Incorporating Generative Feedback for Mitigating Hallucinations in Large Vision-Language Models

Ce Zhang[*1] **Zifu Wan**[*1] **Zhehan Kan**[2] **Martin Q. Ma**[1] **Simon Stepputtis**[1]
Deva Ramanan[1] **Russ Salakhutdinov**[1] **Louis-Philippe Morency**[1] **Katia Sycara**[1] **Yaqi Xie**[1]
[1]School of Computer Science, Carnegie Mellon University
[2]Shenzhen International Graduate School, Tsinghua University

Abstract

While recent Large Vision-Language Models (LVLMs) have shown remarkable performance in multi-modal tasks, they are prone to generating hallucinatory text responses that do not align with the given visual input, which restricts their practical applicability in real-world scenarios. In this work, inspired by the observation that the text-to-image generation process is the inverse of image-conditioned response generation in LVLMs, we explore the potential of leveraging text-to-image generative models to assist in mitigating hallucinations in LVLMs. We discover that generative models can offer valuable self-feedback for mitigating hallucinations at both the response and token levels. Building on this insight, we introduce self-correcting Decoding with Generative Feedback (DeGF), a novel training-free algorithm that incorporates generative feedback into the decoding process to effectively mitigate hallucinations. Specifically, DeGF generates an image from the initial response produced by LVLMs, which acts as an auxiliary visual reference and provides self-feedback to verify and, if necessary, correct the initial response. Extensive experimental results validate the effectiveness of our approach in mitigating diverse types of hallucinations, consistently surpassing state-of-the-art methods across two evaluated LVLMs and five benchmarks.

1 Introduction

Recently, Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across various multi-modal tasks, such as image captioning and visual question answering, by extending the capabilities of powerful large language models (LLMs) to incorporate visual inputs [34, 28, 13, 2, 51]. Despite their proficiency in interpreting both visual and textual modalities, these models often suffer from *hallucinations*, where LVLMs erroneously produce responses that are inconsistent with the visual input [30, 19, 52]. This potential for misinformation raises significant concerns, limiting the models' reliability and restricting their broader deployment in real-world scenarios [33, 3, 6].

Recent research has revealed that a major cause of hallucinations in LVLMs is the over-reliance on language priors due to biased training sets, which can override the visual content in response generation [3, 33]. In response, various strategies have been developed to detect and mitigate these hallucinations by directly introducing additional training [19, 44, 23, 8], demonstrating promising results in reducing over-reliance. However, the need for additional data and costly training processes hinders their deployment in downstream tasks. Recently, a new paradigm of methods has emerged to tackle the hallucination problem in LVLMs by intervening in the decoding process [21, 14]. Among these, recent training-free contrastive decoding-based methods [29] have proven effective in mitigating undesired hallucinations by contrasting token predictions derived from original visual input with bias-inducing counterparts, such as no/distorted visual input [15, 26], disturbed instructions [48], or premature layers [12]. While they address hallucinations arising from language priors, hallucinations

[*]Equal contribution.

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

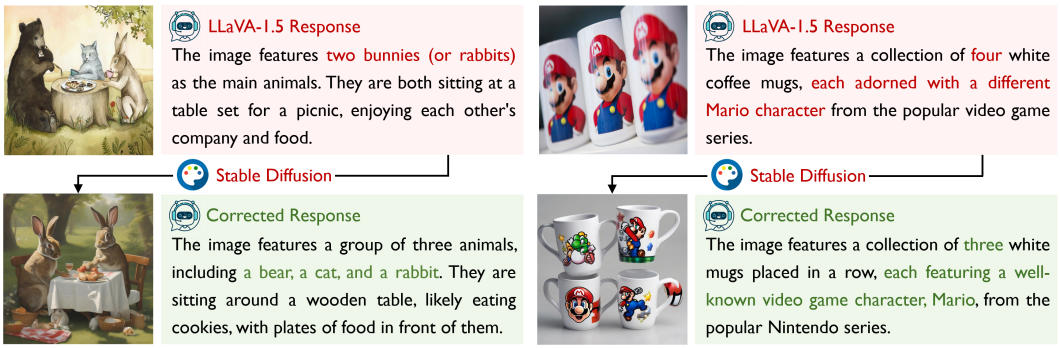

Figure 1: **Generative models can visualize and help correct various types of hallucinations in the initial response**. We query LLaVA-1.5 [34] with the prompt "*Describe this image in detail.*" for two examples from LLaVA-Bench. Based on the initial response, we utilize Stable Diffusion XL [36] to generate a new image, which effectively highlights hallucinations and provides valuable self-feedback. Our approach then incorporates this feedback in the decoding process to successfully correct various types of hallucinations in the original response (as highlighted in red and green).

can also originate beyond language bias, stemming from visual deficiencies in LVLMs [45]. For instance, in counting hallucinations, language does not imply any count information; instead, miscounts largely arise from visual recognition errors of LVLMs, as complex scenes include numerous, similar objects at ambiguous positions which may confuse the LVLMs, leading to incorrect visual understanding and, consequently, hallucinated answers. Therefore, we argue that current contrastive decoding-based methods may struggle to generalize effectively across all types of hallucinations.

