

# Trust but Verify: Programmatic VLM Evaluation in the Wild

Viraj Prabhu<sup>1\*</sup> Senthil Purushwalkam<sup>2†</sup> An Yan<sup>1</sup> Caiming Xiong<sup>1</sup> Ran Xu<sup>1</sup>

<sup>1</sup>Salesforce AI Research <sup>2</sup>OpenAI

## Abstract

Vision-Language Models (VLMs) frequently hallucinate responses to visual queries, undermining their reliability for critical applications. However, quantifying the effect of such hallucinations in free-form responses to open-ended queries requires visually verifying each claim within the response, which is highly challenging. 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 with a high-fidelity scene-graph representation constructed from a detailed image caption, and prompt it to generate i) diverse and challenging question-answer (QA) pairs that test a range of image understanding capabilities, and ii) programs that can be executed over the scene graph object to verify each QA pair. We thus construct a benchmark of 10.6k challenging but grounded visual QA pairs.

Next, we propose a scene graph-based evaluation framework to programmatically measure both the helpfulness and truthfulness of free-form VLM responses to questions from our benchmark that does not rely on subjective LLM judgments. We extensively benchmark a range of VLMs on PROVE, and uncover a concerning tradeoff where models that provide more helpful responses often hallucinate more, whereas truthful models tend to be less informative. PROVE serves as a foundation for developing next-generation VLMs that balance helpfulness with truthfulness. Project page: [prove-explorer.netlify.app](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-counterparts), 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

significant research [14, 17, 25, 47] and regulatory [7] interest, particularly as using such models as the “foundation” of various high-stakes applications becomes imminent [8].

This has led to a flurry of research on *reliably* benchmarking VLM performance [29], by measuring not just the helpfulness but also the *truthfulness* of their responses. Existing benchmarks fall into two categories – *discriminative* [16, 24, 31], which evaluate the model’s responses to close-ended, existence-based queries (“Is there a man in this image?”), and *generative* [14, 25, 26, 38, 39], which evaluate responses to free-form, open-ended questions (“Describe this image.”). While discriminative benchmarks ease evaluation, they do not realistically simulate in-the-wild usage. On the other hand, generative benchmarks, while realistic, are *extremely* challenging to reliably evaluate, as they require verifying both that the model response fully answers the question (*i.e.* is helpful) and does not make any false claims (*i.e.* is truthful).

Evaluating such free-form responses typically relies on external models (usually, a proprietary LLM) to score responses given some image context (usually ground-truth annotations), under the now de-facto “LLM-as-judge” paradigm. However, we find that in several such benchmarks, the context provided is *completely insufficient* to judge if the response contains hallucinations. Consider Fig. 1: a VLM may respond to the query “How many puppies are in the image?” (correct answer = “four”), with “There are four labradoodle puppies”. Evaluating the truthfulness of this statement requires verifying multiple claims about the puppies (<count == four> and <breed == labradoodle>); however, an LLM judge provided only with a brief image caption as context (“four puppies placed on a light blue rug”) will be unable to do so, and will in-fact penalize the model for providing extra (correct) information! Further, the absence of a clear scoring rubric coupled with the sensitivity of LLMs to minor prompt differences often leads to inconsistent and arbitrary scores in such cases. In Fig. 2, we provide real examples from existing benchmarks that illustrate this problem.

We propose Programmatic VLM Evaluation (PROVE), a new evaluation paradigm that performs reliable and interpretable *programmatic* evaluation of free-form VLM re-

\*Correspondence to [viraj.prabhu@salesforce.com](mailto:viraj.prabhu@salesforce.com).

†Work done at Salesforce.

