FAITHSCORE: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

We introduce FAITHSCORE (**Faith**fulness to Atomic Image Facts **Score**), a reference-free and fine-grained evaluation metric that measures the faithfulness of the generated free-form answers from large vision-language models (LVLMs). The FAITHSCORE evaluation first identifies sub-sentences containing descriptive statements that need to be verified, then extracts a comprehensive list of atomic facts from these sub-sentences, and finally conducts consistency verification between fine-grained atomic facts and the input image. Meta-evaluation demonstrates that our metric highly correlates with human judgments of faithfulness. We collect two benchmark datasets (i.e. LLaVA-1k and MSCOCO-Cap) for evaluating LVLMs instruction-following hallucinations. We measure hallucinations in state-of-the-art LVLMs with FAITHSCORE on the datasets. Results reveal that current systems are prone to generate hallucinated content unfaithful to the image, which leaves room for future improvements. We hope our metric FAITHSCORE can help evaluate future LVLMs in terms of faithfulness and provide insightful advice for enhancing LVLMs' faithfulness.

1 Introduction

Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020) and ChatGPT (OpenAI, 2022), have demonstrated various language modeling capabilities. Despite their achievements, they still lack the capacity to handle multimodal inputs effectively. As a result, a significant amount of research has shifted its focus towards Large Vision-Language Models (LVLMs) (Liu et al., 2023e; Ye et al., 2023; Sun et al., 2023) by incorporating powerful LLMs (Touvron et al., 2023; Chiang et al., 2023) and Vision Foundation Models (VFMs) (Dosovitskiy et al., 2021; Bommasani et al., 2021). LVLMs have shown strong performance on various multimodal tasks, such as Visual
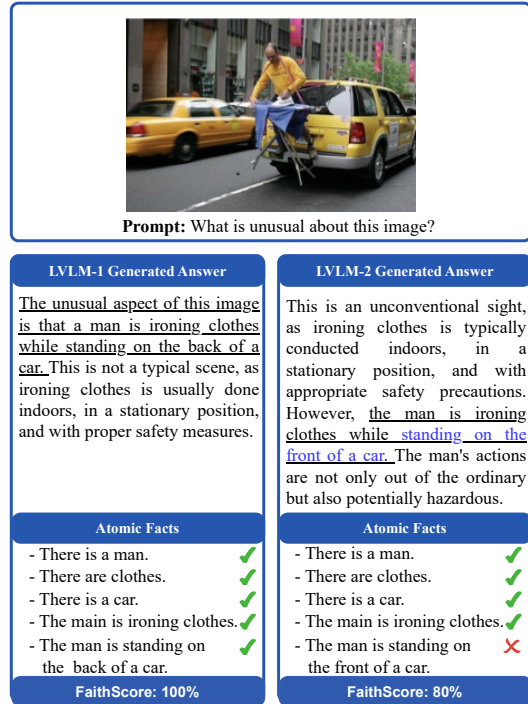

Figure 1: Illustration of how FAITHSCORE evaluation works. Given the answers generated by an LVLM, in step 1, we identify the descriptive content (with an LLM); In step 2, we extract corresponding atomic facts from the identified sentences; In step 3, the faithfulness of all atomic facts is verified according to the input image. In this example, the underlined part denotes objective descriptive content in the answer. The blue contents denote hallucinations in the answers. FAITHSCORE allows a more fine-grained and interpretable evaluation of the factual precision of free-form answers.

Question Answering (Antol et al., 2015), Image Captioning (Lin et al., 2014), and Multimodal Conversation (Liu et al., 2023e).

Despite the effectiveness of LVLMs, the problem of hallucination is pervasive, often leading these models to generate fabricated information that is incongruent with the provided visual input (Rohrbach et al., 2018; Liu et al., 2023b,a; Yin et al., 2023). In the context of LVLM, the problem of hallucination can manifest as answers containing descriptions of

the input image that are incorrect (Li et al., 2023c). As shown in Figure 1, the LVLM-2 generates an answer with an inaccurate description (*i.e., standing on the front of a car*), which is not faithful towards the input image. The phenomenon of hallucination in LVLMs introduces potential hazards that could result in significant consequences such as misinformation and safety concerns, thus degrading the model's reliability in practical applications inevitably (MacLeod et al., 2017). Hence, it is imperative that these issues are thoroughly measured and addressed (Ji et al., 2023).

Nevertheless, there have been limited explorations that measure the hallucination problem in LVLMs. Li et al. (2023c) was among the first to measure the hallucinations of LVLMs with a polling-based object probing evaluation method. In addition, Gunjal et al. (2023) annotated a multimodal hallucination detection dataset tailored for detailed image description evaluation. Lovenia et al. (2023) devised Negative Object Presence Evaluation (NOPE), which used VQA to quantitatively evaluate object hallucination in LVLMs. These approaches, however, exhibit two key weaknesses: (1) they focus on the limited setting of image captioning, and none of them explored evaluating hallucination of the complex and free-form answers to the open-ended questions (OpenAI, 2023) (e.g. multimodal conversations (Liu et al., 2023e; Sundar and Heck, 2022), world knowledge-based VQA (Schwenk et al., 2022) and visual storytelling (Huang et al., 2016)); (2) they ignore fine-grained hallucinations of visual attributes in the generated answer (Liu et al., 2024; Rohrbach et al., 2018).

Evaluating hallucinations present in free-form answer is especially challenging for two primary reasons: (1) **Free-form answers contain a hybrid of descriptive and analytical contents.** Unlike close-domain tasks such as image captioning, answering open-domain questions in a free-form manner does not only require generating the question-relevant descriptive content of the given image. It may also involve analytical content such as rationales that include external commonsense knowledge. As depicted in Figure 1, certain sub-sentences (*i.e.,* those without the underline) do not require verification with the image input due to their analytical nature. Because they encompass subjective analytical content that extends beyond a direct description of the visual inputs. Neglecting to distinguish between analytical and descriptive content inevitably distracts the factual measurement. Thus, pinpointing the descriptive content within the answers generated by LVLMs is significant. (2) **Model outputs are prone to the multiplicity of hallucinations.** Current methodologies offer a constricted view on evaluating hallucinations, primarily concentrating on coarse-grained object existences (Rohrbach et al., 2018; Lovenia et al., 2023), while neglecting other fine-grained elements, such as counts, colors, and the interrelations between objects (*e.g.,* the spatial relation between the person and the car in Figure 1), which also form a significant portion of visual hallucinations (Gunjal et al., 2023). Consequently, devising a method to holistically evaluate fine-grained hallucinations of visual attributes is also important.

To address the aforementioned challenges, we propose the FAITHSCORE metric, which can evaluate *fine-grained hallucinations in various multimodal tasks*, such as image captioning and open-ended questions. This metric comprises three primary components: Descriptive Sub-sentence Identification, Atomic Fact Generation, and Fact Verification, as illustrated in Figure 2. The first component is tasked with discerning descriptive sub-sentences within the composite content of the generated free-form answer. Thereafter, the second component deconstructs this descriptive content into fine-grained elements (*i.e.,* atomic facts) (Min et al., 2023). These atomic facts cover comprehensive types, such as objects attributes and interrelationships. The last component emphasizes verifying the consistency between the visual information and the derived atomic facts via a Visual Entailment Model (VEM) (Xie et al., 2019). Based on the proposed metric, we evaluated several advanced LVLMs, such as LLaVA (Liu et al., 2023e) and MiniGPT-4 (Zhu et al., 2023). From the results, we conclude that current LVLMs still face challenges of answers that are not faithful to the input image, which leaves a large room for improvement.

In summary, our contributions are as follows: (1) We introduce FAITHSCORE, a metric tailored to assess hallucinations in LVLMs free-form answers to open-ended questions, which is not yet addressed by current studies; (2) To the best of our knowledge, we are the first to systematically study the LVLMs free-form answers and evaluate the fine-grained hallucinations of various types in LVLMs; (3) In our quest to understand the hallucinations manifested by LVLMs, we embark on comprehensive experiments with six open source

2

models across diverse tasks and datasets. Our findings underscore that the hallucination phenomenon remains a pressing challenge for current LVLMs.

