Toward More Reliable Multimodal Systems: Mitigating Hallucinations in Large Vision-Language Models

Zifu Wan

CMU-RI-TR-25-47
June 16, 2025

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Professor Katia Sycara, *chair*
Professor Deva Ramanan
Zhiqiu Lin

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

To everyone who supported me all along the way.

Abstract

Recent advances in Large Vision-Language Models (LVLMs) have led to impressive performance across a wide range of multimodal tasks. However, their tendency to produce hallucinated responses—text that is inconsistent with the visual input—poses a significant challenge to their reliability and real-world applicability. In this thesis, we investigate two training-free approaches for mitigating hallucinations during the decoding process. First, we propose Self-Correcting Decoding with Generative Feedback (DeGF), which leverages the inverse nature of text-to-image generation to detect and correct hallucinations. By synthesizing an auxiliary image from the model's initial textual response, DeGF provides visual self-feedback to verify and revise hallucinated outputs via contrastive or complementary decoding. Second, we introduce ONLY, a highly efficient decoding method that requires only a single query and a lightweight one-layer intervention. By selectively amplifying textual signal based on a text-to-visual entropy ratio, ONLY improves response reliability while maintaining real-time efficiency with minimal computational overhead. Extensive experiments across multiple hallucination benchmarks demonstrate that both DeGF and ONLY significantly outperform existing methods, offering practical and effective solutions for enhancing the trustworthiness of LVLMs in real-world applications.

Acknowledgments

"If You Have a Lemon, Make a Lemonade."—Thank you all for helping me make my lemonade.

Life is full of challenges—moving to a completely new environment, switching to a new area of research—yet it is through these challenges that I grow. Over my two years at CMU, I've stumbled, learned, and matured, becoming a better version of myself. None of this would have been possible without the support of the incredible people I've met here. They've given me unforgettable memories and lessons that I will carry throughout my life.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Katia Sycara. Her insight into meaningful research that serves the broader community has profoundly shaped my perspective. She has always encouraged me to think deeply before acting—a lifelong lesson I'll never forget. Her wisdom, dedication, and collaborative spirit have been a guiding light throughout my journey. I'm especially grateful for her career advice and unwavering support during my job search. Her mentorship has shaped me into a more thoughtful researcher and empowered me to make important life decisions with confidence.

I also extend my sincere thanks to my committee members. Professor Deva Ramanan provided invaluable guidance on both research direction and paper writing. His detailed feedback helped me greatly improve my work. Dr. Zhiqiu Lin offered constructive suggestions on my first research project during my master's program and kindly encouraged me to keep refining my paper until it was the best it could be.

I am especially grateful to Yaqi Xie and Simon Stepputtis, who guided me closely throughout my research. Yaqi offered critical insights into both technical questions and career development. Simon was always willing to engage in deep discussions—whether about research problems, his own journey, or personal growth. Their mentorship has been instrumental to my development as a more capable and reflective researcher. I am also deeply grateful to my undergraduate advisor, Professor Pingping Zhang, who introduced me to computer vision research and helped me build a strong foundation in deep learning.

To my labmates Ce Zhang and Silong Yong—thank you for being such amazing friends and colleagues. Whether it's late-night games in RoboLounge or shared dinners, your presence made every moment better. Ce's sharp skills in academic writing and Silong's broad expertise in computer vision and robotics were constant sources of inspiration. I've also cherished meaningful conversations about life and careers with other members of the lab. In alphabetical order, I would like to thank Aryan, Sarthak Bhagat, Srujan Deolasee, Dana Hughes, Yiqi Jin, Woojun Kim, Muhan Lin, Samuel Li, Venkata Nagarjun, Shuyang Shi, Shreya Sharma, Wei-Hao Zeng, Sha Yi, and Renos Zabounidis.

I'm especially thankful to my roommate, Guofei Chen. His hands-on abilities and deep knowledge in robotics make him a truly outstanding engineer and researcher. Beyond that, he has always been a supportive friend in both life and academics. I'll never forget our adventures—traveling together, watching NBA games, shooting hoops, playing pool, and driving around Mt. Washington to enjoy the beautiful night view of Pittsburgh.

To my dear friends—thank you for being there through it all. In alphabetical order: Chen Bao, Cunxi Dai, Yu Chen, Wen-Tse Chen, Jeff Tan, Jialu Gao, Botao He, Hanzhe Hu, Junkai Huang, Xiaohan Liu, Yifei Liu, Tianxiang Lin, Ming-Feng Li, Yulong Li, Zhixuan Liu, Qianli Ma, Zongzhe Xu, Haochen Zhang, Tianyi Zhang, Wenli Xiao, Yanbo Xu, Jiahui Yang, Zilin Zhang, Jiaxi Zheng, and Shen Zheng. Your companionship made every step of this journey more joyful and fulfilling.

Finally, I owe the greatest thanks to my parents. I could never have come this far without their unconditional support and endless love. They've always encouraged me to pursue my dreams and reminded me that they are always behind me, no matter what. I am forever grateful for the education, values, and freedom they have given me.

Words can never fully capture my gratitude for everyone mentioned here—and many more I haven't named. Every encounter, every conversation, and every memory has helped shape who I am today. Thank you all from the bottom of my heart.

Funding

I gratefully acknowledge the financial support from multiple funding sources, which made this research possible. Specifically, I would like to acknowledge the support from the Army Research Laboratory (ARL) award W911NF-23-2-0007 and W911QX-24-F-0049, DARPA award FA8750-23-2-1015, ONR award N00014-23-1-2840, and the Air Force Office of Scientific Research (AFOSR) award FA9550-18-1-0097 and FA9550-18-1-0251.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

Large Vision-Language Models (LVLMs) have rapidly advanced the capabilities of multimodal AI, demonstrating strong performance in tasks such as image captioning, visual question answering, and instruction following. Despite this progress, a fundamental challenge remains: hallucination, where the generated textual output does not faithfully reflect the visual input. These hallucinations undermine the reliability and trustworthiness of LVLMs, particularly in safety-critical or interactive applications. To address this issue, we propose two novel and training-free approaches in this thesis: Self-Correcting Decoding with Generative Feedback (DeGF) and One-Layer Intervention (ONLY).

DeGF is based on an intuitive yet powerful hypothesis—the process of generating an image from text can serve as a mirror to validate and refine the original text. Specifically, DeGF takes the initial response generated by an LVLM, feeds it into a pre-trained text-to-image generative model to produce a synthetic image, and then uses this auxiliary image to guide the LVLM through a second round of decoding. By comparing the model's predictions conditioned on the original and generated images, DeGF can identify and correct hallucinations at both the response and token levels through complementary and contrastive decoding strategies. DeGF has shown strong empirical performance, outperforming prior methods across multiple hallucination benchmarks.

However, DeGF suffers from a critical limitation: computational efficiency. The method requires generating a high-quality image via a diffusion model and performing three additional LVLM forward passes—one for the initial response and two for decoding under the original and synthetic images—effectively incurring approximately 4× the inference cost of standard decoding. This computational burden hinders its practical deployment in latency-sensitive applications, such as interactive assistants or real-time perception systems.

To address this bottleneck, we introduce ONLY, a training-free approach that mitigates hallucinations with minimal overhead during decoding. ONLY identifies attention heads that favor textual over visual information—specifically, those exhibiting a high text-to-visual entropy ratio—to promote textually grounded next-token predictions. These enhanced predictions are then adaptively combined with the original output logits via a single-layer intervention, effectively reducing dominant and irrelevant language biases. Despite its simplicity, ONLY achieves performance comparable to or better than state-of-the-art methods, while incurring only 1.07× the inference cost—a fraction of DeGF's runtime.

With these two methods, we aim to advance the robustness and reliability of Large Vision-Language Models by addressing hallucinations through novel, training-free strategies. DeGF leverages generative feedback to refine outputs with high accuracy, while ONLY offers an efficient solution that reduces hallucinations with minimal computational overhead. Together, they contribute toward building more trustworthy, effective, and deployable multimodal systems.

Chapter 2

Hallucination Mitigation via Generative Feedback

2.1 Introduction

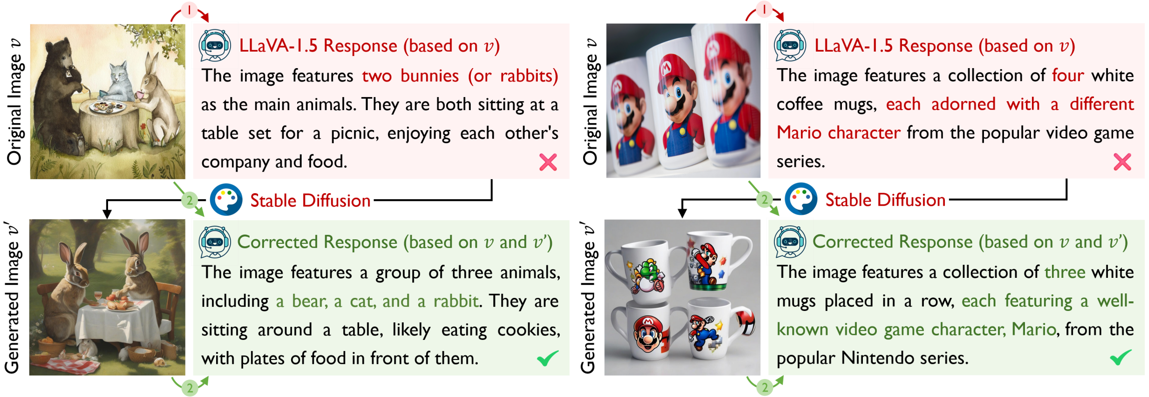
Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across various multi-modal tasks, such as image captioning and visual question answering, by extending the capabilities of powerful Large Language Models (LLMs) to incorporate visual inputs [2, 15, 31, 37, 58]. Despite their 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 [21, 33, 55, 59]. This potential for misinformation raises significant concerns, limiting the models' reliability and restricting their broader deployment in real-world scenarios [3, 6, 36, 67].

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, 29, 36]. In response, various strategies have been developed to detect and mitigate these hallucinations by directly introducing additional training [5, 9, 24, 49, 66], demonstrating promising results in reducing over-reliance. However, the need for additional data and costly training processes hinders their deployment in downstream tasks. More recently, a new paradigm of methods has emerged to tackle the hallucination problem in LVLMs by intervening in the decoding process [16, 23, 28]. Among these, recent training-free contrastive decoding-based methods [32] 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 [17, 29], disturbed instructions [52], or premature layers [13].

While these contrastive decoding-based methods effectively mitigate hallucinations arising from language priors, we recognize that hallucinations can also originate beyond language bias, stemming from visual deficiencies in LVLMs [50]. 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 different types of hallucinations.

In this work, we explore the potential of leveraging powerful text-to-image generative models (*e.g.*, Stable Diffusion [41, 44]) to mitigate various types of 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 response conditioned on the original image is accurate and non-hallucinatory, a text-to-image generative model should be capable of reversing this process to produce a similar image from that response. Alternatively, if there is a discrepancy between the original image and the one generated from the response, this difference 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 2.3.2), 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 self-correcting Decoding with Generative Feedback (DeGF), a novel training-free decoding algorithm that effectively incorporates feedback from text-to-image generative models to recursively enhance the accuracy of LVLM responses. Specifically, for each instance, we generate a new image based on the initial response, which serves as an *auxiliary visual reference* to assess and verify the accuracy of the initial output. We propose self-correcting decoding that either enhances or contrasts predictions from the original and this reference based


Original Image v
1
LLaVA-1.5 Response (based on v)
Stable Diffusion
2
Generated Image v'
Corrected Response (based on v and v')
Original Image v
1
LLaVA-1.5 Response (based on v)
Stable Diffusion
2
Generated Image v'
Corrected Response (based on v and v')

LLaVA-1.5 Response (based on v): The image features two bunnies (or rabbits) as the main animals. They are both sitting at a table set for a picnic, enjoying each other's company and food. ✗

Corrected Response (based on v and v'): The image features a group of three animals, including a bear, a cat, and a rabbit. They are sitting around a table, likely eating cookies, with plates of food in front of them. ✓

LLaVA-1.5 Response (based on v): The image features a collection of four white coffee mugs, each adorned with a different Mario character from the popular video game series. ✗

Corrected Response (based on v and v'): The image features a collection of three white mugs placed in a row, each featuring a well-known video game character, Mario, from the popular Nintendo series. ✓

Figure 2.1: **Generative models can visualize and help correct various types of hallucinations in the initial response.** ① In the first query, we provide LLaVA-1.5 [37] with the prompt "`Describe this image in detail`" to produce captions for two examples from LLaVA-Bench. Based on the initial response, we utilize Stable Diffusion XL [41] to generate a new image v', which effectively highlights hallucinations and provides valuable self-feedback. ② In the second query, our approach incorporates both the original image v and the generated image v' into the decoding process, using the feedback to successfully correct various types of hallucinations.

on the auxiliary visual reference, *confirming* or *revising* the initial LVLM response based on the degree of 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 2.1, 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 text-to-image generative feedback as a self-correcting mechanism for mitigating hallucinations in LVLMs.