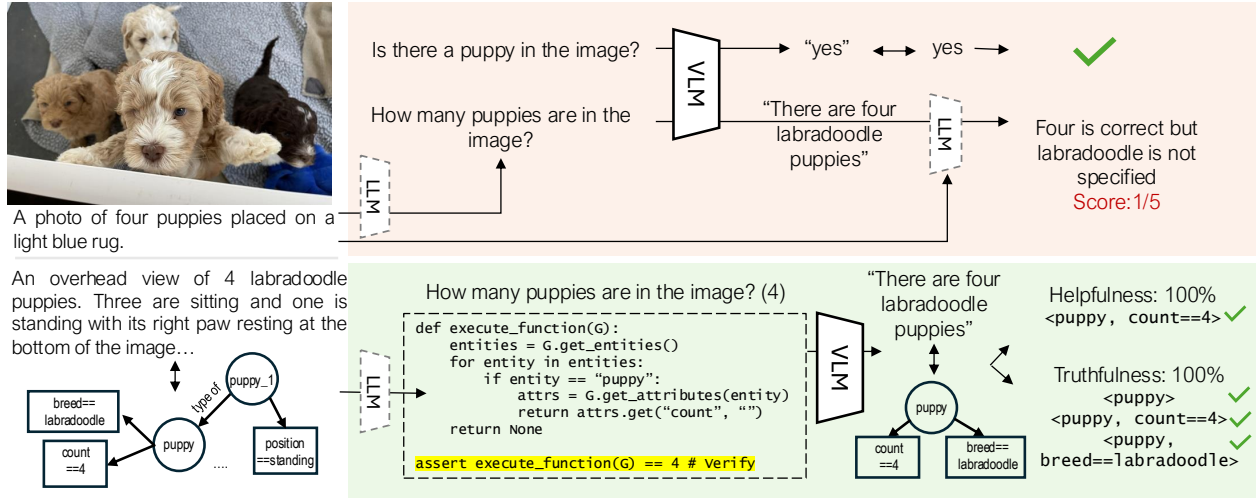


Figure 1. **Top.** Existing benchmarks for evaluating vision-language models (VLMs) have critical limitations: they either (1) restrict query-types to easy-to-evaluate binary questions that don’t reflect real-world use, or (2) use external LLMs to generate open-ended questions without verifying their visual groundedness and score responses with incomplete image context and inconsistent rubrics. **Bottom.** We propose PROVE, a novel benchmark for VLM evaluation that addresses these challenges. PROVE first constructs high-fidelity scene-graph representations from hyper-detailed image captions, which are used to prompt an LLM to generate i) diverse and challenging question-answer (QA) pairs that test a range of image understanding capabilities, and ii) programs that can be executed over the scene graph object to *verify* each QA pair. At test-time, our framework enables interpretable programmatic evaluation of both helpfulness and truthfulness in free-form VLM responses through systematic scene-graph comparisons, avoiding the inconsistencies of LLM-based judges.

sponses to challenging, diverse, and grounded questions. To build this dataset, we first use hyper-detailed image captions to construct a high-recall scene graph image representation. We then use an LLM to generate a diverse set of open-ended question-answer (QA) pairs along with accompanying *verification programs*. While the QA pairs are meant to test a range of model capabilities under real-world use, the verification programs can be executed over a given scene graph object to verify the correctness and groundedness of its corresponding QA pair. We thus only retain the QA pairs that we can programmatically verify and construct a benchmark of 10.6k diverse and challenging examples which are visually grounded by design, that we can use to reliably benchmark VLM responses. Further, unlike LLM judges with limited context, our verification programs can systematically decompose and verify multiple claims within a single response (e.g., both `<count>` and `<breed>` attributes), providing a consistent scoring rubric that avoids penalizing models for correct but additional information.

Next, we benchmark VLM responses to queries in PROVE by comparing scene graph representations. First, we measure the helpfulness of a response by computing its scene graph-based *recall* against the ground truth answer. Next, we measure response truthfulness as its scene graph-based *precision* against both the scene-graph constructed from the full caption or the image itself. Unlike LLM-as-judge approaches that may inconsistently penalize extra information, our precision-recall framework provides a systematic way to

reward models for complete answers (high recall) while ensuring factual correctness (high precision). We benchmark a range of VLM responses using this approach, and study their respective trade-offs between helpfulness and truthfulness. Our findings suggest that much of the recent progress in training “better” VLMs also translate to improved helpfulness on our benchmark, but not necessarily to higher truthfulness.