2 Related Work

Large Vision-Language Model Motivated by the success of the pretraining technique in LLMs and VFMs, the multimodal learning research community has recently shifted the research attention to LVLMs (Awadalla et al., 2023; Li et al., 2023a). Contemporary advanced LVLMs predominantly feature three core components: a text encoder, an image encoder, and a cross-modal alignment module (Rohrbach et al., 2018). Specifically, the text encoder often takes the form of a language model, as seen in examples like LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). Conversely, the image encoder is typically derived from VFMs, such as ViT (Dosovitskiy et al., 2021). The function of the cross-modal alignment module is to bridge visual content with textual representation, enhancing the text encoder's capacity to interpret visual semantics. To accomplish visual understanding, LVLMs typically undergo multiple training phases (Gong et al., 2023; Zhu et al., 2023; Liu et al., 2023d,e; Ye et al., 2023; Dai et al., 2023). For instance, Liu et al. (2023e) first aligns the image features with the word embeddings of a pre-trained LLM during an initial pre-training stage, and subsequently fine-tunes the LVLM using specialized language-image instruction-following datasets. For efficiency enhancement, LVLMs often freeze parameters of the LLM or VFM and are trained with efficient fine-tuned techniques (Ye et al., 2023; Dai et al., 2023), such as LoRA (Hu et al., 2022).

However, in spite of the considerable advancements made by LVLMs, they consistently grapple with hallucination issues. These issues markedly impact their efficacy across a range of vision-language tasks (Rohrbach et al., 2018).

Vision-language Model Hallucinations and Evaluations Though hallucination phenomenons and mitigation methods have been extensively studied in the text generation literature (Ji et al., 2023; Min et al., 2023), it is much less investigated in vision-language models (Dai et al., 2023; Liu et al., 2023e). Although there are a few existing works tackling this issue, they mainly focus on the constraint problem setting such as image captioning (Johnson et al., 2016). For example, Rohrbach et al. (2018) propose caption hallucination assess-
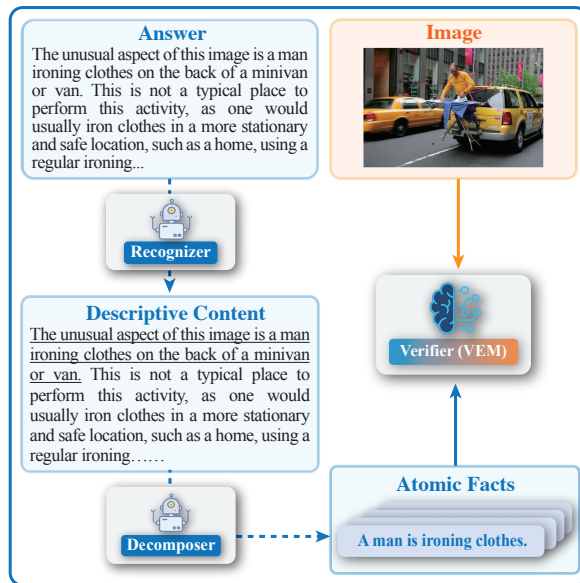

Figure 2: An overview of estimating FAITHSCORE, which mainly consists of three steps: Descriptive Subsentence Identification, Atomic Fact Generation, and Fact Verification. These steps are implemented by three modules: Recognizer, Decomposer, and Verifier. The underlined part denotes recognized descriptive content.

ment with image relevance (CHAIR), which is a popular metric for evaluating object hallucination in sentence-level captions. They also show that popular metrics like METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) do not capture this. Li et al. (2023c) extends CHAIR and proposes "POPE", a polling-based query technique for probing objects. Besides, Lovenia et al. (2023) devised Negative Object Presence Evaluation (NOPE) to quantitatively assess object hallucination through VQA, based on "POPE". Gunjal et al. (2023) further proposed to detect hallucinations in more detailed image captions and investigated utilizing a reward model for mitigating them. Lu et al. (2023) introduced an evaluation benchmark that contains more diverse types of questions, such as Yes-or-No and Fill-in-the-Blank.

Different from all the above, we are the first to propose a reference-free metric for fine-grained evaluating the answers in the open-ended visual question-answering setting, where answers are of free form and can be lengthy.

3 Estimating FAITHSCORE

In this section, we begin by clearly defining the research problem in Section 3.1, followed by a detailed framework of estimating FAITHSCORE in Section 3.2.

3.1 Task and Settings

Suppose we have an image I and a corresponding prompt Q. We then feed them into the LVLM denoted as $\mathcal{M}$, to obtain the generated answer A. Our objective is to design a scoring framework to estimate FAITHSCORE f based on the input prompt Q, the input image I, and the generated answer A. It is defined as: $s = \mathcal{F}(A, Q, I)$. s is a scalar value ranging between 0.0 and 1.0. Notably, the devised evaluation method is reference-free and doesn't require a ground truth answer.

3.2 The Evaluation Framework

In order to estimate FAITHSCOREof the generated answers, we introduce a novel framework to implement the scoring function $\mathcal{F}$. The framework comprises three key steps: descriptive sub-sentence identification, atomic fact generation, and fact verification, as depicted in Figure 2.

Descriptive Sub-sentence Identification. Faithfulness in the context of LVLMs refers to the consistency between the input visual content and the generated answer. Notably, we focus on the details in the answer that describe the input image objectively, to obtain a more precise and fine-grained understanding of the hallucination. As shown in Figure 1, only some sub-sentences (*i.e.,* those with the underline) from the answer require verification with the image input. Hence, we need to identify the descriptive sub-sentences from the answer using a recognizer. The sub-sentences denote the short sentences that are split by punctuation in the answer.

Humans are capable of distinguishing descriptive sub-sentences from other contents (referred to as analytical sub-sentences) by analyzing the semantics of the answers generated by LVLMs. However, manually identifying descriptive sub-sentences is a resource-intensive process, requiring plenty of human labor. To address this problem, we turn to ChatGPT to implement the recognizer as a practical solution, since it has demonstrated remarkable semantics understanding capabilities across a wide range of natural language processing tasks (OpenAI, 2022). Section 4.2 shows that Chat-GPT can achieve promising performance on this task.

To be more specific, our approach first crafts a prompt P that encompasses task instructions and K_1 in-context learning examples. We feed this designed prompt along with the to-be-processed answer A into the ChatGPT, generating the recognized results, defined by the equation $\hat{A} = ChatGPT(A, P)$, where $\hat{A} = \{\{a_1, l_1\}, \cdots, \{a_k, l_k\}\}$ signifies the generated result, in which the answer is split into sun-sentences a, and every sub-sentence is assigned a label l (*i.e.,* descriptive or analytical). Then we extract all descriptive sub-sentences denoted as $A' = \{a'_1, \cdots, a'_t\}$. For a more comprehensive understanding of the specific prompt P utilized in this process, please refer to Section L of the Appendix.

Atomic Fact Generation. Despite we have identified descriptive sub-sentences from the answer, there are still multiple facts hybrid in each sub-sentence. Each descriptive sub-sentence consists of multiple pieces of information (*i.e.,* atomic facts), each of which may contain hallucination. Therefore, to access a fine-grained evaluation, we design a decomposer to further break the sub-sentences into atomic facts. In particular, we define atomic facts as an element belonging to an entity, relation, or attribute, inspired by the existing works (Min et al., 2023; Hu et al., 2023). Importantly, the atomic fact is a minimal unit of information. This handling can ensure the verification of *each* element in the answer without being disturbed by other information. Atomic facts include three types: entity existence, attributes, and relations. An entity fact indicates an object's existence. Attribute facts relate to characteristics like color and shape. Relation facts describe inter-entity relationships, e.g., the spatial relation. In Figure 1, we show some examples of atomic facts.

To achieve this, similar to the process of identifying descriptive sub-sentences, we also utilize the ChatGPT for the generation of atomic facts. This is because ChatGPT has shown a strong ability in information extraction (Wei et al., 2023). More precisely, we annotate a set of K_2 examples for demonstrations and prompt the ChatGPT for atomic fact generation with P' as follows: $E_i = ChatGPT(A', P'), i \in [1, C]$, where A' are all descriptive sentences identified in step 1, $E_i = \{e_i^1, \cdots, e_i^{n_i}\}$ represents all n_i atomic facts belonging to the i-th category, and C stands for the total number of categories (*i.e.,* $C = 5$ in our case) and the category include "entity", "relation", "color", "count", and "others". Further details regarding the specific prompt P' utilized in this process can be found in Section L of the Appendix.

