SKIP \N: A SIMPLE METHOD TO REDUCE HALLUCINATION IN LARGE VISION-LANGUAGE MODELS

Zongbo Han[1], **Zechen Bai**[2], **Haiyang Mei**[2], **Qianli Xu**[3], **Changqing Zhang**[1*], **Mike Zheng Shou**[2*]
College of Intelligence and Computing, Tianjin University [1]
Show Lab, National University of Singapore[2]
Institute for Infocomm Research, A*STAR[3]

ABSTRACT

Recent advancements in large vision-language models (LVLMs) have demonstrated impressive capability in visual information understanding with human language. Despite these advances, LVLMs still face challenges with multimodal hallucination, such as generating text descriptions of objects that are not present in the visual information. However, the underlying fundamental reasons of multimodal hallucinations remain poorly explored. In this paper, we propose a new perspective, suggesting that the inherent biases in LVLMs might be a key factor in hallucinations. Specifically, we systematically identify a semantic shift bias related to paragraph breaks ('\n\n'), where the content before and after '\n\n' in the training data frequently exhibit significant semantic changes. This pattern leads the model to infer that the contents following '\n\n' should be obviously different from the preceding contents with less hallucinatory descriptions, thereby increasing the probability of hallucinatory descriptions subsequent to the '\n\n'. We have validated this hypothesis on multiple publicly available LVLMs. Besides, we find that deliberately inserting '\n\n' at the generated description can induce more hallucinations. A simple method is proposed to effectively mitigate the hallucination of LVLMs by skipping the output of '\n'. Code is available at https://github.com/hanmenghan/Skip-n.

1 INTRODUCTION

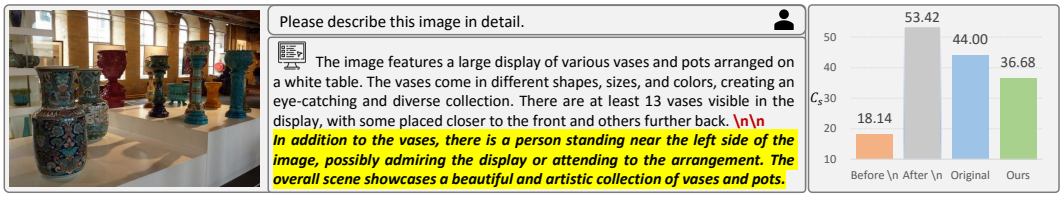

Figure 1: (Left) A hallucination example where the LVLM generates the hallucinatory description after the paragraph breaks '\n\n'. (Right) The severity of hallucinations across different outputs with BakLLaVA (Liu et al., 2023a), including Before \n, After \n, Original output, and Our mitigation results. Higher values indicate more severe hallucinations.

LVLMs have demonstrated remarkable capabilities in describing and analyzing the provided visual information using human language, marking a significant step towards general artificial intelligence (Achiam et al., 2023; Zhu et al., 2023; Li et al., 2023; Liu et al., 2023a;b). However, LVLMs often suffer from multimodal hallucinations, such as object hallucination where non-existent objects in the visual information are described in the generated responses (Wang et al., 2023b; Huang et al., 2023; Zhou et al., 2023; Cui et al., 2023). Such misleading responses can limit the deployment of LVLMs in many safety-critical applications such as autonomous driving (Bojarski et al., 2016) and machine-learning-aided medical diagnosis (Esteva et al., 2017).

*Corresponding author. Work done during an internship of Zongbo Han (zongbo@tju.edu.cn) at Show Lab.

