking into account factors such as visual noise (Leng et al. 2023), image-induced bias (Zhu et al. 2024), detailed visual grounding (Chen et al. 2024), and self-generated description (Kim et al. 2024).

We investigate the following research question: If the model's own perceptual capabilities are very limited, then no matter what contrastive decoding method is used, it will be unable to convey key effective information to the model. Therefore, a key question arises: *can auxiliary models be incorporated to provide precise visual clues during decoding, thereby reducing the likelihood of hallucination*? To address this issue, we propose a novel Visual Clue Guided Decoding strategy. Specifically, we leverage a caption model to provide precise visual clues 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 four dimensions: format, accuracy, informativeness, and redundancy control. As a result, it facilitates the generation of more precise and pragmatically useful image descriptions. To avoid the propagation of hallucinations, we also introduce image confidence constraints during the decoding process, which prevent the propagation of hallucinations. As show in **Figure 1**, the Caption Model can generate more and more critical visual clues, and reduces hallucination propagation through image confidence constraints. Moreover, it outperforms other contrastive decoding methods on MMVP (Tong et al. 2024).

By conducting extensive experiments and analyses on prevailing cutting-edge MLLMs (Liu et al. 2023b; Chen et al. 2023b; Liu et al. 2024a), we demonstrate the effectiveness of our method in reducing hallucination in various benchmarks (Li et al. 2023c; Liu et al. 2024b; Tong et al. 2024). Our contribution can be summarized 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 four dimensions: format, accuracy, informativeness, and redundancy control.
- We introduced image confidence constraints during the decoding process to avoid the propagation of hallucinations.
- We evaluate the proposed decoding approach on various benchmarks using state-of-the-art MLLMs. Experimental results demonstrate that VCGD significantly reduces hallucinations.

Related Work

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). 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. 2024b), Qwen-VL (Bai et al. 2023b), LLaVA-NEXT (Liu et al. 2024a) and InternVL (Chen et al. 2023b) 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.

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. 2023c; Wang et al. 2023; Sun et al. 2023; Zhong et al. 2024; Tong et al. 2024; Cao et al. 2024; Huang et al. 2024b; Wan and Bansal 2022; Zhang, Zuo, and Jing 2024; Min et al. 2023; Yan et al. 2024). Concurrently, significant attention is directed towards ***analyzing*** the underlying causes of hallucinations (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. 2023a) or preference learning (Sun et al. 2023; Yu et al. 2023; Li et al. 2023a; Zhao et al. 2023; Liu et al. 2023a; Yu et al. 2024; Zhou et al. 2024; Jiang et al. 2024; Chen et al. 2025). *Training-free* approaches address potential hallucinations by post-processing the outputs of MLLMs (Leng et al. 2023; Huang et al. 2024a; Manevich and Tsarfaty 2024; Wang et al. 2024b; Kim et al. 2024). For example, VCD (Leng et al. 2023) 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. ICD (Wang et al. 2024b) suppresses hallucinations through disturbance instructions affecting multimodal alignment. Similarly, LCD (Manevich and Tsarfaty 2024) 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 propose to utilize visual clues from Caption Mod-