## 2. Related work

**Benchmarking VLM hallucination.** Existing benchmarks fall into one of two groups (see Fig. 2):

▷ **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 [24] limited their scope to object entities annotated by humans or external off-the-shelf models [49], whereas follow-up works additionally evaluate responses to *negative presence* queries [31], which stress-test the model’s abstention capabilities on questions about entities *absent* from the image, or use an LLM to generate a broader range of existence-based questions covering objects and their attributes [16]. 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 [38] measures the precision and recall of entities mentioned in a generated image description against the ground truth. HaELM [44]





	discriminative (POPE)	generative, templated (CHAIR)	generative, free-form (MMHal-Bench, GAVIE)	
	Label: two girls under a large umbrella in the rain	Label: a woman talking on a cell-phone	Labels: girl, head, hair, dog, person==4, face	
				
Question	Is there a man?	Describe this image.	How many people do you see?	Is the buckle width & height=50?
Answer	"no" == no	"a <u>woman</u> talks on a <u>cell phone</u> sitting on a <u>bench</u> ."	" <u>four</u> , <u>two adults</u> , <u>two children</u> ."	"the <u>size is hard to determine</u> ."
Score	acc=100% ✓	CHAIR <sub>i</sub> (↓) =0.33	4 is true 2 adults, 2 children is false. GPT-4 score: 1/5	The buckle is described to have width and height of 19 GPT-4 Score: 2/5

Figure 2. Existing VLM hallucination benchmarks fall into three categories: (1) "discriminative" approaches testing binary object existence queries [24], (2) "generative, templated" methods measuring entity precision/recall in image captions [38], and (3) recent "generative, free-form" benchmarks with open-ended queries [39]. While the last category better reflects real-world usage, it faces two critical challenges: generating natural questions (right example: GAVIE’s [26] unnatural coordinate-based questions) and reliable evaluation (second from right: GPT-4 incorrectly penalizing a response with additional correct details not in the ground truth in MMHal-Bench [39]). We propose PROVE, a new benchmark of verifiable open-ended questions that we use to reliably evaluate free-form VLM responses in the wild.

additionally uses a large language model (LLM) to judge generations, whereas M-HalDetect [14] has humans annotate hallucinations in model generated descriptions are used to train a predictive model. Recently, AMBER [43] 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-style templates (“Describe this image in detail.”).

Most recently, a few benchmarks with truly open-ended queries have been proposed [19, 25, 26, 39, 46], 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 [39] and HallusionBench [25] rely on a series of off-the-shelf models at various stages which introduce noise (see Fig. 2, col 3). GAVIE’s [26] 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, see Fig. 2, col 4). Finally, GPT-4-based evaluation is both expensive and inherits the model’s own limitations.

**Understanding and mitigating VLM hallucination.** Several works have sought to better understand *why* VLMs hallucinate. One prevalent theory is the model learning spurious correlations between the input and the output: either due to overly strong text priors learned by the LLM backbone [17, 23], or due to distilling synthetic outputs generated by stronger models (such as GPT-4V) that may themselves contain confabulation [27]. This is often exacerbated by the predominant training recipe [2, 27] that learns a shallow projection from the visual input to the text embedding space which limits the expressivity of the model to learn visually

grounded representations.

Recent work has proposed training-based and training-free strategies for mitigating hallucinations. The former involves finetuning [26] or preference optimization [39, 47] of “preferred” ground truth responses against dis-preferred synthetically generated “hallucinations”. Training-free methods instead focus on specialized decoding strategies [17, 20, 23] that seek to correct for potential statistical bias that may lead to hallucination.

However, developing better understanding and mitigation strategies are both contingent on the availability of reliable evaluation benchmarks. In this work, we introduce such a benchmark of challenging but verifiable open-ended visual questions that we use to jointly evaluate both the truthfulness and helpfulness of free-form model responses.

### 3. Approach

We introduce PROVE, a framework for evaluating vision-language models through a programmatically verifiable pipeline that generates challenging visual QA pairs and uses a scene graph-based framework to assess both helpfulness (answer correctness) and truthfulness (absence of hallucinations) in VLM responses.

Vision-language models produce responses  $\hat{\mathcal{A}}=m_{\theta}(Q, \mathcal{I})$  to questions about images. While prior work typically evaluates either correctness (is  $\hat{\mathcal{A}}=\mathcal{A}$ , where  $\mathcal{A}$  is the ground-truth answer) or truthfulness (is  $p(\hat{\mathcal{A}}|\mathcal{I}) > \text{threshold}$ ) separately, we propose a unified framework that jointly evaluates both.

#### 3.1. Generating verifiable Visual Question-Answers

To build PROVE, we leverage image-caption pairs  $(\mathcal{I}, \mathcal{C})$  from the test set of DOCCI [34], containing 5k images with



after filtering contains 10.6k high-quality visual question answers – see Fig. 3.