4

Fact Verification. In this stage, we compare the atomic facts derived above with the image to determine if the facts are faithful to the input visual information. Specifically, to calculate the FAITHSCORE for the derived atomic facts, we first compute the score for each fact and then aggregate them to derive the overall score using the following formula:

$$\hat{s} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n_i} w_i^j \cdot s(e_i^j, I)}{\sum_{i=1}^{C} n_i}, \qquad (1)$$

where $\hat{s}$ represents the overall FAITHSCORE of the answer A. The function $s(e_i^j, I)$ refers to the verification function (*i.e.,* Verifier), which measures whether e_i^j can be supported by the input image I. The parameter w_i^j is a weighted factor that can be used to assign different weights to different atomic facts for various tasks. To implement function $s(e_i^j, I)$, we resort to the Visual Entailment Model (VEM) (*e.g.,* OFA (Wang et al., 2022)), which is able to predict whether the image semantically entails the text. We elaborate on selections of the verifier models in Section 4.3. In particular, when the output of the VEM is positive, indicating that the image I semantically entails the text e_i^j resulting in $s(e, I) = 1$, and negative otherwise. In this work, we set all the weights w_i^j to 1, following the setting of the existing work (Min et al., 2023; Krishna et al., 2023). In addition, we further introduce a sentence-level FAITHSCORE metric as follows, $\hat{s}_s = 1 - S_h/S$, where S is the total number of descriptive sub-sentences in the answer and S_h is the total number of descriptive sub-sentences with hallucinations.

4 Meta-evaluate FAITHSCORE

To verify that our automatic evaluation correlates with human judgment, we conduct human evaluations in terms of hallucination. We select the test dataset from the LLaVA paper (Liu et al., 2023e) (LLAVA-Bench) for human evaluation, which is constructed based on the MSCOCO dataset. This test set is a visual instruction following dataset comprising three distinct question types: detailed description, conversation, and complex question. For each type, this dataset includes 90 questions. We select answers from LLaVA (Liu et al., 2023e) and InstructBLIP (Dai et al., 2023) models for evaluation.

| Recognizer | LVLM | | Overall |
	LLaVA	InstructBLIP	
ChatGPT	89.84	91.58	90.74
LLaMA-7B	68.01	71.39	69.75
LLaMA-7B (w/ context)	72.80	66.76	69.68

Table 1: Comparison of recognizer LLMs' accuracy (%) on identifying descriptive sub-sentences. For LLaMA, we used two different prompt settings, either to input only the sub-sentence or both the sub-sentence and its context into the model (LLaMA-7B w/ context).

4.1 Human Evaluations of Hallucinations

For each test example, we craft an annotation process to assign the faithfulness score to models' generated answers via the subsequent steps.

Step 1: Sub-sentence Identification. Annotators first review the given question, the corresponding answer, and the associated image. Subsequently, they evaluate each sub-sentence extracted from the answer. If a sub-sentence is an objective description of visual information, they mark it as the "descriptive" category; otherwise, it's categorized as "analytical". For the "analytical" sub-sentence, annotators should skip the following steps. Otherwise, they should follow the next steps.

Step 2: Atomic Fact Generation and Revision. In this step, human annotators are asked to decompose the descriptive sub-sentences into a sequence of atomic facts. To optimize the annotation process and reduce the time required, we pre-supply atomic facts derived from ChatGPT. Annotators then have the flexibility to use or modify these facts as needed. In particular, annotators examine each atomic fact to ensure its fidelity to the given sub-sentence. The facts that are either redundant or non-atomic are asked to be removed. Subsequently, the focus shifts to the linguistic aspect, ensuring that each atomic fact is articulated in a coherent manner and that it accurately represents the original entity or concept of the answer by revising facts manually. Additionally, any missing atomic facts from the descriptive sub-sentence are added. For the process of removing and revising atomic facts, please refer to the Interface functionalities in the Appendix. Errors introduced by the ChatGPT in this stage are shown in Appendix G.

Step 3: Fact Verification. In this step, for every individual atomic fact derived from the descriptive sub-sentence, annotators assess its consistency with the given image. If the content of atomic facts

5

Verifier	LVLM		Overall
	LLaVA	InstructBLIP	
OFA-EM	81.07	78.08	79.42
OFA	84.47	80.71	82.39
mPLUG	84.95	83.86	84.35
BLIP-2-flant5xl	78.64	77.42	77.97
BLIP-2-flant5xxl	82.36	83.20	82.83
LLaVA	67.25	67.10	67.17
LLaVA-1.5	85.65	84.49	85.07

Table 2: Comparison of the Verifier LLMs accuracy on verifying the atomic facts (the third step).

Metric	r (%)	ρ (%)	τ (%)
BLEU-4	-1.9	-8.2	-5.8
ROUGE-L	-8.7	-6.2	-4.7
METEOR	-12.2	-8.5	-6.3
CHAIR	16.8	19.2	14.8
CLIP-Score	19.8	16.6	11.7
SPICE	20.2	21.3	25.4
Ours	48.2	38.4	47.6

Table 3: Correlation between each evaluation metric and human judgment on LVLMs (i.e., LLaVA and InstructBLIP) hallucinations, measured by Pearson's r, Spearman's ρ, and Kendall's τ. The p-value of the significant test between our result and the baseline result is less than 0.01.

is not present or contradicts the image, it's identified as a hallucination, and accordingly marked as "yes". Conversely, if the element is in alignment with the image, it's validated and marked as "no". To quantify the human evaluation of faithfulness, we employ the Likert Scale (Likert, 1932). This approach transforms human evaluations into a tangible scale, ranging from 1 (being the poorest) to 5 (being the best). The details about the annotation process are given in Section A of the Appendix.

4.2 Recognizer Accuracy on Descriptive Sub-sentence Identification

To obtain the performance of recognizers (e.g. LLMs) on the sub-sentence identification task, we construct a sub-sentence identification dataset based on our annotated samples. The final label for each sub-sentence is determined by the majority voting scheme. The total number of sub-sentences is $1,382$ and the average number of sub-sentences in the answer is 7.68. We select the superior Chat-GPT (Proprietary) and LLaMA-7B (Public) models for this task and report their accuracy on identifying descriptive sub-sentences. The results are shown in Table 1. ChatGPT outperforms LLaMA-7B on sub-sentence identification. For LLaMA-7B based method, when additional context beyond the sub-sentence itself is included, there is an improvement on LLaVA answers test set, but overall there is no significant improvement.

4.3 Verifier Accuracy on Fact Verification

Another key factor of our automatic method is the reliability of the verifier visual entailment model (VEM). Hence, we also evaluate the accuracy of different VEMs on the annotated samples. Because of the atomic fact revision operation during the annotation process, there may be some differences in atomic facts labeled by different annotators. To improve reliability, we only keep these

atomic facts annotated by all three annotators for VEM evaluation. The final label for each atomic fact is determined by the majority voting scheme. The total number of atomic facts derived from descriptive sub-sentences is $1,380$ and the average number of atomic facts in each descriptive sub-sentence is 2.04. For verifier VEMs, we evaluate OFA-EM, OFA (Wang et al., 2022), mPLUG (Li et al., 2022), BLIP-2-flant5xl, BLIP-2-flant5xxl (Li et al., 2023b), LLaVA, LLaVA-1.5 (Table 2). More details about these models are shown in Section B of the Appendix. Among all models, LLaVA-1.5 performs best on fact verification, so we use it for estimating FAITHSCORE in Section 5. Another potential reason for employing this LVLM as a verifier is that our verification task is a discriminative task that usually generates a shorter response, and tends not to generate hallucination, which has been demonstrated in our Section 5.5 (The Influence of Answer Length on Hallucinations) and the existing work (Min et al., 2023; Hu et al., 2023).