The effectiveness of DeGF is evaluated on LLaVA-1.5, InstructBLIP, and Qwen-VL across six benchmarks: POPE [33], CHAIR [43], MME-Hallucination [18], MM-Bench [39], MMVP [50], and LLaVA-Bench. Extensive experimental results validate the effectiveness of our DeGF in mitigating various types of hallucinations in LVLMs. Qualitative case studies and GPT-4V-aided evaluation on LLaVA-Bench further demonstrate that our approach enhances both the accuracy and detailedness of the LVLM responses.

The contributions of this work are summarized as follows:

- We investigate the potential of text-to-image generative models in mitigating hallucinations in LVLMs and demonstrate that text-to-image generative models can provide valuable self-feedback for mitigating hallucinations at both the response and token levels.
- We propose self-correcting Decoding with Generative Feedback (DeGF), a novel training-free decoding algorithm for LVLMs that recursively enhances the accuracy of responses by integrating feedback from text-to-image generative models with complementary/contrastive decoding.
- Extensive experimental evaluations across six benchmarks demonstrate that our DeGF consistently outperforms state-of-the-art approaches in effectively mitigating hallucinations in LVLMs.

2.2 Related Work

Hallucination in LVLMs. With advances of autoregressive LLMs [11, 12, 51], researchers have extended these powerful models to process visual inputs, leading to the development of LVLMs [2, 15, 37, 58]. 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 [3, 36]. However, LVLMs are prone to hallucinations, where contradictions arise between the visual content and the generated textual response [3, 33, 36].

To mitigate hallucinations in LVLMs, early works have introduced various approaches, including reinforcement learning from human feedback (RLHF) [21, 49], applying auxiliary supervision [9, 24], incorporating negative [35] or noisy data [62], and training post-hoc revisors for correction [59, 68]. 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 [17, 29] and guided decoding with auxiliary information [8, 16, 54]. In this work, we present a novel training-free approach that recursively enhances the accuracy of the LVLM response by incorporating text-to-image generative feedback. To the best of our knowledge, we are the first work to effectively utilize feedback from text-to-image generative models to mitigate hallucinations in LVLMs.

Text-to-Image Synthesis. Text-to-image synthesis aims to create realistic images from textual descriptions [19, 69]. In recent years, significant progress has been achieved in this area, largely due to the advent of deep generative models [20, 64]. These advances include Generative Adversarial Networks (GAN) [26, 46], autoregressive models [4, 60], and diffusion models [22, 27, 40, 44, 45]. Among these, diffusion-based methods have been particularly distinguished due to their ability to generate high-quality, detailed images with fine-grained control over the synthesis process [14, 57]. Pre-trained on large-scale text-image datasets such as LAION [47], diffusion-based methods have demonstrated strong vision-language alignment, making them valuable for downstream tasks such as classification [30] and semantic segmentation [1, 53].

More recently, Jiao et al. [25] incorporate text-to-image generative models to enhance fine-grained image recognition in LVLMs by introducing the Img-Diff dataset, which generates pairs of similar images using Stable Diffusion XL [41]. Their results demonstrate that fine-tuning LVLMs with this additional data leads to improved performance on several VQA tasks. In contrast, in this work, we directly leverage a pre-trained diffusion model to provide valuable self-feedback for refining the generated responses of LVLMs in the decoding process, dynamically improving the accuracy and consistency of the model's response without modifying the underlying LVLMs.

2.3 Method

In this work, we present DeGF, a novel training-free algorithm that recursively improves the accuracy of LVLM responses using text-to-image generative feedback, as illustrated in Figure 2.2.

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

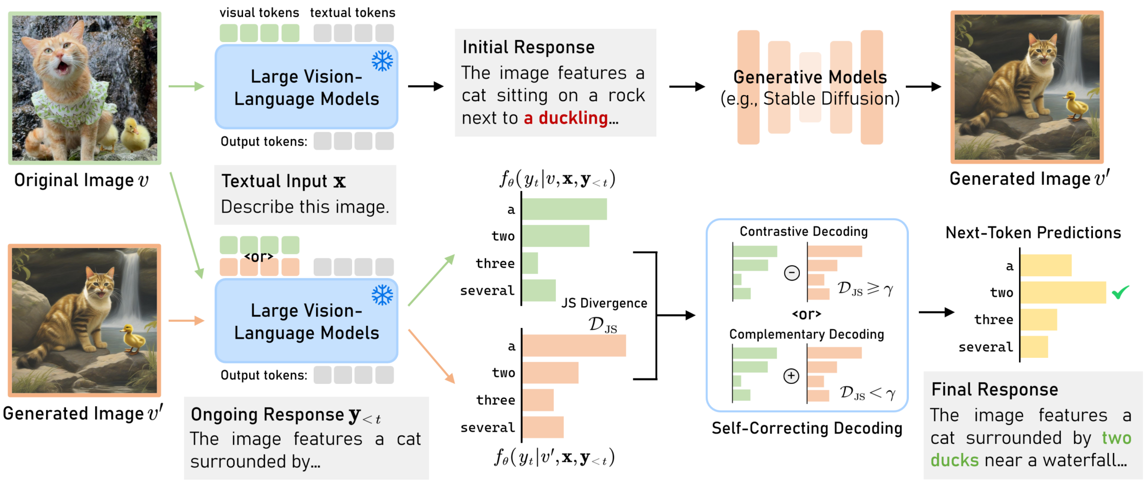

visual tokens
textual tokens
Large Vision-Language Models
Output tokens:
Initial Response
The image features a cat sitting on a rock next to a duckling...
Generative Models
(e.g., Stable Diffusion)
Original Image v
Generated Image v'
Textual Input $\mathbf{x}$
Describe this image.
<or>
Large Vision-Language Models
Output tokens:
Generated Image v'
Ongoing Response $\mathbf{y}_{<t}$
The image features a cat surrounded by...
$f_\theta(y_t|v,\mathbf{x},\mathbf{y}_{<t})$
a
two
three
several
JS Divergence
$\mathcal{D}_{\text{JS}}$
a
two
three
several
$f_\theta(y_t|v',\mathbf{x},\mathbf{y}_{<t})$
Contrastive Decoding
$\mathcal{D}_{\text{JS}} \geq \gamma$
<or>
Complementary Decoding
$\mathcal{D}_{\text{JS}} < \gamma$
Self-Correcting Decoding
Next-Token Predictions
a
two
three
several
Final Response
The image features a cat surrounded by two ducks near a waterfall...

Figure 2.2: **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 either enhances or contrasts the next-token predictions conditioned on the original and generated images to mitigate hallucinations in the LVLM response.

into visual tokens within the textual input space using a vision-language alignment module (*e.g.*, Q-Former [31] or linear projection [37]). These visual tokens, along with the textual query tokens, are then fed into the language encoder for conditioned autoregressive generation. We denote the autoregressive generation process as

$$y_t \sim p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) \propto \exp f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}), \tag{2.1}$$

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.

2.3.2 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 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:

$$v' = \mathcal{G}(\boldsymbol{\tau}, x_T), \quad \text{where } x_T \sim \mathcal{N}(0, \mathbf{I}). \tag{2.2}$$

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 Text-to-Image Generative Models in Reflecting Hallucinations. We validate the effectiveness of generative models in reflecting hallucinations through an empirical study, as shown in Figure 2.3.[1] The experimental results demonstrate that *text-to-image generative models can provide valuable self-feedback for mitigating hallucinations* at both the response and token levels.

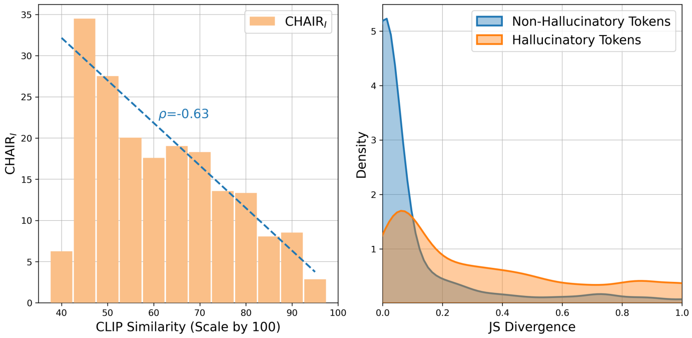


Figure 2.3: **Text-to-image generative models can provide feedback for reflecting hallucinations**. (*Left*) Bar plot of average CHAIR$_I$ scores binned by CLIP similarity (scaled by 100) on the CHAIR benchmark; (*Right*) Density plots of token-level JS divergence for both hallucinatory and non-hallucinatory tokens on the POPE 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 [10] ViT-H/14 backbone. Following prior work, we use the CHAIR [43] benchmark, a rule-based metric on MS-COCO [34] for evaluating object hallucination from generated captions. We report the average per-instance

[1] For Figure 2.3, we evaluate 1,000 CHAIR samples (*Left*) and 3,000 POPE samples (*Right*).

metric CHAIR$_I$ within each bin of CLIP similarity, which evaluates the object hallucination rates in the entire initial response. As shown in Figure 2.3 (*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 the 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 [33] benchmark. In Figure 2.3 (*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.

2.3.3 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:

$$p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})], \;\; p_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})]\,. \tag{2.3}$$

We define and compute the following distance metric based on Jensen-Shannon (JS)

[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.

divergence at each timestep t to quantify the discrepancy between two next-token probability distributions:

$$\begin{aligned} d_t(v, v') = \mathcal{D}_{\mathrm{JS}}\left(p_\theta\left(y_t|v, \mathbf{x}, \mathbf{y}_{<t}\right) \parallel p_\theta\left(y_t|v', \mathbf{x}, \mathbf{y}_{<t}\right)\right), \\ \text{where } \mathcal{D}_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2}\mathcal{D}_{\mathrm{KL}}(P \parallel M) + \frac{1}{2}\mathcal{D}_{\mathrm{KL}}(Q \parallel M), \text{ and } M = \frac{1}{2}(P+Q). \end{aligned} \tag{2.4}$$

Here, $\mathcal{D}_{\mathrm{KL}}$ 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 [54]). (2) Conversely, if there is a 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 [29]). To implement this, we introduce a distance threshold γ and develop two corresponding decoding approaches as follows:

$$y_t \sim p_\theta(y_t) = \begin{cases} \texttt{Softmax}\left[f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) + \alpha_1\, f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})\right], & \text{if } d_t(v, v') < \gamma; \\ \texttt{Softmax}\left[(1+\alpha_2)\, f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) - \alpha_2\, f_\theta(y_t|v', \mathbf{x}, \mathbf{y}_{<t})\right], & \text{if } d_t(v, v') \geq \gamma, \end{cases} \tag{2.5}$$

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 $p_\theta(y_t)$.

2.4 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.