Our primary methodological contribution over prior work [11, 46] lies in our question generation and verification steps: DSG evaluates text-to-image generation using simple binary questions derived from scene graphs, whereas FIHA [46] additionally generates *templated-Wh* (where, who, etc.) questions. We seek to simulate “in-the-wild” VLM usage and harness the world-knowledge and reasoning capabilities of LLMs to generate *challenging, open-ended questions* (eg. “Describe the condition of the cargo carts.”, see Fig 2 – a question neither DSG nor FIHA is capable of generating). While this approach generates diverse, QA pairs that comprehensively test a range of perception and reasoning abilities for realistically stress-testing VLMs, it is susceptible to hallucinations. We remedy this with our novel *trust, but verify* strategy, that overcomes the reliability issues that typically plague LLM-generated QA datasets [24, 39] by ensuring all pairs are verifiable.

**Dataset statistics.** We now present some statistics about PROVE, which comprises of 10.6k QA pairs generated from 5k image-caption pairs from the DOCCI test set. These are obtained after applying both programmatic filtering *i.e.* either the unit test fails (18.3%) or returns the wrong answer (9.8%), and text-based filtering ( $\sim 50\%$  of the total from the previous stage). Note that we opt to filter out such a large percentage of QA pairs in the interest of ensuring high-quality evaluation data. Further, our benchmark curation process is fully automatic and so can be readily scaled to a larger image-caption source.

Fig. 4a, left presents a sunburst visualization of the first 4 words in the questions within PROVE. As seen, the questions are diverse and span a wide range of question types. Further, while nearly 50% of the questions begin with “What”, even this subset spans a range of topics testing numerous model capabilities. On the right, we measure the distribution of question and answer lengths in PROVE and find that questions average 10.3 words whereas answers average 13.4 words (see Fig. 4, right), far exceeding average answer lengths in competing approaches (*e.g.* 1 in DSG and  $\leq 3$  for FIHA.)

Finally, we highlight the flexibility of our approach in generating verifiable QAs for *any* information captured in the scene graph, many of which require understanding reasoning and context. Some real examples: **Q1.** What is the geographic feature the man engages with on the relief? (*depiction of the State of Texas*), **Q2.** Why might the airplane be landing as inferred from the scene? (... *because its landing gear is deployed*), **Q3.** What changes would the scene experience if the windows were tinted dark? (...*it might affect perceived visibility into the vehicle’s interior...*). Additional examples can be viewed on our project page: <https://prove-explorer.netlify.app>.

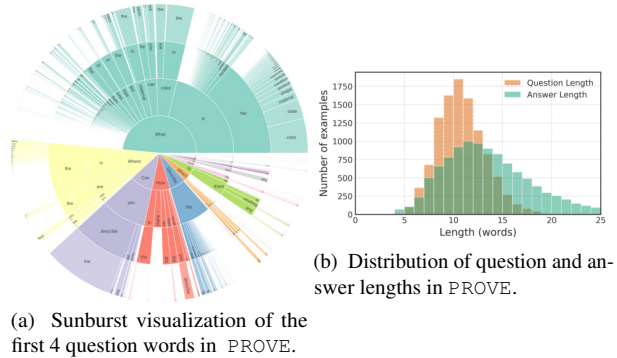


Figure 4. PROVE: Additional dataset statistics.

### 3.2. Programmatic VLM Evaluation (PROVE)

Our evaluation framework uniquely measures both helpfulness (does the model answer the question correctly?) and truthfulness (does the model avoid hallucinations?) through a scene graph-based approach. This dual assessment provides a more complete picture of VLM performance than previous metrics.

After ensuring the validity of our QA pairs, we evaluate free-form VLM responses  $\hat{\mathcal{A}} = m_{\theta}(\mathcal{Q}, \mathcal{I})$  by:

1. Extracting tuples (using an LLM [12] with in-context prompting) from both the model response  $\hat{\mathcal{A}}$  and ground truth answer  $\mathcal{A}$ , and building corresponding scene graph representations  $g(\hat{\mathcal{A}})$  and  $g(\mathcal{A})$ .
2. Refining the ground truth answer’s scene graph to exclude “premise” tuples entailed by the question  $g(\mathcal{A}) - g(\mathcal{Q})$ .
3. Computing helpfulness and truthfulness metrics based on these graphs.