In this work, we explore the potential of leveraging powerful text-to-image generative models (*e.g.*, Stable Diffusion [39, 36]) to mitigate hallucinations in LVLMs. Our work is based on a simple yet intuitive hypothesis: given a visual input and a textual prompt to an LVLM, if the generated image-conditioned response is accurate and non-hallucinatory, then a text-to-image generative model should be capable of reversing this process to produce a similar image from that response. Therefore, the discrepancy between the original image and the generated image can serve as valuable self-feedback, guiding the decoding process to correct potential hallucinations in the initial response. To verify this hypothesis, we conduct an empirical study (in Section 3.2), demonstrating that generative models can provide valuable self-feedback for mitigating hallucinations at both the response and token levels.

Building on this insight, we introduce the self-correcting Decoding with Generative Feedback (DeGF), a corresponding training-free decoding algorithm, which effectively incorporates feedback from generative models to recursively enhance the accuracy of generated responses. Specifically, we generate a new image based on the initial response for each instance, which acts as an *auxiliary visual reference* to verify the correctness of the initial response. We propose self-correcting decoding that selectively enhances or contrasts predictions from the original and this reference, *confirming* or *revising* the initial response from the LVLMs based on the measured divergence between the two predictions. By integrating this additional visual reference and generative feedback, LVLMs can gain enhanced visual insights and verify the initial response to ensure accurate visual details in the text outputs. In Figure 1, we demonstrate that incorporating generative feedback in our approach can reduce various types of hallucinations, including object existence, visual appearance, counting, *etc*. To the best of our knowledge, we are the first work to explore the use of generative models for mitigating hallucinations in LVLMs.

The effectiveness of DeGF is evaluated on LLaVA-1.5 [34] and InstructBLIP [13] across five benchmarks: POPE [30], CHAIR [38], MME-Hallucination [16], MMVP [45], and LLaVA-Bench. The experimental results validate the the effectiveness of our DeGF in reducing hallucinations in LVLMs, with performance improvements of up to 5.24% on POPE, 3.0% on CHAIR, and 21.11 points on MME-Hallucination compared to existing state-of-the-arts. A qualitative case study further demonstrates that our approach enhances both the accuracy and detailedness of the generated responses.

The contributions of this paper are summarized as follows:

- We discover the potential of generative models in mitigating hallucinations in LVLMs and demonstrate that generative models can provide valuable self-feedback for mitigating hallucinations at both the response and token levels.

- We propose DeGF, a novel training-free decoding algorithm for LVLMs that recursively enhances the accuracy of responses by integrating generative feedback to confirm or revise the initial output.
- Extensive experimental evaluations across five benchmarks demonstrate that our DeGF consistently outperforms state-of-the-art approaches in effectively mitigating hallucinations in LVLMs.

2 Related Work

Hallucination in LVLMs. With advances of autoregressive LLMs [46, 11, 4, 10], researchers have extended these powerful models to process visual inputs [34, 13, 2, 51]. These models typically train a modality alignment module to project visual tokens into the textual embedding space of the LLM, demonstrating impressive performance in various multi-modal tasks such as visual question answering and image captioning [33, 3]. However, LVLMs are prone to hallucinations, where contradictions arise between the visual content and the generated textual response [30, 33, 3]. To mitigate hallucinations in LVLMs, early works have introduced various approaches, including employing reinforcement learning from human feedback (RLHF) [19, 44], applying auxiliary supervision [23, 8], incorporating negative [32] or noisy data [54, 47], and training post-hoc revisors for correction [57, 52]. Despite promising results, these methods often lack practicality due to their reliance on additional data and costly training processes. To address this, another line of work focuses on training-free methods that can be seamlessly integrated into existing LVLMs. Such methods encompass contrastive decoding [26, 15, 58] and guided decoding with auxiliary information [7, 56, 14, 50]. In this work, we present a novel training-free decoding method that recursively enhance the accuracy of the generated response by incorporating generative feedback. To our best knowledge, we are the first work to effectively utilize generative models for mitigating hallucinations in LVLMs.

Text-to-Image Synthesis. Text-to-image synthesis aims to create realistic images from textual descriptions [59, 17]. In recent years, significant progress has been achieved in this area, largely due to the advent of deep generative models [55, 18]. These advances include Generative Adversarial Networks (GAN) [41, 24], autoregressive models [5, 53], and diffusion models [20, 25, 35, 40, 39]. Pre-trained on large-scale text-image datasets [42], diffusion-based methods have shown strong vision-language alignment, making them valuable for downstream tasks such as classification [27] and semantic segmentation [1, 49]. In this work, we leverage a pre-trained diffusion model to provide useful feedback for refining the generated response of LVLMs.

3 Method

In this work, we present DeGF, a novel training-free decoding algorithm for LVLMs that recursively improves the accuracy of generated responses using generative feedback, as illustrated in Figure 2.

3.1 Preliminary: Decoding of LVLMs

We consider an LVLM parameterized by θ, which processes an input image v and a textual query $\mathbf{x}$, aiming to autoregressively generate a fluent sequence of textual responses $\mathbf{y}$. The visual input v is first processed by a vision encoder and then projected into visual tokens within the textual input space using a vision-language alignment module (e.g., Q-Former [28] or linear projection [34]). These visual tokens, along with the textual query tokens, are then fed into the language encoder for conditioned autoregressive generation. Formally, we denote the autoregressive generation process as

$$y_t \sim p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) \propto \exp f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}), \tag{1}$$

where y_t represents the token at time step t, $\mathbf{y}_{<t} \triangleq [y_0, \ldots, y_{t-1}]$ denotes the sequence of tokens generated before time step t, and f_θ is the logit distribution (unnormalized log-probabilities) produced by the LVLM over a vocabulary of textual tokens $\mathcal{V}$. At each step $t \in [0, \ldots, T]$, the response token y_t is sampled from the probability distribution $p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})$, and this generative process continues iteratively until the response sequence $\mathbf{y} \triangleq [y_0, \ldots, y_T]$ is complete.