4.4 Correlations with Human Evaluations

To prove the superiority of our proposed metric FAITHSCORE, we compare it with several multimodal generation evaluation metrics: 1) reference-based: BLEU-{4} (Papineni et al., 2002), Rouge-{L} (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CHAIR (Rohrbach et al., 2018), SPICE (Anderson et al., 2016) and 2) reference-free: CLIP-Score (Hessel et al., 2021). Table 3 delineates the correlation between various evaluation metrics and human judgment regarding LVLM faithfulness. Among all metrics, our metric FAITHSCORE achieved the best correlation with human correlation. More details and analysis about human correlation can be found in Appendix F and J.
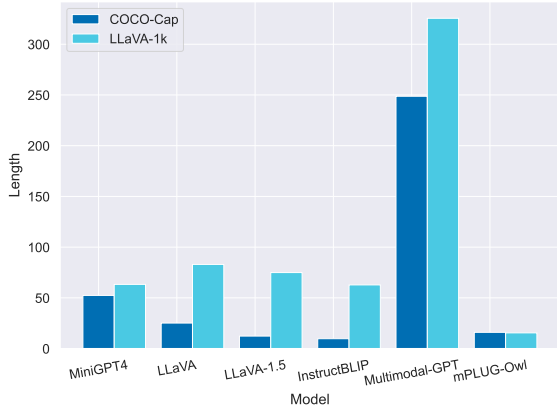
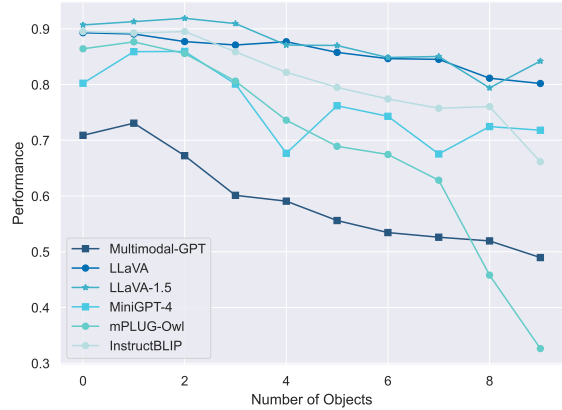Figure 3: Answer lengths distributions of different models on two benchmark datasets.

Figure 4: The relation between FAITHSCORE and numbers of objects (*i.e.,* entities) in the answers (LLaVA-1k dataset). As the number of entities increases, model performance (*i.e.,* FAITHSCORE) drops significantly.

5 Evaluating Vision-Language Model Hallucinations with FAITHSCORE

5.1 Models

We selected six open-source LVLMs for evaluation. 1) MiniGPT-4 (Zhu et al., 2023); 2) LLaVA (Liu et al., 2023e); 3) InstrucBLIP (Dai et al., 2023); 4) Multimodal-GPT (Gong et al., 2023); 5) mPLUG-Owl (Ye et al., 2023); 6) LLaVA-1.5.

5.2 Datasets

To assess the performance of existing LVLMs, we conducted experiments using two datasets. Here is a description of each dataset: (1) MSCOCO-Cap: This dataset is designed for the image captioning task. We randomly select 1,000 images from the MSCOCO (Lin et al., 2014) validation set and devised the prompt as "Generate a concise caption for the given image"; (2) LLaVA-1k: We extract 1,000 images from the MSCOCO validation set and generated three types of prompt-answer pairs (*i.e.,* detailed description, conversation, and complex question) for each image by ChatGPT, following the data generation method in (Liu et al., 2023e).

5.3 Hallucination Evaluation

Table 4 presents a comprehensive performance comparison of various models in terms of FAITH-SCORE when benchmarked on the LLaVA-1k and MSCOCO-Cap datasets. We observe that: (1) LLaVA-1.5 outperforms their counterparts in most situations. This demonstrates their preeminent capability in achieving and maintaining faithfulness during generation processes. (2) It's worth noting that different models have similar performance across tasks. For instance, MiniGPT achieved 0.5679, 0.5768, and 0.5691 FAITHSCORE on the

"Conversation", "Detailed Description", and "Complex Question" tasks, respectively. (3) For most models, the performance on the MSCOCO-Cap dataset is better than their performance on the LLaVA-1K dataset. The potential reason may be that model answers to the MSCOCO-Cap questions are usually shorter than their answers to the LLaVA-1K questions (see Figure 3).

5.4 Sentence-level Hallucination Evaluation

To further understand the faithfulness of LVLMs, we evaluate them with the FAITHSCORE (sentence-level). Table 5 shows the sentence-level FAITHSCORE evaluation across different LVLMs. Multimodal-GPT achieves poor performance in FAITHSCORE it also performs less favorably in terms of sentence-level hallucination evaluation. In addition, LLaVA-1.5 performs well in terms of FAITHSCORE and FAITHSCORE (sentence-level). This indicates the consistency between FAITH-SCORE and sentence-level FAITHSCORE.

5.5 Other Analysis

The Influence of Answer Length on Hallucinations. To further elucidate the impact of answer length on hallucinations, we analyze answer lengths across various LVLMs on different datasets. As illustrated in Figure 3, there's a significant variation in the distribution of answer lengths produced by different models. Multimodal GPT consistently generates the lengthiest responses, potentially compromising its performance across tasks. In contrast, mPLUG-Owl tends to produce shorter answers than its counterparts, hence it may generate fewer hallucinations. Meanwhile, the image cap-

7

	LLaVA-1k				MSCOCO-Cap
	Conversation	Detailed Description	Complex Question	Overall	-
Multimodal-GPT	0.5321	0.5299	0.5385	0.5335	0.5440
MiniGPT-4	0.5679	0.5768	0.5691	0.5713	0.6359
mPLUG-Owl	0.7246	0.7240	0.7015	0.7167	0.8546
InstructBLIP	0.8061	0.8161	0.8049	0.8091	0.9392
LLaVA	0.8302	0.8386	0.8392	0.8360	0.8729
LLaVA-1.5	**0.8569**	**0.8611**	**0.8516**	**0.8566**	**0.9425**

Table 4: FAITHSCORE evaluation results (↑) of different LVLMs on the LLaVA-1k and MSCOCO-Cap datasets.

	LLaVA-1k				MSCOCO-Cap
	Conversation	Detailed Description	Complex Question	Overall	-
Multimodal-GPT	0.4615	0.4827	0.5131	0.4858	0.6277
MiniGPT-4	0.6441	0.6489	0.6499	0.6476	0.6017
LLaVA	0.7106	0.6979	0.7038	0.7041	0.6681
InstructBLIP	0.7231	0.7327	0.7149	0.7236	0.7970
mPLUG-Owl	0.7369	0.7163	0.7344	0.7292	0.6447
LLaVA-1.5	**0.7722**	**0.7717**	**0.7699**	**0.7713**	**0.8258**

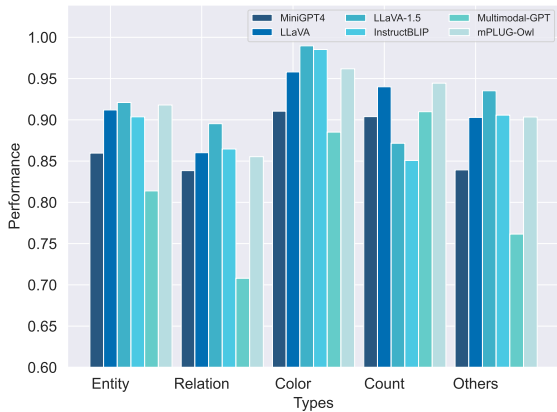Table 5: FAITHSCORE (sentence-level) evaluation results (↑) of different LVLMs.

Figure 5: FAITHSCORE on each type of atomic facts on the LLaVA-1k benchmark. The types are ENTITY, RELATION, COLOR, COUNT, and OTHERS.

tioning task showed better faithfulness in generated content than the other task for most LVLMs. This may be attributed to the fact that captioning sentences mainly are brief descriptions and shorter.

The Influence of Multiple Objects. Figure 4 shows how the number of objects in the answer generated by different models affects the FAITH-SCORE The model's faithfulness varies with the number of objects. While all models start with relatively high scores when there are few objects in the answer, their performance generally drop as the number of objects increases. For example, Instruct-BLIP starts with a high FAITHSCORE of 0.895 for 1 object and sustains a relatively low score of 0.662 for 10 objects.