Evaluated LVLMs. We evaluate the effectiveness of our method on three state-of-the-art open-source LVLMs: LLaVA-1.5 [38], InstructBLIP [15], and Qwen-VL [2]. Both LLaVA-1.5 and InstructBLIP utilize Vicuna-7B [11] as the language encoder, which is instruction-tuned from LLaMA [51]. In contrast, Qwen-VL [2] is based on the Qwen 7B backbone. Specifically, we implement our approach using weights of the Qwen-VL-Chat model.

Benchmarks. We conduct extensive experiments on six benchmarks: (1) **POPE [33]** is a widely used benchmark for assessing object hallucinations in LVLMs, which tests the models with yes-or-no questions regarding the presence of specific objects, such as, "`Is there a {object} in the image?`" (2) **CHAIR [43]** 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; (3) **MME-Hallucination [18]** 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; (4) **MM-Bench [39]** serves as a comprehensive benchmark designed to assess the multi-modal understanding capabilities of LVLMs across 20 dimensions; (5) **MMVP [50]** collects CLIP-blind pairs and evaluates the fine-grained visual recognition capabilities of LVLMs. It consists of 150 image pairs, each accompanied by a binary-option question; (6) **LLaVA-Bench** provides 24 images featuring complex scenes, memes, paintings, and sketches, along with 60 challenging questions.

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 with three state-of-the-art decoding approaches: VCD [29], M3ID [17], and RITUAL [54]. For evaluations on the CHAIR [43] and MME-Hallucination [18] benchmark, we further include comparisons with Woodpecker [59], HALC [8], DoLa [13] and OPERA [23]. 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 [38] and InstructBLIP [15]. Additionally, we set $\alpha_1 = 3$, $\alpha_2 = 1$, and $\gamma = 0.1$ by default in our decoding process.

We follow VCD [29] to implement adaptive plausibility constraints [32], where we set $\beta = 0.1$ in open-ended CHAIR benchmark and $\beta = 0.25$ for other tasks. 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.

2.4.1 Results and Discussions

Results on POPE. In Table 2.1, we compare the performance of our method against other baselines on the POPE benchmark under three different negative sampling settings, across three datasets. As shown in the table, our method consistently outperforms other decoding methods on three LVLMs, achieving state-of-the-art accuracies across all settings, with improvements of up to 5.24% in accuracy, 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 self-correcting decoding 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 [33], 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 2.2. The results, evaluated across two different LVLMs, consistently demonstrate performance improvements achieved by our method over the compared approaches. Specifically, our method

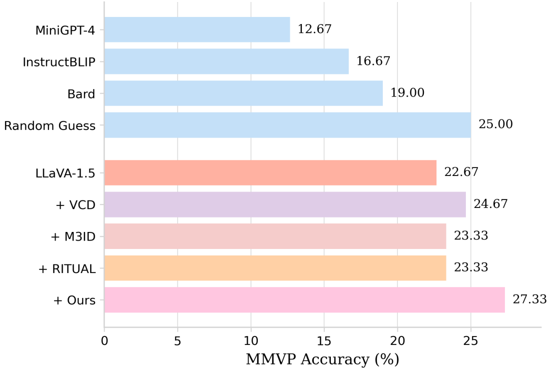


Figure 2.4: **Results on MMVP [50].** We apply our approach to LLaVA-1.5 [38] and compare its performance against other hallucination mitigation methods.

Method	**LLaVA-1.5**		**InstructBLIP**	
	Acc. ↑	Det. ↑	Acc. ↑	Det. ↑
Regular	2.88	3.29	3.42	3.96
Ours	**4.29**	**4.54**	**4.38**	**4.79**
VCD	3.62	3.83	3.71	4.21
Ours	**4.04**	**4.38**	**4.17**	**4.58**
M3ID	3.88	4.08	4.00	4.33
Ours	**4.04**	**4.29**	**4.08**	**4.50**

Table 2.4: **GPT-4V-aided evaluation on LLaVA-Bench**. Higher accuracy and detailedness (↑) indicate better performance. The evaluation is performed on LLaVA-1.5 [38].

outperforms the second-best approach by 3.0% and 2.6% on the CHAIR$_S$ metric, while also enhancing the detailedness of generated responses compared to regular decoding, as indicated by the higher recall and increased response length. These results demonstrate that by incorporating generative feedback into the decoding process of LVLMs, our method effectively mitigates object hallucinations in open-ended captioning tasks.

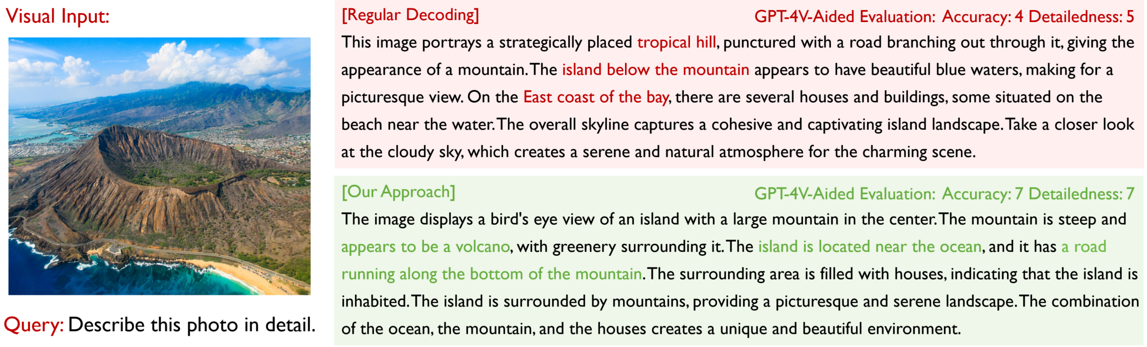
Results on MME-Hallucination and MMBench. Beyond object hallucinations, we further compare the performance of our method with other approaches using the more comprehensive MME-Hallucination benchmark, which includes both object-level and attribute-level hallucinations. The results in Table 2.3 demonstrate 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, which is particularly challenging as it requires models to accurately capture subtle attribute information. This further illustrates the effectiveness of our approach in addressing a wide range of hallucinations, both at the object existence level and in finer-grained attribute recognition. Additionally, our proposed DeGF enhances the general multi-modal understanding capabilities of LVLMs, as evidenced by its superior performance on the MMBench benchmark.

Results on MMVP. We conduct experiments on the MMVP benchmark to assess the fine-grained visual recognition capabilities of LVLMs. As shown in Figure 2.4, applying our self-correcting decoding approach to LLaVA-1.5 significantly improves performance from 22.67% to 27.33%. Our approach also demonstrates notable advantages over other hallucination mitigation baselines, further showcasing its superiority in handling nuanced visual recognition tasks. These results suggest that our approach significantly enhances the model's capacity to discern and correctly interpret fine-grained distinctions between images with similar appearances but different contents. By integrating generative feedback, our approach effectively reduces misinterpretations and improves the precision of visual recognition tasks, contributing to more reliable and accurate performance in complex scenarios.

Results on LLaVA-Bench. In Figure 2.5, 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. 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. The GPT-4V-aided evaluation shown in Table 2.4 further confirms that our method enhances both the accuracy and detailedness of the generated response, outperforming other hallucination mitigation approaches such as VCD and M3ID.

2.4.2 Ablation Studies

Analysis of Distance Threshold γ. In Section 2.3.3, we introduce a distance threshold γ to determine the appropriate decoding algorithm for each generated token. Table 2.5 presents an analysis of our method's performance with various values of γ across three benchmarks. For simplicity, we report the performance on the MS-COCO dataset with *random* setting for all POPE results in the ablation studies. 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



Visual Input:

Query: Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 4 Detailedness: 5

This image portrays a strategically placed tropical hill, punctured with a road branching out through it, giving the appearance of a mountain. The island below the mountain appears to have beautiful blue waters, making for a picturesque view. On the East coast of the bay, there are several houses and buildings, some situated on the beach near the water. The overall skyline captures a cohesive and captivating island landscape. Take a closer look at the cloudy sky, which creates a serene and natural atmosphere for the charming scene.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 7 Detailedness: 7

The image displays a bird's eye view of an island with a large mountain in the center. The mountain is steep and appears to be a volcano, with greenery surrounding it. The island is located near the ocean, and it has a road running along the bottom of the mountain. The surrounding area is filled with houses, indicating that the island is inhabited. The island is surrounded by mountains, providing a picturesque and serene landscape. The combination of the ocean, the mountain, and the houses creates a unique and beautiful environment.

Figure 2.5: **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 2.5: **Sensitivity analysis of distance threshold γ.** We present the performance of our approach, based on the LLaVA-1.5 backbone, across three benchmarks for varying values 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 2.6: **Effects of different generative models.** We report the performance of different variants of our method, utilizing various 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**

decline, by 0.6% and 1.1% in POPE accuracy, respectively. Moreover, our default setting of $\gamma = 0.1$ achieves the optimal performance in 3 out of 4 evaluated metrics.

Effects of Different Generative Models. Table 2.6 presents the performance of various variants of our method that incorporate different generative models (*i.e.*, different versions of Stable Diffusion) while using the same LLaVA-1.5 backbone. The results indicate that the effectiveness of our DeGF is robust to the choice of generative models, as performance remains largely unaffected by the specific model used, and all variants demonstrate consistent improvements over the original regular decoding approach. Although utilizing SD-XL-v1.0 [41] 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).

2.4.3 Efficiency Comparison

In Table 2.7, we compare the efficiency of our approach with other methods on the CHAIR benchmark using the LLaVA-1.5 model, with the maximum token length set to 128. Our approach involves two queries and incorporates a text-to-image generative model to mitigate hallucinations, resulting in a 4.04× increase in latency and a 1.21× increase in GPU memory usage. Specifically, our method consists of three stages: initial response generation, image generation, and response self-correction, which take an average of 3.4 seconds, 3.8 seconds, and 6.6 seconds per instance, respectively. Compared to other approaches, while our method is slower than regular decoding and contrastive decoding-based methods, it demonstrates efficiency advantages over OPERA and HALC. Note that our approach also achieves the lowest hallucination rates among all compared methods.

Table 2.7: **Efficiency comparison**. For each method, we present the average inference latency per instance and peak GPU memory. Experiments are conducted on a single RTX A6000 Ada GPU.

Method	Avg. Latency ↓	GPU Memory ↓	CHAIR$_S$ ↓
Regular	3.44 s (×1.00)	15778 MB (×1.00)	55.0
VCD	6.91 s (×2.01)	16634 MB (×1.05)	54.4
OPERA	24.70 s (×7.18)	22706 MB (×1.44)	52.6
Woodpecker	10.68 s (×3.10)	22199 MB (×1.41)	57.6
HALC	22.61 s (×6.51)	23084 MB (×1.46)	51.0
Ours	13.89 s (×4.04)	19119 MB (×1.21)	48.8

Table 2.1: **Results on POPE [33] benchmark**. Higher (↑) accuracy, precision, recall, and F1 indicate better performance. The best results are **bolded**, and the second-best are underlined.

