Mitigating Large Vision-Language Model Hallucination at Post-hoc via Multi-agent System

Chung-En (Johnny) Yu[1], **Brian Jalaian**[1], **Nathaniel D. Bastian**[2]
[1]University of West Florida
[2]United States Military Academy
cy31@students.uwf.edu, bjalaian@uwf.edu, nathaniel.bastian@westpoint.edu

Abstract

This paper addresses the critical issue of hallucination in Large Vision-Language Models (LVLMs) by proposing a novel multi-agent framework. We integrate three post-hoc correction techniques: self-correction, external feedback, and agent debate, to enhance LVLM trustworthiness. Our approach tackles key challenges in LVLM hallucination, including weak visual encoders, parametric knowledge bias, and loss of visual attention during inference. The framework employs a Plug-in LVLM as the base model to reduce its hallucination, a Large Language Model (LLM) for guided refinement, external toolbox models for factual grounding, and a agent debate system for consensus-building. While promising, we also discuss potential limitations and technical challenges in implementing such a complex system. This work contributes to the ongoing effort to create more reliable and trustworthy multimodal multi-agent systems.

Introduction

The advent of Large Language Models (LLMs) has marked a significant advancement towards Artificial General Intelligence (AGI). Recent research has primarily focused on enhancing various capabilities of LLMs, including text generation, complex reasoning, and tool utilization (Zhao et al. 2023). While initial studies were predominantly text-centric, the field rapidly evolved with the introduction of Multimodal Large Language Models (MLLMs), which are capable of processing images, videos, and other diverse forms of data. MLLMs are typically architected with a modality encoder coupled with a powerful LLM, enabling them to tackle complex multimodal tasks such as visual dialogue and captioning (Yin et al. 2023a).

Despite the remarkable potential these models offer, ensuring their trustworthiness remains a critical and pressing challenge in the field of artificial intelligence (AI). The trustworthiness of LLMs is a multifaceted and intricate issue, encompassing various dimensions including truthfulness, fairness, and robustness, etc., as defined in (Sun et al. 2024). A crucial aspect of trustworthiness is truthfulness, particularly concerning the phenomenon of hallucination in LLMs. Mitigating hallucination not only enhances trustworthiness but also indirectly contributes to improving reasoning abilities, thereby expanding the potential applications of LLMs (Zhang et al. 2023). Furthermore, with the advent of MLLMs, the research community has increasingly focused on the trustworthiness of Large Vision-Language Models (LVLMs) (Yin et al. 2023a). Addressing hallucination in LVLMs can yield substantial benefits for a wide range of downstream visual tasks. It is worth noting that hallucination mitigation techniques developed for LVLMs have the potential to be extended and adapted to other modalities like video, further emphasizing the importance of this research direction (Bai et al. 2024).

Hallucination in LLMs is defined as the generation of outputs that either contradict or cannot be verified from the source content (Zhang et al. 2023). In contrast, hallucination in LVLMs primarily manifests as a discrepancy between the model's outputs and the provided visual content (Bai et al. 2024). This fundamental difference in the nature of hallucination between LLMs and LVLMs means that most LLM hallucination mitigation techniques cannot be directly applied to address LVLM hallucination. However, studies (Zhang et al. 2023; Bai et al. 2024) have revealed that, despite the different modalities, LVLM hallucination mitigation techniques often share similar approaches to those used in LLMs. Among the various mitigation methods, post-hoc correction, which addresses hallucination at the inference phase, stands out due to its effectiveness in enhancing reasoning ability and reducing mitigation efforts while saving time and resources comparing to training paradigms. Moreover, post-hoc correction typically employs a plug-and-play framework, allowing for flexible adaptation to different models. This flexibility and efficiency make it an particularly attractive approach for addressing hallucination in both LLMs and LVLMs.

Interestingly, many post-hoc correction strategies, such as self-correction and external feedback, inherently embody the principles of multi-agent systems in both LLMs and MLLMs. In the context of AI, a multi-agent system leverages additional AI agents to complete complex tasks. Recent research (Xi et al. 2023) has demonstrated the efficacy of multi-agent systems in mitigating hallucination and their adaptability to diverse scenarios. Furthermore, a comprehensive survey (Wang et al. 2024) highlights the emerging trend of multi-agent systems in both research and industry communities across various downstream tasks and domains.