3.2 Visual Reference Generation

In our method, we incorporate generative feedback from diffusion models to guide the decoding process. Specifically, given a visual input v and a textual query $\mathbf{x}$, we first prompt the LVLMs to

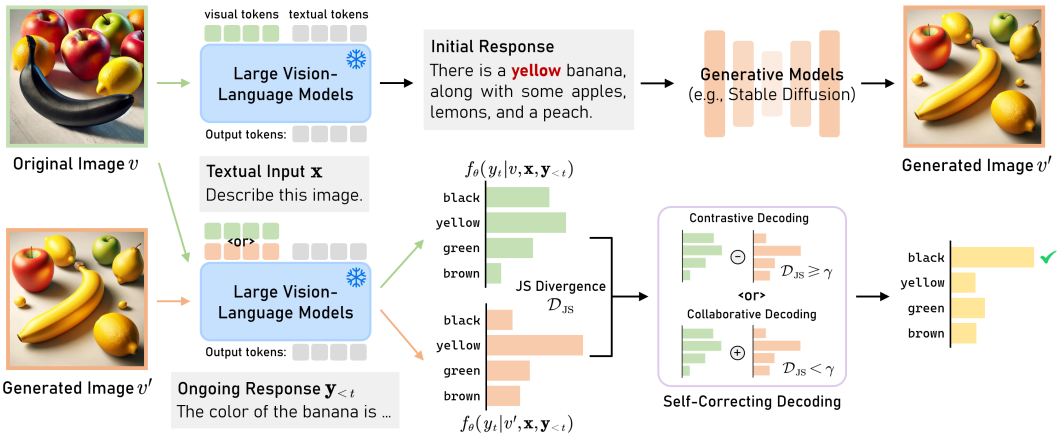

Figure 2: **Overview of our proposed DeGF**. Our method follows a two-step process: first, a generative model produces a high-quality image based on the initial response; second, this image acts as an auxiliary visual reference, providing feedback to refine the next-token predictions. Additionally, we introduce self-correcting decoding, which selectively enhances or contrasts the next-token predictions conditioned on the original and generated images to mitigate hallucinations in the generated response.

generate an initial response $\boldsymbol{\tau}$, which includes relevant descriptions of the visual input with potential hallucinations. Subsequently, we leverage a pre-trained diffusion model $\mathcal{G}$ to generate a new image v' based on the initial response:

$$v' = \mathcal{G}(\boldsymbol{\tau}, x_T), \quad \text{where } x_T \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

Here, x_T denotes a sample from the standard Gaussian distribution, which serves as the initial noisy input to the diffusion model. Starting from this pure noise image x_T, the diffusion model $\mathcal{G}$ iteratively applies T steps of the denoising process to obtain $x_T, x_{T-1}, \ldots, x_0$, where the final output x_0 corresponds to the final generated image v'. Through this diffusion process, the generative model visualizes the initial response, providing a visual reference that helps mitigate potential hallucinations and produce a more accurate and consistent output.

Effectiveness of Generative Models in Reflecting Hallucinations. We validate the effectiveness of generative models in reflecting hallucinations through an empirical study, as shown in Figure 3. The experimental results verify that *generative models can provide valuable self-feedback for mitigating hallucinations* at both the response and token levels.

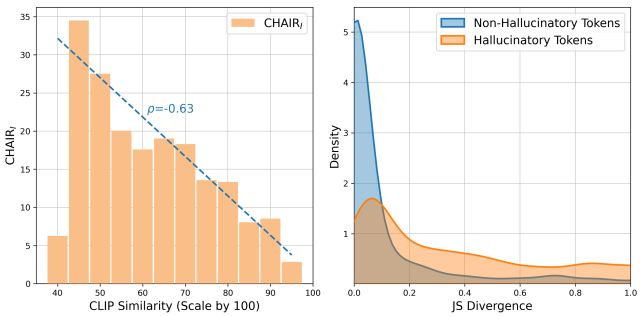

Figure 3: **Generative models can provide feedback for reflecting hallucinations**. (*Left*) Density plot of CLIP similarities and bar plot of average CHAIR$_I$ in each bin on the CHAIR [38] benchmark; (*Right*) Density plots of token-level JS divergence for both hallucinatory and non-hallucinatory tokens on the POPE [30] benchmark.

We conduct the following two experiments: (1) We generate an image v' using diffusion model based on the initial caption provided by LLaVA-1.5 and compute the CLIP image similarities between the original image v and the generated image v' using OpenCLIP [9] ViT-H/14 backbone. Following prior work, we use the CHAIR [38] benchmark, a rule-based metric on MS-COCO [31] for evaluating object hallucination from generated captions. We report the average per-instance metric CHAIR$_I$ within each bin of CLIP similarity, which evaluates the object hallucination rates in the entire initial response. As shown in Figure 3 (*Left*), a clear negative correlation between hallucination rates and CLIP similarities is observed (with a correlation coefficient of $\rho = -0.63$). This indicates that *lower similarity between original image and generated image corresponds to higher rates of hallucinations at the response level*. (2) Similarly, we generate an image v' based on the initial response given by LLaVA-1.5 for each instance on

the POPE [30] benchmark. In Figure 3 (*Right*), we present the density plot of Jensen-Shannon (JS) divergence between the predicted probabilities for both images, *i.e.*, $p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})$ and $p_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})$, for hallucinatory and non-hallucinatory tokens.[2] The results show that the density of JS divergence follows a long-tail distribution, with hallucinatory tokens exhibiting significantly longer tails and higher JS divergence. This shows *JS divergence between probabilities derived from the original and the generated image corresponds well to hallucinations at the token level.* These observations provide insights into the effectiveness of generative models in reflecting hallucinations, and motivate us to incorporate the generative feedback during the decoding process.

