Reducing Hallucinations of Large Vision Language Models
Project Final Report

Group Member: Yiteng Zhang, Jiaxuan Niu, Kuang Jiang, Jiarong Wang

Abstract

Large Vision-Language Models (LVLMs) have made remarkable progress in understanding and combining visual and textual information. However, they often generate hallucinations, descriptions of objects that do not appear in the input image, limiting their reliability, especially in fields such as healthcare and autonomous systems. In this project, we explored ways to address this issue by analyzing the hallucination tendencies of the LLaVA model using the POPE evaluation framework on a subset of the COCO val 2014 dataset. To reduce hallucination, we utilized VCD with Cosine-Beta Noise Scheduling. Our methods demonstrated improvements in Accuracy, Precision, Recall, and F1 Score, especially in adversarial settings.

1 Introduction

The development of Large Vision-Language Models (LVLMs) has made significant advances in understanding the combination of visual and text inputs, enabling models to generate responses from context-relevant descriptions from a zero-shot, world-knowledge, and complex-reasoning setting. However, these models are still faced with a major challenge: hallucinations, that is models can generate textual outputs that include objects not in the given input image. This defect can cause much greater risk when it comes to LVLM applications in domains like healthcare, autonomous systems, and robotics. Therefore, to further contribute to solving this problem, we first set up a dataset with 53 images from COCO val 2014 dataset, and then evaluate object hallucination tendencies in LLaVa using POPE, and finally mitigate hallucinations with Visual Contrastive Decoding (VCD) and with Cosine-Beta Noise Scheduling.

2 Related Work

2.1 Hallucinations in LVLMs

Hallucinations are a problem that starts with LLMs themselves. Two categories of hallucination problems have been scientifically identified by the NLP community: 1) Fulfillment hallucination refers to the divergence of generated content from user instructions or the context provided by the input, as well as self-consistency within generated content; 2) Factuality hallucination highlights the disparity between generated content and verifiable real-world facts, usually presenting as factual inconsistency or fabrication. Research on hallucination in MLLMs, as opposed to pure LLMs, primarily focuses on the cross-modal inconsistency—the difference between the generated text answer and the presented visual content. This discrepancy implies that research conducted in LLMs cannot be apparently applied to MLLMs. Consequently, there is an increasing need to thoroughly examine current developments in MLLMs.

2.2 Polling-based Object Probing Evaluation

The Polling-based Object Probing Evaluation (POPE) frames hallucination as a binary classification task, asking models simple “Yes” or “No” questions about object presence in an image. This approach minimizes variability introduced by prompt forms, resulting in more consistent evaluations. POPE also enables evaluation on unannotated datasets by using tools like SEEM for automated object detection. POPE uses three sampling strategies–random, popular, and adversarial – to test LVLMs’ hallucination tendencies across difficulty levels, with adversarial sampling targeting commonly co-occurring objects to reveal bias in generating familiar pairs. Evaluated on 500 images from the MSCOCO validation set, POPE measures hallucination through balanced questions on ground-truth and nonexistent objects. Metrics like accuracy, precision, recall, and F1 score demonstrate POPE's effectiveness in reliably revealing hallucination patterns in LVLMs, providing a more stable and flexible evaluation compared to CHAIR. Therefore, we use POPE to serve as an evaluation benchmark by providing a structured, standardized approach to measure hallucination tendencies in large vision-language models (LVLMs).

2.3 Visual Contrastive Decoding

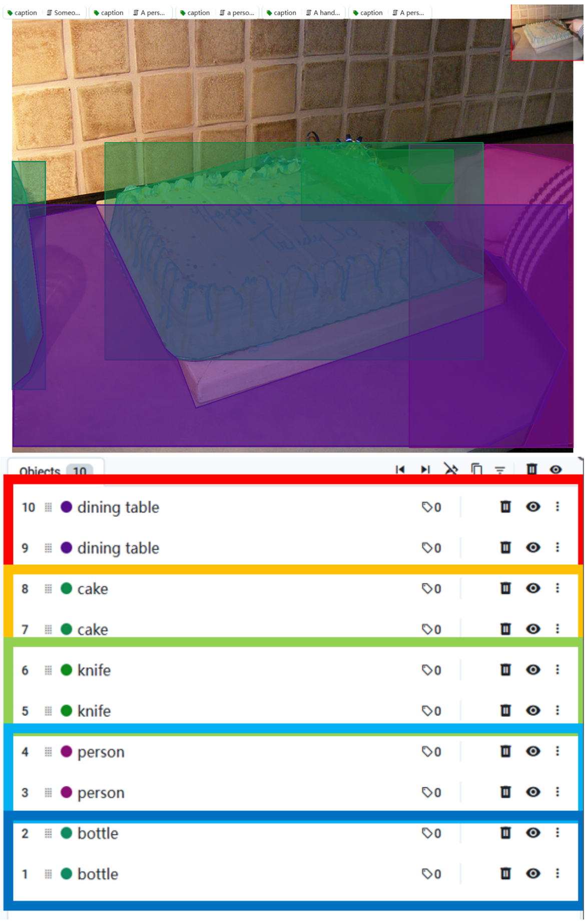
The Visual Contrastive Decoding (VCD) is a training-free technique developed to mitigate object hallucinations in LVLMs. The method functions in reducing object hallucinations by reducing reliance on statistical bias and unimodal priors without the need for extensive retraining or external tools. The paper is built upon previous research on object hallucinations in both LLMs and LVLMs, leveraging various mitigation strategies, and evaluates the contributions and limitations in this task.