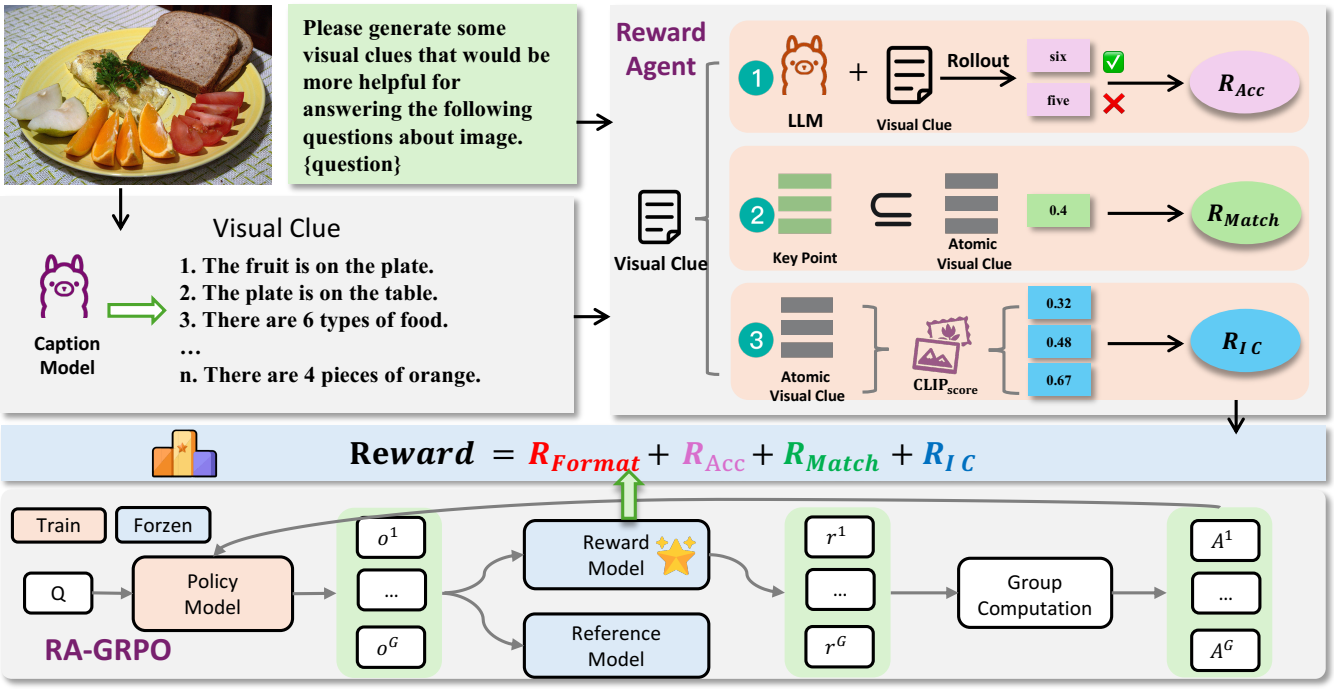

Figure 2: **Caption Model Training framework:** (1) Our designed Reward Agent contains Format Reward, Accuracy Reward, Matching Reward, and Image-Consistency Reward. "Atomic Visual Clue" denotes splitting the Visual Clue by index. (2) Training: The Reward Agent evaluates the corresponding answer candidate and assigns a reward score. The parameters are then optimized using our proposed RA-GRPO framework.

els as a contrasting visual counterpart to correct hallucinatory responses.

Methodology

Our VCGD framework consists of two main components: *Caption Model Training* and *Visual Clue Guided Decoding*.

Caption Model

Overview To provide more precise visual clues (*vc*) during the model's decoding process, we propose a novel Caption Model training method to generate accurate and informative visual clues based on our designed Reward-Agent framework (RA-GRPO). We sample data from ShareGPT4V (Chen et al. 2023a) and LLaVA-CoT (Xu et al. 2024), and utilize GPT-4 to filter and augment some inappropriate original data to meet our training task requirements. We also use GPT-4 to inductively summarize the original questions (*Q*) and answers (*A*), extracting key points (*Key*) that are crucial for problem-solving, which are used for subsequent Matching Reward calculation in the Reward Agent. The final curated and processed data constitute ourtraining dataset D_s.

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

Format Reward. Format reward is primarily used to constrain the model to output visual clues in a fixed format, thereby affecting readability and the validation of other rewards.

Accuracy Reward. In order that the *vc* can effectively capture visual information and assist the model in understanding image content, we use GPT-4 (No visual input) as 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:

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

Matching Reward. This reward signal is designed to encourage the Caption Model to generate the *vc* that minimize the inclusion of redundant information related to the image content but irrelevant to the given question. To achieve this, we still use GPT-4 to calculate the matching degree between the *vc* and key points (*Key*). This mechanism suppresses the generation of redundant content irrelevant to answering the question and encourages the generation of key content.

$$R_{\text{Match}} = \frac{1}{|K|} \sum_{i=1}^{|K|} (K_i \subseteq vc) \quad (2)$$

Image Consistency Reward. The image consistency reward aims to reduce hallucination phenomena induced by

