Mitigating Hallucination in Vision-Language Models With Fine-Tuning and Self-Revision

Sadman Sakib, Danyal Maqbool, Rishika Ahuja, Muhammad Musa, Apoorva Mittal

1 Introduction

In recent years, Large Multimodal Models (LMMs) have shown remarkable capabilities in generating responses across various modalities, such as text and images. However, these models are prone to hallucinations [6], where their outputs are not grounded in the provided multimodal context, leading to unreliable or incorrect answers. This project aims to reduce hallucinations in vision-language models by implementing two recent techniques: Fact-RLHF [6] and feedback-guided self-revision [1]. Fact-RLHF focuses on training the model with reinforcement learning to prioritize factual accuracy, while feedback-guided self-revision employs a critique-feedback-revise loop to refine generated responses. Similar to Fact-RLHF, we have trained a vision-language model, BLIP2, with supervised fine-tuning and DPO. Taking inspiration from Fact-RLHF, we have augmented the revision process in feedback-guided self-revision with factual information to further mitigate hallucinations. Our evaluation were conducted on vision-language models from HuggingFace. Our code is publicly available on: `https://github.com/DanyalMaq/Foundation-Model-Project`

2 Related Work

2.1 Vision-Language Models (VLMs)

Vision-Language Models (VLMs) integrate visual and textual data, enabling them to perform tasks that require an understanding of both modalities. These models bridge the gap between visual and textual perception, supporting applications in multi-modal AI [5] [3].

Key characteristics of VLMs include:

1. **Multi-Modal Input:** VLMs process and learn from both image and text data, aligning textual descriptions with corresponding visual content.
2. **Feature Extraction:**
   - *Visual Features:* Extracted using architectures like Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs).
   - *Textual Features:* Derived from models such as Transformers, pretrained on large-scale text corpora.
3. **Cross-Modal Alignment:** VLMs align visual and textual embeddings in a shared representation space. Common techniques include:
   - *Contrastive Learning:* Models like CLIP align image-text pairs by maximizing similarity for matched pairs and minimizing it for mismatched pairs.
   - *Attention Mechanisms:* These architectures refine the alignment by focusing on salient features across modalities.
4. **Training Objectives:**
   - *Contrastive Loss:* Optimizes similarity for aligned image-text pairs while penalizing mismatches.

- *Masked Token Prediction:* Extends masked language modeling to image or text tokens (e.g., ViLT).
- *Image-Text Matching:* Predicts the relatedness of an image-text pair.
- *Instruction Tuning:* The language model, conditioned on one or more images, learns to generate response for an instruction. (e.g. LLaVa [5])

2.2 Fact-RLHF

Fact-RLHF [6] first performs supervised fine-tuning on LLaVa model and then trains the model with RLHF. The supervised fine-tuning step was performed on high-quality human-annonated datasets. For RLHF training, first a reward model was trained, which was initialized with LLaVa-SFT model weights and a scalar head. The preference data was prepared from crowdworkers who were asked to choose the more honest and helpful response among LLaVa-SFT generated response pairs. When the reward model was trained with preference data, it was found that the model was prone to reward hacking, without focusing on hallucination. So, Fact-RLHF augments the reward model training data with factual information to help the model reason about the preferred data. RLHF training was performed using PPO with a length and a correctness penalty.

2.3 Volcano

Volcano [1] is a multimodal model designed to mitigate hallucinations by employing a self-feedback-guided revision process. The model generates initial responses to visual inputs, critiques them using natural language feedback grounded in the visual data, and then revises the responses accordingly. This iterative critique-revision-decide loop enables VOLCANO to align its outputs more closely with the provided visual information.

3 Methodology

3.1 Models and techniques

For training and evaluation, we have used open-source VLMs available on HuggingFace. We performed supervised fine-tuning and DPO training on BLIP-2 model. We performed fact-augmented self-revision on BLIP-2 [3], and LLaVA-OneVision [2] .

3.1.1 DPO Training on BLIP-2

1. Generating Response Pairs

Objective: To generate image-text pairs that form the base dataset for preference alignment.

Dataset: We took the image and questions from LLaVa-Human-Preference dataset. The dataset has 10K questions about COCO images.

Temperature Settings:

- **0.7:** Balanced creativity and reliability in generated outputs.
- **1.0:** Diverse outputs with higher variance.

2. Transforming to a Preference Dataset

Model Used: Used Qwen2-VL as judge.

Procedure:

1. Image-text pairs generated by BLIP-2 were evaluated by Qwen-2VL.
2. The judge was asked to choose which response was better.

3. Alignment Using Direct Preference Optimization (DPO)

Methodology:

- We used DPOTrainer from TRL library to perform DPO training on BLIP-2 with the preference dataset.
- DPO simplifies alignment by directly optimizing for preferred outputs without requiring a separate reward model.

3.1.2 Supervised Fine-tuning

We fine-tuned BLIP2 model on LLaVa-Instruct dataset using SFTTrainer from TRL library. The image and question tokens were masked and the loss was computed only from the generated answer. We trained with 5000 visual questions from the dataset. The model was trained in float16 precision and with LoRA adapters.

3.1.3 Self-guided-revision with captions

We adopt Volcano's strategy of having the VLM engage in a feedback loop with itself to course correct any mistakes that might be happening. We also combine this effort with Fact-RLHF's approach of providing image captions as ground-truth facts for an image. In our case, we first have the VLM generate captions for itself and then initiate the feedback loop. Our goal is to achieve performance boosts similar to what LLaVA-RLHF model was able to achieve. We compare the results with and without these captions.