Many approaches have been proposed for mitigating hallucinations in LVLMs. These methods are primarily categorized into two types, including retraining-based and post-hoc processing-based approaches. Retraining-based approaches include redesigning the vision encoder (Tong et al., 2024), collecting high-quality data for finetuning (Wang et al., 2023a), and employing reinforcement learning for LVLM fine-tuning (Zhao et al., 2023; Sun et al., 2023; Yu et al., 2023). Besides, post-hoc processing-based methods involve designing decoding strategies (Huang et al., 2023; Leng et al., 2023) and training an additional reviser model to detect and mitigate hallucinations (Zhou et al., 2023). Even though these approaches are effective in some circumstances, they still lack an exploration of key factors in LVLM hallucinations. Recent studies have found that imperfections in the vision encoder (Tong et al., 2024) and inherent uncertainties in the models (Zhou et al., 2023) can lead to hallucinations. In contrast, our work investigates the occurrence of hallucinations in LVLMs from the perspective of inherent biases within the models.

As shown in Fig. 1, we identify a special semantic shift bias triggered by paragraph breaks, where training data often show significant semantic changes before and after '\n\n'. This leads to a tendency for LVLMs to deviate from the previous non-hallucinatory description after '\n\n', resulting in hallucinations. We validate this hypothesis across several LVLMs (Liu et al., 2023b;a; Li et al., 2023; Zhu et al., 2023; Bavishi et al., 2023). Besides, we explore the use of '\n\n' as a method to induce hallucinations in existing LVLMs. We find that inserting '\n\n' in generated sentences significantly increases the probability of hallucinations, which further supports our findings that '\n\n' increases the probability of hallucinations. Based on this observation, we propose two simple yet effective methods to reduce hallucinations, including changing the prompt on the input side and modifying logits on the output side, both aimed at avoiding the output of '\n\n'. Experimental results show that the proposed method significantly reduces the occurrence of hallucinations.

Overall, our main contributions are as follows. Firstly, we identify that the inherent bias in LVLMs may be a key factor leading to hallucinations. Secondly, we discover a method to induce multimodal hallucinations, serving as an effective attack mechanism. Finally, we propose two effective and efficient solutions to reduce hallucinations in LVLMs without requiring additional costs.

2 METHOD

Our proposed method aims to reduce hallucinations by preventing the model from generating paragraph breaks ('\n\n'). Therefore, we can mitigate the semantic shift bias in LVLMs, which may cause the description to stray from the initial focus, leading to hallucinatory content in the following outputs. This goal can be efficiently achieved through two orthogonal methods: modifying the prompt given to the LVLMs during input and changing the decoding strategies during output.

Mitigating Hallucinations during Input (MiHI). LVLMs can usually follow human instructions well through instruction tuning. Therefore, we try to modify the prompt for LVLMs and encourage them to fulfill the original instructions while avoiding the output of '\n', thereby maintaining the continuity and coherence of the generated text. Specifically, taking the task of describing an input image as an example, the commonly used prompt is *"Please describe this image in detail"*. The proposed method modify the above prompt to *"Please describe this image in detail in one paragraph."* This modification emphasizes the generation of a single, continuous paragraph, thereby avoiding the output of paragraph breaks '\n\n'. Note that the prompt (*in one paragraph*) provided above can be adjusted as needed based on practical performance.

Mitigating Hallucinations during Output (MiHO). From the perspective of modifying the output decoding strategies, we can avoid the output of '\n' by reducing the logits corresponding to the '\n' token. Formally, considering the next token logits is L. We can adjust the next token logits to avoid outputting '\n'. Specifically, the adjusted next token logits $\hat{L}$ can be obtained with $\hat{L} = L - \lambda \cdot \mathbf{1}_{\backslash n}$, where λ is a hyperparameter used to control the penalty strength. $\mathbf{1}_{\backslash n}$ represents a one-hot encoding vector in which the dimension corresponding to the '\n' token is set to 1 while all other dimensions are set to 0. In our implementation, we set λ to positive infinity to effectively eliminate the prediction probability of '\n' token. In contrast, when we want to intentionally insert '\n' to attack the model[1], we can adjust λ at the specific position accordingly .

[1]Unlike traditional attacks, the attack described here aims to better observe the impact of the '\n' token.

