Trust but Verify: Reliable VLM evaluation in-the-wild with program synthesis

Viraj Prabhu, Senthil Purushwalkam, An Yan, Jieyu Zhang, Caiming Xiong & Ran Xu
Salesforce AI Research
{viraj.prabhu,spurushwalkam,an.yan,cxiong,xurantju}@salesforce.com

Abstract

Vision-Language Models (VLMs) often generate plausible but incorrect responses to visual queries. However, reliably quantifying the effect of such hallucinations in free-form responses to open-ended queries is challenging as it requires visually verifying each claim within the response. We propose Programmatic VLM Evaluation (PROVE), a new benchmarking paradigm for evaluating VLM responses to open-ended queries. To construct PROVE, we provide a large language model (LLM) with a high-fidelity scene-graph representation constructed from a hyper-detailed image caption, and prompt it to generate diverse question-answer (QA) pairs, as well as programs that can be executed over the scene graph object to *verify* each QA pair. We thus construct a benchmark of 10k challenging but visually grounded QA pairs. Next, to evaluate free-form model responses to queries in PROVE, we propose a *programmatic* evaluation strategy that measures both the helpfulness and truthfulness of a response within a unified scene graph-based framework. We benchmark the helpfulness-truthfulness trade-offs of a range of VLMs on PROVE, finding that very few are in-fact able to achieve a good balance between the two. Project page: `https://prove-explorer.netlify.app/`.

1 Introduction

Vision-language models (VLMs) have emerged as an effective solution for generating responses to queries about visual content. However despite impressive progress (and much like their LLM-cousins) VLMs are still known to hallucinate – to generate plausible but incorrect answers that are either inconsistent or unverifiable against the provided visual context. This crucial shortcoming has the potential to erode trust in such systems and has already begun to attract research [24, 13, 7, 10] and regulatory [3] interest, particularly as using such models as the "foundation" of various high-stakes applications becomes imminent [4].

This has led to a flurry of research on reliably benchmarking VLMs by measuring not just the helpfulness but also the *truthfulness* of responses [19, 11, 22, 9, 16, 12, 20, 14, 13, 7]. Existing benchmarks either evaluate the model's discriminative responses to existence-based queries, or generative responses to open-ended questions. Discriminative benchmarks comprise of *close-ended questions* (typically yes/no questions) that ease evaluation but do not simulate real-world use. Generative benchmarks, on the other hand, include open-ended queries (*eg.* "describe this image") but resort to *open-ended evaluation*, using external models (typically a proprietary LLM) to score responses given some context (typically ground-truth image annotations). However we find that in several such benchmarks, the context provided is completely insufficient to judge if the response contains hallucinations. For example, for a query like "How color is the dog?" and with a ground truth answer "brown", a VLM might respond with "The dog has a dark brown shiny coat". This response contains several details ("dark" and "shiny") that cannot be easily verified against the ground-truth by an LLM. Furthermore, the absence of a clear scoring rubric coupled with the sensitivity of LLMs to minor prompt differences, often leads to inconsistent and arbitrary scores – see Fig. 1.

We propose PROVE , a new benchmark for evaluating VLM hallucinations that performs reliable and interpretable *close-ended evaluation* of responses to diverse, grounded, and unambiguous *open-ended*

