Towards Dynamic and Realistic Evaluation of Multi-modal Large Language Model

Yijiang Li*, Huiqi Zou*, Ziang Xiao
UC San Deigo, Johns Hopkins University
yijiangli@ucsd.edu, {hzou11, ziang.xiao}@jhu.edu

Abstract

Hallucinations in the multimodal domain occur when a model provides information that contradicts the content of an image. Existing benchmarks for evaluating hallucinations in current Multi-modal Large Language Models (MLLMs) often adopt a static captioning or question-answering format, which deviates from the realistic use of the MLLMs in downstream tasks. To address these limitations, we propose to evaluate MLLMs in a dynamic and multi-turn way to close the realism gap. Our approach involves an evaluator agent engaging in evaluative conversation, generating diverse contextually relevant questions and their follow-ups. The framework features a context module and a question generation module, enabling the evaluator to mimic human-like questioning in real work tasks while assuring the quality of the evaluation. Comparison with single-turn question-answering or captioning evaluation methods demonstrates that our approach identified more hallucination issues in VLMs and covered more aspects of the visual content.

1 Introduction

The training paradigms for aligning visual and textual features unlock greater potential for vision-language models (VLMs) in multimodal tasks such as captioning, visual question answering (VQA) and visual reasoning (Dai et al., 2023; Driess et al., 2023; Gong et al., 2023; Liu et al., 2023; Zhu et al., 2023). Despite the impressive performance of VLMs, models tend to either produce contents that conflict with the image or fail to identify specific elements. Hallucination, an issue with generative models (Liu et al., 2024), limits the broader applicability of VLMs. To enhance their reliability, there is a need for hallucination evaluation that can reliably identify errors in the models' responses.

The current method for VLM hallucination relies on static benchmarks which are divided into two categories. The model is evaluated on either discriminative tasks, i.e., asking the model if an object is present in the image (Li et al., 2023), or generative tasks, i.e., asking the model to generate image descriptions (Gunjal et al., 2024; Qiu et al., 2024). Compared to the discriminative task, the generative task, although more costly in the human annotation, offers a more in-depth exploration of relationships between objects. However, both methods suffer from the following issues (Liao and Xiao, 2023). First, there is a gap between the questions asked in the benchmarks and how the model is being deployed in downstream tasks. Second, the current benchmarks often represent single-turn interaction, which can not capture real-world multi-turn user interaction. Third, the static nature of these benchmarks raises concerns about data contamination, which can undermine their reliability.

To address the issues above, we propose a dynamic and realistic evaluation framework based on multimodal large language models (MLLM). Our method dynamically evaluates VLMs by conducting an interrogative multi-turn conversation with the goal of emulating user interactions. The evaluator probes model hallucinations through a series of questions that are contextualized in real-world downstream tasks and follow-ups generated on the fly. To assess the effectiveness of our approach, we conduct evaluations on LLaVA (Liu et al., 2023). Preliminary results indicate that compared to caption generation, our methods identified more hallucination issues based on the CHAIR (Rohrbach et al., 2018), coverage and human evaluations.

2 Methodology

To enable a dynamic and realistic evaluation of VLM hallucinations through multi-turn conversations, our framework involves two main components to create an evaluator agent: a context module that takes consideration of down-stream application

*These authors contributed equally to this work.

and user interaction and a question generation module that dynamically generates context-relevant and challenging questions to probe for model hallucination. The evaluator agent needs to consider every aspect of a given image and process texts with contextual information. Given the strong capabilities of MLLMs in cross-modal understanding, we adopt one as the back-end model.

2.1 Evaluator Design

The evaluator design includes a system message that introduces a formatted instruction outlining the task requirements. During each evaluation round, the captions and bounding boxes of the test image are provided in the instructions as reference information. The basic requirements, which describe the evaluation workflow, emphasize naturalness and contextual awareness. To address potential hallucinations related to objects, attributes and relationships, we include a coverage requirement that encourages the MLLM to ask diverse and detailed questions. Additionally, examples of poor questions are provided to help the evaluator avoid undesirable cases, such as directly revealing ground truth information to the evaluated model.