3.3 Self-Correcting Decoding with Generative Feedback

In this section, we focus on effectively utilizing generative feedback during the decoding process to mitigate potential hallucinations. Specifically, we propose a self-correcting decoding approach that leverages generative feedback to *confirm* or *revise* the initial response by selectively enhancing or contrasting the logits for each generated token based on the measured divergence between the two predicted probability distributions.

Specifically, to predict a specific token y_t, we utilize LVLMs to generate two output distributions, each conditioned on either the original image v or the synthesized visual reference v', expressed as:

$$p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})], \quad p_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})]\,. \tag{3}$$

We define and compute the following distance metric based on Jensen-Shannon (JS) divergence at each timestep t to quantify the discrepancy between two next-token probability distributions:

$$\begin{aligned} d_t(v, v') &= \mathcal{D}_{\text{JS}}\left(p_\theta\left(y_t|v, \mathbf{x}, \mathbf{y}_{<t}\right) \parallel p_\theta\left(y_t|v', \mathbf{x}, \mathbf{y}_{<t}\right)\right), \\ \text{where } \mathcal{D}_{\text{JS}}(P \parallel Q) &= \frac{1}{2}\mathcal{D}_{\text{KL}}(P \parallel M) + \frac{1}{2}\mathcal{D}_{\text{KL}}(Q \parallel M), \text{ and } M = \frac{1}{2}(P + Q). \end{aligned} \tag{4}$$

Here, $\mathcal{D}_{\text{KL}}$ represents the Kullback-Leibler (KL) divergence. Note that $d_t(v, v') \in [0, 1]$ is a symmetric metric, providing a fine-grained measure of how closely the two distributions align as the model predicts each subsequent token.

We consider two scenarios based on the token-level generative feedback: (1) If the two predictions are aligned and both images agree on a specific token prediction, we *confirm* the original prediction as correct, and the auxiliary prediction from the generated image can be combined with the original prediction for enhancement (complementary decoding [50]). (2) Conversely, if there is significant discrepancy between the predictions, indicating that the original prediction is likely hallucinatory, we *revise* the original response by using the generated visual input as a contrasting reference to refine the initial next-token prediction (contrastive decoding [26]). To implement this, we introduce a distance threshold γ and develop two corresponding decoding approaches as follows:

$$y_t \sim p_\theta(y_t) = \begin{cases} \texttt{Softmax}\left[f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) + \alpha_1\, f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})\right], & \text{if } d_t(v, v') < \gamma; \\ \texttt{Softmax}\left[(1 + \alpha_2)\, f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) - \alpha_2\, f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})\right], & \text{if } d_t(v, v') \geq \gamma, \end{cases} \tag{5}$$

where α_1 and α_2 are hyperparameters that control the influence of the generated visual reference in the final prediction. Note that setting $\alpha_1 = 0$ or $\alpha_2 = 0$ degrades this process to regular decoding. The final generated token y_t is sampled from the multinomial distribution with probabilities $p_\theta(y_t)$.

4 Experiments

In this section, we evaluate the effectiveness of our method in mitigating hallucinations in LVLMs across a range of benchmarking scenarios, comparing it with existing state-of-the-art approaches.

4.1 Experimental Settings

Evaluated LVLMs. We evaluate the effectiveness of our method on two state-of-the art open-source LVLMs: LLaVA-1.5 [34] and InstructBLIP [13]. Both LVLMs utilize Vicuna-7B [10] as the language encoder, which is instruction-tuned from LLaMA [46]. LLaVA-1.5 [34] employs a

[2]Note that POPE benchmark contains yes-or-no questions about object existence. In this experiment, we evaluate only the first response token (*i.e.*, *yes* or *no*) to determine the presence of hallucinations.

pre-trained CLIP ViT-L/14 [37] as the vision encoder, and trains a linear mapping layer to connect the vision and language modalities. In contrast, InstructBLIP [13] builds on a pre-trained BLIP-2 [28] and incorporates an instruction-aware Q-Former module to bridge the modalities.