Concretely, we assess response helpfulness  $h\text{score}(\cdot)$  based on the response graph’s *recall* of the refined ground truth answer graph, *i.e.* the fraction of tuples (nodes, attributes, and relationships) in  $g(\mathcal{A}) - g(\mathcal{Q})$  that are recovered by  $g(\hat{\mathcal{A}})$ . To do so, we compute average cosine similarity between each ground truth tuple and its closest response tuple in embedding [37] space. Finally, to aid fine-grained discrimination between methods, we additionally threshold scores using a parameter  $\theta$ .

$$h\text{score}(\hat{\mathcal{A}}) = \frac{\sum_{t \in g(\mathcal{A}) - g(\mathcal{Q})} \mathbb{I}[\max_{t' \in g(\hat{\mathcal{A}})} \text{sim}(t, t') > \theta]}{|g(\mathcal{A}) - g(\mathcal{Q})|} \quad (1)$$

Next, we compute  $t\text{score}(\cdot)$  as the *precision* of the response graph *i.e.* the fraction of response tuples that are consistent with either the original image scene graph or the image itself. We find that also considering the image itself helps reduce false-positive hallucination detections, as no

caption can capture every aspect of an image. We define:

$$\text{tscore}(\hat{\mathcal{A}}) = \frac{1}{|\mathcal{g}(\hat{\mathcal{A}})|} \sum_{t' \in \mathcal{g}(\hat{\mathcal{A}})} \mathbb{I}[\max(\max_{t \in \mathcal{g}(C)} (\text{sim}(t', t), p(\mathcal{I} \models t')) > \theta]; \quad (2)$$

where  $\models$  denotes visual entailment, and  $p(\mathcal{I} \models t')$  is approximated using a state-of-the-art visual grounding model [30].

Note that hscore and tscore are not necessarily correlated – a response can be helpful (by answering the query) but not entirely truthful (might contain hallucinations), and vice versa. Our approach stands apart from previous evaluation methods by providing fine-grained, interpretable metrics that can reveal where and how models are failing. By measuring both helpfulness and truthfulness together, PROVE uniquely enables analysis of the trade-offs different models make between these sometimes competing objectives, offering deeper insights into VLM performance characteristics.

## 4. Experiments

We now present our benchmarking experiments on PROVE. We include a broad set of models spanning a range of sizes and learning strategies and extensively analyze their performance, including their performance trade-offs. We also conduct a human study to validate both the quality of our benchmark and how well our proposed metrics correlate with human judgement.

### 4.1. Setup

**Baselines.** Considering the rapidly evolving VLM landscape, we seek to present a representative rather than comprehensive snapshot of its current state. To this end, we benchmark VLMs of three sizes – small (<5B parameters), medium (5-10B parameters), and large (>10B parameters) – and include both open-source and proprietary models.

**Data.** PROVE is constructed from images, tuples, and captions released under a CC by 4.0 license as the test split of the DOCCI [34] dataset<sup>2</sup>. DOCCI images were reviewed both by human and automatic methods to remove or obfuscate PII (faces, phone numbers, and URLs) and unsafe content. Images underwent a rigorous 3-stage human annotation phase resulting in hyper-detailed and high-recall captions averaging 136 words. As ground truth we use the scene graphs included with the DOCCI test set (generated using Palm-2 (340B) [4] following DSG [11]), which have been human-validated to be of extremely high quality (92.2% precision and 100% recall on TIFA160).

**Implementation details.** We use GPT-4o [3] for generating structured question, answers, and verification programs

<sup>2</sup>Our method is also readily applicable to other dense captioning / scene graph datasets, both crowd-sourced [5, 13, 22, 36] and synthetic (e.g. using pretrained detectors/segmenters) [35, 48].

using the batch API and prompting it with a detailed task description, examples, and a Python definition of the Scene-Graph class. We also use GPT-4o for the first round of text-based post-processing described in Sec. 3.1. During evaluation, we use LLaMA-3.1-Instruct (8B) [12] with in-context prompting for tuple extraction, Grounding DINO [30] for image-tuple entailment (Eq. 2), and Sentence-Bert [37] to extract text embeddings. Premise exclusion is performed by direct set differencing. We set  $\theta = 0.75$  across experiments.