We've explored post-hoc correction methods and suggest a new strategy to reduce LVLM hallucination, utilizing a multi-agent framework in line with principles for mitigating LLM hallucination. Our approach seeks to connect mitigation techniques across LLM and LVLM models, providing a versatile solution to improve their trustworthiness.

Related Work

Post-hoc correction techniques are primarily divided into three categories: self-correction, external feedback, and agent debate (Pan et al. 2023). These methods tackle hallucination from data and model perspectives, with advanced multi-agent frameworks such as Self-Checker (Li et al. 2023a) utilizing both aspects concurrently. *Self-correction* involves a base model generating an initial output and then refining it. This technique typically requires a powerful LLM to generate follow-up questions based on the model's initial outputs, guiding the base model in refining its responses. The approach has demonstrated effectiveness in mitigating hallucination in both LLMs and LVLMs, as evidenced by studies such as CoVe (Dhuliawala et al. 2023) and LogicCheckGPT (Wu et al. 2024). However, this method often sacrifices efficiency due to its iterative nature. Additionally, the quality of refinement is constrained by the inherent limitations of LLMs and LVLMs, such as the inability to access up-to-date or factual information, or the risk of inheriting parametric knowledge biases from more powerful LLMs in the feedback loop. *External feedback* leverages an agent system that utilizes external tools such as code interpreters, logic reasoners, external knowledge sources, or task-specific well-trained models in the feedback loop. This approach corrects the base model's outputs using retrieved factual evidence. It has emerged as one of the mainstream techniques for mitigating hallucination at the post-hoc stage in both LLMs and LVLMs. Related studies, including CRITIC (Gou et al. 2023) and Woodpecker (Yin et al. 2023b), have demonstrated the efficacy of external feedback in enhancing model trustworthiness. *Agent debate* employs multiple LLM-based agents and facilitates a debate among them regarding their individual answers over several rounds, aiming to reach a consensus. Studies such as LM vs LM (Cohen et al. 2023) and MARDA (Wang et al. 2023) have showcased the success of this approach in enhancing reasoning ability and reducing LLM hallucination through debate.

Although much research has centered on self-correction and external feedback (Pan et al. 2023), to the best of our knowledge, agent debate has been underutilized in mitigating LVLM hallucination. Studies such as (Li et al. 2022) and (Zeng et al. 2022) show that agent debate can enhance visual reasoning, closely linked to reduced hallucination. We believe that agent debate could be effectively used to address LVLM hallucination.

A recent study (Xu, Jain, and Kankanhalli 2024), which formulates the LLM hallucination problem using mathematical equations, argues that self-correction alone is insufficient to eliminate hallucination for all tasks by simply modifying prompts and expecting the LLM to automatically prevent hallucination. External feedback leveraging knowledge and tools is potentially an effective mitigator of hallucination. Agent debate, while potentially reducing hallucination, cannot eliminate it entirely. These insights inspire us to propose a multi-agent system that leverages the strengths of all three post-hoc correction techniques.

By combining self-correction, external feedback, and agent debate within a single multi-agent framework, we expect significant reductions in LVLM hallucination. This integrated approach has the potential to address the limitations of each individual technique while capitalizing on their respective strengths. Such a comprehensive framework could offer a more robust and effective solution for enhancing the trustworthiness of MLLMs and their underlying LLMs, potentially setting a new standard for hallucination mitigation in both language and vision-language models.