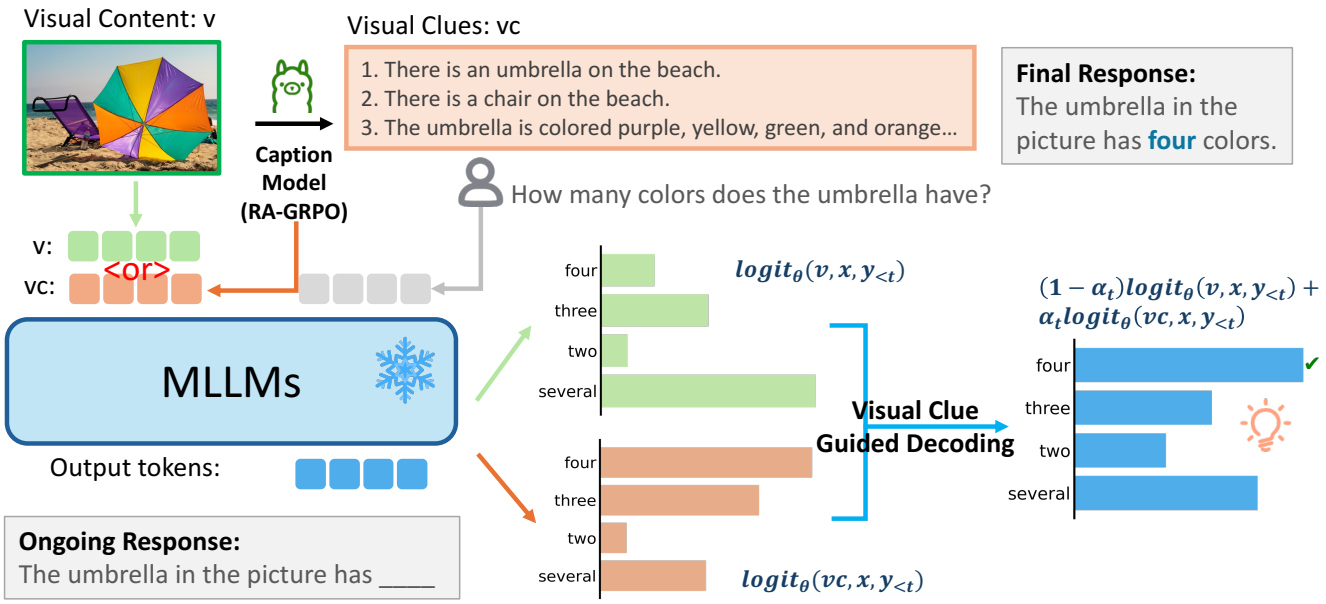

Figure 3: **The overall decoding procedure of VCGD.** After the Caption Model generates Visual Clues for the image, the model recursively outputs logits from each v and vc. By contrasting between two log-likelihoods, VCGD produces more contextual and correct responses that match the given visual content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

the vc. Specifically, we employ the FG-CLIP (Xie 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.

$$R_{\text{IC}} = \frac{\sum_{i=1}^{N} \text{CLIP-Score}(vc_i, \text{Image})}{N} \tag{3}$$

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

$$\text{Reward} = R_{\text{Format}} + R_{\text{Acc}} + R_{\text{Match}} + R_{\text{IC}} \tag{4}$$

Reward-Agent framework (RA-GRPO) To further enhance caption quality, inspired by Dai et al. (2024) and Guo et al. (2025), 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 Q, we sample G outputs $\{o^1, ..., o^G\} \sim \pi_\theta(\cdot|Q)$, evaluate them with a reward function $r(Q, o)$ 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:

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

where the mean group reward serves as the baseline, and 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 the loss defined as:

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

where D_s denotes the training dataset, and $\mathbf{o}^i$ denotes a group of outputs sampled for a given query Q ($Q \in D_s$) and β is a hyperparameter in $[0, 1]$. 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):

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

Visual Clue Guided Decoding

An overview of our proposed Visual Clue Guided Decoding (VCGD) framework is shown in **Figure 3**.

Problem Setup and Preliminaries Let M_θ denote a MLLM parameterized by θ that auto-regressively generates responses for the given *visual content* 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.

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:

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

$$\alpha_t = \frac{\sum_{i=1}^{N}(\text{CLIP-Score}(vc_i, \text{Image}))^{\frac{1}{\gamma}}}{N}, \tag{9}$$

where α_t denotes the dynamic constraint, and γ is an adjustable parameter. To address the hallucination propagation from *vc* in the Caption Model, we impose an image confidence constraint on α_t, which utilizes CLIP-score of FG-CLIP(Xie et al. 2025) to calculate the confidence of *vc* with respect to the image. When confidence becomes low, we consider that severe hallucinations have occurred in *vc*. To prevent hallucination propagation, α_t approaches 0. This reduces the dependency on *vc*.

