Mitigating Hallucination Caused by Excessive Reliance on LLM within MLLM instead of Images

Anonymous ACL submission

Abstract

In the domain of multimodal generation and comprehension, multimodal large language models (MLLMs), which integrate visual encoders with large language models, have garnered significant success. However, solely relying on modal connection layers/modules to unify these models can lead to a neglect of image information, resulting in visual hallucinations. This manifests as generated text that is independent of the image content, such as descriptions of objects not present within the image. To mitigate this issue, we introduce a fine-tuning approach: Adversarial Contrast Dual Fine-tuning (ACD for short). This approach leverages the MLLM itself and employs the Fast Gradient Sign Method (FGSM) to generate adversarial image samples. During fine-tuning, both the original and adversarial images are utilized to perform dual contrastive fine-tuning on the MLLM. The experimental results show that our method significantly reduces hallucinations without any external annotations.

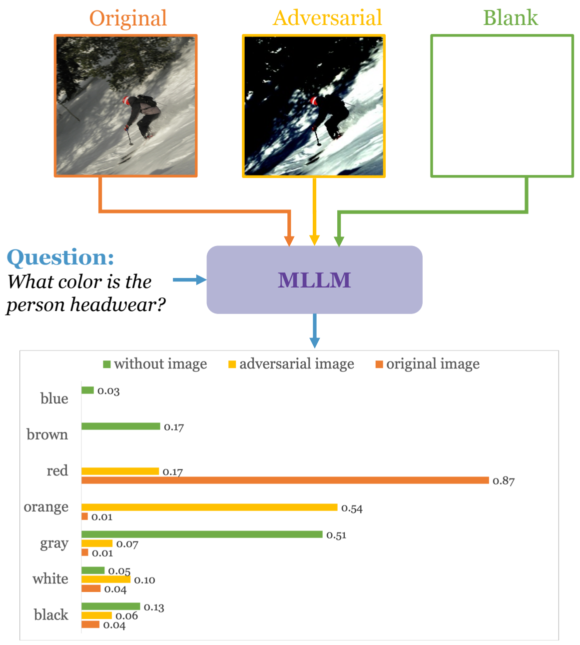

Figure 1: The impact of different visual inputs on the distribution of token logits values in the model's output. The ground truth of "What color is the person headwear?" is "red".

1 Introduction

In the realm of Natural Language Processing, Large Language Models (LLMs) have emerged as frontrunners (OpenAI, 2023a,b; Touvron et al., 2023), excelling across a range of tasks encompassing language understanding (Hendrycks et al., 2020), generation (Zhang et al., 2024), and reasoning (Ji et al., 2023; Yu et al., 2023a; Qiao et al., 2022). As a notable advancement of LLMs, the multimodal large language model (MLLM) combines LLMs with visual cues to demonstrate excellent performance in tasks related to multimodal understanding, reasoning, and interaction (Yang et al., 2023; Lu et al., 2023). However, MLLMs sometimes generate hallucinations during the reasoning process, e.g., the generated content does not exist in the image or cannot accurately describe the image. This phenomenon severely impacts the reliability and security of MLLMs.

In LLMs, the use of a pre-training mechanism causes the model to overly rely on prior knowledge obtained from the pre-training data, leading to hallucinations. Similar challenges exist in multimodal language models (MLLMs), such as overreliance on statistical bias (Gong et al., 2023; Goyal et al., 2017) and unimodal priors (Niu et al., 2021; Gupta et al., 2022). To mitigate hallucination, one of the direct methods is to use a stronger LLM (e.g., GPT-4) as an auxiliary model and then directly correct the inference content (Huang et al., 2023; Yin et al., 2023). Another approach is to mitigate hallucination during model decoding (Leng et al., 2023; Zhu et al., 2024). Due to the current MLLM being a combination of a visual pre-trained