Proposed Approach

Our proposed multi-agent framework integrates three post-hoc correction techniques: self-correction, external feedback, and agent debate, as illustrated in Fig. 1. It is designed to effectively mitigate LVLM hallucination through a series of interconnected components. *Plug-in LVLM*: This base model generates initial and follow-up responses based on the initial message and follow-up questions from the Supervisor. *Supervisor*: A powerful LLM that generates follow-up questions based on the Plug-in LVLM's initial responses, facilitating a self-correction loop. Additionally, the Supervisor also formulates validation questions for the toolbox models. *Toolbox Models*: These serve as the external feedback mechanism in the system: *Open-set Object Detector*: A model that recognizes open-world object categories and its presence in an image, e.g., GroundingDINO (Liu et al. 2023). *VQA Model*: A LVLM that validates object attributes, such as colors and quantities, and object relations within an image. *Data Pool*: It stores responses from both self-correction and external feedback mechanisms, which then serve as information for subsequent debates. *Agent Debate System*: It comprises two LLM agents, *Debater 1* and *Debater 2*, which autonomously discuss the information in the data pool to reach a consensus and finalize the response that best describes the image.

The finalized responses in our system cover three crucial aspects: object recognition, attributes, and relations in the image, consistent with LVLM hallucination definitions (Bai et al. 2024). This comprehensive response addresses all potential hallucination issues in LVLMs. Importantly, our framework is versatile, capable of generating outputs that suitable for both discriminative and generative LVLM hallucination benchmarks like POPE (Li et al. 2023b) and FAITHScore (Jing et al. 2023), allowing extensive evaluation and validation across different metrics.

Discussion

Our multi-agent framework is meticulously designed to address several key challenges in LVLM hallucination, as identified in (Bai et al. 2024). In LVLMs, the visual encoder is typically weaker than the backbone language model, which is a primary contributor to LVLM hallucination. This dis-

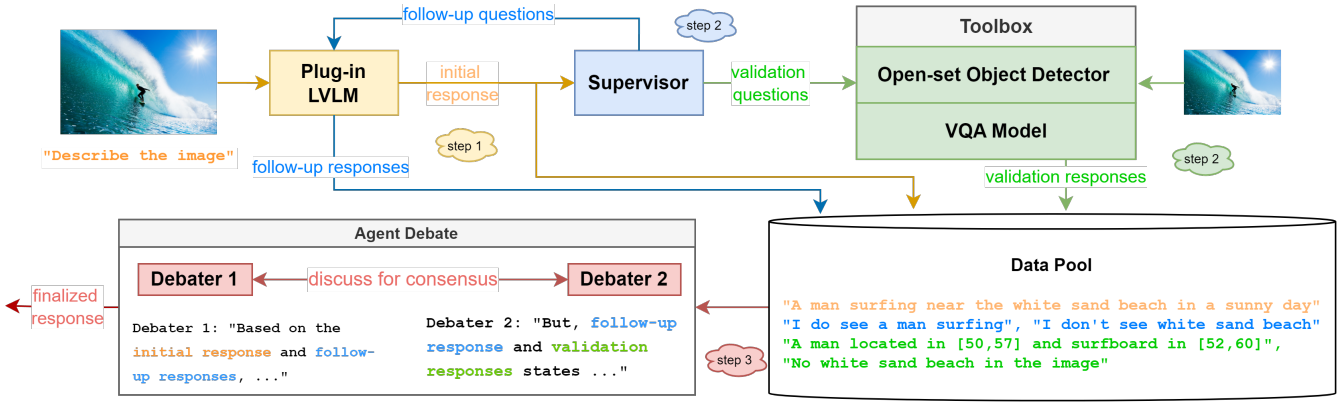

Figure 1: Our proposed multi-agent framework.

parity often results in the LVLM suffering from parametric knowledge bias inherent in the larger language model. For instance, when presented with an image of a "yellow cherry," the language model might hallucinate during decoding due to its prior knowledge that cherries are typically red. To address this, we employ the Supervisor to guide the backbone language model in enhancing its visual reasoning abilities. By generating fine-grained follow-up questions based on initial responses, the Supervisor prompts the Plug-in LVLM to produce more detailed and visually-grounded outputs. This iterative process helps mitigate statistical biases (e.g., typical car has four wheels), reduces the influence of language model priors (e.g., typical cherry is red), and maintains long visual attention throughout the inference process.