	Setup	Method	LLaVA-1.5			InstructBLIP			Qwen-VL		
			Acc. ↑	Prec. ↑	F1 ↑	Acc. ↑	Prec. ↑	F1 ↑	Acc. ↑	Prec. ↑	F1 ↑
MS-COCO	Random	Regular	83.13	81.94	83.44	83.07	83.02	83.08	87.43	93.56	86.48
		VCD	87.00	86.13	87.15	86.23	88.14	85.88	88.80	93.89	88.11
		M3ID	87.50	87.38	87.52	86.67	88.09	86.41	**89.83**	95.44	89.17
		RITUAL	88.87	89.23	**88.81**	**88.83**	90.48	**88.60**	89.47	**96.32**	88.62
		Ours	**89.03**	**91.20**	88.74	**88.83**	**93.73**	87.71	89.73	93.19	**89.31**
	Popular	Regular	81.17	78.28	82.08	77.00	73.82	78.44	84.70	88.24	83.96
		VCD	83.10	79.96	83.94	80.07	77.67	80.89	85.13	87.27	84.69
		M3ID	84.30	81.58	84.95	80.97	77.93	81.85	86.27	89.19	**85.73**
		RITUAL	85.83	84.17	86.17	81.97	78.90	**82.87**	84.57	84.09	84.67
		Ours	**86.63**	**87.75**	**86.28**	**82.73**	**84.02**	82.10	**86.50**	**89.87**	85.71
	Adversarial	Regular	77.43	73.31	79.26	74.60	71.26	76.45	79.83	80.13	79.73
		VCD	77.17	72.18	79.47	77.20	74.29	78.49	81.33	80.60	81.55
		M3ID	78.23	73.51	80.22	77.47	73.68	79.14	82.03	81.47	82.19
		RITUAL	78.80	74.43	80.54	78.73	74.57	**80.39**	82.80	83.15	82.71
		Ours	**81.63**	**80.59**	**81.94**	**80.30**	**80.90**	80.11	**83.47**	**84.49**	**82.98**
A-OKVQA	Random	Regular	81.90	76.63	83.53	80.63	76.82	81.92	86.27	90.66	85.48
		VCD	83.83	78.05	85.34	84.20	80.90	85.00	87.87	90.06	87.53
		M3ID	84.67	79.25	85.97	85.43	81.77	86.23	**88.13**	92.06	87.55
		RITUAL	85.17	79.79	86.40	87.13	83.92	**87.71**	87.73	**92.49**	87.01
		Ours	**86.93**	**84.28**	**87.42**	**87.40**	**87.67**	87.26	87.90	89.16	**87.58**
	Popular	Regular	75.07	68.58	78.77	75.17	70.15	77.91	84.60	87.99	83.88
		VCD	76.63	69.59	80.19	78.63	73.53	80.72	86.23	87.30	86.03
		M3ID	77.80	70.98	80.91	78.80	73.38	81.00	**86.50**	89.59	85.95
		RITUAL	78.83	71.99	81.68	78.73	72.83	81.17	86.36	88.73	86.20
		Ours	**80.90**	**75.68**	**82.66**	**81.47**	**78.61**	**82.35**	86.43	**90.74**	**86.52**
	Adversarial	Regular	67.23	61.56	73.70	69.87	64.54	74.54	76.90	75.59	77.48
		VCD	67.40	61.39	74.21	71.00	65.41	75.45	79.13	76.04	80.30
		M3ID	68.60	62.22	75.11	70.10	64.28	75.16	79.50	77.54	80.21
		RITUAL	68.57	62.26	74.99	70.27	64.15	75.55	80.20	79.08	**80.58**
		Ours	**72.70**	**66.70**	**76.86**	**73.93**	**69.36**	**76.67**	**80.75**	**80.37**	80.46
GQA	Random	Regular	82.23	76.32	84.03	79.67	76.05	80.99	84.90	89.51	83.96
		VCD	83.23	76.73	85.05	82.83	80.16	83.56	85.21	92.05	84.21
		M3ID	84.20	78.00	85.77	83.07	80.06	83.87	85.69	93.11	84.67
		RITUAL	86.10	80.30	87.31	84.87	82.52	**85.39**	**86.13**	93.78	84.81
		Ours	**87.40**	**83.51**	**88.09**	**85.40**	**85.64**	85.12	85.95	**94.22**	**85.08**
	Popular	Regular	73.47	66.83	77.84	73.33	68.72	76.26	81.33	83.38	80.74
		VCD	72.37	65.27	77.58	76.13	71.10	**78.68**	81.97	82.82	81.73
		M3ID	73.87	66.70	78.49	75.17	69.94	78.04	**82.13**	84.58	81.48
		RITUAL	74.80	67.50	79.15	74.50	69.17	77.61	81.13	85.48	81.03
		Ours	**78.10**	**71.56**	**80.98**	**76.90**	**73.89**	78.27	82.10	**86.39**	**81.85**
	Adversarial	Regular	68.60	62.43	74.84	68.60	63.94	73.10	79.03	80.43	78.54
		VCD	68.83	62.26	75.43	71.00	65.75	75.14	80.87	81.07	80.80
		M3ID	68.67	62.16	75.28	71.17	65.79	**75.36**	81.03	82.93	**80.94**
		RITUAL	68.23	61.75	75.10	70.17	64.76	74.78	81.07	83.29	80.41
		Ours	**74.07**	**67.42**	**78.22**	**73.63**	**70.08**	75.11	**81.13**	**84.18**	80.57

Table 2.2: **Results on CHAIR [43] 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				InstructBLIP			
	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall ↑	Length ↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall ↑	Length ↑
Regular	26.2	9.4	58.5	53.4	31.2	11.1	59.0	53.6
VCD	24.4	7.9	63.3	54.2	30.0	10.1	61.8	54.2
M3ID	21.4	6.3	**64.4**	53.5	30.8	10.4	62.6	53.4
RITUAL	22.4	6.9	63.0	**54.9**	26.6	8.9	63.4	55.3
Woodpecker	24.9	7.5	60.8	49.7	31.2	10.8	62.3	51.3
HALC	21.7	7.1	63.4	53.4	24.5	8.0	63.8	55.1
Ours	**18.4**	**6.1**	62.7	54.1	**24.0**	**7.7**	**67.2**	**55.5**

Table 2.3: **Results on MME-Hallucination [18] and MMBench [39] benchmark.** We report the average MME scores along with the standard deviation across three random seeds for each subset. We also report the overall accuracy achieved by the different methods on the MMBench benchmark in the final column. Higher scores (↑) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Method	Object-level		Attribute-level		MME Score ↑	MMBench ↑
	Existence ↑	Count ↑	Position ↑	Color ↑		
Regular	173.75 (±4.79)	121.67 (±12.47)	117.92 (±3.69)	149.17 (±7.51)	562.50 (±3.96)	64.1
DoLa	176.67 (±2.89)	113.33 (±10.41)	90.55 (±8.22)	141.67 (±7.64)	522.22 (±16.78)	63.8
OPERA	183.33 (±6.45)	137.22 (±6.31)	122.78 (±2.55)	155.00 (±5.00)	598.33 (±10.41)	64.4
VCD	186.67 (±5.77)	125.56 (±3.47)	128.89 (±6.73)	139.45 (±12.51)	580.56 (±15.13)	64.6
M3ID	186.67 (±5.77)	128.33 (±10.41)	131.67 (±5.00)	151.67 (±20.88)	598.11 (±20.35)	64.4
RITUAL	187.50 (±2.89)	139.58 (±7.64)	125.00 (±10.27)	164.17 (±6.87)	616.25 (±20.38)	63.8
Woodpecker	187.50 (±2.89)	125.00 (±0.00)	126.66 (±2.89)	149.17 (±17.34)	588.33 (±10.00)	64.0
HALC	183.33 (±0.00)	133.33 (±5.77)	107.92 (±3.69)	155.00 (±5.00)	579.58 (±9.07)	64.2
Ours	**188.33** (±2.89)	**150.00** (±7.64)	**133.89** (±3.85)	**172.22** (±3.47)	**644.44** (±9.18)	**65.5**

Chapter 3

Hallucination Mitigation via One-Layer Intervention

3.1 Introduction

Recent advances in large vision-language models (LVLMs), which expand the capabilities of large language models (LLMs) to visual understanding and reasoning [2, 31, 58], have demonstrated exceptional performance across various vision-language tasks, such as object detection [56, 63] and image captioning [34, 43]. However, a persistent challenge with current LVLMs is their tendency to generate hallucinated content, where the generated responses do not align accurately with the actual image input [36]. This can significantly impact

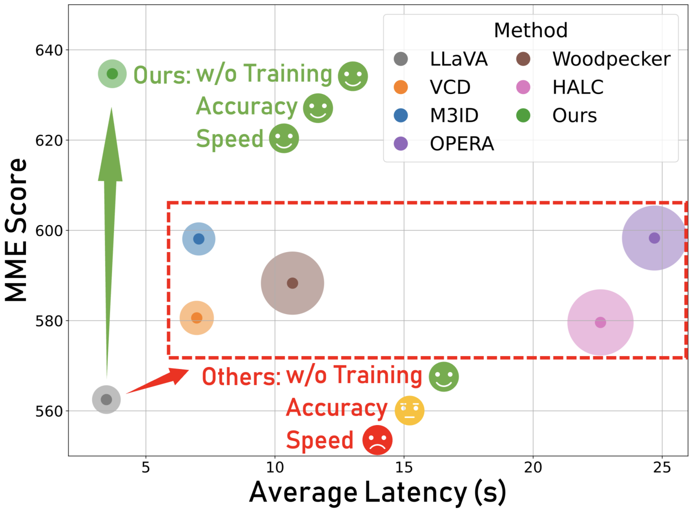


Figure 3.1: **Comparisons of accuracy and speed of multiple hallucination mitigation methods.** The size of bubbles stands for the GPU memory consumption. Our method mitigates hallucination with only 0.07× extra time.

the reliability of LVLMs in real-world applications where precise visual interpretation is essential [3, 6, 36]. Therefore, addressing hallucinations in LVLMs is crucial to ensuring their safe and effective deployment in critical domains.

To alleviate the hallucination problem, early work identified biased training sets as a critical cause and, as a result, attempted to establish curated training datasets and adopt robust fine-tuning techniques [9, 49, 66]. However, their reliance on additional data and the need for fine-tuning large-scale models make these approaches time-consuming and impractical for individual users. Another common approach is contrastive decoding [32], which eliminates the need for costly training by directly intervening in the decoding process during inference. Specifically, these methods typically introduce a distorted set of inputs, and contrast their respective token predictions with the predictions from original data to mitigate undesired hallucinations [13, 17, 29, 52]. Although existing contrastive decoding-based approaches achieve notable performance improvements, they require multiple LVLM queries to process both the original and distorted inputs, resulting in response times that are twice as long, or more, making them less suitable for real-time applications [7, 17, 29].

To illustrate this, we analyze the performance-efficiency trade-off of existing approaches for mitigating hallucinations in LVLMs and present the results in Figure 3.1. As we can see, while other hallucination mitigation methods achieve higher accuracy on the hallucination evaluation benchmark, they come at a significant cost, requiring 2× or more inference time and higher GPU memory consumption. We recognize this overhead is impractical given the limited performance improvements, highlighting the urgent need for more efficient approaches that can effectively mitigate hallucinations in LVLM.

In this work, we introduce ONLY, a training-free approach that requires only a single query and a one-layer intervention during decoding, offering an efficient solution for mitigating hallucinations in LVLMs. Our ONLY approach selects attention heads that prioritize textual information over visual information—specifically, those with a high text-to-visual entropy ratio—to stimulate textually enhanced next-token predictions. The enhanced output is then adaptively contrasted/collaborated with the original output logits using a single-layer intervention, aiming to reduce predominant and irrelevant language bias. Our ONLY approach is both simple and effective, requiring just one additional attention layer computation. It incurs a modest 1.07×

increase in inference time with negligible GPU memory overhead, significantly lower than the 2× or more increase seen in previous contrastive decoding methods. Moreover, ONLY achieves superior performance across multiple benchmarks, outperforming the current state-of-the-art by 3.14% on POPE and 1.6% on CHAIR.

To validate the effectiveness of our proposed ONLY approach, we evaluate it on three LVLMs (*i.e.*, LLaVA-1.5 [38], InstructBLIP [15], and Qwen-VL [2]) across various benchmarks, including POPE [33], CHAIR [43], MME-Hallucination [18], MMBench [39], MM-Vet [61], and MMVP [50]. Extensive experimental results demonstrate that our ONLY approach consistently outperforms state-of-the-art methods across these benchmarks while requiring minimal implementation effort and computational cost. Additionally, qualitative case studies and GPT-4V-aided evaluations on LLaVA-Bench further validate the effectiveness of our ONLY approach in enhancing the coherence and accuracy of LVLM responses.