3 EXPERIMENTS

We conduct extensive experiments on multiple LVLMs to address the following questions. Q1 Hypothesis verification: Does the description generated after '\n' exhibit more serious hallucinations? Q2 Attackability: Does the insertion of '\n\n' in the generated description trigger hallucinations? Q3 Effectiveness: Can our proposed method (MiHO and MiHI) effectively mitigate hallucinations?

Experimental settings. Our evaluation are conducted on the six publicly available LVLMs, including BakLLaVA, LLaVA-v1.5-7B, LLaVA-v1.5-13B (Liu et al., 2023a), InstructBLIP-7B (Li et al., 2023), MiniGPT-v2 (Chen et al., 2023), and Fuyu-8B (Bavishi et al., 2023). We primarily focus on the occurrence of object hallucination within the generated descriptions. To this end, we randomly select 5,000 images from the MSCOCO validation set (Lin et al., 2014) and prompt the LVLMs to generate detailed descriptions of these images. Then we employ the CHAIR evaluation framework (Rohrbach et al., 2018) for our analysis. The corresponding metrics are formulated as follows:

$$C_s = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, C_i = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}. \quad (1)$$

Higher C_s and C_i indicate more serious hallucinations. In terms of decoding strategy selection, we adopt two commonly used decoding strategies, including greedy decoding and random multinomial sampling decoding (Wolf et al., 2019). Since the proposed MiHO and MiHI are two orthogonal methods, we report the results of MiHO, MiHI, and the combined use of MiHO and MiHI.

Q1 Hypothesis verification. In Table 1, we report the hallucination evaluation performance for content before '\n' compared to content generated after '\n'. We can see that the content produced after '\n' has a significant probability of hallucination.

Table 1: Q1 Hypothesis verification. Sentences generated after '\n' have more hallucinations.

Model	BakLLaVA (Liu et al., 2023a)				InstructBLIP-7B (Li et al., 2023)				LLaVA-v1.5-7B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Before \n	18.14	6.09	22.28	7.83	48.12	14.47	55.96	17.66	21.24	7.37	29.98	11.30
After \n	53.42	23.33	53.84	24.46	55.88	29.45	63.29	33.33	57.23	28.06	59.99	30.62
Model	Fuyu-8B (Bavishi et al., 2023)				MiniGPT-v2 (Chen et al., 2023)				LLaVA-v1.5-13B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Before \n	17.36	6.84	27.04	10.31	24.26	8.91	30.92	11.95	20.12	6.60	26.72	9.20
After \n	58.22	29.23	61.53	30.16	58.83	31.49	62.09	32.40	52.11	23.77	57.66	26.93

Q2 Attackability. We also validate that inserting '\n\n' at appropriate positions can induce LVLMs to generate more hallucinations. Specifically, when the sentence outputs the period ('.') token for the k-th time, we manually insert the '\n\n' to initiate the attack, where k is a manually specified position hyperparameter. The experimental results are shown in Table 2. The table indicates that inserting '\n\n' later in the sentence increases hallucinations.

Table 2: Q2 Attackability. Performance comparison when attack LVLMs at different position, where Attack-k refers to initiating the attack upon encountering the period ('.') token for the k-th time.

Model	BakLLaVA (Liu et al., 2023a)				InstructBLIP-7B (Li et al., 2023)				LLaVA-v1.5-7B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	48.56	13.00	51.58	14.37	48.44	14.53	56.66	17.81	51.74	15.35	57.16	18.47
Attack-1	47.42	12.64	51.72	14.33	58.28	17.29	63.76	19.80	55.28	16.42	60.60	19.90
Attack-2	49.36	13.23	55.16	15.55	59.52	18.63	65.74	20.81	56.42	16.72	62.48	20.46
Attack-3	54.54	13.83	58.28	15.56	58.18	18.13	65.76	20.69	60.58	16.95	65.96	20.71
Model	Fuyu-8B (Bavishi et al., 2023)				MiniGPT-v2 (Chen et al., 2023)				LLaVA-v1.5-13B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	59.74	17.27	63.08	19.18	62.62	19.50	64.36	21.08	47.52	13.08	53.04	15.69
Attack-1	51.26	15.35	59.28	18.48	61.26	18.31	64.14	20.88	47.56	13.08	53.16	15.96
Attack-2	54.60	17.16	63.18	19.80	63.16	20.30	67.78	22.47	50.50	14.27	57.32	17.51
Attack-3	63.14	17.98	68.74	20.69	65.50	20.17	71.64	22.79	54.62	14.15	61.12	17.08