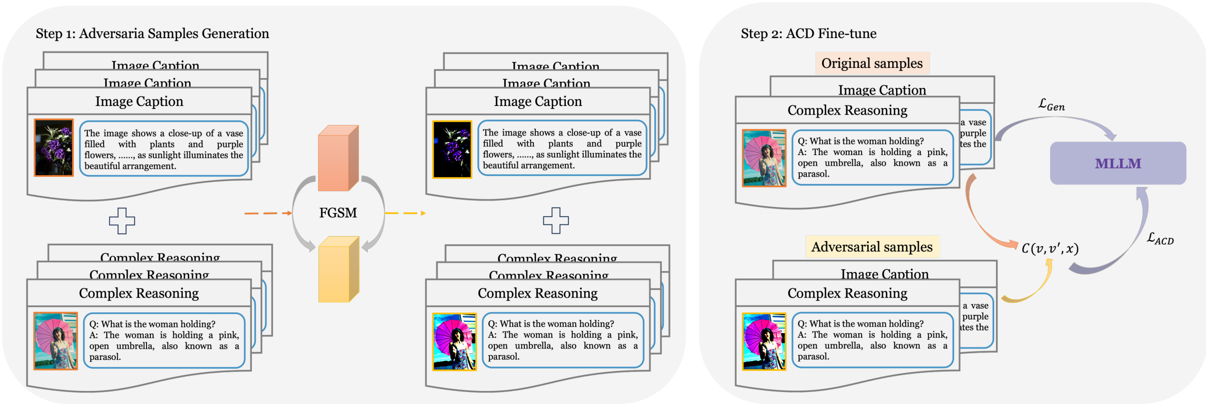

Figure 2: The illustration for ACD fine-tuning. Divided into two steps:(1) Visual Adversarial Samples Generation: Based on the original dataset(left), use FGSM to generate adversarial visual samples(right), the complete dataset including two types of visual information (original and adversarial); (2) ACD fine-tuning: Use original data to perform the first update on the model $\mathcal{L}_{Gen}$, and use new dataset to complete the ACD update $\mathcal{L}_{ACD}$.

model and a pre-trained language model, recent research has attempted to enhance modality alignment consistency and reduce hallucinations. Preference fine-tuning techniques are the most common method, such as direct preference optimization(DPO) (Zhao et al., 2023), or human feedback reinforcement learning (RLHF) (Sun et al., 2023). Zhou et al. (2024) proposed the Preference Optimization in MLLM with AI-Generated Dispreferences (POVID) framework based on DOP, which aims to exclusively generate dispreferred feedback data using AI models; Yu et al. (2023b) proposed RLHF-V, which enhances MLLM trustworthiness via behavior alignment from fine-grained correctional human feedback. However, they often focus on the hallucinations caused by modal component alignment, without considering the impact of language models in the backbone network, and rely on human feedback or external annotations when generating fine-tuning datasets.

In this paper, we delve into how language models in MLLM influence the generation of visual hallucinations. As shown in Figure 1, for a given scene-related question (i.e., "What color is the human headwear?"), by inputting the original image, adversarial image, and blank image into the model, respectively, it can be observed that the distribution of logits varies across different tokens. Specifically, due to the influence of prior knowledge of LLM and different input visual information, the tokens corresponding to the maximum logits may vary. This suggests that incomplete or incorrect image information in MLLMs acts like perturbed images, indicating that the source of the hallucination still comes from the influence of the LLM's prior knowledge.

Inspired by this, we propose a novel method called Adversarial Contrastive Dual Fine-Tuning (ACD). Based on the mechanism of hallucination generation in MLLMs, ACD fine-tuning consists of two primary steps: first, using FGSM to generate adversarial visual samples; second, calculate the contrastive distribution between the original and the adversarial samples and construct a new ACD loss function and fine-tune the model.

Our main contributions are summarized as follows: (1) We propose Adversarial Contrastive Dual (ACD) fine-tuning, a new method that combines adversarial and contrastive techniques to mitigate hallucination in MLLMs (Sec.2). (2) We conducted hallucination and comprehensive experiments to demonstrate the effectiveness of the ACD fine-tuning method in mitigating model hallucinations while retaining the comprehensive ability of MLLM (Sec.4.1).

2 Method

As shown in Figure 2, we propose the Adversarial Contrastive Dual fine-tuning method (ACD) mainly includes two steps: (1) adversarial sample generation, where we use FGSM (Goodfellow et al., 2014) to generate visual adversarial samples. (2) The ACD fine-tuning utilizes original and adversarial data to construct ACD fine-tuning data pairs and calculate the contrastive distribution between pairs to construct a new loss function – ACD loss.