VCD suggests a lightweight solution to the hallucination problem. Identifying that visual hallucinations come from over relying on the language prior from the training data and the inherited potential statistical biases, VCD contrasts output distributions generated from original and distorted visual inputs. Then the posterior distribution of the model is corrected in the generative process by such means, allowing the model to effectively mitigate hallucinated objects. The method mitigates the effect of the two primary causes of object hallucination problem in LVLM by correcting the generative process, with less dependency on language model priors and fewer statistical biases inherited from training data.

2.4 Baseline Model

We selected the LLaVA model as the baseline model and used POPE to assess the tendency toward object hallucinations in Large Vision-Language Models (LVLMs) on COCO val 2024 dataset . Based on this, we utilized Visual Contrastive Decoding (VCD) and Cosine-Beta Noise Scheduling function to reduce hallucinations.

3 Dataset



10	●	dining table	0
9	●	dining table	0
8	●	cake	0
7	●	cake	0
6	●	knife	0
5	●	knife	0
4	●	person	0
3	●	person	0
2	●	bottle	0
1	●	bottle	0

Figure 1: For example, in this image, it contains 5 ground-truth annotated objects.

To evaluate LLaVA's hallucination tendencies using POPE on the MSCOCO dataset, The dataset includes diverse objects and annotations for robust evaluation. We selected 53 images from it and make sure that each image contains at least three annotated ground-truth objects (Figure 1). To make alignment with POPE's evaluation standards, for each image, we generate binary "Yes" or "No" questions

Dataset	Setting	Model	Decoding	Accuracy↑	Precision	Recall	F1 Score↑
MSCOCO	*Random*	LLaVA1.5	Regular	$83.29_{(\pm 0.35)}$	$92.13_{(\pm 0.54)}$	$72.80_{(\pm 0.57)}$	$81.33_{(\pm 0.41)}$
			VCD	$87.73_{(\pm 0.40)}$	$91.42_{(\pm 0.55)}$	$83.28_{(\pm 0.42)}$	$87.16_{(\pm 0.41)}$
			VCD with CosBeta	**88.27**	**91.89**	**83.95**	**87.74**
	Popular	LLaVA1.5	Regular	$81.88_{(\pm 0.48)}$	$88.93_{(\pm 0.60)}$	$72.80_{(\pm 0.57)}$	$80.06_{(\pm 0.05)}$
			VCD	$85.38_{(\pm 0.38)}$	$86.92_{(\pm 0.53)}$	$83.28_{(\pm 0.42)}$	$85.06_{(\pm 0.37)}$
			VCD with CosBeta	**86.72**	**88.89**	**83.95**	**86.34**
	Adversarial	LLaVA1.5	Regular	$78.96_{(\pm 0.52)}$	$83.06_{(\pm 0.58)}$	$72.75_{(\pm 0.59)}$	$77.57_{(\pm 0.57)}$
			VCD	$80.88_{(\pm 0.33)}$	$79.45_{(\pm 0.29)}$	$83.29_{(\pm 0.43)}$	$81.33_{(\pm 0.34)}$
			VCD with CosBeta	**85.18**	**86.08**	**83.95**	**85.00**

Table 1: Results on POPE. Regular decoding denotes direct sampling, whereas VCD refers to sampling from our proposed contrastive distribution pvcd. VCD with CosBeta refers to our method.

based on the presence of the objects. Specifically, we will use ground-truth objects to create "Yes" responses and sampling strategies to create "No" responses.