Q3 Effectiveness. As shown in Table 3, we compare the proposed methods with the original outputs of LVLMs. Furthermore, to eliminate the influence of sentence length on the output, we conduct a comparison between the proposed and original methods at equal output lengths, achieved by truncating the end of each sentence. The experimental results are shown in Table 4. Based on these experimental results, we can draw the following conclusions: (1) As shown in Table 3, both MiHO and MiHO+MiHI significantly reduce the occurrence of hallucinations across all models. MiHI also significantly reduces hallucinations in all models except Fuyu-8B, possibly because Fuyu-8B is not fine-tuned with instructions, resulting in a poorer understanding of prompts. (2) According to Table 4, when comparing original descriptions of the same length, MiHO demonstrates significant improvements in almost all models. However, MiHI and MiHO+MiHI sometimes exhibit performance decreases, possibly because the modified prompts negatively impacts the descriptions of LVLMs. (3) Compared to greedy decoding, sampling decoding strategy is more prone to producing hallucinations. Our method shows better performance when used with greedy decoding strategy.

Table 3: Q3 Effectiveness. Performance comparison of the proposed method on different LVLMs.

Model	BakLLaVA (Liu et al., 2023a)				InstructBLIP-7B (Li et al., 2023)				LLaVA-v1.5-7B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	48.56	13.00	51.58	14.37	48.44	14.53	56.66	17.81	51.74	15.35	57.16	18.47
MiHO	38.96	10.35	42.66	11.87	48.30	14.50	57.06	18.27	38.62	11.35	47.70	15.35
MiHI	42.04	11.66	47.10	13.39	45.70	12.91	56.26	17.11	39.40	12.54	45.32	16.02
MiHO+MiHI	36.68	10.04	42.16	11.93	45.70	12.91	57.40	17.34	39.38	12.53	45.36	16.03
Model	Fuyu-8B (Bavishi et al., 2023)				MiniGPT-v2 (Chen et al., 2023)				LLaVA-v1.5-13B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	59.74	17.27	63.08	19.18	62.62	19.50	64.36	21.08	47.52	13.08	53.04	15.69
MiHO	38.14	10.60	45.92	14.44	33.02	11.38	45.66	15.60	35.32	9.61	43.58	13.15
MiHI	59.08	17.37	63.40	19.54	49.28	14.99	58.16	19.13	37.64	10.54	47.26	14.25
MiHO+MiHI	40.76	11.16	48.36	14.94	42.38	13.24	51.34	17.62	34.64	9.73	44.72	13.45

Table 4: Q3 Effectiveness. Performance comparison of the proposed method on different LVLMs when the output sentence lengths are equal.