Our contributions are summarized as follows:

- We investigate and challenge the performance-efficiency trade-off of existing contrastive decoding approaches for mitigating hallucinations in LVLMs, highlighting the efficiency issues.
- We present ONLY, a novel training-free decoding algorithm that leverages a single additional Transformer layer to improve the accuracy of LVLM responses.
- We conduct comprehensive experiments across various benchmarks and demonstrate that our proposed ONLY consistently outperforms existing approaches with minimal implementation effort and computational cost.

3.2 Related Work

Large Vision-Language Models (LVLMs). Recently, large-scale LLMs have demonstrated remarkable proficiency in handling human queries and exhibit robust linguistic capabilities [11, 51]. Leveraging these powerful models, researchers are exploring ways to align the visual modality with language, unlocking advanced visual recognition and reasoning capabilities across various multi-modal tasks [3, 36]. For example, LLaVA-1.5 [37] employs a pre-trained CLIP ViT-L/14 [42] as the vision encoder, and trains a linear mapping layer to connect the vision and language modalities. In contrast, InstructBLIP [15] builds on a pre-trained BLIP-2 [31] and incorporates an instruction-aware Q-Former module to bridge the modalities. Despite their exceptional multi-modal performance, these LVLMs still suffer from hallucinations, often generating text responses that do not accurately reflect the given image input [5, 6, 49, 65]. Such hallucinations pose significant challenges for deploying these models in real-world applications. In this work, we propose a novel, training-free algorithm designed to mitigate hallucinations, thereby enhancing the practical deployment of LVLMs in real-world scenarios.

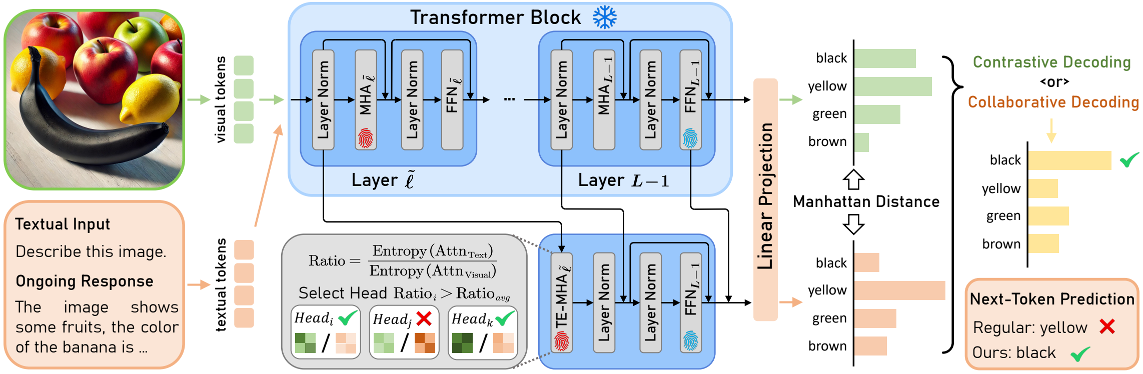
Hallucination in LVLMs. Recent studies have revealed that LVLMs may generate cross-modal inconsistencies between visual inputs and their corresponding responses, *i.e.*, hallucinations, which can lead to misinformation and performance degradation [29, 36]. To mitigate these hallucinations, early works have explored the use of additional robust instruction tuning on curated datasets [5, 24, 49]. While effective, these methods require extensive and costly training, making them impractical for individual users. More recently, researchers have explored an alternative approach by developing variant methods based on contrastive decoding strategies, which mitigate hallucinations and enhance coherence by contrasting logits from counterpart outputs [8, 17, 29]. However, we recognize that these methods require two or even multiple queries, which slows down LVLM response generation, making them less suitable for real-time applications. In response, we propose ONLY, a contrastive decoding-based approach that requires only a one-time query and a one-layer intervention during decoding, achieving competitive performance while effectively minimizing implementation efforts and computational costs.

3.3 Method

In this work, we present ONLY, a training-free algorithm that uses only one Transformer layer to improve the accuracy of LVLM responses, as illustrated in Figure 3.2.

3.3.1 Preliminaries

LVLM Decoding. Recent LVLMs effectively process both visual and linguistic data using three key components: vision encoders, connectors, and a Large Language



Textual Input

Describe this image.

Ongoing Response

The image shows some fruits, the color of the banana is ...

Figure 3.2: **Overview of our proposed ONLY**. Our method retains the core decoding process of LVLMs but incorporates a textual-enhanced multi-head attention layer with a residual connection to the last layer's output. This adjustment aims to produce an output with a greater focus on textual information. The resulting textual-enhanced logits are then adaptively decoded alongside the original output, employing either contrastive or collaborative decoding strategies to optimize performance.

Model (LLM). An LVLM, parameterized by θ, autoregressively generates a fluent textual response sequence $\mathbf{y}$ from an input image v and a textual query $\mathbf{x}$. Initially, v is processed by a vision encoder and transformed into visual tokens via a vision-language alignment module (*e.g.*, Q-Former [31] or linear projection [37]). These tokens, combined with the query tokens, are input to the LLM. The generation of each token y_t in the sequence $\mathbf{y}$ is modeled as:

$$y_t \sim p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) = \text{softmax}(f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}))_{y_t}, \tag{3.1}$$

where $y_t \in \mathcal{S}$ is the current token, $\mathbf{y}_{<t} = [y_0, \dots, y_{t-1}]$ are the previously generated tokens, and f_θ represents the logits over a vocabulary set $\mathcal{S}$.

Transformer Decoder. The language model is structured as a Transformer, where a sequence of tokens $\{x_1, x_2, \dots, x_{t-1}\}$ are initially embedded into a sequence of hidden states $\mathcal{H}_{t-1}^0 = \{h_1^0, \dots, h_{t-1}^0\}$. The Transformer comprises L layers, each layer incorporates a Multi-Head Attention (MHA) module and a Multi-Layer Perceptron (MLP). At time step t, the output of each layer $\mathcal{H}_t^{\ell+1}$ is derived from the hidden states input $\mathcal{H}_t^\ell$, employing two primary residual connections:

$$\bar{\mathcal{H}}_t^\ell = \text{MHA}_\ell(\mathcal{H}_t^\ell) + \mathcal{H}_t^\ell, \quad \mathcal{H}_t^{\ell+1} = \text{MLP}_\ell(\bar{\mathcal{H}}_t^\ell) + \bar{\mathcal{H}}_t^\ell. \tag{3.2}$$

Each MHA module consists of H attention heads that compute self-attention, where the attention score is derived from query, key, and value matrices. Specifically, for the i-th head in layer ℓ, the operation is given by:

$$\text{Head}_{\ell,i}(\mathcal{H}_t^\ell) = \text{Attention}(Q_{\ell,i}, K_{\ell,i}, V_{\ell,i}) = \text{softmax}\left(\frac{Q_{\ell,i} \cdot K_{\ell,i}^\top}{\sqrt{d_k}}\right) V_{\ell,i}, \tag{3.3}$$

where $Q_{\ell,i}/K_{\ell,i}/V_{\ell,i} = \mathcal{H}_t^\ell W_{\ell,i}^{Q/K/V}$, are query/key/value matrices obtained from learned weights. The outputs from all H heads are then concatenated and projected using an projection matrix W_ℓ^O:

$$\text{MHA}_\ell(\mathcal{H}_t^\ell) = \text{Concat}(\text{Head}_{\ell,1}(\mathcal{H}_t^\ell), \text{Head}_{\ell,2}(\mathcal{H}_t^\ell), \ldots, \text{Head}_{\ell,H}(\mathcal{H}_t^\ell))W_\ell^O. \tag{3.4}$$

At last, a projection head $\phi(\cdot)$ predicts the logits of the next token x_t over the vocabulary set $\mathcal{S}$:

$$f_\theta(y_t|y_{<t}) = \phi(\mathcal{H}_t^L), y_t \in \mathcal{S}. \tag{3.5}$$

Combining Eq. 3.5 with Eq. 3.1, we finally obtain:

$$p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) = \text{softmax}(\phi(\mathcal{H}_t^L))_{y_t}. \tag{3.6}$$

3.3.2 One-Layer Intervention for Textual Enhancement

Previous contrastive decoding methods focus primarily on the visual modality or the effect of visual input on the textual modality: *e.g.*, VCD [29] contrasts the outputs obtained with original vs. visual distorted input, and M3ID [17] amplifies the influence of the reference image over the language prior. However, the effect of textual modality has been less studied. In this work, we propose to address hallucination by directly producing textually-enhanced outputs with minimal additional computational overhead. Specifically, inspired by information theory [48], we introduce an attention-head selection strategy guided by the text-to-visual entropy ratio. As illustrated in Figure 3.3, we observe that when the distortion escalates (similar to the diffusion steps in VCD), textual entropy increases while visual entropy decreases. Guided by this observation, we propose to directly select attention heads with a higher text-to-visual

Algorithm 1 Predict Textual-Enhanced (TE) Logits

```
Require: Initial hidden states $\mathcal{H}_t^0$, total transformer layers $L$, total attention heads $H$,
    layer index for textual enhancement $\tilde{\ell}$.
procedure PREDICT_TE_LOGITS($A$)
    for $\ell \in \{0, 1, 2, \ldots, L-1\}$ do
        for $i \in \{0, 1, \ldots, H-1\}$ do
            Step 1: Calculate TE attention output
            if $\ell = \tilde{\ell}$ then
                $\tilde{\mathcal{H}}_t^{\tilde{\ell}} \leftarrow \text{TE-MHA}_{\tilde{\ell}}(\mathcal{H}_t^{\tilde{\ell}})$        ▷ Equation 3.13
            end if
        end for
        Step 2: Calculate Transformer output for each layer
        $\bar{\mathcal{H}}_t^{\ell} \leftarrow \text{MHA}_{\ell}(\mathcal{H}_t^{\ell}) + \mathcal{H}_t^{\ell}$        ▷ Equation 3.2
        $\mathcal{H}_t^{\ell+1} \leftarrow \text{MLP}_{\ell}(\bar{\mathcal{H}}_t^{\ell}) + \bar{\mathcal{H}}_t^{\ell}$        ▷ Equation 3.2
        if $\ell = L - 1$ then
            Step 3: Calculate TE Transformer output
            $\bar{\tilde{\mathcal{H}}}_t^{L-1} \leftarrow \tilde{\mathcal{H}}_t^{\tilde{\ell}} + \mathcal{H}_t^{L-1}$        ▷ Equation 3.15
            $\hat{\mathcal{H}}_t^{L} \leftarrow \text{MLP}_{L-1}(\bar{\tilde{\mathcal{H}}}_t^{L-1}) + \bar{\tilde{\mathcal{H}}}_t^{L-1}$        ▷ Equation 3.16
        end if
    end for
    Step 4: Calculate original logits and TE logits
    Logits $= f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})) \leftarrow \texttt{Linear}(\mathcal{H}_t^L)$
    Logits_TE $= \hat{f}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})) \leftarrow \texttt{Linear}(\hat{\mathcal{H}}_t^L + \mathcal{H}_t^L)$
    return Logits, Logits_TE
end procedure
```

entropy ratio to stimulate textually-enhanced outputs to avoid double queries while extracting language bias.