Adaptive Plausibility Constraint In the VCGD method described in Formula (8), there is a possibility of wrongly penalizing reasonable tokens or wrongly rewarding unreasonable ones. To address this issue, we draw inspiration from the adaptive plausibility constraints method used in open-ended text generation (Li et al. 2023b) and add adaptive plausibility constraints to the VCGD method:

$$\begin{aligned} &\mathcal{V}_{\text{head}}\left(y_{<t}\right) = \{y_t \in \mathcal{V}: \\ &p_\theta\left(y_t \mid v, x, y_{<t}\right) \geq \lambda \max_w p_\theta\left(w \mid v, x, y_{<t}\right)\}, \\ &p_{vcgd}\left(y_t \mid v, vc, x\right) = 0, \text{ if } y_t \notin \mathcal{V}_{\text{head}}\left(y_{<t}\right), \end{aligned} \tag{10}$$

where $\mathcal{V}$ is the output vocabulary of MLLMs and λ is a hyperparameter in $[0, 1]$ for controlling the truncation of the next token distribution. Larger λ indicates more aggressive truncation, keeping only high-probability tokens.

Combining the visual contrastive decoding and the adaptive plausibility constraint, we obtain the full formulation:

$$\begin{aligned} y_t \sim \text{Softmax}\left[(1-\alpha_t)\,\text{logit}_\theta\left(y_t \mid v, x, y_{<t}\right)\right. \\ \left.+\alpha_t\,\text{logit}_\theta\left(y_t \mid vc, x, y_{<t}\right)\right], \\ subject\ to\ y_t \in \mathcal{V}_{\text{head}}\left(y_{<t}\right) \end{aligned} \tag{11}$$

Experiment

In this section, we evaluate the efficacy of our method for mitigating hallucinations in MLLMs.

Experimental Setup

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

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. 2024a), the VCD method (Leng et al. 2023), and the CODE method (Kim et al. 2024) as comparative decoding approaches.

Implementation Details We train Qwen2.5-VL-7B-Instruct as Caption Model and apply VCGD on three MLLMs in different sizes, LLaVA-1.5-13B (Liu et al. 2023b), LLaVA-NeXT-34B (Liu et al. 2024a), and InternVL-26B (Chen et al. 2023b).

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).

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. 2024b). As shown in Table 1, VCGD achieves competitive performance compared to other contrastive decoding methods.

Results on MMHal-Bench. Additionally, we compare our models in MMHal-Bench (Sun et al. 2023) 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.

Ablation Analysis

We conduct analysis on VCGD considering the following questions: (**Q1**) Are all four reward functions in the Reward Agent necessary? (**Q2**) Is it truly necessary to train the caption model using reinforcement learning? (**Q3**) Does the size of the caption model have an impact? What would happen if a stronger proprietary caption model is used? (**Q4**) Does the caption model contribute to hallucination propagation? (**Q5**) Does this decoding strategy introduce inference latency?

A1: All four 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

Model	MMVP										POPE		LLaVA QA90	MMHal Bench	
										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)	42.3	**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.

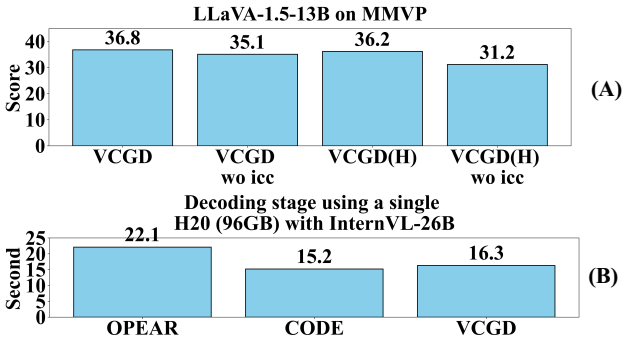

Figure 4: In Figure (A), "VCGD(H)" indicates that we manually injected hallucination information into visual clues and "*wo icc*" indicates that we do not use image confidence constraints. Figure (B) shows the inference time (seconds) for each method.

Table 2. The results demonstrate that all reward functions are essential for achieving the final objectives. To ensure the verifiability of other reward terms, we kept the format reward term unchanged during ablation experiments.