Benchmarks and Metrics. We conduct extensive experiments on the following five benchmarks:

- **POPE [30]** is a widely used benchmark for assessing object hallucinations in LVLMs. It tests the models with yes-or-no questions regarding the presence of specific objects, such as, "`Is there a {object} in the image?`" The benchmark draws data from three existing datasets: MSCOCO [31], A-OKVQA [43], and GQA [22], and comprises three distinct subsets—*random*, *popular*, and *adversarial*—based on how the negative samples are generated. For each dataset setting, the benchmark provides 6 questions per image, resulting in 3,000 test instances. We evaluate the performance of different methods using four metrics: accuracy, precision, recall, and F1 score.
- **CHAIR [38]** evaluates object hallucinations in open-ended captioning tasks. It prompts the LVLMs to describe specific images selected from a random sample of 500 images from the MSCOCO validation set and assesses performance based on two metrics:

$$\text{CHAIR}_I = \frac{\text{\# hallucinated objects}}{\text{\# all objects mentioned}}, \quad \text{CHAIR}_S = \frac{\text{\# sentences with hallucinated object}}{\text{\# all sentences}}. \tag{6}$$

  Additionally, we assess the recall and the average length of the generated responses.
- **MME-Hallucination [16]** is a comprehensive benchmark for LVLMs consisting of four subsets: *existence* and *count* for object-level hallucinations, and *position* and *color* for attribute-level hallucinations. Each subset includes 30 images and 60 questions. Similar to POPE [30], these questions are structured as yes-or-no queries, and performance is assessed based on binary accuracy. Following the official implementation, the reported score is calculated by combining accuracy and accuracy+, where accuracy is based on individual questions, and accuracy+ is based on images where both questions are answered correctly.
- **MMVP [45]** collects CLIP-blind pairs and evaluate the fine-grained visual recognition capabilities of LVLMs. It consists of 150 image pairs, each accompanied by a binary-option question. Each image is queried independently, and for a given pair, the LVLM's response is considered correct only if both associated questions are answered accurately.
- **LLaVA-Bench**[3] provides 24 images featuring complex scenes, memes, paintings, and sketches, along with 60 challenging questions. We select examples from this dataset to provide qualitative comparisons between the responses generated by different decoding methods. We also follow Yin *et al*. [52] to evaluate the accuracy and detailedness of generated responses of different methods using the advanced LVLM, GPT-4V[4].

Baselines. As a simple baseline, we include results from regular decoding, where the next token is sampled directly from the post-softmax probability distribution. Additionally, we compare the performance of our method three state-of-the-art decoding approaches: VCD [26], M3ID [15], and RITUAL [50]. For evaluations on the MME-Hallucination benchmark, we further include comparisons with DoLa [12] and OPERA [21]. We report the performance of these baselines based on our re-implementation using their released code bases.

Implementation Details. In our experiments, we adhere to the default query format for the input data used in both LLaVA-1.5 [34] and InstructBLIP [13]. Additionally, we set $\alpha_1 = 3$, $\alpha_2 = 1$, and $\gamma = 0.1$ by default in our decoding process. We follow VCD [26] to implement adaptive plausibility constraints [29], with β set to 0.1 by default. To ensure the reliability of our results, we conduct MME experiments three times with different initialization seeds and report the mean accuracy along with the standard deviation. All experiments are conducted on a single 48GB NVIDIA RTX 6000 Ada GPU.

4.2 Results and Discussions

Results on POPE. In Table 1, we compare the performance of our method against other baselines on POPE benchmark under three different negative sampling settings, across three datasets. As shown in the table, our method consistently outperforms other decoding methods on both LVLMs, achieving state-of-the-art accuracies across all 18 settings, with improvements of up to 5.24% in accuracy,

[3] https://huggingface.co/datasets/liuhaotian/llava-bench-in-the-wild.
[4] https://openai.com/index/gpt-4v-system-card.

Table 1: **Results on POPE [30] benchmark**. Higher (↑) accuracy, precision, recall, and F1 indicate better performance. The best results in each setting are **bolded**, and the second-best are underlined.