Attention Head Selection Using Text-to-Visual Entropy Ratio. Suppose a token is generated at time step t, and the initial hidden states input to the Transformer decoder for this token is $\mathcal{H}_t^0$. For layer ℓ, the input hidden states can be denoted as $\mathcal{H}_t^{\ell}$. We distinguish between text and visual attention within the attention matrix by computing the raw attention scores $a_{\ell,i}$ for each head:

$$a_{\ell,i} = \text{softmax}(Q_{\ell,i} \cdot K_{\ell,i}^{\top} / \sqrt{d_k}), \tag{3.7}$$

where $Q_{\ell,i}$ and $K_{\ell,i}$ are the query and key matrices for head i in layer ℓ. To isolate

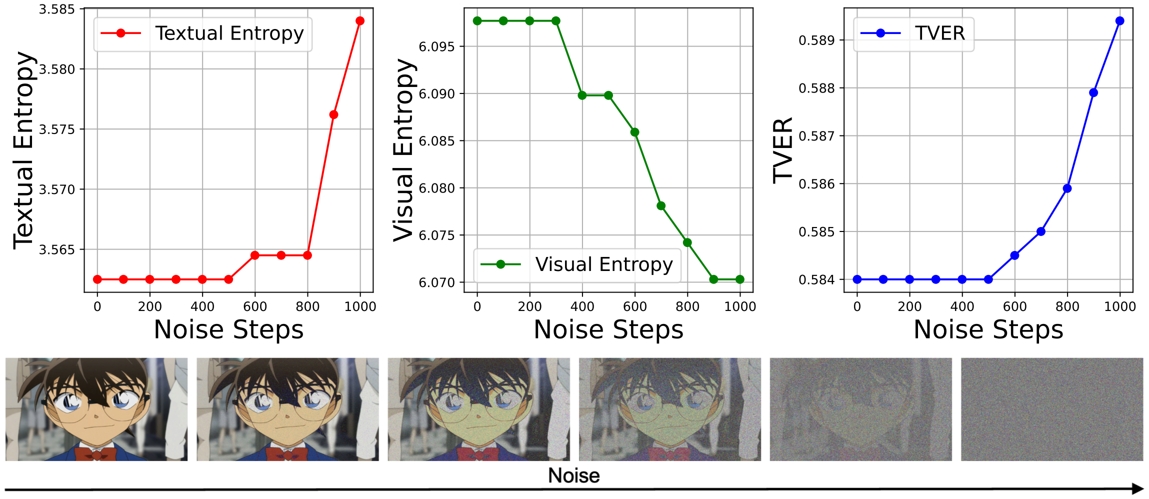


Figure 3.3: **Impact of applying diffusion noise on textual and visual attention entropy**. We perform an analysis on all COCO samples from the POPE benchmark and observe that as distortion increases, textual entropy rises whereas visual entropy decreases.

text and visual attentions, we utilize indices corresponding to textual or visual tokens:

$$a^{\mathcal{T}}_{\ell,i} = \{a_{\ell,i,j} \mid j \in \text{indices}_{\mathcal{T}}\}, a^{\mathcal{V}}_{\ell,i} = \{a_{\ell,i,j} \mid j \in \text{indices}_{\mathcal{V}}\}, \tag{3.8}$$

where $\text{indices}_{\mathcal{T}}$ and $\text{indices}_{\mathcal{V}}$ specify positions of textual and visual tokens, respectively. The entropy for these attention sets is computed as follows:

$$\text{Entropy}(a^{\mathcal{T}}_{\ell,i}) = -\sum_k p^{\mathcal{T}}_{\ell,i,k} \log p^{\mathcal{T}}_{\ell,i,k}, \quad \text{Entropy}(a^{\mathcal{V}}_{\ell,i}) = -\sum_k p^{\mathcal{V}}_{\ell,i,k} \log p^{\mathcal{V}}_{\ell,i,k}, \tag{3.9}$$

where $p^{\mathcal{T}}_{\ell,i,k}$ and $p^{\mathcal{V}}_{\ell,i,k}$ represent the normalized attention probabilities, computed from the softmax of each subset:

$$p^{\mathcal{T}}_{\ell,i,k} = \text{softmax}(a^{\mathcal{T}}_{\ell,i,k}), p^{\mathcal{V}}_{\ell,i,k} = \text{softmax}(a^{\mathcal{V}}_{\ell,i,k}). \tag{3.10}$$

The Text-to-Visual Entropy Ratio (TVER) for each attention head is calculated as:

$$\text{TVER}_{\ell,i} = \frac{\text{Entropy}(a^{\mathcal{T}}_{\ell,i})}{\text{Entropy}(a^{\mathcal{V}}_{\ell,i})}. \tag{3.11}$$

To optimize the attention output for enhanced textual relevance while reducing visual information, we selectively deactivate heads with a TVER below the average for that layer, setting their attention weights to zero. This approach prioritizes heads with relatively higher text-to-visual entropy ratios, providing a clue where uncertainty in the textual modality is higher:

$$\tilde{a}_{\ell,i} = \begin{cases} a_{\ell,i}, & \text{if } \mathrm{TVER}_{\ell,i} \geq \text{average}(\mathrm{TVER}_{\ell}), \\ 0, & \text{otherwise.} \end{cases} \tag{3.12}$$

With this, we obtain the output of the Textual-Enhanced Multi-Head Attention (TE-MHA) module:

$$\begin{aligned} &\text{TE-MHA}_{\ell}(\mathcal{H}_t^{\ell}) = \\ &\mathrm{Concat}(\tilde{a}_{\ell,1}V_{l,1}, \tilde{a}_{\ell,2}V_{l,2}, \ldots, \tilde{a}_{\ell,H}V_{l,H})W_{\ell}^{O}. \end{aligned} \tag{3.13}$$

3.3.3 Adaptive Decoding

In this section, we utilize the logits obtained from textual-enhanced attention outputs for adaptive decoding.

Suppose layer $\tilde{\ell} \in \{0, 1, \cdots, L-1\}$ is the selected layer for textual enhancement, where we calculate a textual-enhanced attention output as discussed in Eq. 3.13. To ensure that the output logits do not deviate excessively from the original LVLM outputs, we implement two residual connections. These connections are defined as

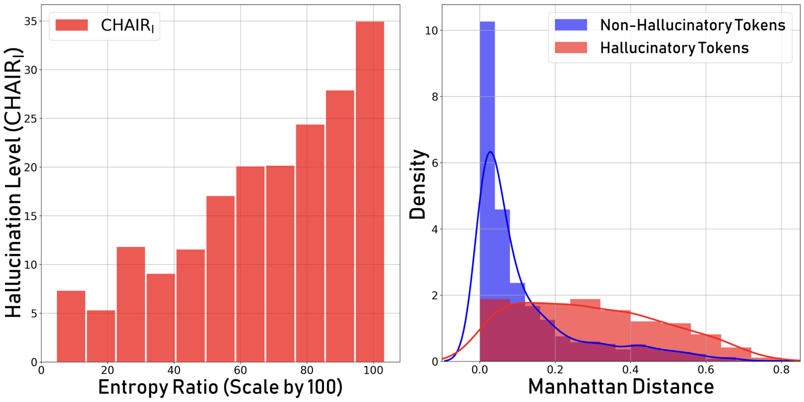


Figure 3.4: **Text-to-visual entropy ratio is correlated with hallucinations.** (*Left*) Density plot of token-wise average textual-to-visual entropy ratio and bar plot of average CHAIR$_I$ in each bin on the CHAIR benchmark; (*Right*) Density plots of token-level Manhattan distance between original and textual-enhanced logits for both hallucinatory and non-hallucinatory tokens on POPE.

follows:

$$
\begin{aligned}
\tilde{\mathcal{H}}_t^{\tilde{\ell}} &= \text{TE-MHA}_{\tilde{\ell}}(\mathcal{H}_t^{\tilde{\ell}}), && (3.14)\\
\bar{\tilde{\mathcal{H}}}_t^{L-1} &= \tilde{\mathcal{H}}_t^{\tilde{\ell}} + \mathcal{H}_t^{L-1}, && (3.15)\\
\hat{\mathcal{H}}_t^{L} &= \text{MLP}_{L-1}(\bar{\tilde{\mathcal{H}}}_t^{L-1}) + \bar{\tilde{\mathcal{H}}}_t^{L-1}. && (3.16)
\end{aligned}
$$

Finally, the textual-enhanced predicted probability can be obtained by:

$$
\begin{aligned}
\tilde{p}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) &= \text{softmax}(\tilde{f}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}))_{y_t}\\
&= \text{softmax}(\phi(\hat{\mathcal{H}}_t^{L}))_{y_t}. \qquad (3.17)
\end{aligned}
$$

To adaptively contrast the original and textual-enhanced logits, we measure the Manhattan distance between the two probability distributions at each timestep t:

$$
d_t = \sum_{y_t \in \mathcal{S}} |p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) - \tilde{p}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t})|, \qquad (3.18)
$$

where d_t provides a measure of the difference between the distributions. Based on this distance, we adjust the original logits either collaboratively or contrastively:

$$
y_t \sim p_\theta(y_t) = \text{softmax}\left(f_\theta^{\text{final}}\right) \qquad (3.19)
$$

$$
f_\theta^{\text{final}} = \begin{cases} f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) + \alpha_1\, \tilde{f}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}), & \text{if } d_t < \gamma \text{ (collaborative);} \\ (1+\alpha_2)\, f_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) - \alpha_2\, \tilde{f}_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}), & \text{if } d_t \geq \gamma \text{ (contrastive),} \end{cases} \qquad (3.20)
$$

where γ is a predefined threshold that determines the decoding strategy based on the measured distance.

Effectiveness of text-to-visual entropy ratio for textual information enhancement. We further conduct an empirical study to validate the effectiveness of applying text-to-visual entropy ratio for language bias reflection, as shown in Figure 3.4. The experimental results demonstrate that *the entropy ratio is strongly correlated to the hallucination level* at both the response and token levels.

3.4 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.

3.4.1 Experimental Settings

Evaluated LVLMs. We evaluate the effectiveness of our method on three state-of-the-art open-source LVLMs: LLaVA-1.5 [38], InstructBLIP [15] and Qwen-VL [2].

Benchmarks. We conduct extensive experiments on six benchmarks: (1) **POPE [33]** is a benchmark commonly used to assess object hallucinations in LVLMs, which evaluates model accuracy through yes-or-no questions about the presence of specific objects in images; (2) **CHAIR [43]** evaluates object hallucinations through image captioning, where the LVLMs are prompted to describe 500 randomly selected images from the MSCOCO validation set; (3) **MME-Hallucination [18]** 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; (4) **MMBench [39]** is a benchmark for evaluating LVLMs' multi-modal understanding ability across 20 dimensions; (5) **MMVP [50]** comprises 150 CLIP-blind image pairs, each paired with a binary-option question to evaluate the fine-grained visual recognition capabilities of LVLMs; (6) **MM-Vet [61]** utilizes LLM-based evaluator to evaluate LVLMs on 6 capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math; (7) **LLaVA-Bench** provides 24 images in complex scenes, memes, and sketches, along with 60 challenging questions.

Baselines. We compare the performance of our ONLY approach with the following state-of-the-art approaches: VCD [29], M3ID [17], Woodpecker [59], HALC [8], DoLa [13] and OPERA [23]. We apply sampling-based decoding in default, where the next token is sampled directly from the post-softmax probability distribution.

Implementation Details. We follow the default query format for all LVLMs. Besides, we set $\alpha_1 = 3$, $\alpha_2 = 1$, and $\gamma = 0.2$ for LLaVA-1.5 [38], and $\gamma = 0.4$ for InstructBLIP [15] / Qwen-VL [2]. Following VCD [29], we implement adaptive plausibility constraints [32] with $\beta = 0.1$ across all tasks. All experiments are performed on a single 48GB NVIDIA RTX 6000 Ada GPU.

3.4.2 Results and Discussions

Results on POPE. In Table 3.1, we compare our method's performance against various baselines on the POPE benchmark. As shown in the table, our approach consistently outperforms previous state-of-the-art methods across various LVLM models and settings, demonstrating its robustness across different evaluation scenarios. Specifically, in the MS-COCO (Random) setting with the LLaVA-1.5 backbone, our method surpasses VCD by 2.20% and M3ID by 1.70% in accuracy. Even in the more challenging adversarial setting, our approach maintains its superior performance, outperforming VCD by 2.23% and M3ID by 1.17%. Overall, these consistent gains across different datasets and LVLM models highlight the effectiveness of our method as a strong and generalizable solution for mitigating hallucinations in LVLMs.