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

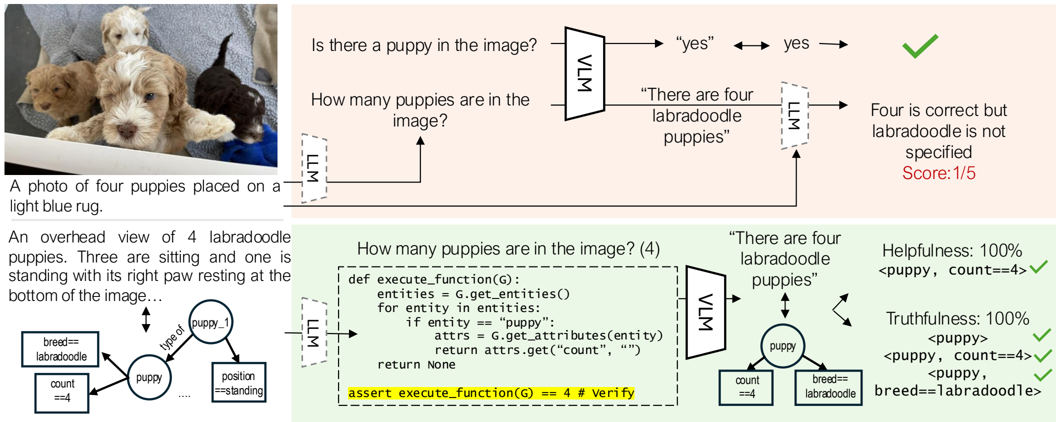

Figure 1: **Top.** Existing VLM benchmarks either limit query-types to easy-to-evaluate but restrictive binary questions, or use external LLMs to generate open-ended questions (without verifying their validity) and score answers (often without complete image context or a clear scoring rubric). **Bottom.** We propose PROVE, a new benchmark that constructs high-fidelity scene-graph representations from hyper-detailed image captions, that are queried via an LLM-generated program to verify a free-form generated question-answer pair. At test-time, we perform an interpretable and close-ended evaluation of the helpfulness and truthfulness of VLM responses by comparing scene-graphs.

questions. To do so, we first use hyper-detailed image captions to construct a high-fidelity scene graph representation of the image. We then use an LLM to generate a diverse set of open-ended question-answer pairs that test a range of model capabilities while simulating real-world use. To ensure that the questions are grounded in visual content, we prompt an LLM to generate Python *code* that can be executed to verify the QA pair (using predefined scene graph functions). We only retain the QA pairs that we can programmatically verify. We thus construct a benchmark of 10k examples that we use to perform close-ended and interpretable evaluation of the helpfulness and truthfulness of responses from a range of VLMs. We demonstrate the superiority of PROVE over existing benchmarks in terms of reliability, interpretability, and scalability.

2 Related work

Discriminative benchmarks generate a series of binary questions to verify the presence (or absence) of various entities (or distractors) in the image. Early benchmarks like POPE [12] limited their scope to object entities annotated by humans or external off-the-shelf models [25], while generating distractor entities using various strategies. Follow-up works expand the scope to additionally evaluate responses to *negative presence* queries [16] or using an LLM to generate a broader range of existence-based questions covering objects and their attributes [9]. However, while the binary questions that typify such benchmarks simplify evaluation, they do not realistically simulate in-the-wild use.

Generative benchmarks instead evaluate model hallucinations in response to free-form questions. CHAIR [19] measures the precision and recall of entities mentioned in a generated image description against the ground truth. HaELM [23] additionally uses a large language model (LLM) to judge generations, whereas M-HalDetect [7] has humans annotate hallucinations in model generated descriptions and a predictive model. Recently, AMBER [22] combines a POPE style evaluation with a generative evaluation over an open-ended split. While these benchmarks are indeed more realistic, they still restrict the query instruction to image captioning ("Describe this image in detail.") and do not stress-test performance in response to truly free-form queries.

Most recently, a few benchmarks with truly open-ended queries have been proposed [20, 13, 11, 14], which either hand-design or use an LLM to generate free-form questions, and use external models to judge the corresponding responses. However, these too have limitations: MMHal [20] and HallusionBench [13] rely on a series of off-the-shelf models which introduce noise. GAVIE's [14] reliance on dense captions and bounding boxes leads to a majority of questions querying localized image regions and spatial relationships, many of which have unnatural-sounding responses (*eg.* mentioning image coordinates). Finally, GPT-4-based evaluation is both expensive and confounded by the model's own limitations.

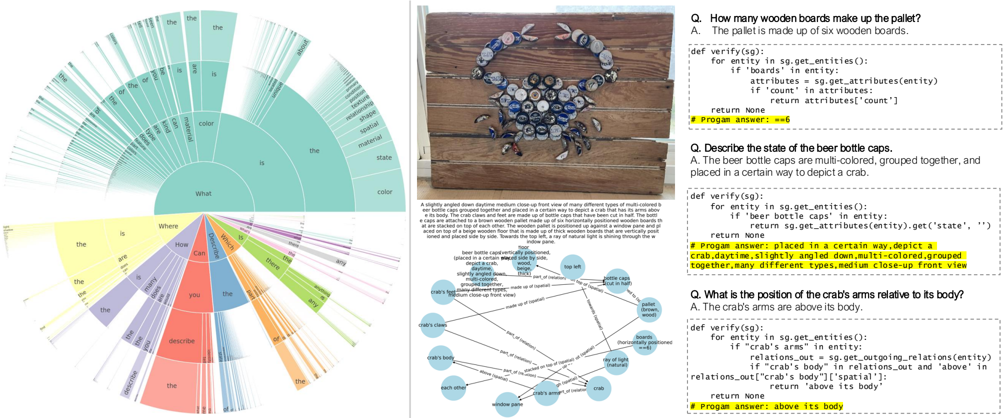

Figure 2: The PROVE benchmark (v0.1). **Left.** Sunburst visualization of the first 4 question words. **Right.** Example scene-graph, QA-pairs, and verification programs generated for a sample image.