The incorporation of external tools is crucial in solidifying the truthfulness of visual descriptions, compensating for the typically weaker visual encoder in LVLMs. Our approach leverages state-of-the-art (SOTA) models in object detection to retrieve more accurate and factual visual information. This external feedback mechanism serves as a reliable counterbalance to the language model's prior knowledge, ensuring that the system's outputs are firmly grounded in the actual visual content. By providing concrete, tool-derived visual information, we mitigate the LVLM potentially losing attention on the relevant visual elements throughout the inference process, thus reducing the likelihood of hallucination.

Even when provided with references, LLMs have a tendency to generate responses that may not closely adhere to the given information, especially when the references are diverse or complex. In our framework, the references consist of visual information stored in the Data Pool. To address this challenge and further enhance visual reasoning capabilities, we employ an agent debate technique. This approach involves multiple LLM agents engaging in a structured discussion about the visual information, working towards a consensus. By fostering this debate, we aim to mitigate the impact of individual language model priors and encourage a more comprehensive and accurate interpretation of the visual data.

Our multi-agent framework offers potential in addressing LVLM hallucination but introduces challenges, notably increased inference times due to the system's complexity. The agent debate process, essential for consensus, may extend computation time significantly. Additionally, prompt quality crucially affects the system, potentially triggering hallucinations. We suggest employing formal methods to develop more robust prompts (Jha et al. 2023a,b).

Implementing a multi-agent system of this complexity presents several technical challenges that must be addressed. One key issue is ensuring seamless communication and coordination between the various components, particularly between the Plug-in LVLM, Supervisor, toolbox models, and debate agents. This requires implementing efficient agent development framework to manage the flow of information and tasks across the system. Another significant challenge lies in maintaining consistency and coherence in the final output, given the diverse sources of information and potential discrepancies between agent perspectives. To tackle this, we propose implementing a sophisticated consensus mechanism that not only aggregates but also reconciles conflicting information from different agents. Additionally, the system must be designed with scalability in mind, allowing for the integration of new tools or models as they become available, without requiring a complete overhaul of the existing architecture. This flexibility is crucial for the long-term viability and adaptability of the framework in the rapidly evolving field of machine learning.

Conclusion

We believe that our proposed plug-and-play framework represents a significant step forward in enhancing the trustworthiness of LVLMs. As the field of AI continues to evolve rapidly, approaches like ours will play a crucial role in building more adaptable, reliable and trustworthy multimodal AI systems. Future work should focus on optimizing the system's efficiency, refining the consensus mechanisms, and ensuring adaptability to accommodate emerging tools and models. By addressing these challenges, we can further improve the framework's effectiveness in mitigating hallucination and enhancing the overall performance of LVLMs across a wide range of applications.

Acknowledgments

This work was supported in part by the U.S. Military Academy (USMA) under Cooperative Agreement No. W911NF-23-2-0108. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of Defense, or U.S. Government.

References

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Gou, Z.; Shao, Z.; Gong, Y.; Shen, Y.; Yang, Y.; Duan, N.; and Chen, W. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Jha, S.; Jha, S. K.; Lincoln, P.; Bastian, N. D.; Velasquez, A.; and Neema, S. 2023a. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, 149–152. IEEE.

Jha, S. K.; Jha, S.; Lincoln, P.; Bastian, N. D.; Velasquez, A.; Ewetz, R.; and Neema, S. 2023b. Counterexample guided inductive synthesis using large language models and satisfiability solving. In *MILCOM 2023-2023 IEEE Military Communications Conference (MILCOM)*, 944–949. IEEE.

Jing, L.; Li, R.; Chen, Y.; Jia, M.; and Du, X. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Li, M.; Peng, B.; Galley, M.; Gao, J.; and Zhang, Z. 2023a. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Li, S.; Du, Y.; Tenenbaum, J. B.; Torralba, A.; and Mordatch, I. 2022. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; and Wang, W. Y. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Wang, H.; Du, X.; Yu, W.; Chen, Q.; Zhu, K.; Chu, Z.; Yan, L.; and Guan, Y. 2023. Apollo's Oracle: Retrieval-Augmented Reasoning in Multi-Agent Debates. *arXiv preprint arXiv:2312.04854*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wu, J.; Liu, Q.; Wang, D.; Zhang, J.; Wu, S.; Wang, L.; and Tan, T. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023a. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023b. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.