Results on CHAIR. On the open-ended CHAIR benchmark, our ONLY method achieves superior performance with lower hallucination rates. Table 3.2 presents a comparison against four state-of-the-art approaches, evaluating hallucination rates with CHAIR_S and CHAIR_I under maximum token generation limits of 64 and 128 across three LVLM backbones. Notably, in the LLaVA-1.5 (Max Token = 128) setting, our approach reduces CHAIR_S by 5.2 points and CHAIR_I by 2.0 points compared to regular decoding.

Results on MME. In Table 3.3, we compare our approach against other methods on the MME benchmark. The results show that our method consistently outperforms all baselines, achieving the highest scores across both object-level (Existence, Count) and attribute-level (Position, Color) evaluations. Notably, our method attains an MME score of 634.67, outperforming the second-best method, M3ID, by 36.34 points, demonstrating its superior capability in mitigating various types of hallucinations.

Results on MMVP. To evaluate the effectiveness of our approach on fine-grained visual recognition tasks, we conduct experiments on the MMVP benchmark and present the results in Figure 3.5. With our ONLY approach, the LVLM is able to handle more nuanced visual recognition tasks, improving the performance from 22.67% to 28.00%.

Results on MMBench and MMVet. We also report the performance of all compared methods on the MMBench and MMVet benchmarks in Table 3.4. Our approach continues to outperform existing state-of-the-art methods, demonstrating that it also enhances the general multi-modal understanding capabilities of LVLMs.

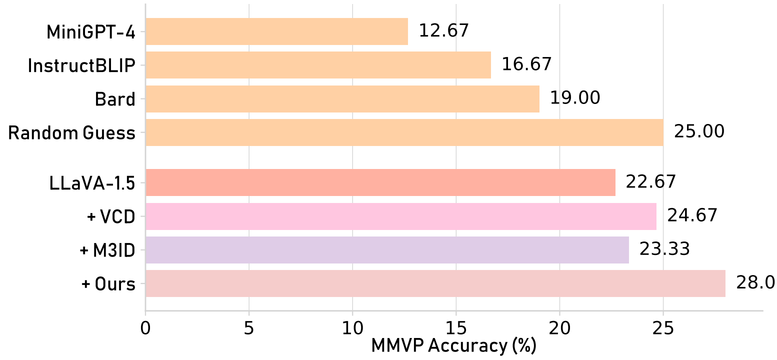


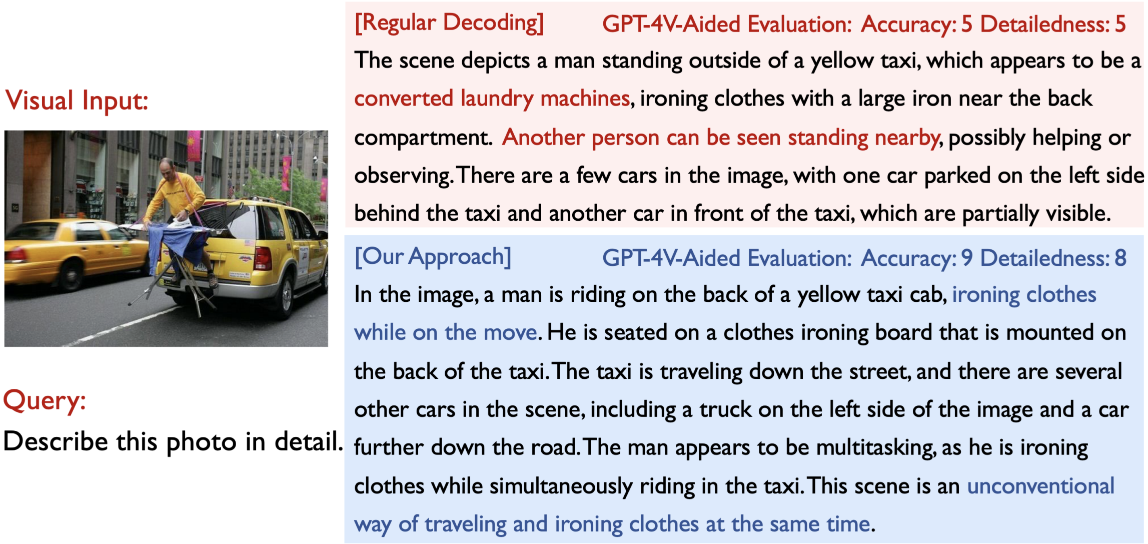
Figure 3.5: **Results on MMVP [50].** We apply our approach to LLaVA-1.5 [38] and compare its performance against other hallucination mitigation methods.

Results on LLaVA-Bench. In Figure 3.6, 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. Specifically, regular decoding often leads to hallucinated or inaccurate content, such as describing "`taxi appears to be converted laundry machines`" and "`another person can be seen standing nearby`". In contrast, our response is more detailed, focusing on the fact that "`a person is ironing clothes while on the move, which is an unconventional way`".

3.4.3 Efficiency Comparison

In Table 3.4, we evaluate the efficiency of our approach using the LLaVA-1.5 model on the CHAIR benchmark, with a maximum token length of 128. We also report the performance of all compared methods across 5 benchmarks. Our approach demonstrates consistently superior performance, with only a 1.07× increase in time consumption and negligible additional GPU memory usage. These results validate that our approach is both efficient and effective, offering a favorable performance-cost trade-off.

3.4.4 Ablation Study



Visual Input:

Query:

Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 5 Detailedness: 5

The scene depicts a man standing outside of a yellow taxi, which appears to be a converted laundry machines, ironing clothes with a large iron near the back compartment. Another person can be seen standing nearby, possibly helping or observing. There are a few cars in the image, with one car parked on the left side behind the taxi and another car in front of the taxi, which are partially visible.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 9 Detailedness: 8

In the image, a man is riding on the back of a yellow taxi cab, ironing clothes while on the move. He is seated on a clothes ironing board that is mounted on the back of the taxi. The taxi is traveling down the street, and there are several other cars in the scene, including a truck on the left side of the image and a car further down the road. The man appears to be multitasking, as he is ironing clothes while simultaneously riding in the taxi. This scene is an unconventional way of traveling and ironing clothes at the same time.

Figure 3.6: **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 blue.

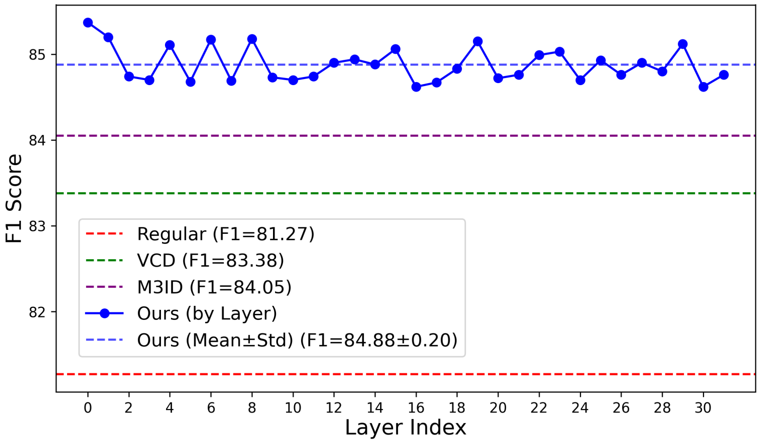


Figure 3.7: **Impacts of different selected layers**. We present the results obtained by selecting different layers for textual enhancement on the POPE benchmark using all 9,000 samples from COCO.

Selection of Layer for Textual Enhancement. To investigate the impacts of choosing different layers for textual enhancement, we conduct ablation experiments on the POPE benchmark. Results in Figure 3.7 demonstrate that by selecting the initial layer for textual enhancement, our ONLY method achieves optimal performance on the POPE benchmark. Additionally, we observe that the performance of our approach is robust across different layers chosen for intervention, with ONLY exhibiting minimal variation and consistently outperforming VCD and M3ID. This robustness is due to our attention-head selection strategy, which dynamically selects different sets of heads across multiple layers, efficiently and effectively capturing language bias.

Other Strategies for Textual Enhancement. In Table 3.5, we compare the performance achieved by various textual enhancement strategies. Our approach of attention head selection using TVER achieves the best performance. In contrast, directly modifying attention weights—such as zeroing out or adding noise to visual attention weights, or doubling textual attention weights—results in suboptimal outcomes. Additionally, selecting attention heads based on the ratio of the sum of attention weights also leads to a performance decrease of 0.71% on POPE and 3.1% on CHAIR.

	Setup	Method	LLaVA-1.5				InstructBLIP				Qwen-VL			
			Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑
MS-COCO	Random	Regular	83.13	81.94	85.00	83.44	83.07	83.02	83.13	83.08	85.23	97.23	72.53	83.09
		VCD	87.00	86.13	88.20	87.15	86.23	88.14	83.73	85.88	87.03	97.36	76.13	85.45
		M3ID	87.50	87.38	87.67	87.52	86.67	88.09	84.80	86.41	86.40	98.23	74.13	84.50
		Ours	**89.70**	**89.95**	**88.27**	**89.10**	**89.23**	**91.83**	**86.13**	**88.89**	**88.90**	**98.52**	**79.27**	**87.85**
	Popular	Regular	81.17	78.28	86.27	82.08	77.00	73.82	83.67	78.44	84.53	94.50	73.33	82.58
		VCD	83.10	79.96	88.33	83.94	80.07	77.67	84.40	80.89	85.87	94.98	75.73	84.27
		M3ID	84.30	81.58	88.60	84.95	80.97	77.93	86.40	81.85	86.07	96.56	74.80	84.30
		Ours	**86.00**	**84.44**	88.27	**86.31**	**83.27**	**81.46**	86.13	**83.73**	**87.47**	95.63	**79.48**	**86.81**
	Adversarial	Regular	77.43	73.31	86.27	79.26	74.60	71.26	82.47	76.45	83.37	91.47	73.60	81.57
		VCD	77.17	72.18	88.40	79.47	77.20	74.29	83.20	78.49	83.73	89.84	76.07	82.38
		M3ID	78.23	73.51	88.27	80.22	77.47	73.68	85.47	79.14	83.37	91.19	73.87	81.62
		Ours	**79.40**	**75.00**	88.20	**81.07**	**80.10**	**76.89**	**86.07**	**81.22**	**83.80**	**92.33**	**76.14**	**83.46**
A-OKVQA	Random	Regular	81.90	76.63	91.80	83.53	80.63	76.82	87.73	81.92	86.40	94.32	77.47	85.07
		VCD	83.83	78.05	94.13	85.34	84.20	80.90	89.53	85.00	87.93	94.59	80.47	86.96
		M3ID	84.67	79.25	93.93	85.97	85.43	81.77	91.20	86.23	87.50	95.33	78.87	86.32
		Ours	**86.07**	**80.91**	**94.40**	**87.14**	**88.57**	**86.13**	**91.93**	**88.94**	**89.47**	**95.34**	**83.84**	**89.22**
	Popular	Regular	75.07	68.58	92.53	78.77	75.17	70.15	87.60	77.91	85.77	92.82	77.53	84.49
		VCD	76.63	69.59	94.60	80.19	78.63	73.53	89.47	80.72	87.33	93.68	80.07	86.34
		M3ID	77.80	70.98	94.07	80.91	78.80	73.38	90.40	81.00	87.37	95.31	78.60	86.15
		Ours	**79.00**	**72.17**	94.40	**81.80**	**80.83**	**75.23**	**91.93**	**82.75**	**89.47**	94.77	**84.43**	**89.30**
	Adversarial	Regular	67.23	61.56	91.80	73.70	69.87	64.54	88.20	74.54	80.37	82.56	77.00	79.68
		VCD	67.40	61.39	93.80	74.21	71.00	65.41	89.13	75.45	81.90	83.07	80.13	81.57
		M3ID	68.60	62.22	94.73	75.11	70.10	64.28	90.47	75.16	81.90	84.25	78.47	81.26
		Ours	**68.70**	**62.35**	94.40	**75.70**	**72.47**	**66.19**	**91.87**	**76.94**	**82.07**	**85.02**	**81.09**	**83.01**
GQA	Random	Regular	82.23	76.32	93.47	84.03	79.67	76.05	86.60	80.99	85.10	91.42	77.47	83.87
		VCD	83.23	76.73	95.40	85.05	82.83	80.16	87.27	83.56	87.00	92.11	80.93	86.16
		M3ID	84.20	78.00	95.27	85.77	83.07	80.06	88.07	83.87	87.07	92.64	80.53	86.16
		Ours	**86.70**	**80.94**	**96.00**	**87.83**	**86.17**	**83.84**	**89.60**	**86.63**	**88.03**	93.59	**82.68**	**87.80**
	Popular	Regular	73.47	66.83	93.20	77.84	73.33	68.72	85.67	76.26	80.87	82.65	78.13	80.33
		VCD	72.37	65.27	95.60	77.58	76.13	71.10	88.07	78.68	82.53	83.52	81.07	82.27
		M3ID	73.87	66.70	95.33	78.49	75.17	69.94	88.27	78.04	82.68	83.74	80.85	82.27
		Ours	**74.03**	66.70	**96.00**	**78.71**	**77.20**	**71.79**	**89.60**	**79.72**	**82.87**	**83.88**	**82.55**	**83.21**
	Adversarial	Regular	68.60	62.43	93.40	74.84	68.60	63.94	85.33	73.10	78.77	79.33	77.80	78.56
		VCD	68.83	62.26	95.67	75.43	71.00	65.75	87.67	75.14	81.17	81.48	80.67	81.07
		M3ID	68.67	62.16	95.40	75.28	71.17	65.79	88.20	75.36	81.90	83.07	80.13	81.57
		Ours	**69.23**	**62.55**	**95.87**	**75.70**	**71.93**	**65.98**	**87.93**	**75.84**	81.33	82.38	**81.50**	**81.94**