A2: Our RA-GRPO training method demonstrates significant effectiveness in Caption Model. As shown in **Table 3**, we present the performance of InternVL-26B on four benchmark. Comparing the results between the "No Train" versions and the "Train with RA-GRPO" versions using Qwen2.5-VL-3B and Qwen2.5-VL-7B, we observe improvements across all evaluation metrics to varying degrees, indicating that our RA-GRPO training method significantly enhances the capabilities of the Caption Model, thereby validating the effectiveness of this approach.

A3: Model scale does indeed impact performance, but the gains diminish gradually as the scale increases. As shown in **Table 3**, we observe significant performance improvements when scaling from 3B to 7B models, but the performance gains tend to saturate with further increases in model scale. Additionally, the 7B version trained with RA-GRPO surpasses current open-source and closed-source SOTA models on multiple datasets, further validating the effectiveness and scalability of the proposed training method.

A4: Our method demonstrates good robustness in suppressing hallucination propagation. Under VCGD with injected hallucinations manually (**Figure 4**(A)), using image confidence constraints maintains stable performance, while their removal causes significant degradation. The results highlight that our confidence-aware approach provides reliable protection against hallucination propagation while sustaining high overall performance.

A5: Although our approach shows some decline in inference speed compared to the original method, it remains within an acceptable range and outperforms some comparative methods. As shown in **Figure 4**(B), among

R_F	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 2: Ablation results of different reward in Reward Agent. R_F denote R_{Format}, which is always enabled, R_A denote R_{Acc}, R_M denote R_{Match}. The results demonstrate that each component plays an indispensable role in the effectiveness of the Reward Agent.

various decoding strategies, our method achieves the best overall performance, while maintaining competitive inference efficiency, demonstrating a favorable trade-off between effectiveness and computational cost.

Case Study

To provide a more intuitive demonstration of VCGD's performance in mitigating hallucinations, we conduct case studies of the VCGD. As shown in **Figure 5**, we compare the performance differences of the InternVL-26B model in generative tasks under two settings: without and with the VCGD method. From the perspective of Discriminative Task, the Caption Model can provide auxiliary visual clues for question answering, thus enabling the VCGD method to function effectively.

Caption Model	MMVP	POPE		LLaVA QA90	MMHal Bench	
	Avg	Acc	F1	Oa	Oa	Hal ↓
No Train						
Qwen2.5-VL-3B	44.8	81.7	84.0	82.3	2.53	35.3
Qwen2.5-VL-7B	48.0	85.8	86.4	86.6	3.15	33.3
Qwen2.5-VL-32B	48.2	85.8	86.7	86.6	3.21	32.8
Qwen2.5-VL-72B	52.6	87.9	88.3	91.6	**3.55**	32.6
OpenAI o3	52.9	87.8	88.2	91.6	3.50	31.6
Train with RA-GRPO						
Qwen2.5-VL-3B	51.4	85.6	86.4	87.8	2.82	32.9
Qwen2.5-VL-7B	**55.9**	**88.2**	**89.8**	**94.8**	3.48	**31.2**

Table 3: Ablation results with the performance of InternVL-26B on four benchmarks, comparing the effectiveness of different Caption Model, including "No Train" and a version trained using our designed RA-GRPO.

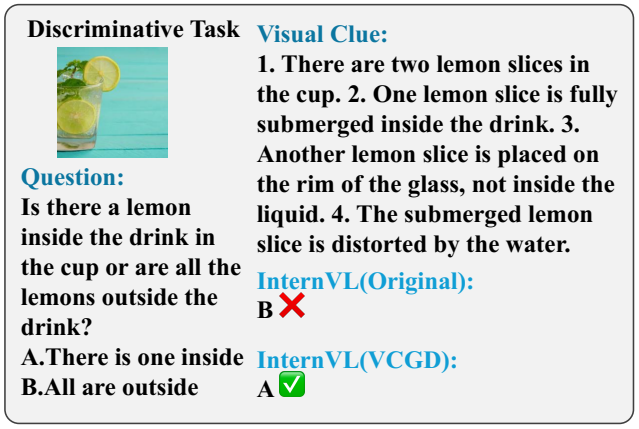

Figure 5: Case study on InternVL-26B with VCGD.

Conclusion

In this paper, we propose VCGD, a novel strategy that uses high-quality visual clues 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. And we introduced image confidence constraints during the decoding process to avoid the propagation of hallucinations. Experiments on multiple benchmarks show that VCGD enhances cross-modal consistency, lowers hallucination rates, and integrates seamlessly into existing multimodal systems.