Model	BakLLaVA (Liu et al., 2023a)				InstructBLIP-7B (Li et al., 2023)				LLaVA-v1.5-7B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	44.00	11.93	45.78	13.11	48.38	14.49	52.52	16.71	45.04	13.24	49.40	15.98
MiHO	38.96	10.35	42.66	11.87	48.30	14.50	53.08	17.21	38.62	11.35	47.70	15.35
Original	44.34	12.20	45.88	13.20	45.70	13.35	51.82	16.55	36.16	10.98	41.82	14.26
MiHI	42.04	11.66	47.10	13.39	45.70	12.91	52.18	16.24	39.40	12.54	45.32	16.02
Original	41.64	11.44	43.30	12.55	45.70	13.35	52.52	16.71	36.16	10.98	41.58	14.32
MiHO+MiHI	36.68	10.04	42.16	11.93	45.70	12.91	53.66	16.49	39.38	12.53	45.36	16.03
Model	Fuyu-8B (Bavishi et al., 2023)				MiniGPT-v2 (Chen et al., 2023)				LLaVA-v1.5-13B (Liu et al., 2023a)			
Decoding	Greedy		Sampling		Greedy		Sampling		Greedy		Sampling	
Method	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$	$C_s \downarrow$	$C_i \downarrow$
Original	44.36	13.10	51.06	16.00	35.04	12.41	49.34	17.12	42.10	11.49	46.50	13.95
MiHO	38.14	10.60	45.92	14.44	33.02	11.38	45.66	15.60	35.32	9.61	43.58	13.15
Original	58.30	16.90	59.08	18.00	54.90	16.77	56.90	18.86	42.42	11.57	47.02	14.12
MiHI	59.08	17.37	63.40	19.54	49.28	14.99	58.16	19.13	37.64	10.54	47.26	14.25
Original	47.62	13.90	53.00	16.29	49.44	15.59	52.56	17.58	41.04	11.30	45.80	13.76
MiHO+MiHI	40.76	11.16	48.36	14.94	42.38	13.24	51.34	17.62	34.64	9.73	44.72	13.45

4 Related Work

Hallucination mitigation in LVLMs can be primarily categorized into two types including retraining-based methods and post-hoc processing-based methods. Retraining-based methods include collecting high-quality data (Wang et al., 2023a), modifying the architecture of LVLMs (Tong et al., 2024; Zhai et al., 2023), and adjusting the training strategy for LVLMs (Yu et al., 2023; Zhao et al., 2023; Sun et al., 2023; Ben-Kish et al., 2023). Although these methods are effective, they

often require additional computational overhead to alleviate hallucinations of LVLMs. On the other hand, post-hoc processing methods aim to mitigate hallucination without retraining the LVLMs. These methods involve modifying the decoding method of LVLMs (Huang et al., 2023; Leng et al., 2023), enriching the visual context of LVLMs by integrating existing open-source vision models (Zhao et al., 2024), and training an additional reviser model (Zhou et al., 2023).

Bias in machine learning. Bias in machine learning models, caused by biased training data, can lead to distrust and serious consequences (Elazar et al., 2023; Longpre et al., 2023). It has been widely explored include bias against minority subpopulations (Han et al., 2022; 2023), bias against datasets (Torralba & Efros, 2011; Liu & He, 2024), bias in alignment of image and text information Tong et al. (2024); Lin et al. (2023). In this paper we focus on the semantic shift bias in LVLMs triggered by paragraph breaks.

5 Conclusions and Takeaways

In this paper, we identify a phenomenon of '\n\n'-induced hallucinations in some existing LVLMs, attributed to semantic shift bias. Besides, we find that inserting '\n\n' during the description generation process can induce hallucinations in LVLMs. Based on this, we propose a method to alleviate hallucinations by reducing the probability of '\n' from the input and output perspectives. Extensive experiments on multiple publicly available LVLMs are conducted to verify the performance of our method. It should be highlighted that the '\n\n'-induced hallucination problem is not found in some LVLMs, *e.g.*, GPT-4 (Achiam et al., 2023). What causes the hallucination problem remains an open question. Finally, it remains to be explored whether this bias can be overcome when the model scale continues to increase.

6 Acknowledgement

This work is supported in part by the scholarship from China Scholarship Council (No.202306250107). We also appreciate discussions with Shiwei Wu and his suggestions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2023.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35: 37704–37718, 2022.

Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Qinghua Hu, Bingzhe Wu, Changqing Zhang, et al. Reweighted mixup for subpopulation shift. *arXiv preprint arXiv:2304.04148*, 2023.