2.1 Adversarial Samples Generation

From the previous analysis, it can be concluded that using adversarial samples with small perturba-

MODEL	Hallucination Benchmark				Comprehensive Benchmark		
	POPE	MMHal	CHAIRs↓	CHAIRi↓	MMbench	MM-Vet	GQA
InstructBLIP	77.83	2.10	40.00	8.00	36.00	26.20	49.20
Qwen-VL-Chat	<u>87.07</u>	**2.89**	48.20	9.10	60.60	**41.20**	57.50
mPLUG-Owl2	86.20	2.17	54.40	12.00	64.50	<u>36.20</u>	56.10
LLaVA-1.5	85.90	2.42	66.80	12.70	64.30	30.50	**62.00**
RLHF-V	86.20	2.59	44.60	7.90	63.60	30.90	-
POVID	86.90	<u>2.69</u>	**31.80**	**5.40**	<u>64.90</u>	31.80	-
ACD	**88.47**	2.47	<u>39.80</u>	<u>5.90</u>	**71.15**	30.60	<u>61.00</u>

Table 1: Compare the performance of the ACD fine-tuning model with other state-of-the-art models and fine-tuning methods for hallucinations. Evaluate their performance on hallucination and comprehension benchmarks. We **bold** the best result and <u>underline</u> the second-best result.

tions stimulates LLMs to generate hallucinations based on prior knowledge. Therefore, we first need to construct visual adversarial samples. To create these samples from MLLMs, we adopted the FGSM, which is related to the model gradient.

Given the visual input v, use FGSM to generate adversarial visual input v', where θ represents the hyper-parameters of MLLM, and ϵ represents the disturbance level of FGSM. A smaller ϵ value was used to minimize the perturbation.

$$v' = v + \epsilon \cdot \text{sign}\left(\nabla_v M(\theta)\right) \quad (1)$$

After generating the adversarial samples, we combine the input text x and output text y, representing each dataset item as $< v, v', x, y >$.

2.2 Adversarial Contrastive Dual Fine-tune

Adversarial Contrastive Dual fine-tuning uses adversarial and contrastive methods to fine-tune the model. Our method merges the original data with adversarial data and performs two rounds of fine-tuning. The original data is used for the first update to prevent the model from forgetting past knowledge and generating new hallucinations:

$$\mathcal{L}_{Gen} = -\textstyle\sum_{i=1}^{N} y_i \log p(y_i'|x, v) \quad (2)$$

The second update is ACD fine-tuning. Specifically, given a text query x and visual input, two distributions are generated: one conditioned on the original visual input v and the other on the adversarial visual input v'. The difference between these distributions yields a contrast distribution $C(v, v', x)$ between the two visual inputs:

$$\begin{aligned} C(v, v', x) = (1+\delta) \cdot \text{logits}_\theta(y|(x, v)) \\ - \delta \cdot \text{logits}_\theta(y|(x, v')) \end{aligned} \quad (3)$$

Here, δ controls the significance of the adversarial samples during the decoding process of the LLM. A smaller δ value indicates a weaker influence of the adversarial samples on the LLM. Then, a new contrastive probability distribution $C(v, v', x)$ is computed by leveraging the difference between the two initially obtained distributions:

$$p(C(v, v', x)_i) = \frac{\exp(C(v, v', x)_i)}{\sum_{j=1}^{N} \exp(C(v, v', x)_j)} \quad (4)$$

Finally, the ACD loss is obtained by calculating the cross-entropy between the adversarial contrastive probability distribution and the ground truth y:

$$\mathcal{L}_{ACD} = -\textstyle\sum_{i=1}^{N} y_i \log(p(C(v, v', x)_i)) \quad (5)$$

3 Evaluation Metrics

Visual Hallucination Benchmark To evaluate object hallucinations, we used commonly adopted benchmarks: POPE (Li et al., 2023) and CHAIR (Rohrbach et al., 2018).Here, POPE uses a set of binary classification tasks to prompt MLLM with simple "yes" or "no" questions about the existence of certain objects in the image. CHAIR, including CHAIRs and CHAIRi, compares the objects mentioned in the title with those appearing in the image. To evaluate the degree of hallucination and informative of the model's generated content, we evaluate on MMHal (Sun et al., 2023), using GPT-4 for evaluation.

Comprehensive Benchmark To demonstrate that our method can enhance the model's comprehensive ability while mitigating hallucinations, we evaluated the model using MMbench (Liu et al., 2023b), MM-Vet (Yu et al., 2023c), and GQA (Hudson and Manning, 2019). Here, MMbench evaluates the model's capabilities in detail across 20 dimensions; MM-Vet utilizes GPT-4 to assess the model based on six core vision-related functions (e.g., recognition, OCR); GQA evaluates the models' real-world visual reasoning abilities.