Acknowledgments

The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057).

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning. In *ICLR*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Cao, Q.; Cheng, J.; Liang, X.; and Lin, L. 2024. VisDiaHalBench: A Visual Dialogue Benchmark For Diagnosing Hallucination in Large Vision-Language Models. In *ACL*.

Chen, G.; Zhang, F.; Lin, J.; Lu, C.; and Cheng, J. 2025. RRHF-V: Ranking Responses to Mitigate Hallucinations in Multimodal Large Language Models with Human Feedback. In *COLING*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Muyan, Z.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. *ICML*.

Dai, S.; Ye, K.; Zhao, K.; Cui, G.; Tang, H.; and Zhan, L. 2024. Constrained Multiview Representation for Self-supervised Contrastive Learning. *arXiv preprint arXiv:2402.03456*.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding. *arXiv preprint arXiv:2402.15300*.

Fadeeva, E.; Rubashevskii, A.; Shelmanov, A.; Petrakov, S.; Li, H.; Mubarak, H.; Tsymbalov, E.; Kuzmin, G.; Panchenko, A.; Baldwin, T.; Nakov, P.; and Panov, M. 2024. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *ACL*.

Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Li, H.; Hu, H.; et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *CVPR*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*.

Huang, W.; Liu, H.; Guo, M.; and Gong, N. 2024b. Visual Hallucinations of Multi-modal Large Language Models. In *ACL*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; and Zhang, S. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*.

Kim, J.; Kim, H.; Yeonju, K.; and Ro, Y. M. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In *NeurIPS*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023b. Contrastive Decoding: Open-ended Text Generation as Optimization. In *ACL*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; and Wen, J.-R. 2023c. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. In *NeurIPS*.

Lu, Z.; Tian, J.; Wei, W.; Qu, X.; Cheng, Y.; Xie, W.; and Chen, D. 2024. Mitigating Boundary Ambiguity and Inherent Bias for Text Classification in the Era of Large Language Models. In *ACL*.

Manevich, A.; and Tsarfaty, R. 2024. Mitigating Hallucinations in Large Vision-Language Models (LVLMs) via Language-Contrastive Decoding (LCD). In *ACL*.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

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

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *EMNLP*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sui, P.; Duede, E.; Wu, S.; and So, R. 2024. Confabulation: The Surprising Value of Large Language Model Hallucinations. In *ACL*.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, D.; and Bansal, M. 2022. Evaluating and improving factuality in multimodal abstractive summarization. *arXiv preprint arXiv:2211.02580*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Wang, L.; He, J.; Li, S.; Liu, N.; and Lim, E.-P. 2024a. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, 32–45.

Wang, X.; Pan, J.; Ding, L.; and Biemann, C. 2024b. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In *ACL*.

Wang, Y.; Fu, T.; Xu, Y.; Ma, Z.; Xu, H.; Du, B.; Lu, Y.; Gao, H.; Wu, J.; and Chen, J. 2024c. Twin-gpt: Digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Xie, C.; Wang, B.; Kong, F.; Li, J.; Liang, D.; Zhang, G.; Leng, D.; and Yin, Y. 2025. FG-CLIP: Fine-Grained Visual and Textual Alignment. *arXiv preprint arXiv:2505.05071*.

Xu, G.; Jin, P.; Hao, L.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yan, B.; Zhang, Z.; Jing, L.; Hossain, E.; and Du, X. 2024. FIHA: Autonomous Hallucination Evaluation in Vision-Language Models with Davidson Scene Graphs. *arXiv preprint arXiv:2409.13612*.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Zhai, B.; Yang, S.; Xu, C.; Shen, S.; Keutzer, K.; and Li, M. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, arXiv–2310.

Zhang, Y.; Zuo, J.; and Jing, L. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Zhong, W.; Feng, X.; Zhao, L.; Li, Q.; Huang, L.; Gu, Y.; Ma, W.; Xu, Y.; and Qin, B. 2024. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. In *ACL*.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Zhu, L.; Ji, D.; Chen, T.; Xu, P.; Ye, J.; and Liu, J. 2024. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding. In *CVPR*.

Zhuo, L.; Fu, Y.; Chen, J.; Cao, Y.; and Jiang, Y.-G. 2024. Unified View Empirical Study for Large Pretrained Model on Cross-Domain Few-Shot Learning. *ACM Transactions on Multimedia Computing, Communications and Applications*.