2.2 Context Module

To emulate user interaction and real-world application scenarios, we design the context module for contextual role-playing. The module dynamically assigns the evaluator a context and a user persona with a specific image-querying objective and an initial question, outlining the evaluator's representative traits and motivations. For example, a persona might be: "`As a blind person, I want to navigate using images for orientation.`" Accordingly, the evaluation conversation will focus on identifying landmarks to facilitate navigation.

2.3 Question Generation Module

This module is responsible for generating the unanswerable and adversarial questions that can probe potential hallucinations.

Unanswerable questions, which cannot be resolved using the given image, are expected to be rejected by the evaluated VLM. However, VLMs tend to respond based on statistical correlations rather than grounded reasoning (Zhang et al., 2023). To test the robustness of VLMs, we prompt the evaluator to generate such questions by replacing objects and adjectives in the ground-truth image information with their antonyms and avoiding existence or yes-no questions. At the prompt's end, we include some unanswerable questions, along with explanations for their selection, to serve as demonstrations.

Evaluator	CHAIRs	CHAIRi	Coverage	Human
Caption	46	22.9	83.4%	72%
Dynamic	77.5	31.4	96.1%	76%

Table 1: Model performances with different evaluators. The results indicate that hallucination rate is higher and more objects are mentioned when evaluated by our dynamic evaluator, compared to the caption-based method.

Adversarial questions are answerable but are specifically designed to elicit hallucinations deliberately. We incorporate words associated with image objects from the WordNet synset (Princeton University, 2010) into the prompt as additional information, and instruct the evaluator to use them for question generation. The instructions also include examples of challenging questions intended to induce mistakes.

3 Preliminary Experiments and Results

Considering the need to test the evaluator across different application scenarios, we select testing samples from the validation set of COCO (Lin et al., 2014), an object recognition dataset known for its diverse object types and labeled instances. GPT-4o (OpenAI, 2024) serves as the evaluator's back-end model. Appendix A shows a simulated conversation example with LLaVA (Liu et al., 2023).

Table 1 shows that the model achieves higher scores on CHAIR (Rohrbach et al., 2018) and coverage (where coverage refers to the percentage of instances from the ground truth annotations that appear in the evaluated VLM's responses), as well as in human evaluations when prompted with our method. This suggests an increase in hallucinations and a broader range of objects mentioned when the model is evaluated using our dynamic evaluator.

4 Future Works

In this work, we propose a novel framework that dynamically evaluates model hallucination through an evaluative multi-turn conversation. To close the realism gap, our methods account for downstream task scenarios and user interaction. In the next phase, we will conduct a more comprehensive evaluation of our framework by testing multiple aspects of its validity and reliability. By doing so, we could offer a more effective evaluation method that can better inform downstream applications.

References

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49250–49267.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488.

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. *Preprint*, arXiv:2305.04790.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

OpenAI. 2024. Hello GPT-4o.

Princeton University. 2010. About WordNet.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv preprint arXiv:2404.13874*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Yuwei Zhang, Chih-Hui Ho, and Nuno Vasconcelos. 2023. Toward unsupervised realistic visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15613–15624.

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

A Simulated Conversation Sample

Evaluator: Could you provide a brief description of this image?

LLaVA: The image depicts a busy city street with several people walking around. A woman is walking down the sidewalk, and a man is pushing a shopping cart filled with items. There are other people walking in the background, and a few individuals are standing on the sidewalk. A traffic light is visible in the scene, indicating that the street is likely a busy intersection. A handbag can be seen in the image, possibly belonging to one of the pedestrians. The overall atmosphere of the scene is bustling and lively, typical of a city street.

Evaluator: And can you tell more about the large red bus that is seen parked on the street?

LLaVA: The large red bus is parked on the side of the street, and it appears to be a public transit bus. It is likely used for transporting passengers within the city, providing a convenient and affordable means of travel for residents and visitors. The bus's presence in the image suggests that the area is well-connected by public transportation, which can help reduce traffic congestion and promote a more sustainable urban environment.

Evaluator: I see. Can you also describe the activities of the kids that are playing soccer in the park seen in the image?

LLaVA: In the image, there are two kids playing soccer in the park. They are actively engaged in the game, kicking the soccer ball around and enjoying their time outdoors. The presence of the soccer ball and the children's involvement in the game suggest that they are having fun and staying physically active, which is an important aspect of their overall well-being.

Table 2: An example of a simulated conversation with three turns.