Review of: "Trust but Verify: Programmatic VLM Evaluation in the Wild"

Xiangtai Li[1]

1 Nanyang Technological University, Singapore

Potential competing interests: No potential competing interests to declare.

Review: Trust but Verify: Programmatic VLM Evaluation in the Wild

Strengths:

1. The paper is well-motivated and addresses a critical gap in the evaluation of Vision-Language Models (VLMs). The proposed benchmark (Programmatic VLM Evaluation, PROVE) is useful and explainable to our community.

2. The paper is clearly written. Additionally, the inclusion of illustrative examples, such as Figure 4, aids in clarifying these concepts.

3. The proposed method shows effectiveness on various inputs.

4. The benchmark methods are solid and good enough to measure current state-of-the-art MLLM methods.

Weaknesses:

1. The major weakness is that there are no comparisons with existing metrics/benchmarks, in particular for explainable VLM benchmarks. There are so many VLM hallucination benchmarks. I do not find any comparison with existing hallucination benchmarks.

2. The second major concern is that the authors do not present a simple or effective method to improve the truthfulness and helpfulness of the model. Moreover, there are no discussions on further directions. Both make the submission weak.

3. The related works section is not good. The authors should cite and discuss more scene graph generation works.

[-] Unbiased scene graph generation from biased training, CVPR-2020

[-] Auto-encoding scene graphs for image captioning, CVPR-2019

[-] Panoptic Scene Graph Generation, ECCV-2022

[-] 4D Panoptic Scene Graph Generation, NeurIPS-2024

[-] Panoptic Video Scene Graph Generation, CVPR-2023