Table 3.1: **Results on POPE [33] benchmark.** Higher (↑) accuracy, precision, recall, and F1 indicate better performance. The best results are bolded, and the second-best are underlined.

Method	**LLaVA-1.5**				**InstructBLIP**				**Qwen-VL**			
	Max Token 64		Max Token 128		Max Token 64		Max Token 128		Max Token 64		Max Token 128	
	CHAIR_S ↓	CHAIR_I ↓	CHAIR_S ↓	CHAIR_I ↓	CHAIR_S ↓	CHAIR_I ↓	CHAIR_S ↓	CHAIR_I ↓	CHAIR_S ↓	CHAIR_I ↓	CHAIR_S ↓	CHAIR_I ↓
Regular	26.2	9.4	55.0	16.3	31.2	11.1	57.0	17.6	33.6	12.9	52.0	16.5
VCD	24.4	7.9	54.4	16.6	30.0	10.1	60.4	17.8	33.0	12.8	50.2	16.8
M3ID	21.4	6.3	56.6	15.7	30.8	10.4	62.2	18.1	32.2	11.5	49.5	17.2
Woodpecker	24.9	7.5	57.6	16.7	31.2	10.8	60.8	17.6	31.1	12.3	51.8	16.3
HALC	21.7	7.1	51.0	14.8	24.5	**8.0**	53.8	15.7	28.2	9.1	49.6	15.4
Ours	**20.0**	**6.2**	**49.8**	**14.3**	**23.5**	8.2	**52.2**	**15.5**	**27.3**	**8.4**	**48.0**	**14.3**

Table 3.2: **Results on CHAIR [43] benchmark.** We limit the maximum number of new tokens to 64 or 128. Lower (↓) CHAIR_S, CHAIR_I indicate better performance. The best results in each setting are **bolded**, and the second-best are underlined.

Method	**Object-level**		**Attribute-level**		**MME Score ↑**
	Existence ↑	Count ↑	Position ↑	Color ↑	
Regular	173.75	121.67	117.92	149.17	562.50
DoLa	176.67	113.33	90.55	141.67	522.22
OPERA	183.33	137.22	122.78	155.00	598.33
VCD	186.67	125.56	128.89	139.45	580.56
M3ID	186.67	128.33	131.67	151.67	598.11
Woodpecker	187.50	125.00	126.66	149.17	588.33
HALC	183.33	133.33	107.92	155.00	579.58
Ours	**191.67**	**145.55**	**136.66**	**161.66**	**635.55**

Table 3.3: **Results on MME-Hallucination [18] with LLaVA-1.5 [38].** We report the average MME scores for each subset. Higher scores (↑) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Method	Avg. Latency ↓	GPU Memory ↓	CHAIR_S ↓	MME ↑	POPE ↑	MMBench ↑	MM-Vet ↑
Regular	3.47 s (×1.00)	14945 MB (×1.00)	55.0	562.5	83.44	64.1	26.1
VCD	6.97 s (×2.01)	15749 MB (×1.05)	54.4	580.6	87.15	64.6	30.9
M3ID	7.05 s (×2.03)	15575 MB (×1.04)	54.4	598.1	87.52	64.4	29.9
OPERA	24.70 s (×7.12)	22706 MB (×1.52)	52.6	598.3	88.85	64.4	32.0
Woodpecker	10.68 s (×3.08)	22199 MB (×1.49)	57.6	588.3	86.45	64.0	30.6
HALC	22.61 s (×6.52)	23084 MB (×1.54)	51.0	579.6	87.68	64.2	30.8
Ours	3.70 s (×1.07)	14951 MB (×1.00)	**49.8**	**635.55**	**89.10**	**65.0**	**32.8**

Table 3.4: **Efficiency comparison**. For each method, we present the average inference latency per instance and peak GPU memory. Experiments are conducted on a single RTX A6000 Ada GPU.

Strategy	**POPE ↑**				**CHAIR ↓**	
	Acc.	Prec.	Rec.	F1	CHAIR_S	CHAIR_I
Regular	80.42	78.20	84.59	81.27	26.2	9.4
$a^{\mathcal{V}}_{\ell,i} \leftarrow 0$	84.26	82.13	87.69	84.82	21.2	6.9
$a^{\mathcal{V}}_{\ell,i} \leftarrow a^{\mathcal{V}}_{\ell,i} + \varepsilon$	83.95	81.67	88.16	84.79	22.1	7.6
$a^{\mathcal{T}}_{\ell,i} \leftarrow a^{\mathcal{T}}_{\ell,i} * 2$	84.37	82.52	87.55	84.96	21.6	6.8
Ratio $\leftarrow \sum a_{\mathcal{T}} / \sum a_{\mathcal{V}}$	84.20	81.57	87.56	84.46	23.1	8.2
Ours	**84.91**	**82.84**	**88.07**	**85.37**	**20.0**	**6.2**

Table 3.5: **Different Strategies for textual enhancement**. We conduct experiments with different textual enhancement strategies.

Chapter 4

Conclusions

Hallucination remains a persistent and critical challenge in the deployment of Large Vision-Language Models (LVLMs), particularly in applications that demand high reliability and fidelity to visual inputs. In this thesis, we introduced two novel, training-free approaches—Self-Correcting Decoding with Generative Feedback (DeGF) and One-Layer Intervention (ONLY)—to address this problem from different angles. DeGF enhances output fidelity by incorporating generative feedback from a text-to-image model, enabling refined decoding based on visual alignment, albeit with higher computational demands. In contrast, ONLY offers a lightweight and efficient mechanism that selectively amplifies textually grounded signals during decoding, providing a practical solution with minimal inference overhead.

Both methods demonstrate strong empirical performance across standard hallucination benchmarks, with ONLY achieving comparable or superior results to existing approaches at a fraction of the computational cost. Together, these contributions provide practical, effective tools for mitigating hallucinations in LVLMs and underscore the potential of decoding-time interventions in enhancing model reliability. Future work may explore integrating these techniques into broader multimodal systems or extending them to more complex real-world tasks requiring grounded, trustworthy generation.

Bibliography

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2.2

[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. 2.1, 2.2, 2.4, 3.1, 3.4.1, 3.4.1

[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. 2.1, 2.2, 3.1, 3.2

[4] 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. 2.2

[5] Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 2024. 2.1, 3.2

[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. 2.1, 3.1, 3.2

[7] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3.1

[8] 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, pages 7824–7846. PMLR, 2024. 2.2, 2.4, 3.2, 3.4.1

[9] 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. 2.1, 2.2, 3.1

[10] 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. 2.3.2

[11] 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/`. 2.2, 2.4, 3.2

[12] 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(1):1–113, 2023. ISSN 1532-4435. 2.2

[13] 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`. 2.1, 2.4, 3.1, 3.4.1

[14] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2.2

[15] 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. 2.1, 2.2, 2.4, 2.4, 3.1, 3.2, 3.4.1, 3.4.1

[16] 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. 2.1, 2.2

[17] 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. 2.1, 2.2, 2.4, 3.1, 3.2, 3.3.2, 3.4.1

[18] 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. (document), 2.1, 2.4, 2.4, 2.3, 3.1, 3.4.1, 3.3

[19] 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. 2.2

[20] 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. 2.2

[21] 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. 2.1, 2.2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2.2

[23] 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. 2.1, 2.4, 3.4.1

[24] 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. 2.1, 2.2, 3.2

[25] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*, 2024. 2.2

[26] 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. 2.2

[27] 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. 2.2

[28] Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 2024. 2.1

[29] 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. 2.1, 2.2, 2.3.3, 2.4, 3.1, 3.2, 3.3.2, 3.4.1

[30] 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. 2.2

[31] 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. 2.1, 2.3.1, 3.1, 3.2, 3.3.1

[32] 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. 2.1, 2.4, 3.1, 3.4.1

[33] 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. (document), 2.1, 2.1, 2.2, 2.3.2, 2.4, 2.4.1, 2.1, 3.1, 3.4.1, 3.1

[34] 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. 2.3.2, 3.1

[35] 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`. 2.2

[36] 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. 2.1, 2.2, 3.1, 3.2

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916,

2023. (document), 2.1, 2.1, 2.2, 2.3.1, 3.2, 3.3.1

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. (document), 2.4, 2.4, 2.4, 3.1, 3.4.1, 3.4.1, 3.5, 3.3

[39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2024. (document), 2.1, 2.4, 2.3, 3.1, 3.4.1

[40] 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. 2.2

[41] 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`. (document), 2.1, 2.1, 2.2, 2.4.2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3.2

[43] 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. (document), 2.1, 2.3.2, 2.4, 2.4, 2.2, 3.1, 3.4.1, 3.2

[44] 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. 2.1, 2.2

[45] 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. 2.2

[46] 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. 2.2

[47] 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. 2.2

[48] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 3.3.2

[49] 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. *arXiv preprint arXiv:2309.14525*, 2023. 2.1, 2.2, 3.1, 3.2

[50] 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. (document), 2.1, 2.1, 2.4, 2.4, 3.1, 3.4.1, 3.5

[51] 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. 2.2, 2.4, 3.2

[52] 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 ACL 2024*, pages 15840–15853, 2024. 2.1, 3.1

[53] 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. 2.2

[54] 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. 2.2, 2.3.3, 2.4

[55] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in lvlms. In *International Conference on Machine Learning*, pages 53553–53570. PMLR, 2024. 2.1

[56] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang,

Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024. 3.1

[57] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 2.2

[58] 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. 2.1, 2.2, 3.1

[59] 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. 2.1, 2.2, 2.4, 3.4.1

[60] 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. ISSN 2835-8856. URL `https://openreview.net/forum?id=AFDcYJKhND`. 2.2

[61] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=KOTutrSR2y`. 3.1, 3.4.1

[62] 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. 2.2

[63] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, pages 1–19, 2024. 3.1

[64] 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*, 45(12):15098–15119, 2023. 2.2

[65] Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva

Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tTBXePRKSx. 3.2

[66] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024. 2.1, 3.1

[67] Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. Fact-and-reflection (FaR) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8702–8718, 2024. 2.1

[68] 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. 2.2

[69] 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. 2.2