Analysis on Types of Hallucination To deduce the model strengths and vulnerabilities of each in maintaining faithfulness, we compared the faithfulness performance of various models across different categories of hallucination. We mainly investigated the five distinct categories: ENTITY, COUNT, COLOR, RELATION, and OTHER attributes, motivated by the existing works. From Figure 5, we can observe that while LLaVA-1.5 consistently excels across most categories, other models also showcase strengths in specific domains. The bad performance of some types may provide insightful information for model improvement. Importantly, achieving consistently high faithfulness across a diverse range of categories remains a formidable challenge for LVLMs.

6 Conclusion

We introduce a novel metric called FAITHSCORE for evaluating free-form and open-domain answers generated by large vision-language models. Compared to previous metrics, FAITHSCORE offers a finer level of granularity, interpretability, and closer alignment with human judgments. Our quantitative analysis demonstrates that current LVLMs are prone to visual hallucination problems. We also find that the answer length and number of objects could affect the faithfulness of LVLMs. In addition, the faithfulness performance of LVLMs on different types of atomic facts varies. We expect that FAITHSCORE will be of great value for evaluating forthcoming advanced LVLMs.

8

Limitations

FAITHSCORE focuses primarily on factual precision, ensuring that each piece of information in a generated text is supported by the visual input. However, it does not account for factual recall, meaning it doesn't penalize models for generating fewer facts. This can be seen as unfair, as there's often a trade-off between precision and recall. Additionally, the distinction between precision and recall can sometimes be unclear, as a generation may contain supported information but still miss important details. To address this, we suggest reporting FAITHSCORE and the average length of generated text. We leave a more holistic approach for future work.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433. IEEE Computer Society.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, pages 65–72. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. vicuna: An open-source chatbot impressing gpt-4 with 90

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

9

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. *CoRR*, abs/2305.04790.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The International Conference on Learning Representations*. OpenReview.net.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. *CoRR*, abs/2303.11897.

Ting-Hao 'Kenneth' Huang, Francis Ferraro, N. Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 1642–1661. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *ArXiv*, abs/2305.10355.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. 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*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023c. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023d. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning. *CoRR*, abs/2304.08485.

10

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Zi-wei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023. Evaluation and mitigation of agnosia in multimodal large language models. *CoRR*, abs/2309.04041.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Denver*, pages 5988–5999. ACM.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *CoRR*, abs/2305.14251.

David S Moore, William I Notz, and William Notz. 2006. *Statistics: Concepts and controversies*. Macmillan.

OpenAI. 2022. Chatgpt blog post. https://openai.com/blog/chatgpt.

OpenAI. 2023. Gpt-4v(ision) system card. *OpenAI Blog Post*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Anirudh Sundar and Larry Heck. 2022. Multimodal conversational AI: A survey of datasets and approaches. In *Proceedings of the Workshop on NLP for Conversational AI, ConvAI@ACL*, pages 131–147. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045.

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

11

A Likert Scale Guideline

For human evaluation, we utilize the Likert Scale to get the final faithfulness score for each testing sample. Specifically, suppose the generated answer consists of n atomic facts, out of which x atomic facts are determined as hallucinations. Both n and x are labeled by the annotators. The benchmark scoring guideline is outlined as follows:

- Score 1: All atomic facts are hallucinations, symbolized as $x == n$;

- Score 2: More than half of the atomic facts are hallucinations, represented as $x > n/2$;

- Score 3: Half or fewer atomic facts are hallucinations, represented as $n/3 <= x < n/2$;

- Score 4: Less than one-third of the atomic facts are hallucinations, which translates to $x < n/3$;

- Score 5: All atomic facts accurately represent the visual content, meaning $x = 0$.

B   Details about VEMs

We select OFA-EM, OFA[1] (Wang et al., 2022), mPLUG[2] (Li et al., 2022), BLIP-2-flant5xl, BLIP-2-flant5xxl[3] (Li et al., 2023b), LLaVA (Liu et al., 2023e), and LLaVA-1.5[4] (Liu et al., 2023d) as VEM and evaluate them based on our annotated dataset.. OFA-EM is an open-source model which was finetuned on the SNLE-VE dataset (Xie et al., 2019). Hence, this model can tackle visual entailment tasks directly. For the OFA-EM model, the "neutral" is categorized as hallucination because the OFA can't decide whether the verified content appears in the input image. For the other models, they are also open-source and finetuned on the visual question answering dataset. To enable them to tackle the visual entailment task, we get an input a prompt "Statement: {*atomic facts*} Is this statement right according to the image? Please output yes or no.", into models.

C   Testing Examples of GPT-4Vision

Hallucination in Advanced GPT-4Vision   Here we test the GPT-4Vision model on four examples. Based on the results, we can come to the conclusion

that the GPT-4Vision answers still contain various hallucinations despite it may have very large parameters and have been trained with a large corpus, as shown in Figure 6.

D   More benchmarks

We further compute our metric on one dataset: LRV-Instruction (Liu et al., 2023c). The results are shown as follows, which are consistent with the experimental results on datasets LLaVA-1k and MSCOCO-Cap: InstructBLIP 0.6626, Multimodal-GPT 0.4903, mPLUG-Owl 0.6433, MiniGPT-4 0.4638, LLaVA 0.7017, LLaVA-1.5 0.7855.

E   Examples of Evaluation

Here we show three examples of how FAITH-SCORE is computed and the existing best reference-free CLIP-Score value in Figure 7, Figure 8, and Figure 9. Additionally, we present an example (see Figure 10) where the proposed metric score diverges from human judgment, illustrating a discrepancy attributed to an error generated by the recognizer system.

F   More Details about Human Evaluation

We employ 3 workers for annotation and each person annotated 180 testing samples, via Amazon Mechanical Turk[5]. Every worker is a native English speaker. They are paid 15-20 USD per hour. Every worker went through a qualification test of 2 hours and was tested to be highly qualified. We designed one HIT to consist of one question-answer pair. The average time to complete one HIT (including all steps of the annotation process) is 212.8 seconds. After the annotation process, we calculate the inter-annotator agreement rate by the Fleiss' κ. Firstly, we computed the Fleiss' κ values across all annotators for the sub-sentence identification task, arriving at a value of 75.97%. This signifies a robust consensus among the annotators (Moore et al., 2006). Additionally, for the definitive faithfulness score (1-5 Likert Scale), we computed the values involving all annotators and achieved a result of 60.0%. This concordance among the evaluation participants suggests the human evaluation results are reliable.

We show our human evaluation results and automatic evaluation results in Table 6. From this Table, we find that models that perform better in the

---

[1]https://github.com/OFA-Sys/OFA.
[2]https://github.com/X-PLUG/mPLUG.
[3]https://github.com/salesforce/LAVIS/tree/main/projects/blip2.
[4]https://github.com/haotian-liu/LLaVA.

[5]https://requester.mturk.com/.

12

Model	Human	Automatic
LLaVA	0.7708	0.6997
InstructBLIP	0.7804	0.7165

Table 6: Human evaluation results and automatic evaluation results of different LVLMs on the LLaVA dataset.

Metric	Pearson's r %	Spearman's ρ %	Kendall's τ %
BLEU-1	-15.1	-10.3	-7.5
BLEU-2	-12.7	-9.0	-6.6
BLEU-3	-7.2	-10.6	-7.6
BLEU-4	-1.9	-8.2	-5.8
ROUGE-1	-6.6	-3.0	-2.7
ROUGE-2	-5.7	-4.4	-3.4
ROUGE-L	-8.7	-6.2	-4.7
METEOR	-12.2	-8.5	-6.3
CHAIR	16.8	19.2	14.8
CLIP-Score	19.8	16.6	11.7
SPICE	20.2	21.3	25.4
Ours	48.17	38.44	47.61

Table 7: Correlation between each evaluation metric and human judgment on LVLM hallucinations, measured by Pearson's r, Spearman's ρ, and Kendall's τ.

manual evaluations also have better performance in the automated evaluations. This indicates the high correlation between objective and subjective evaluation.

To facilitate the annotator's working process, we designed a user interface, as shown in Figure 13. Annotators have the option to start by reading the instructions located at the top of the interface, and they can access more detailed instructions through a link (refer to Figure 14). Following this, annotators can proceed to review the task description. In the third section, annotators can utilize buttons for sub-sentence identification and atomic fact verification. Simultaneously, they are able to add, modify, or delete atomic facts to enhance the quality of the atomic information. For example, the annotator should remove the duplicated atomic and add entity category fact "There are suitcases." in Figure 11.