### 4.2. Results

Figure 5 presents evaluation results. We find that:

▷ **Few models strike a good balance between helpfulness and truthfulness.** As Fig. 5b (left) shows, models tend to exhibit a range of trade-offs between helpfulness and truthfulness, with only one model from the ones we survey (GPT-4o) managing to strike a good balance between the two. In fact, we find that many recent models that rank highly on perception and reasoning-focused aggregate benchmarks [32], such as Intern-VL2 (26B) [10] *do not* necessarily translate to high truthfulness on PROVE, lagging behind simpler and smaller models like LLaVA-1.5 [28] in tscore. In fact, we find the LLaVA-1.5 model series to obtain the best tscore overall. Overall, we observe a weak linear correlation of **0.1** between hscore and tscore averaged across models, suggesting that the impressive recent gains in model helpfulness have not necessarily translated to higher truthfulness.

▷ **Increasing model size improves hscore but not necessarily tscore.** Across both model families that we benchmark at multiple sizes – InternVL2 [10] (2B, 8B, and 26B), and LLaVA [27] (1.5-7B, Next-7B, and 1.5-13B), we find that larger or more recent variants tend to outperform smaller ones in terms of helpfulness but not necessarily truthfulness.

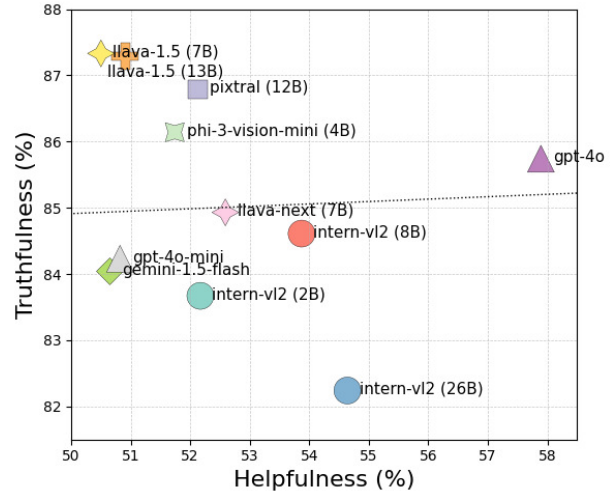
▷ **Models fail in different ways.** In Fig. 6, we compare responses from models that excel at different metrics: GPT-4o (high hscore) and LLaVA-1.5-7B (high tscore). Both struggle with perceptual tasks like OCR, counting, and reading analog clocks. However, GPT-4o generally produces more detailed and accurate descriptions, leading to higher helpfulness scores despite occasionally including incorrect details. LLaVA-1.5, while less descriptive, demonstrates greater restraint in its assertions, contributing to its higher truthfulness. These differences highlight the fundamental trade-off between providing comprehensive responses and avoiding hallucination. Additional fine-grained analysis of GPT-4o’s performance across question types is presented in supplementary. As seen, the model performs particularly well on questions that require reasoning about spatial relationships (where are/is), object attributes (what color), and generating image descriptions. We also visualize our full evaluation pipeline for various models and examples at [prove-explorer.netlify.app](https://prove-explorer.netlify.app).

Finally, while results in Table 5a might suggest a rela-

Method	#params	hscore (↑)	tscore (↑)	avg (↑)
Qwen2-VL [6]	2B	45.70	82.79	64.25
InternVL2 [10]	2B	<b>52.17</b>	83.67	67.92
Phi-3.5-Vision [2]	4B	51.74	<b>86.14</b>	<b>68.94</b>
LLaVA-1.5 [28]	7B	50.50	<b>87.33</b>	68.91
LLaVA-Next [27]	7B	52.59	84.93	68.76
InternVL2 [10]	8B	<b>53.87</b>	84.61	<b>69.24</b>
Pixtral [33]	12B	52.13	86.79	<b>69.46</b>
LLaVA-1.5 [28]	13B	50.91	<b>87.29</b>	69.10
InternVL2 [10]	26B	<b>54.64</b>	82.24	68.44
GPT-4o-mini† [3]	-	50.82	84.23	67.52
Gemini-1.5-Flash† [41]	-	50.65	84.04	67.35
GPT-4o† [3]	-	<b>57.89</b>	<b>85.74</b>	<b>71.81</b>

(a) Benchmarking VLMs on PROVE

Figure 5. Performance evaluation of various VLMs on the PROVE benchmark. (a) We report helpfulness (hscore), truthfulness (tscore), and their average. (b) We plot hscore and tscore for VLMs on PROVE – as seen, models with higher helpfulness tend to lag behind on truthfulness, with very few striking a good trade-off between the two. Averaged across models, we observe a weak linear correlation of **0.1** between hscore and tscore .