4 Experiment

Dataset We use FGSM (Goodfellow et al., 2014) to construct adversarial samples for fine-tuning based on the LLaVA Instruct-150K dataset[1] (Liu et al., 2024), which across various task types including image captioning, simple VQA, and complex logical reasoning.

Baseline We compare our model with state-of-the-art baselines. (1) General baselines: InstructBLIP (Dai et al., 2024), QwenVL-Chat (Bai et al., 2023), mPLUG-Owl2 (Ye et al., 2023) and LLaVA-1.5 (Liu et al., 2023a). (2) Different fine-tuning methods for LLaVA-1.5(7B): RLHF-V (Yu et al., 2023b) and POVID (Zhou et al., 2024), which leverage human feedback and external data annotation, respectively.

4.1 Main results

We use LLaVa-1.5 (7B)[2] as the backbone model, with a hyper-parameter ϵ of 1e-5. During the ACD fine-tuning process, we use a warmup learning rate of 1e-7 and learning rate of 1e-5. This fine-tuning process requires only one A100 80G GPU.

The main experimental results are shown in Table 1[3]: After ACD fine-tuning, the model achieved comparable results to the current state-of-the-art, surpassing it with 71.15% on the MMbench and 88.47% on the POPE. In addition, according to CHAIR, it significantly reduces object hallucinations, with CHAIRs of 39.8% and CHAIRi of 5.90%.

Compared to RLHF-V and POVID which rely on human feedback or AI data annotation, our method performs comparably across multiple benchmarks and is more effective at reducing object hallucination on POPE and CHAIR benchmarks.

[1] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K

[2] https://huggingface.co/liuhaotian/llava-v1.5-7b/tree/main

[3] The results related to GPT-4 may vary due to different versions.

Figure 3: The impact of different ϵ values on scene independence accuracy, hallucination accuracy, error rate, and overall accuracy during the ACD fine-tuning.

4.2 Analysis results

To analyze the impact of ϵ value, We examined the effects of ACD fine-tuning from four aspects: (1) Scene-independent accuracy: reveals the model's robustness in understanding scenes; higher ϵ values may introduce more significant perturbations. (2) Hallucination accuracy: measures whether the model's answers are consistent and correct with or without visual information. (3) Error rate: refers to the model's inconsistent and incorrect answers with or without visual information. (4) Overall accuracy: evaluates the model's general performance.

Figure 3 shows that the ϵ value is not directly proportional to the experimental results, and an optimal ϵ value exists[4]. Although the overall accuracy difference is minor between ϵ values of 1e-5 and 1e-7, when ϵ is 1e-5, scene-independent accuracy (22.06%) is highest, hallucination accuracy (9.18%) and error rate (27.28%) are lowest, and overall accuracy (63.55%) is highest.

5 Conclusion

In this work, we introduce a new method called Adversarial Contrastive Dual fine-tuning (ACD). First, we use MLLM and the Fast Gradient Symbolic Method (FGSM) to generate adversarial visual samples, building the ACD fine-tuning dataset. Then, by calculating the contrastive distribution between the original and adversarial samples, we construct the ACD loss function to fine-tune the model. Experimental results demonstrate that without any external annotations, ACD effectively reduces hallucinations without compromising the model's understanding ability.

[4] The model used in the experiment is Instructblip, and the fine-tuning dataset is VQAv2

Limitation

Although our work explores LLM hallucinations from a visual perspective, it has some limitations. We only focus on the impact of visual information on MLLM hallucinations and do not consider the influence of inputs from other modalities, such as text. And despite our method's significant improvements over the backbone model, a gap remains compared to other fine-tuning methods that use supervised learning or external data annotation. In the future, we plan to evaluate our method's performance using more MLLMs as backbones and further explore LLM hallucinations from a multi-modal perspective.

References

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5078–5088.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

OpenAI. 2023a. Chatgpt: A language model for conversational ai. Technical report, OpenAI. [Online]. Available: https://www.openai.com/research/chatgpt.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774. [Online]. Available: https://arxiv.org/abs/2303.08774.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

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

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

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023a. Natural language reasoning, a survey. *ACM Computing Surveys*.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

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

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

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