Besides, we show a comprehensive correlation comparison in Table 7. Traditional metrics that require references (*i.e.,* BLEU, ROUGE, and METEOR), have a poor correlation with human evaluation. For the open-ended question, it is hard to get a ground truth answer. For the reference answer, we use the answers provided by the LLaVA paper. This leads to a poor correlation between these metrics and human evaluation.

Surprisingly, CLIP-Score shows a similar correlation with CHAIR which is specifically devised for object hallucination evaluation. This demonstrates

the robustness and generalization of CLIP-Score. The original CHAIR show reflects the severity of the hallucinations. The larger the value of CHAIR, the more serious the hallucination problem of the model. The original CHAIR exhibits a pronounced negative correlation with human evaluation. Hence, we use the negative of CHAIR to compute the correlation.

Compared with FAITHSCORE, CHAIR achieves a sub-optimal degree of correlation. A potential reason for CHAIR's deviation from human evaluation could be rooted in its inherent design, which narrows its focus predominantly to a limited range of objects. This constrained evaluation scope may not adeptly deal with fine-grained and open-domain hallucinations, thus diminishing its validity and resonance with more comprehensive human evaluations. To justify our viewpoint, we compute the average number of objects with CHAIR for each answer and the result is 2.4, which is far less than the average number of atomic facts (*i.e.,* 11.3) found in our human evaluation. Amid the varied metrics landscape, our metric FAITHSCORE achieved best correlation with human correlation.

We further conduct an ablation study to investigate the overall effect of different VE models and the error introduced by the ChatGPT on FAITHSCORE. Table 8 reports the correlation between FAITHSCORE calculated by different VE models answers and human answers. We observed that the higher VE model performance is directly related to the human correlation. Table 9 reports the correlation between FAITHSCORE and different VE models calculated with the annotated atomic facts. Similarly, the higher VE model performance is directly related to the human correlation.

G   Error in Atomic Fact Decomposing

To further learn the degree of hallucination in decomposing phrases, we further sample 100 instances from the human evaluation samples and then use ChatGPT to generate atomic facts. Finally, we found the hallucination ratio (hallucinated atomic facts in proportion to all atomic facts) is just 2%. This verifies the effectiveness of our method.

H   Experimental Detail

We run all VLMs on an NVIDIA A100 GPU. The recognizer accuracy on descriptive sub-sentence identification task is defined as $a = N_s/N$ where N is the total number of sentences in the evaluated

Metric	Pearson's r %	Spearman's ρ %	Kendall's τ %
OFA_EM	31.85	21.27	29.03
BLIP-2-flant5xxl	41.80	28.52	36.81
LLaVA-1.5	48.17	38.44	47.61

Table 8: Correlation between our ablation methods and human judgment on LVLM hallucinations, measured by Pearson's r, Spearman's ρ, and Kendall's τ.

Metric	Pearson's r %	Spearman's ρ %	Kendall's τ %
OFA_EM	32.34	22.28	30.12
BLIP-2-flant5xxl	45.84	31.62	40.09
LLaVA-1.5	58.46	42.67	56.23

Table 9: Correlation between vem models with the annotated atomic facts and human judgment on LVLM hallucinations, measured by Pearson's r, Spearman's ρ, and Kendall's τ.

dataset and N_s is the number of sentences that are classified correctly.

I Proportions of the descriptive sub-sentences and analytical sub-sentences

To prove the necessity of the sentence identification step, we calculate the proportion of descriptive and analytical sub-sentences in answers to different classes of input questions (Figure 12). We can observe that the distribution of sub-sentences is significantly different in different category questions. For example, detailed description questions only have a small portion of analytical sub-sentences, while complex questions have the opposite. In addition, analytical sub-sentences account for nearly half of the distribution of clauses in the overall annotated dataset, illustrating the importance of identifying analytical sub-sentences and excluding them from the fact checking step.

J Cost and Time Analysis

Gathering accurate hallucination evaluation manually for each response is both costly and time-consuming, making it unrealistic. Therefore, although our reference-free metric typically takes more time than traditional metrics such as BLEU and ROUGE, it is much more important for evaluating model output. There is a trade-off between evaluation results and speed. Although some metrics (such as BLEU) can achieve higher speed, our metric can achieve a superior correlation with human evaluation. Compared with human evaluation, our method can speed up the evaluation process. In addition, we can verify multiple facts in parallel to

Model	LLaVA-1k	Model	LLaVA-1k
Multimodal-GPT	0.5335	LLaVA	0.8360
MiniGPT-4	0.5713	LLaVA-1.5	0.8566
mPLUG-Owl	0.7167	GPT-4V	0.8788
InstructBLIP	0.8091	Gemini	0.8865

Table 10: Model performance on LLaVA-1k benchmark.

speed up the time. The average consuming time for a long response is 4s, which is much less than the manual time of 212 seconds, reducing about 1.2 USD cost.

K Samples of Description/Analytics Sub-sentence

We provide three examples of "description"/"analytics" sub-sentence, where [A] denotes the analytical sub-sentence label and [D] denotes the descriptive sub-sentence label. (1) The skateboard is positioned on a ramp, [D] with the skateboarder standing on it. [D] (2) The image features a white bird, [D] which is likely a swan. [A] (3) The image features a young boy standing on a skateboard, [D] which is placed on a wooden ramp. [D] The boy is wearing a green shirt and brown shorts, [D] and he is positioned on the ramp, ready to ride down. [D] The wooden ramp is placed on a sidewalk, [D] which is a common location for skateboarding. [A] The presence of the ramp and the sidewalk suggests that the boy is likely practicing or enjoying skateboarding in a public space. [A] The overall composition of the image highlights the boy's focus and determination as he prepares to ride down the ramp, [A] showcasing the excitement and thrill of skateboarding. [A]

L Prompts

We detailed the prompts of sub-sentence identification and atomic fact generation in Figure 15 and Figure 16, respectively.

M Results on Closed-source Model

To comprehensively evaluate the existing LVLMs, we further evaluate the closed-source model GPT-4V and Gemini (Anil et al., 2023) with several open-source models. The results are shown in Table 10. The results show the closed-source models surpass the existing open-source model.

Prompt:
What is the position of the skateboard in the image?
Image:

Answer:
The image primarily focuses on a person standing on an upside-down skateboard.

Prompt:
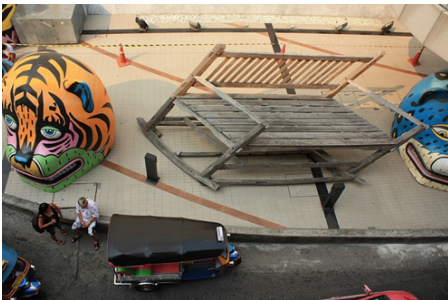Is the person riding a motorbike or scooter?

Image:

Answer:
The person is riding a scooter.

Prompt:
What kind of objects are included in the art installation?
Image:

Answer:
The art installation in the image includes:

1. Large spherical objects painted with vibrant and colorful designs. Specifically:
   - One sphere is painted with a design that resembles a stylized tiger face with a mix of bright colors, including orange, black, pink, and green.
   - Another sphere, partially visible to the right, has a design that appears to be another animal face, with blue as one of the dominant colors……

Prompt:
Can you describe the main features of this image for me?
Image:

Answer:
Certainly! Here's a detailed description of the main features of the image……
There are two main individuals in the foreground:
  1. **Man**: He is wearing a red shirt and blue pants. On his head, he has an umbrella hat with multicolored panels, predominantly in red, yellow, and blue.
  2. **Woman**: She is wearing a blue dress with a design on the front, possibly a yellow symbol or character…….

Figure 6: Illustration of answers generated from GPT-4Vision. Similar to existing open-source VLMs, GPT-4Vision is still prone to the hallucination problem in the generated answer. The blue contents denote hallucinations.

15

Prompt:

What is the main object in the image?