4 Evaluation framework

For evaluation, we chose POPE, which employs three sampling methods—random, popular, and adversarial—to select objects absent from the image, ensuring varied difficulty in hallucination detection. The balanced questions are then inputted to LLaVA, which responds with an answer per question based on the image content. By comparing LLaVA's responses to expected answers, we calculate accuracy, precision, recall, and F1 score, which together provide a nuanced assessment of hallucination tendencies. This perspective also enables us to validate that LLaVA is stable and accurate for object recognition with no prompt variability, providing POPE as a trustworthy benchmark of hallucination detection here. Accuracy represents the overall proportion of correctly answered questions, indicating general performance. Precision measures the ratio of correctly answered "Yes" questions, focusing on the accuracy of affirmative responses. Recall reflects the ratio of correctly answered "No" questions, showing how well the model avoids hallucinating nonexistent objects. F1 Score combines Precision and Recall, providing a balanced measure of model accuracy and selected as our primary evaluation metric.

5 Methodology

5.1 Cosine-beta noise scheduling

Noise scheduling is important for final performance in diffusion models. Though we are working with methods on LVLMs, the noise adding process in the original VCD work is quite similar to the forward passing phase in training a diffusion model. Therefore we used the same idea and tested the performance of different noise schedule functions in the VCD method, including sigmoid scheduling and cosine scheduling. Then we follow similar way in the noising process of data and define the final image to be

$$x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\varepsilon$$

5.2 Input scaling

We also scaled the input by a constant to indirectly adjust noise scheduling. The scaled effect can be organized as:

$$x_t = \sqrt{\gamma(t)}bx_0 + \sqrt{1-\gamma(t)}\varepsilon$$

in the noising process. As we reduce the factor b from 1 to 0, the image becomes darker and increases the noise level. Input scaling achieves a slightly different trajectory in the logSNR in the forward passing phase when compared to the effect of cosine and sigmoid schedules, particularly when t is closer to 0. Using input scaling, we can achieve a more rapid drop in logSNR during the start and the end of the image noising process. Considering the resolution of our images is close to 500×500 and that we can get a smaller FID score with a lower input scaling factor and higher resolution image, I tuned the scale factor b to be 0.27.

6 Result

To quantitatively verify our findings, we evaluated the methods across different metrics, including Accuracy, Precision, Recall, and F1 Score, for the MSCOCO dataset under the three dataset settings. The results are presented in figure 1.

Performance Gains: In all settings, the VCD with the CosBeta method consistently outperformed Regular Decoding and VCD alone, achieving the highest Accuracy and F1 Scores. This demonstrates its effectiveness in reducing hallucinations while maintaining the integrity of LVLM outputs.

Robustness Under Adversarial Conditions: The Adversarial setting showed the greatest improvement with VCD with CosBeta, highlighting its ability to address challenging input cases effectively. Compared to Regular Decoding, Accuracy increased by 6.22 percentage points, and F1 Score improved by 7.43 percentage points.

Balanced Metrics: The proposed methods not only reduced hallucinations but also maintained high Precision and Recall, ensuring output relevance and correctness.

These results strongly support our method, VCD with CosBeta, is a robust and effective approach for mitigating hallucinations in LVLMs.

7 Conclusion

This project focused on the challenges related to hallucinations in Large Vision-Language Models (LVLMs), as this problem reduces LVLMs reliability in applications that need high precision (e.g., healthcare, autonomous systems, and robotics) We explored and deployed the COCO val 2014 dataset subset with POPE evaluation framework to investigate and reduce hallucinations in LLaVA. Our methods, VCD with Cosine-Beta Noise Scheduling, showed consistent improvements in Accuracy, Precision, Recall, and F1 Score across multiple dataset settings.

8 Limitations

Our current work is limited by dataset, model scale, and task specialization. Although the datasets used provide useful insights, these are not broad enough and do not cover a sufficient range of the number of objects, scene complexity, and domains to allow us to generalize our method to offer competitive performance in real applications. Moreover, our experiments utilize the standard 7B version of LLaVA, which, even if computationally attainable, does not take advantage of the performance gains obtainable by using larger models. This limitation emphasizes the potential to explore more performance gains from higher model capacity. Moreover, the baseline model used in our study was not fine-tuned on datasets specifically designed for multi-object-aware tasks. As a result, its ability to handle scenes with numerous objects or intricate relationships is limited. These limitations provide clear directions for future work to enhance the robustness and applicability of our approach.

References

[1] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint* arXiv:2404.18930.

[2] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2023). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint* arXiv:2311.16922.

[3] Li, X., Zhang, Y., & Wang, Q. (2023). Polling-based object probing evaluation. *IEEE Transactions on Multimedia*. Retrieved from https://arxiv.org/pdf/2305.10355.

[4] Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint* arXiv:2102.09672. Retrieved from https://arxiv.org/pdf/2102.09672.

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2023). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint* arXiv:2301.10972. Retrieved from https://arxiv.org/pdf/2301.10972.