(b) Tradeoffs between helpfulness and truthfulness

tively small performance gap between models, a deeper look paints a different picture. In Table 1 we segment performance by question type, over “describe” type questions (eg. “Can you describe the condition of the bench?”, 4.1k queries in total) and “query” type questions (eg. “What is the color of the bike next to the bench?”, 6.2k queries in total), for the model with the best hscore and tscore respectively. Across both question types, we find that while GPT-4o only slightly lags LLaVA-1.5 (7b) in tscore(-1% absolute) for more targeted “query” type questions, it does so more significantly (-2.61% absolute) for descriptive questions, wherein the open-ended nature of the questions leads to a higher degree of hallucination. Conversely, GPT-4o outperforms LLaVA-1.5 (7b) by 7.13 and 7.85 percentage points absolute on hscore for query and describe questions.

Our evaluation reveals three key insights: (1) current VLMs exhibit a clear trade-off between helpfulness and truthfulness, with few models excelling at both; (2) increasing model size generally improves helpfulness but has an inconsistent effect on truthfulness; and (3) different models demonstrate distinct error patterns that reflect their underlying training objectives and architectures. These findings highlight the importance of multi-faceted evaluation metrics that can capture these nuanced performance differences.

### 4.3. Analysis

**Human evaluation of PROVE and proposed metrics.** We conducted two human studies of our benchmark. First, annotators (3 per example) evaluated question relevance and answer correctness of QA pairs from the qual-test split of DOCCI (100 images, 170 QA pairs). After majority vot-

Model	Query		Describe	
	hscore	tscore	hscore	tscore
LLaVA-1.5 (7b)	50.95	<b>87.55</b>	49.70	<b>87.14</b>
GPT-4o	<b>58.08</b>	86.54	<b>57.55</b>	84.53

Table 1. Fine-grained performance comparisons on PROVE .

ing, 95.9% of questions were judged relevant and 98.2% of answers correct. Manual inspection revealed the few examples deemed irrelevant / incorrect to typically be either particularly challenging or subjective.

In our second study, subjects rated responses from four models (GPT-4o, LLaVA-1.5-7B, LLaVA-Next-7B, GPT-4o-mini) on helpfulness (0=unhelpful, 1=helpful) and truthfulness (0=fully false, 0.5=partially false, 1.0=fully true). We then computed the Pearson correlation between human judgments and our automated metrics. For comparison, we included two off-the shelf LLM-as-judge baselines (GPT-4o-mini and LLaMA-3.1-Instruct) and a *specialized*, state-of-the-art evaluation LLM judge [21]. As shown in Table 2, our metrics correlate substantially better with human judgment than all three competing approaches.

**Evaluating intermediate steps.** We studied potential errors in our evaluation pipeline, focusing on truthfulness where we observed lower correlation with human judgment. On average, our tscore *over-estimates* truthfulness by 11%, primarily due to image-text matching with an external model. However, removing this component causes tscore to underestimate truthfulness by 15% and reduces the Pearson



Figure 6. Example responses from high hscore (GPT-4o) and tscore (LLaVA-1.5 (7B)) VLMs. While both struggle with sub-tasks such as OCR, counting, and reading an analog clock, GPT-4o’s errors tend to be less egregious which leads to a higher hscore.

Approach	Helpfulness	Truthfulness
LLaMA-3.1 (8b)	0.41	0.20
GPT-4o-mini	0.58	0.30
Prometheus-2 (7B)	0.57	0.32
Ours	<b>0.81</b>	<b>0.57</b>

Table 2. Pearson correlation between human judgments and automated evaluation methods for helpfulness and truthfulness. Our proposed metrics show substantially higher correlation with human judgment compared to LLM-as-judge approaches.

correlation with human scores to 0.39, highlighting the complementary strengths of our approach.