Image:

LLaVA Answer:

<u>The main object in the image is a colorful beach umbrella.</u>

Atomic Facts

- There is a beach umbrella. ✔
- The beach umbrella is colorful. ✔
- The beach umbrella is the main object in the image. ✔

FaithScore: 100%

CLIP-Score: 0.34

Figure 7: Illustration of how FAITHSCORE is computed for a testing sample. The underlined contents denote recognized descriptive content.

Prompt:

Can you elaborate on the elements of the picture provided?

Image:

Answer:

<u>The image features a desk with a newspaper, a pair of sunglasses, and a book. The desk is cluttered with various items, including a pen, a pencil, and a notebook. The newspaper is spread out on the desk, and the sunglasses are placed on top of it. The book is also present on the desk,</u> possibly indicating that the person using the desk is reading or studying. The overall scene suggests a workspace or study area where the person has various items and resources available for their tasks

---

Atomic Facts

- There is a desk. ✔
- There is a newspaper. ✔
- The newspaper is on the desk. ✔
- There is a pair of sunglasses. ✔
- There is a book. ✔
- There are various items. ✔
- The items are cluttered on the desk. ✔
- The desk is cluttered. ✔
- There is a pen. ✘
- There is a pencil. ✘
- There is a notebook. ✔
- The newspaper is spread out on the desk. ✔
- The sunglasses are placed on top of newspaper. ✔
- The book is on the desk. ✔

FaithScore: 86%

CLIP-Score: 0.37

Figure 8: Illustration of how FAITHSCORE is computed for a testing sample. The underlined contents denote recognized descriptive content.

Prompt:

Explain the visual content of the image in great detail.

Image:

Answer:

The image depicts a group of people walking down a dirt road in a rural area. One person is riding a motorcycle, while others are walking alongside the road. There are several cars parked along the side of the road, likely belonging to the people who are walking or riding on the motorcycle. In addition to the people and vehicles, there are several trees scattered throughout the scene, adding to the natural setting

Atomic Facts	
- There is a group of people.	✔
- There is a dirt road.	✔
- There is a rural area.	✔
- People are walking down a dirt road.	✔
- The dirt road is in a rural area.	✔
- There is a person.	✔
- There is a motorcycle.	✔
- The person is riding a motorcycle.	✔
- There are others.	✔
- Others are walking alongside the road.	✔
- Others are walking.	✔
- There are cars.	✘
- The cars are parked.	✘
- There are trees.	✔
- The trees are scattered.	✘
FaithScore: 80%	

CLIP-Score: 0.31

Figure 9: Illustration of how FAITHSCORE is computed for a testing sample. The contents highlighted by the underline denote recognized descriptive content.

Prompt:

What is the position of the skateboard in the image?

Image:

Answer:

The image primarily focuses on a person standing on an upside-down skateboard.

Human Annotated	Automatic
- There is a person. ✔	- There is a person. ✔
- There is a skateboard. ✔	- There is a skateboard. ✔
- The skateboard is upside-down. ✔	- The skateboard is upside-down. ✔
- A pearson is standing on an upside-down skateboard. ✘	- A pearson is standing on an upside-down skateboard. ✔
FaithScore: 80%	**FaithScore: 100%**

Figure 10: Illustration of the case that the proposed metric score disagrees with human judgement

Generated Answer
The two suitcases in the image are black.
Atomic Facts
There are two suitcases.
The suitcases are black.
There are two suitcases.
A person is standing on a sketaboard.

Figure 11: Illustration of atomic facts generated by ChatGPT. The red contents denote the duplicated atomic fact.
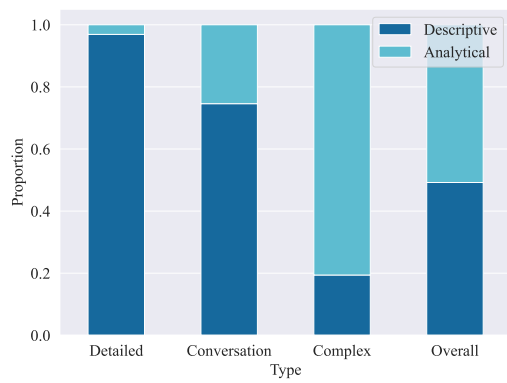
17

Figure 12: Illustration of the proportions of the descriptive sub-sentences and analytical sub-sentences in the answers. "Detailed" and "Complex" denote the "detailed description" and "complex question" categories, respectively. The results are obtained from the 180 annotated samples.

Annotation Instructions

If there is any definition that you cannot understand, please refer to the google doc .

Annotation Procedures:

1.Read the question, answer, and image.

2.Read each sub-sentence that is extracted from the answer. If it is a description sentence, check the "description" box. Otherwise, check the "analytical" box.

3.If you check the "analytical" box, please skip the following steps and repeat step 2 on other sub-sentences.

4.Read all elements in the sub-sentence. To ensure elements are faithful to the above image, you should check them by the following process:

    a.Check whether each element is reasonable according to the sub-sentence. If the element is repeated or doesn't appear in the corresponding sub-sentence, click "remove" to delete it. If the element is not atomic, click the "remove" to delete it.

    b.Check whether the element is a natural sentence or the sentence correctly describes the element/entity. If not, please rephrase/revise them.

    c.Check whether there is any element in the sub-sentence that is not described in the elements part. If so, click "Add an Element" to add it.

        i.If you find the index of an element is not correct, please ignore it.

5.For each element, check whether it contains a hallucination. If so, click "yes". Otherwise, click "no".

Task

We would like to request your feedback on the performance of an AI assistant in response to the user question displayed below. We are evaluating the quality of the generated answer by Vision-Language Models (VLMs). The VLMs can generate a response for multimodal input. The VLMs seem to generate the content (e.g., "person" in the above image) which don't exist in the image input. There are various types of hallucinations, such as entities, relations, and attributes. In addition, some content in the answer may not be a hallucination despite the fact that the content doesn't appear in the input image. Because they are reasonable analyses within the context. Our task is to identify hallucinations that appear in the answers

---

Elapsed Time: 0:13:3

Question:

What is the position of the skateboard in the image?

Image:

Answer:

The skateboard is positioned on a ramp, with the skateboarder standing on it.

Whether this sub-sentence is a descriptive sentence?

sub-sentence 1: The skateboard is positioned on a ramp.   descriptive analytical

Do these elements contain hallucination?

🟥Remove element 1: There is a skateboard. yes no

🟥Remove element 2: There is a ramp. yes no

🟥Remove element 3: The skateboard is positioned on a ramp. yes no

🟥Remove element 4: The skateboard is on a ramp. yes no

🟩Add an Element

Whether this sub-sentence is a descriptive sentence?

sub-sentence 2: with the skateboarder standing on it. descriptive analytical

Do these elements contain hallucination?

🟥Remove element 1: There is a skateboarder. yes no

🟥Remove element 2: The skateboarder is standing on a skateboard. Yes no

🟩Add an Element

Figure 13: System software User Interface (UI) for annotators. Annotators can read the instructions at the top of the interface and get detailed instructions (see Figure 14) via a link. Then the annotator can read the task description. In the third part, the annotator can click buttons for sub-sentence identification and atomic fact verification. Meanwhile, they can add, edit, and remove atomic facts to get high-quality atomic information.

Definitions:

•Atomic Elements: Atomic information derived from the image-related sub-sentence. It is in natural language format. The types of atomic elements contain the entity, relation between entities, color, counting, and other attributes. The derivation process is shown as follows,

> Answer: The image features a red velvet couch with a cat lying on it.
> Entities: There is a couch. There is a cat.
> Relations: A cat is lying on a couch.
> Colors: There is a red couch.
> Counting:
> Other attributes: There is a velvet couch.

An element is **atomic** that needs to meet the following requirements for different types.

> Entities: Only contain **one** entity.
> Relations: Only can be decomposed into **two atomic elements** of entity type.
> Colors: Color information of **one** kind of entity.
> Counting: Counting information of **one** kind of entity.
> Other attributes: Attribute information of **one** kind of entity.

•Descriptive sub-sentence: Objective descriptions of visual information.
•Analytics sub-sentence: Scene or object analysis including complex reasoning or interpretations about the image. These are portions of the data that are more subjective and not grounded visually within the image.
•Hallucination: there is something **described in the sub-sentence but does not appear in the image**. In other words, if an element's content is inconsistent with the image, it is a hallucination.