3 Progammatic VLM Evaluation (PROVE)

Vision-language models are trained to respond to a question $\mathcal{Q}$ about an image $\mathcal{I}$ with a ground-truth answer $\mathcal{A}$. Let $\mathrm{m}_\theta(.)$ denote a VLM model trained on a large dataset of such ($\mathcal{I}$, $\mathcal{Q}$, $\mathcal{A}$) triplets. At test time, we wish to evaluate the model response $\hat{\mathcal{A}}$=$\mathrm{m}_\theta(\mathcal{Q}, \mathcal{I})$. Specifically, while prior work typically evaluates either the response's correctness (is $\hat{\mathcal{A}}$=$\mathcal{A}$) or truthfulness (is $p(\hat{\mathcal{A}}|\mathcal{I}) >$ threshold), we propose a unified framework that jointly evaluates both and captures the tradeoff between the two.

To do so, we first download image-caption pairs ($\mathcal{I}$, $\mathcal{C}$) from the recently proposed DOCCI [17] dataset. This dataset contains 15k manually curated images with comprehensive human-annotated descriptions. DOCCI is particularly well-suited source for VLM evaluation because: i) its captions are extremely detailed, with a higher median caption length than competing datasets, which correlates with high image recall ii) its comprehensive and rigorous 3-stage human annotation protocol leads to high-fidelity captions that are suitable to test a range of image understanding challenges including spatial reasoning, counting, text rendering, and compositionality, and iii) its images are newly curated and so are truly held-out data for existing VLMs.

Building a robust scene-graph representation. Following Cho *et al.* [6], we first prompt an LLM to extract entity (<entity>), attribute (<entity, attribute>), and relationship (<entity_1, attribute, entity_2>) tuples from the image caption, that we use to construct a scene graph representation $\mathrm{g}(\mathcal{C})$ as a directed graph with attributed entities as nodes and relationships as edges. The scene graph is implemented as a Python class with methods to query the graph for its entities, attributes, and relationships, as well as to extract and describe subgraphs in natural language.

Generating open-ended questions with verifiable answers. Next, we prompt a pre-trained LLM to generate challenging, diverse, and unambiguous question-answer (QA) pairs from a provided caption and scene graph, alongwith an accompanying Python program that accepts the scene graph as input and can be executed to verify the generated QA pair [8, 21]. The prompt includes a few examples of such scene-graph and QA+program input/output pairs to guide the model. Finally, we execute the generated program on the scene graph as a unit test to verify the QA pair – if the program fails or returns an answer that is semantically different from the ground truth answer, it is discarded. We also keep track of the subgraph visited by the program for each succesful verification, the size of which we use as a proxy for the *complexity* of the QA pair. We repeat this procedure to create a benchmark of open-ended image+QA pairs $\{(\mathcal{I}_i, \mathcal{Q}_i, \mathcal{A}_i)\}_{i=1}^{N}$ that are amenable to closed-form evaluation.

Dataset statistics. We now present some statistics about PROVE v0.1, which comprises of 10k QA pairs generated from 908 image-caption pairs from the DOCCI test set, with an average of 11 QA pairs per image. These are obtained after filtering out QA pairs with invalid verification programs (18.9% of the total generated) or whose programmatic answers differ semantically from the ground truth answer (9.9% of the total generated). Questions average 10.1 words in length whereas answers average 11.8 words. In Fig. 2 we present a sunburst visualization of the first 4 words in the questions; as seen, our benchmark is diverse and spans a wide range of question types.

Method	Simple			Complex			Full		
	$\mathbb{H}$	$\mathbb{T}$	Avg.	$\mathbb{H}$	$\mathbb{T}$	Avg.	$\mathbb{H}$	$\mathbb{T}$	Avg.
Phi-3 Vision (4B) [1]	60.6	67.7	64.1	56.8	65.3	61.0	60.0	67.3	63.6
LLaVA-1.5 (7B) [15]	60.9	**68.1**	64.5	57.1	**66.4**	61.7	60.3	**67.8**	64.1
GPT-4o-mini* [2]	61.9	63.8	62.9	60.4	63.8	62.1	61.7	63.8	62.7
GPT-4o* [2]	**68.5**	66.8	**67.7**	**63.9**	65.5	**64.7**	**67.7**	66.6	**67.2**

Table 1: Results of benchmarking four different VLMs on `PROVE` v0.1 (*=closed source model).