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. *arXiv preprint arXiv:2311.17911*, 2023.

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. *arXiv preprint arXiv:2311.16922*, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yiqi Lin, Conghui He, Alex Jinpeng Wang, Bin Wang, Weijia Li, and Mike Zheng Shou. Parrot captions teach clip to spot text. *arXiv preprint arXiv:2312.14232*, 2023.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? *arXiv preprint arXiv:2403.08632*, 2024.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

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

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.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023a.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

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

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

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. *arXiv preprint arXiv:2310.00754*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A More Experimental Results

In this section, we present further experimental evidence to validate the effectiveness of our proposed methodology. We conduct comparisons with two contemporary approaches: DoLa (Chuang et al., 2023) and OPERA (Huang et al., 2023). Given that both referenced methods employ the Beam Search decoding strategy, we also incorporate the Beam Search strategy into our proposed method to ensure a fair comparison. The results of these experiments are shown in Table 5. The experimental results demonstrate that our proposed method achieves the best performance.

Table 5: Comparison with current state-of-the-art methods. The experimental results of Beam Search, DoLa and OPERA are from OPERA (Huang et al., 2023).

Model	Llava-1.5-7B		InstructBLIP	
Method	C_s	C_i	C_s	C_i
Beam Search (Graves, 2012)	48.8	13.9	55.6	15.8
DoLa (Chuang et al., 2023)	47.8	13.8	48.4	15.9
OPERA (Huang et al., 2023)	44.6	12.8	46.4	14.2
MiHO	**34.6**	**10.2**	47.9	13.8
MiHI	37.1	11.4	44.7	11.7
MiHI+MiHO	37.1	11.4	**44.7**	**11.7**

To further analyze the impact of the prompt in MiHI, we conducted additional experiments with different prompts generated by GPT-4. The prompts are shown as follows.

P1 *Please describe this image in detail in one paragraph."*

P2 *Please describe this image in detail in a single, continuous text."*

P3 *"Please describe this image in detail , with no separation into paragraphs."*

P4 *Please describe this image in detail without \n."*

P5 *Please describe this image in detail without using paragraph breaks."*

The experimental results are shown in Table 6. It can be seen from the experimental results that prompts may have a significant impact on hallucinations.

Table 6: Results of the proposed method using various prompts.

Model	BakLLaVA		InstructBLIP-7B		LLaVA-v1.5-7B		Fuyu-8B		MiniGPT-v2		LLaVA-1.5-13B	
Method	C_s	C_i	C_s	C_i	C_s	C_i	C_s	C_i	C_s	C_i	C_s	C_i
P1	42.04	11.66	45.70	12.91	39.40	12.54	59.08	17.37	49.28	14.99	37.64	10.54
P2	47.44	12.71	30.46	9.88	45.08	12.62	61.66	18.30	56.08	18.03	49.24	14.62
P3	48.02	12.74	40.06	11.40	46.62	13.03	58.16	17.24	55.78	18.17	50.86	15.16
P4	46.24	12.38	48.86	14.88	44.28	12.40	59.72	17.95	62.70	19.76	48.82	14.58
P5	47.22	12.58	45.00	12.93	46.08	12.85	57.70	16.76	55.96	18.38	51.02	15.26
P1+MiHO	36.68	10.04	45.70	12.91	39.38	12.53	40.76	11.16	42.38	13.24	34.64	9.73
P2+MiHO	38.76	10.25	30.46	9.88	34.84	9.53	40.26	11.08	32.76	11.64	40.08	11.71
P3+MiHO	39.46	10.39	39.98	11.38	35.54	9.66	38.24	10.71	33.62	11.82	39.86	11.73
P4+MiHO	38.50	10.22	48.40	14.74	34.92	9.65	38.60	10.50	31.72	11.33	38.04	11.38
P5+MiHO	39.16	10.53	44.96	12.92	36.44	9.90	38.90	10.72	30.26	11.26	40.90	11.99