Annotation Procedures:

1.Read the question, answer, and image.
2.Read each sub-sentence that is extracted from the answer. If it is a description sentence, check the "description" box. Otherwise, check the "analytics" box.
3.If you check the "analytics" box, please skip the following steps and repeat step 2 on other sub-sentences.
4.Read all elements in the sub-sentence. To ensure elements are faithful to the above image, you should check them by the following process:

1. Check whether each element is reasonable according to the **sub-sentence**. If the element is repeated or doesn't appear in the corresponding sub-sentence, click "remove" to delete it. If the element is not atomic, click the "remove" to delete it.
2. Check whether the element is a natural sentence or the sentence correctly describes the element/entity. If not, please rephrase/revise them.
3. Check whether there is any element in the **sub-sentence** that is not described in the elements part. If so, click "Add an Element" to add it.
   1. If you find the index of an element is not correct, please ignore it.

5.For each element, check whether it contains a hallucination. If so, click "yes". Otherwise, click "no".

Figure 14: Instructions for data annotation. The instruction includes some definitions (e.g. atomic facts and descriptive sub-sentence) to help annotators understand this task. Meanwhile, it also details the annotation procedures.

Give you the description and analysis of a image, please distinguish between sub-sentences that provide an actual description of the image content and those that offer commonsense associations and analysis based on the image.
Please label the text with [D] or [A] in the end of sub-sentence, where [D] denotes the actual description of the image and [A] denotes the analysis and commonsense associations based on the image and context.

Example:

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
Labeled text: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. [D] This is not a typical place to perform this activity, [A] as one would usually iron clothes in a more stationary and safe location[A], such as a home, [A] using a regular ironing board. [A] The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, [D] which can be both unsafe and unconventional. [A] Additionally, [A] it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment. [A]

The image depicts a classroom full of children working together on laptops. There are several kids in the room, with some of them sharing a laptop in pairs. The students are focused on their tasks with laptops placed on desks or tables.\n\nAside from the laptops, there are multiple chairs in the room, accommodating the students as they work. Other objects in the classroom include a bottle, a cell phone, a book, and a keyboard. Some children can also be seen using additional electronic devices such as tablets or cell phones. The overall atmosphere indicates a modern, technology-filled learning environment.
Labeled text: The image depicts a classroom full of children working together on laptops. [D] There are several kids in the room, [D] with some of them sharing a laptop in pairs. [D] The students are focused on their tasks with laptops placed on desks or tables. [D] \n\nAside from the laptops, [D] there are multiple chairs in the room, [D] accommodating the students as they work. [A] Other objects in the classroom include a bottle, [D] a cell phone, [D] a book, [D] and a keyboard. [D] Some children can also be seen using additional electronic devices such as tablets or cell phones. [D] The overall atmosphere indicates a modern, [A] technology-filled learning environment. [A]

The image shows a man in a black shirt and shorts standing on a tennis court, holding a tennis racket, and celebrating with a raised fist. A camera operator is nearby, recording the tennis player's actions, which might be for a competition or production.\n\nSeveral chairs are situated around the tennis court, with one closely placed behind the celebrating player and three others at the edges of the image. Additionally, there are four more individuals located around the court, one close to the camera operator and the others at different spots in the scene. They appear to be onlookers, possibly watching the event or supporting the tennis player.
Labeled text: The image shows a man in a black shirt and shorts standing on a tennis court, [D] holding a tennis racket, [D] and celebrating with a raised fist. [D] A camera operator is nearby, [D] recording the tennis player's actions, [D] which might be for a competition or production. [A] \n\nSeveral chairs are situated around the tennis court, [D] with one closely placed behind the celebrating player and three others at the edges of the image. [D] Additionally, [A] there are four more individuals located around the court, [D] one close to the camera operator and the others at different spots in the scene. [D] They appear to be onlookers, [A] possibly watching the event or supporting the tennis player. [A]

They are skiing in a wooded environment, following a trail through the trees while surrounded by snow.
Labeled text: They are skiing in a wooded environment, [D] following a trail through the trees while surrounded by snow. [D}

The airplane is on the tarmac at the airport, and it's being resupplied with food by the food service truck.
Labeled text: The airplane is on the tarmac at the airport, [D] and it's being resupplied with food by the food service truck. [D]

To perform the frisbee trick shown in the image, where the man is passing a frisbee between or underneath his legs, a person would need a combination of skills. These skills include good hand-eye coordination, agility, balance, flexibility, and dexterity. Additionally, the ability to throw and catch the frisbee accurately while maintaining control of bodily movements would also be essential. To perfect the trick, practicing these skills and building up muscle memory through repetition would be beneficial.
Labeled text: To perform the frisbee trick shown in the image, [D] where the man is passing a frisbee between or underneath his legs, [D] a person would need a combination of skills. [A] These skills include good hand-eye coordination, [A] agility, [A] balance, [A] flexibility, [A] and dexterity. [A] Additionally, [A] the ability to throw and catch the frisbee accurately while maintaining control of bodily movements would also be essential. [A] To perfect the trick, [A] practicing these skills and building up muscle memory through repetition would be beneficial. [A]

The skateboarder, performing a trick in the air, is trying to flip with his skateboard in a park. This activity involves a certain level of risk, especially given the complexity of the trick. Potential risks include falling off the skateboard, which could result in injuries, such as broken bones, sprains, or bruises. Additionally, the skateboarder may risk colliding with nearby objects or other park users if he loses control of the skateboard during the trick. To minimize these risks, the skateboarder should make sure to practice in a safe environment, use proper protective gear, such as a helmet and pads, and gradually develop their skills before attempting more complicated tricks. Being mindful of their surroundings and maintaining a safe distance from others is also essential to ensure the safety of the skateboarder and others around him.
Labeled text: The skateboarder, [D] performing a trick in the air, is trying to flip with his skateboard in a park. [D] This activity involves a certain level of risk, [A] especially given the complexity of the trick. [A] Potential risks include falling off the skateboard, [A] which could result in injuries, [A] such as broken bones, [A] sprains, [A] or bruises. [A] Additionally, [A] the skateboarder may risk colliding with nearby objects or other park users if he loses control of the skateboard during the trick. [A] To minimize these risks, [A] the skateboarder should make sure to practice in a safe environment, [A] use proper protective gear, [A] such as a helmet and pads, [A] and gradually develop their skills before attempting more complicated tricks. [A] Being mindful of their surroundings and maintaining a safe distance from others is also essential to ensure the safety of the skateboarder and others around him. [A]

{Testing sample}
Labeled Text:

Figure 15: A prompt given to ChatGPT to identify descriptive sub-sentence from answers of VLMs.

Given an answer output by a vision-language model, breakdown it into independent atomic facts from it.
First extract elements from the answer. Then classify each element into a category (object, relation, human, animal, food, attribute, counting, color, material, spatial, location, shape, other). Finally, generate atomic facts for each element.

Answer: A man posing for a selfie in a jacket and bow tie.
Entities: There is a man. There is a selfie. There is a jacket. There is a bow tie.
Relations: A man is in a jacket. A man is in a bow tie. A man posing for a selfie.
Colors:
Counting:
Other attributes:

Answer: The image features a red velvet couch with a cat lying on it.
Entities: There is a couch. There is a cat.
Relations: A cat is lying a couch.
Colors: There is a red couch.
Counting:
Other attributes: There is a velvet couch.

Answer: The photo is about a close-up image of a giraffe's head.
Entities: There is a head.
Relations:
Colors:
Counting:
Other attributes: There is a giraffe's head.

Answer: A horse and several cows feed on hay.
Entities: There is a horse. There are cows. There is a hay.
Relations: A horse feed on hay. Cows feed on hay.
Colors:
Counting: There are several cows.
Other attributes:

Answer: A red colored dog.
Entities: There is a dog.
Relations:
Colors: There is a red dog.
Counting:
Other attributes:

Answer: Here are motorcyclists parked outside a Polish gathering spot for women
Entities: There are motorcyclists. There is a gathering spot. There is women.
Relations: The woman is in a spot. Motorcyclist parked outside a spot.
Colors:
Counting:
Other attributes: There is a Polish gathering spot, There is a spot for woman.

Answer: {testing sample}

Figure 16: A prompt given to ChatGPT to generate atomic facts of VLMs answers.