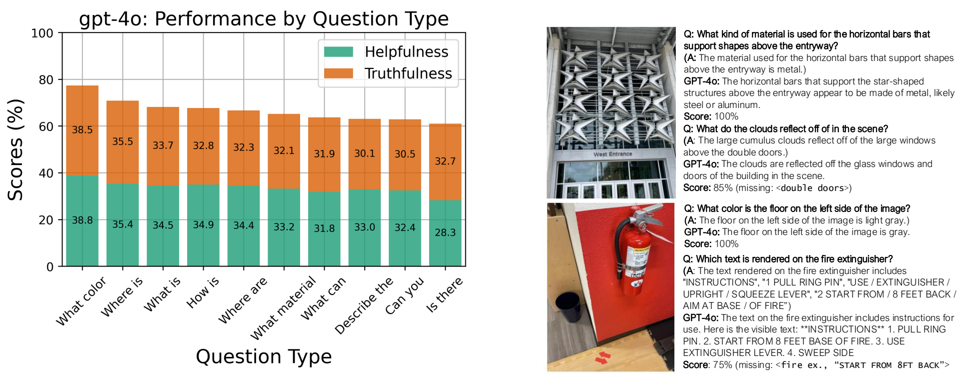

Figure 3: GPT-4o Performance Analysis. **Left.** Score by question type. **Right.** Example responses.

4 Benchmarking VLMs with `PROVE`

Closed-form response evaluation. After ensuring the validity of the generated QA pairs, we proceed to evaluating each VLM response $\hat{\mathcal{A}}=\mathrm{m}_\theta(\mathcal{Q}, \mathcal{I})$. We first extract tuples and build a scene graph representation $\mathrm{g}(\hat{\mathcal{A}})$. We then measure response *helpfulness* $\mathbb{H}(.)$ based on *recall* of the ground truth answer, by computing the fraction of ground truth answer tuples that are entailed by the model response (using a text-entailment model [18]). We also report *truthfulness* $\mathbb{T}(.)$ by computing response *precision i.e.* the fraction of response tuples that are entailed either by the original caption *or* the image itself (using a visual entailment model [5])[1] Let $\models$ denote entailment. We define:

$$\mathbb{H}(\hat{\mathcal{A}}) = \frac{|\{t \in \mathrm{g}(\mathcal{A}) \mid \hat{\mathcal{A}} \models t\}|}{|\mathrm{g}(\mathcal{A})|}; \quad \mathbb{T}(\hat{\mathcal{A}}) = \frac{|\{t \in \mathrm{g}(\hat{\mathcal{A}}) \mid \mathcal{C} \models t \vee \mathcal{I} \models t\}|}{|\mathrm{g}(\hat{\mathcal{A}})|} \tag{1}$$

We report both these metrics as well as their average across the full dataset as well as simple (requiring visiting <=3 nodes of the graph to answer) and complex subsets. Note that the two metrics are not necessarily correlated – a response can be helpful (by answering the query) but not entirely truthful (might contain hallucinations), and vice versa. Naturally, different models and mitigation strategies may lead to varying tradeoffs between the two – an aspect that `PROVE` is uniquely suited to analyze.

Findings. Table 1 presents evaluation results. We compare the performance of a few representative VLMs of varying sizes. We find that while strong models such as GPT-4o [2] do indeed perform best across data splits, this is driven by significantly higher helpfulness – less powerful models such as LLaVA-1.5 [15] in fact score higher on truthfulness. In Fig. 3 we analyze fine-grained performance and find that GPT-4o's performance is particularly strong on questions that require reasoning about colors ("what color") and spatial relationships ("where is") but poorer on verification ("Can you", "Is there"). Finally, we also conduct a human study to evaluate the question relevance and answer correctness of QA pairs in our benchmark. Overall, out of 794 QA pairs, only 1.4% of questions are judged to be irrelevant to the image, 3.5% of answers are judged to be incorrect, and 0.6% are marked as both. We will focus on addressing these issues in subsequent versions of `PROVE`.

[1] This reduces false-positive hallucination detections, as no caption can capture every aspect of an image.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.

[6] Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *The Twelfth International Conference on Learning Representations*.

[7] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 16, pages 18135–18143, 2024.

[8] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

[9] Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[10] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[11] Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.

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

[13] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023.

[14] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[16] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*, 2023.

[17] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024.

[18] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

[20] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[21] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.

[22] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[23] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

[24] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[25] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.