Further, to validate the effectiveness of our choice of Grounding DINO for image-text entailment, we additionally benchmark OFA [45], a strong visual entailment model and compare how well its tscore’s correlate with human judgment: we find that Grounding DINO does significantly better, achieving a Pearson correlation of **0.57** v/s **0.45**.

Finally, we comment that our method relies on external models only for well-defined subtasks (tuple extraction, text matching) where these models excel and continue to improve. Unlike LLM-as-judge scoring, our approach provides an interpretable, auditable trace of how responses are scored, enhancing transparency. While external models may intro-

duce some errors, our robust approach largely mitigates these issues, as validated by its near-perfect human ratings.

## 5. Discussion

Our work evaluates helpfulness-truthfulness trade-offs in vision-language models using an LLM with scene graphs to create grounded QA pairs that enable programmatic evaluation. The reliability of our benchmark stems from high-recall human-annotated captions, programmatic verification ensuring visual grounding, and holistic and interpretable evaluation metrics. Ultimately, our evaluation paradigm that allows for the creation of diverse, challenging, open-ended questions at scale while ensuring they are grounded and evaluable, overcoming key limitations of prior work.

While PROVE offers reliable VLM evaluation, our high-precision approach reduces recall by excluding hard-to-verify questions. Even comprehensive captions may miss image details, and our evaluation depends on potentially imperfect off-the-shelf models. Future improvements could include region-level entailment and enriched scene graphs with additional annotations, moving toward an idealized “Platonic” representation [18] where text-matching suffices for evaluation. We hope future work will explore hallucination mitigation strategies [17, 20, 23, 26, 39, 47] on PROVE, and advance agentic models with planning [15, 40], reasoning, and self-reflection [42] capabilities to improve both helpfulness and truthfulness.

## References

- [1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 4
- [2] Marah Abidin, 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. 3, 7
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 7
- [4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 6
- [5] Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*, 2024. 6
- [6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 7
- [7] Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023. 1
- [8] 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. 1
- [9] 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. 4
- [10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6, 7
- [11] Jaemin Cho, Yushi Hu, Jason Michael Baldrige, 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*. 4, 5, 6
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5, 6
- [13] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunnor, Ranjay Krishna, Jason Baldrige, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*, 2024. 6
- [14] 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. 1, 3
- [15] 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. 4, 8
- [16] 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*. 1, 2
- [17] 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. 1, 3, 8
- [18] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 8
- [19] 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. 3
- [20] Junho Kim, Yeon Ju Kim, and Yong Man Ro. What if...?: Counterfactual inception to mitigate hallucination effects in large multimodal models. *arXiv preprint arXiv:2403.13513*, 2024. 3, 8
- [21] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024. 7
- [22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6
- [23] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 3, 8
- [24] 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. 1, 2, 3, 5

- [25] 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. 1, 3
- [26] 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. 1, 3, 8
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 6, 7
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 6, 7
- [29] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1
- [30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 6
- [31] 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. 1, 2
- [32] Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 6
- [33] Mistral. Announcing pixtral 12b. <https://mistral.ai/news/pixtral-12b/>, 2024. [Online; accessed 21-September-2024]. 7
- [34] 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. 3, 6
- [35] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, et al. Synthetic visual genome. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9073–9086, 2025. 6
- [36] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 6
- [37] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 4, 5, 6
- [38] 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. 1, 2, 3
- [39] 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 multi-modal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1, 3, 5, 8
- [40] 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. 4, 8
- [41] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7
- [42] Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Llms still can’t plan; can lrms? a preliminary evaluation of openai’s o1 on planbench. *arXiv preprint arXiv:2409.13373*, 2024. 8
- [43] 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. 3
- [44] 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. 2
- [45] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 8
- [46] Bowen Yan, Zhongsong Zhang, Liqiang Jing, Eftekhar Hossein, and Xinya Du. Fiha: Autonomous hallucination evaluation in vision-language models with davidson scene graphs. *arXiv preprint arXiv:2409.13612*, 2024. 3, 5
- [47] 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. 1, 3, 8
- [48] Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Silvio Savarese, et al. Provision: Programmatically scaling vision-centric instruction data for multimodal language models. *arXiv preprint arXiv:2412.07012*, 2024. 6
- [49] 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. 2