	Setup	Method	LLaVA-1.5 [34]				InstructBLIP [13]			
			Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑
MS-COCO [31]	Random	Regular	83.13	81.94	85.00	83.44	83.07	83.02	83.13	83.08
		VCD [26]	87.00	86.13	88.20	87.15	86.23	88.14	83.73	85.88
		M3ID [15]	87.50	87.38	87.67	87.52	86.67	88.09	84.80	86.41
		RITUAL [50]	88.87	89.23	**88.40**	**88.81**	**88.83**	90.48	**86.80**	**88.60**
		Ours	**89.03**	**91.20**	86.40	88.74	**88.83**	**93.73**	82.41	87.71
	Popular	Regular	81.17	78.28	86.27	82.08	77.00	73.82	83.67	78.44
		VCD [26]	83.10	79.96	88.33	83.94	80.07	77.67	84.40	80.89
		M3ID [15]	84.30	81.58	**88.60**	84.95	80.97	77.93	86.40	81.85
		RITUAL [50]	85.83	84.17	88.27	86.17	81.97	78.90	**87.27**	**82.87**
		Ours	**86.63**	**87.75**	84.86	**86.28**	**82.73**	**84.02**	80.27	82.10
	Adversarial	Regular	77.43	73.31	86.27	79.26	74.60	71.26	82.47	76.45
		VCD [26]	77.17	72.18	**88.40**	79.47	77.20	74.29	83.20	78.49
		M3ID [15]	78.23	73.51	88.27	80.22	77.47	73.68	85.47	79.14
		RITUAL [50]	78.80	74.43	87.73	80.54	78.73	74.57	**87.20**	**80.39**
		Ours	**81.63**	**80.59**	83.33	**81.94**	**80.30**	**80.90**	79.33	80.11
A-OKVQA [43]	Random	Regular	81.90	76.63	91.80	83.53	80.63	76.82	87.73	81.92
		VCD [26]	83.83	78.05	94.13	85.34	84.20	80.90	89.53	85.00
		M3ID [15]	84.67	79.25	93.93	85.97	85.43	81.77	91.20	86.23
		RITUAL [50]	85.17	79.79	**94.20**	86.40	87.13	83.92	**91.87**	**87.71**
		Ours	**86.93**	**84.28**	90.80	**87.42**	**87.40**	**87.67**	86.85	87.26
	Popular	Regular	75.07	68.58	92.53	78.77	75.17	70.15	87.60	77.91
		VCD [26]	76.63	69.59	**94.60**	80.19	78.63	73.53	89.47	80.72
		M3ID [15]	77.80	70.98	94.07	80.91	78.80	73.38	90.40	81.00
		RITUAL [50]	78.83	71.99	94.40	81.68	78.73	72.83	**91.67**	81.17
		Ours	**80.90**	**75.68**	91.07	**82.66**	**81.47**	**78.61**	86.47	**82.35**
	Adversarial	Regular	67.23	61.56	91.80	73.70	69.87	64.54	88.20	74.54
		VCD [26]	67.40	61.39	93.80	74.21	71.00	65.41	89.13	75.45
		M3ID [15]	68.60	62.22	**94.73**	75.11	70.10	64.28	90.47	75.16
		RITUAL [50]	68.57	62.26	94.27	74.99	70.27	64.15	**91.87**	75.55
		Ours	**72.70**	**66.70**	90.67	**76.86**	**73.93**	**69.36**	85.67	**76.67**
GQA [22]	Random	Regular	82.23	76.32	93.47	84.03	79.67	76.05	86.60	80.99
		VCD [26]	83.23	76.73	95.40	85.05	82.83	80.16	87.27	83.56
		M3ID [15]	84.20	78.00	95.27	85.77	83.07	80.06	88.07	83.87
		RITUAL [50]	86.10	80.30	**95.67**	87.31	84.87	82.52	**88.47**	**85.39**
		Ours	**87.40**	**83.51**	93.20	**88.09**	**85.40**	**85.64**	84.60	85.12
	Popular	Regular	73.47	66.83	93.20	77.84	73.33	68.72	85.67	76.26
		VCD [26]	72.37	65.27	95.60	77.58	76.13	71.10	88.07	**78.68**
		M3ID [15]	73.87	66.70	95.33	78.49	75.17	69.94	88.27	78.04
		RITUAL [50]	74.80	67.50	**95.67**	79.15	74.50	69.17	**88.40**	77.61
		Ours	**78.10**	**71.56**	93.27	**80.98**	**76.90**	**73.89**	83.20	78.27
	Adversarial	Regular	68.60	62.43	93.40	74.84	68.60	63.94	85.33	73.10
		VCD [26]	68.83	62.26	95.67	75.43	71.00	65.75	87.67	75.14
		M3ID [15]	68.67	62.16	95.40	75.28	71.17	65.79	88.20	**75.36**
		RITUAL [50]	68.23	61.75	**95.80**	75.10	70.17	64.76	**88.47**	74.78
		Ours	**74.07**	**67.42**	93.13	**78.22**	**73.63**	**70.08**	82.47	75.11

6.33% in precision, and 2.79% in F1 score compared to the second-best approach. This suggests that incorporating a generative reference enables the LVLMs to perceive more fine-grained visual details, thereby effectively addressing object hallucinations. Moreover, while most decoding methods tend to be overconfident in their responses, the double-check mechanism in our method makes it more conservative in responding `Yes`, as evidenced by significantly higher precision across all settings. This highlights its enhanced performance in filtering out false positives and suppressing misinformation.

Another notable finding is that our method shows significantly improved performance in the *popular* and *adversarial* settings, which are more challenging than the *random* setting. In the *popular* and *adversarial* settings, non-existent negative objects frequently appear and co-occur with other objects, making them more susceptible to hallucination by LVLMs, as evidenced by the varying degrees of performance degradation across all baselines. However, our method exhibits a lower performance drop compared to other baselines, demonstrating its effectiveness in addressing hallucinations arising from object co-occurrence.

Results on CHAIR. We also compare the performance of our methods and other state-of-the-art methods in the open-ended captioning task and report the CHAIR scores, recall, and the average length of responses in Table 2. The results across two LVLMs demonstrate consistent performance improvements from our method over the compared methods. Specifically, our method outperforms others by 3.0% and 2.6% on the CHAIR_S metric, while also improving the detailedness of generated

Table 2: **Results on CHAIR [38] benchmark.** We limit the maximum number of new tokens to 64. Lower (↓) CHAIR_S, CHAIR_I and higher (↑) recall and length indicate better performance. The best results in each setting are **bolded**, and the second-best are underlined.