3.2 Evaluation benchmarks

We use two existing datasets to evaluate how resistant our chosen models are to hallucination.

3.2.1 MMHAL-Bench

The MMHAL-Bench [6] dataset is a specialized benchmark developed to evaluate hallucinations in large multimodal models (LMMs), focusing specifically on assessing whether hallucination exists in generated responses rather than general response quality. Unlike prior benchmarks, MMHAL-Bench avoids oversimplified yes/no questions and instead uses open-ended, realistic queries to better reflect practical user interactions. The dataset comprises 96 meticulously designed image-question pairs spanning eight question categories and 12 object topics, addressing common LMM hallucination patterns, including false claims about object attributes, spatial relations, comparisons, and more. Images are sourced from the validation and test sets of OpenImages to avoid data leakage, with questions designed to trigger hallucinations even in advanced LMMs like LLaVA13BX336. It is evaluated by using GPT-4 as an LLM judge to both rate whether or not the responses from the LLM has any hallucinations as well as rate the quality of the response on a scale of 1-6.

3.2.2 POPE

POPE [4] is a dataset that targets to identify object hallucination rate of a VLM with of 9000 yes/no questions on images sampled from the COCO dataset. It is designed to span a wide variety of object types in the COCO dataset. The goal is for the model to answer whether an objects exists or not in an image.

4 Results

4.1 DPO Training and Fine-tuning

Table 1 shows the accuracy and other metrics for BLIP-2 model and its BLIP-2-DPO. The BLIP-2 SFT model is available on HuggingFace at checkpoint 'sadmankiba/blip2-sft'. In SFT, the initial loss was

4.20 and it reduced to 2.25 at the end. Due to limited resource, we could not evaluate the SFT trained model. Table 2 shows the MMHal-benchmark of BLIP-2 and DPO-trained models. The DPO model performance was identical to base model on POPE benchmark and slightly worse in MMHal-bench. This is likely because the amount of training with DPO was insufficient.

Model	Adversarial	Random	Popular	All
BLIP-2 Base	0.71	0.73	0.74	0.73
BLIP-2 DPO	0.71	0.73	0.74	0.73

Table 1: Evaluation accuracy of BLIP-2 and DPO trained model on POPE benchmark

Model	Avg. Score (GPT4V)	Hallucination Rate
BLIP-2 Model	1.31	0.65
BLIP-2 DPO	1.25	0.66

Table 2: Performance Metrics for BLIP-2 on MMHAL-Bench

4.2 Self guidance

Model	Response Type	Accuracy
BLIP-2	Without Captions	0.31
BLIP-2	With Captions	0.47
OneVision	Without Captions	0.31
OneVision	With Captions	0.50

Table 3: Evaluation results on POPE with and without captions.

For the POPE dataset, we find that the captions do indeed help boost results. Captions improve performance for both the Blip and OneVision models. We think the likely reason for this is that the POPE benchmark only checks for if an object exists or not, and so captions often help the model ground the image better with the descriptions that it provides. A similar boost can be seen for the MMHal-Bench dataset, with the highest results being achieved by LLaVA-OneVision with captions, and also the highest score as rated by the LLM judge.

Response Type	Avg. Score (GPT4V)	Hallucination Rate
With Captions	1.85	0.51
Without Captions	1.54	0.58

Table 4: Performance Metrics for BLIP-2 Settings on MMHAL-Bench

Response Type	Avg. Score (GPT4V)	Hallucination Rate
With Captions	2.30	0.50
Without Captions	1.80	0.52

Table 5: Performance Metrics for LLaVA-OneVision (OV) Settings on MMHAL-Bench

5 Conclusion

In this project, we investigated methods to reduce hallucinations in vision-language models (VLMs) by implementing Fact-RLHF and feedback-guided self-revision techniques. Our experiments with models

such as BLIP-2 and BLIP-2, evaluated on benchmark datasets like MMHAL-Bench and POPE, demonstrated the potential of these methods to improve factual accuracy and alignment in multimodal outputs.

The Direct Preference Optimization (DPO) approach provided a straightforward and effective means of aligning outputs to human-like preferences, while the self-guided revision loop inspired by Volcano allowed models to iteratively refine their responses. Our augmentation of the revision process with generated factual captions showed promising improvements in mitigating hallucinations. Despite resource constraints, our custom adversarial dataset provided valuable insights into the limitations and robustness of the tested models.

6 Future Work

While our methods showed promising results, several opportunities for further research and improvements remain:

- **Scaling to Larger Models:** Apply the proposed techniques to larger models like GPT-4 Vision or multimodal LLaMA for broader applicability.
- **Exploring Adversarial Robustness:** Extend the adversarial dataset to include more complex examples and test its impact on hallucination resistance.
- **Fine-Tuning Techniques:** Investigate alternative fine-tuning approaches such as LoRA or parameter-efficient tuning to optimize resource usage.
- **Dynamic Feedback Mechanisms:** Develop adaptive feedback mechanisms that incorporate real-time critique from human annotators or external verification systems.
- **Expanding Evaluation Benchmarks:** Include more comprehensive evaluation datasets that test not only factual alignment but also reasoning and multimodal coherence.
- **Real-World Applications:** Integrate these methods into practical systems, such as visual question answering or interactive educational tools, to assess performance in real-world scenarios.

By addressing these areas, we hope to advance the development of reliable, robust, and hallucination-resistant vision-language models for diverse applications.

References

[1] S. Lee, S. H. Park, Y. Jo, and M. Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision, 2024.

[2] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer, 2024.

[3] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[4] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models, 2023.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

[6] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell. Aligning large multimodal models with factually augmented rlhf, 2023.