Method	LLaVA-1.5 [34]				InstructBLIP [13]			
	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑	Length ↑	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑	Length ↑
Regular	26.2	9.4	58.5	53.4	31.2	11.1	59.0	53.6
VCD [26]	24.4	7.9	63.3	54.2	30.0	10.1	61.8	54.2
M3ID [15]	21.4	6.3	**64.4**	53.5	30.8	10.4	62.6	53.4
RITUAL [50]	22.4	6.9	63.0	**54.9**	26.6	8.9	63.4	55.3
Ours	**18.4**	**6.1**	62.7	54.1	**24.0**	**7.7**	**67.2**	**55.5**

Table 3: **Results on MME-Hallucination [16] benchmark.** We report the average MME scores for each subset, along with the standard deviation across three random seeds. Higher MME scores (↑) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Model	Method	Object-level		Attribute-level		Total Score ↑
		Existence ↑	Count ↑	Position ↑	Color ↑	
LLaVA-1.5 [34]	Regular	173.75 (±4.79)	121.67 (±12.47)	117.92 (±3.69)	149.17 (±7.51)	562.50 (±3.96)
	DoLa [12]	176.67 (±2.89)	113.33 (±10.41)	90.55 (±8.22)	141.67 (±7.64)	522.22 (±16.78)
	OPERA [21]	183.33 (±6.45)	137.22 (±6.31)	122.78 (±2.55)	155.00 (±5.00)	598.33 (±10.41)
	VCD [26]	186.67 (±5.77)	125.56 (±3.47)	128.89 (±6.73)	139.45 (±12.51)	580.56 (±15.13)
	M3ID [15]	186.67 (±5.77)	128.33 (±10.41)	131.67 (±5.00)	151.67 (±20.88)	598.11 (±20.35)
	RITUAL [50]	187.50 (±2.89)	139.58 (±7.64)	125.00 (±10.27)	164.17 (±6.87)	616.25 (±20.38)
	Ours	**188.33** (±2.89)	**150.00** (±7.64)	**133.89** (±3.85)	**172.22** (±3.47)	**644.44** (±9.18)
InstructBLIP [13]	Regular	160.42 (±5.16)	79.17 (±8.22)	**79.58** (±8.54)	130.42 (±17.34)	449.58 (±24.09)
	DoLa [12]	175.00 (±5.00)	55.00 (±5.00)	48.89 (±3.47)	113.33 (±6.67)	392.22 (±7.88)
	OPERA [21]	175.00 (±3.33)	61.11 (±3.47)	53.89 (±1.92)	120.55 (±2.55)	410.56 (±9.07)
	VCD [26]	158.89 (±5.85)	**91.67** (±18.34)	66.11 (±9.76)	121.67 (±12.58)	438.33 (±16.07)
	M3ID [15]	160.00 (±5.00)	87.22 (±22.63)	69.44 (±9.18)	125.00 (±7.64)	441.67 (±17.32)
	RITUAL [50]	182.50 (±6.45)	74.58 (±5.99)	67.08 (±10.31)	139.17 (±0.96)	463.33 (±12.40)
	Ours	**186.67** (±2.89)	89.44 (±8.22)	58.33 (±4.41)	**150.00** (±1.89)	**484.44** (±11.34)

responses compared to regular decoding, as evidenced by higher recall and length. These results demonstrate that by enhancing the fine-grained visual recognition capabilities of LVLMs, our method effectively mitigates object hallucinations in captioning tasks.

Results on MME-Hallucination. Beyond object hallucinations, we further compare the performance of our method with other approaches using the more comprehensive MME-Hallucination benchmark, which includes both object-level and attribute-level hallucinations. The results in Table 3 demonstrate that our method significantly outperforms the compared methods, with substantial margins in the total score metric (*e.g.*, +18.19 on LLaVA-1.5 and +21.11 on InstructBLIP) and consistently superior performance across various evaluation settings, achieving the best results in 6 out of 8 settings. Moreover, our method shows notable improvements on the attribute-level *color* subset, further demonstrating its effectiveness in addressing a wide range of hallucinations beyond object existence.

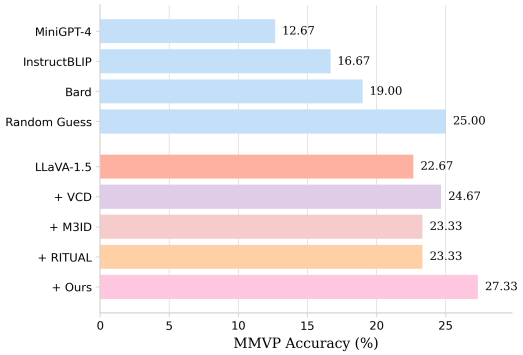

Figure 4: **Results on MMVP [45] benchmark**. We apply our approach to LLaVA-1.5 [34] and compare its performance against other hallucination mitigation methods. For reference, we also report the performance of other LVLMs.

Results on MMVP. We conduct experiments on the MMVP benchmark to assess the fine-grained visual recognition capabilities of LVLMs. As shown in Figure 4, applying our approach to LLaVA-1.5 significantly improves performance from 22.67% to 27.33%, and demonstrates advantages over other hallucination mitigation methods [26, 15, 50]. This suggests that our approach enhances the visual recognition capabilities of LVLMs in distinguishing similar images with different contents.

Case Study on LLaVA-Bench. In Figure 5, we present a case study on LLaVA-Bench comparing our method's response with the response generated by regular decoding using the LLaVA-1.5 model.

Figure 5: **Case study on the LLaVA-Bench benchmark**. We compare the responses generated by regular decoding and our method using LLaVA-1.5. GPT-4V-aided evaluation results are also provided alongside the responses. Hallucinated and accurate content is highlighted in red and green.

Table 4: **Sensitivity analysis of hyperparameter γ.** We report the performance variations on the LLaVA-1.5 backbone by adjusting the value of γ.

Values of γ	POPE Acc.	CHAIR_S	CHAIR_I	MME Score
$\gamma = 0$	87.93	21.0	7.2	622.50
$\gamma = 0.01$	88.07	21.0	6.8	632.22
$\gamma = 0.05$	88.67	19.4	6.3	637.50
$\gamma = 0.1$	**89.03**	**18.4**	**6.1**	644.44
$\gamma = 0.5$	88.73	19.8	6.4	**646.67**
$\gamma = 1$	88.43	21.6	6.6	638.33

Table 5: **Effects of different generative models.** We report the performance using different stable diffusion models on the LLaVA-1.5 backbone.

Models	POPE Acc.	CHAIR_S	CHAIR_I	MME Score
Regular	83.13	26.2	9.4	562.50
SD-v1.1	88.37	19.3	6.5	638.33
SD-v1.5	**89.03**	18.4	6.1	644.44
SD-v2.1	88.70	18.8	6.7	632.22
SD-XL-v0.9	88.87	18.6	6.1	642.50
SD-XL-v1.0	88.60	**17.9**	**5.8**	**648.33**

Specifically, regular decoding often leads to hallucinated or inaccurate content, such as describing "*the island below the mountain*". Besides, the response generated by regular decoding tends to focus on elements like the "*cloudy sky*" and "*cohesive and captivating island landscape*" without providing specific information about the central features of the image. In contrast, our response is more detailed, mentioning the volcano, the road, the surrounding greenery, and the inhabited areas, which gives a clearer understanding of the image's content. GPT-4V-aided evaluation further confirms that our method enhances both the accuracy and detailedness of the generated response.

4.3 Ablation Studies

Impacts of Distance Threshold γ. In Section 3.3, we introduce a distance threshold γ to determine the appropriate decoding algorithm for each generated token. Table 4 presents an analysis of our method's performance across various values of γ. Notably, when γ is set to either 0 or 1—corresponding to the exclusive use of contrastive or complementary decoding for all tokens—the performance exhibits a significant decline, by 0.6% and 1.1% in POPE accuracy, respectively. Moreover, our default setting of $\gamma = 0.1$ achieves the best performance in 3 out of 4 metrics.

Effects of Different Generative Models. In Table 5, we examine the effects of using different generative models (*i.e.*, various versions of Stable Diffusion) with the same LLaVA-1.5 backbone. The results show that different generative models do not result in significant performance variations, and all demonstrate clear improvements compared to the original regular decoding. Although utilizing SD-XL-v1.0 [36] yields slightly better performance, we opt for SD-v1.5 as the default due to its faster image generation speed (3.8 s/image *vs*. 11.3 s/image).

5 Conclusion

In this work, we present self-correcting Decoding with Generative Feedback (DeGF), a novel training-free decoding algorithm that leverages feedback from generative models to recursively improve the accuracy of generated responses. Specifically, we generate a new image based on the initial response given by LVLMs, which serves as a visual reference and provides token-level feedback for mitigating hallucinations. Building on this, we propose a token-level adaptive decoding algorithm that measures the discrepancy between next-token predictions conditioned on the original and generated images, selecting either contrastive or complementary decoding to reduce the likelihood of hallucinatory responses. Extensive experimental results across five benchmarks demonstrate that our proposed DeGF consistently outperforms state-of-the-art methods in mitigating hallucinations in LVLMs.

Acknowledgements

This work has been funded in part by the Army Research Laboratory (ARL) award W911NF-23-2-0007, DARPA award FA8750-23-2-1015, and ONR award N00014-23-1-2840. MM and LPM are partially supported by Meta and National Institutes of Health awards R01MH125740, R01MH132225, and R21MH130767. RS is supported in part by the ONR grant N00014-23-1-2368.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1, 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. 3

[5] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023. 3

[6] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024. 1

[7] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *International Conference on Machine Learning*, 2024. URL: https://openreview.net/forum?id=EYvEVbfoDp. 3

[8] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023. 1, 3

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 4

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/. 3, 5

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. 3

[12] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL: https://openreview.net/forum?id=Th6NyL07na. 1, 6, 8

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023. 1, 2, 3, 5, 6, 7, 8

[14] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024. 1, 3

[15] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 1, 3, 6, 7, 8

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 6, 8

[17] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 3

[19] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024. 1, 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[21] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, 2024. 1, 6, 8

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 6, 7

[23] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 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, 2024. 1, 3

[24] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 3

[26] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 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, 2024. 1, 3, 5, 6, 7, 8

[27] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 3

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 1, 3, 6

[29] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 12286–12312, 2023. 1, 6

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 1, 2, 3, 4, 5, 6, 7

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 4, 6, 7

[32] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, 2024. URL: https://openreview.net/forum?id=J44HfH4JCg. 3

[33] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1, 3

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. 1, 2, 3, 5, 6, 7, 8

[35] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 3

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL: https://openreview.net/forum?id=di52zR8xgf. 2, 9

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[38] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 2, 4, 6, 8

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[41] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning*, pages 30105–30118. PMLR, 2023. 3

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[43] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 6, 7

[44] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics*, pages 13088–13110, 2024. 1, 3

[45] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 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, 2024. 2, 6, 8

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 5

[47] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer, 2024. 3

[48] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics*, pages 15840–15853, 2024. 1

[49] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022. 3

[50] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024. 3, 5, 6, 7, 8

[51] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 1, 3

[52] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 1, 3, 6

[53] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. URL: https://openreview.net/forum?id=AFDcYJKhND. 3

[54] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11766–11781, 2024. 3

[55] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[56] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024. 3

[57] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. URL: https://openreview.net/forum?id=oZDJKTlOUe. 3

[58] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024. 3

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 3